The quadratic isoperimetric inequality for mapping tori of free group automorphisms.

Martin R. Bridson

Daniel Groves

Author address:

Martin R. Bridson, Mathematical Institute, 24–29 St Giles', Oxford, OX1 3LB, UK
E-mail address: bridson@maths.ox.ac.uk

Daniel Groves, University of Illinois at Chicago, 851 S. Morgan St., Chicago, IL 60607-7045, USA
E-mail address: groves@math.uic.edu

For Julie and Anne

Contents

Abstract

We prove that if F is a finitely generated free group and ϕ is an automorphism of F then $F \rtimes_\phi \mathbb{Z}$ satisfies a quadratic isoperimetric inequality.

Our proof of this theorem rests on a direct study of the geometry of van Kampen diagrams over the natural presentations of free-by-cylic groups. The main focus of this study is on the dynamics of the time flow of t-corridors, where t is the generator of the $\mathbb{Z}$ factor in $F \rtimes_\phi \mathbb{Z}$ and a t-corridor is a chain of 2-cells extending across a van Kampen diagram with adjacent 2-cells abutting along an edge labelled t. We prove that the length of t-corridors in any least-area diagram is bounded by a constant times the perimeter of the diagram, where the constant depends only on ϕ. Our proof that such a constant exists involves a detailed analysis of the ways in which the length of a word $w \in F$ can grow and shrink as one replaces w by a sequence of words w_m, where w_m is obtained from $\phi(w_{m-1})$ by various cancellation processes. In order to make this analysis feasible, we develop a refinement of the improved relative train track technology due to Bestvina, Feighn and Handel.

Received by the editor 10 October, 2006.

2000 *Mathematics Subject Classification.* 20F65, (20F06, 20E36, 57M07).

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

Introduction

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 the word problem in 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 [**4**], [**6**] [**7**], [**23**], [**29**], [**35**]. (See [**2**] 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 [**3**], [**18**], i.e. its Dehn function is linear. Epstein and Thurston [**22**] 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.

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

Papasoglu [**33**] proved that if a finitely presented group satisfies a quadratic isoperimetric inequality, then all of its asymptotic cones are simply connected.

Corollary A. *If F is a finitely generated free group and ϕ is an automorphism of F then every asymptotic cone of $F \rtimes_\phi \mathbb{Z}$ is simply connected.*

Ol'shanskii and Sapir [**32**, Theorem 2.5] proved that if a multiple HNN extension of a free group has Dehn function less than $n^2 \log n$ (with a somewhat technical definition of 'less than') then it has a solvable conjugacy problem. Our Main Theorem shows that free-by-cyclic groups fall into this class.

Corollary B. *If F is a finitely generated free group and ϕ is an automorphism of F, then the conjugacy problem for $F \rtimes_\phi \mathbb{Z}$ is solvable.*

Corollary B was first proved in [**8**] using different methods.

Gromov [**26**] proved that a finitely presented group is hyperbolic if and only if its Dehn function is linear. He also proved that if a Dehn function is subquadratic then it must be linear. Thus if one ranks groups according to the complexity of their Dehn functions, the groups that have a quadratic Dehn function demand particular attention. The nature of these groups is far from clear for the moment; in particular it is unclear what they have in common. It is not known, for example, whether they all have a solvable conjugacy problem; nor is it known whether the isomorphism problem is solvable amongst them. Our Main Theorem provides a rich source of new examples on which to test such questions.

Much of our modern understanding of the automorphisms of free groups has been guided by the analogies with automorphisms of free-abelian groups and surface groups [**17**]. The former analogy will prove useful is our analysis of how elements of a free group grow when one repeatedly applies an automorphism, but it offers offers us poor guidance at the level of Dehn functions: the Dehn function of $\mathbb{Z}^d \rtimes_\phi \mathbb{Z}$ can be polynomial of degree $2, 3, \dots, d+1$ or it can be exponential; it depends on the growth rate of ϕ and is quadratic only if $\phi \in \mathrm{GL}(n, \mathbb{Z})$ has finite order [**13**].

The analogy with surface automorphisms is more apt. 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 [**22**] were able to prove that $\pi_1 S \rtimes_\phi \mathbb{Z}$ is an automatic group; 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 [**11**], [**28**].

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 [**9**], [**16**] 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 [**25**] 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 [**9**], while Macura [**31**] 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 work 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.

The focus of this analysis is on the dynamics of the *time flow of t-corridors*, which is closely related to the dynamics of the given free group automorphism. Here, t is the generator of the $\mathbb{Z}$ factor in $F \rtimes_\phi \mathbb{Z}$ and a t-corridor is a chain of 2-cells extending across a van Kampen diagram with adjacent 2-cells abutting along an edge labelled t (see Subsection 1.1.4).

The key estimate – a linear bound on the length of t-corridors (Theorem 3.3.1) – admits the following algebraic formulation. This clarifies the manner in which our results concerning the geometry of van Kampen diagrams give rise to a non-deterministic quadratic time algorithm for the word problem in free-by-cyclic groups (for an alternative approach see [**34**]).

Fix a set of generators $\mathcal{A}$ for F and let d_F be the corresponding word metric. We consider words over the $e_i \in (\mathcal{A} \cup \{t\})^{\pm 1}$, where t is a generator of the righthand factor of $F \rtimes_\phi \mathbb{Z}$. A *bracket* β in a word w is a decomposition $w \equiv w_1(w_2)w_3$; the subword w_2 is the *content* of β, and the initial and terminal letters of w_2 are its *sentinels*. A second bracket β', giving $w \equiv w_1'(w_2')w_3'$ is *compatible* with β if $w_2' \subset w_i$ for some $i \in \{1,2,3\}$ or $w_2 \subset w_i'$. A *t-complete* bracketing is a set of pairwise compatible brackets $\beta_1, \ldots, \beta_m$ such that the sentinels of each β_i are $\{t, t^{-1}\}$ and every $t^{\pm 1}$ in w is a sentinel of a unique bracket. In such a bracketing, the content of each bracket is equal in $F \rtimes_\phi \mathbb{Z}$ to an element of F.

Bracketing Theorem. *There exists a constant $K = K(\phi, \mathcal{B})$ such that any word $w \equiv e_1 \ldots e_n$ that represents the identity in $F \rtimes_\phi \mathbb{Z}$ admits a t-complete bracketing $\beta_1, \ldots, \beta_m$ such that the content c_i of each β_i satisfies $d_F(1, c_i) \leq Kn$.*

In order to prove the above theorems one has to delve deeply into the nature of free-group automorphisms. In particular, one needs a precise understanding of how the iterated images $\phi^n(w)$ of an arbitrary element $w \in F$ can evolve. This delicate task is made possible by the existence of informative geometric representatives for ϕ.

We already alluded to the fact that the study of automorphisms of free groups is informed greatly by the analogies with automorphisms of free-abelian groups and surface groups. However, one often has to work considerably harder in the free group case in order to obtain the appropriate analogues of familiar results from these other contexts. Nowhere is this more true than in the quest for suitable normal forms and geometric representatives. One can gain insight into the nature of individual elements of $\mathrm{GL}(n, \mathbb{Z})$ by realizing them as diffeomorphisms of the n-torus. Likewise, one analyzes individual elements of the mapping class group by realizing them as diffeomorphisms of a surface. The situation for $\mathrm{Aut}(F)$ and $\mathrm{Out}(F)$ is more complicated: the natural choices of classifying space $K(F_n, 1)$ are finite graphs of genus n, and no element of infinite order in $\mathrm{Out}(F)$ is induced by the action on $\pi_1(Y)$ of a homeomorphism of Y. Thus the best that one can hope for in this situation is to identify a graph Y_ϕ that admits a *homotopy equivalence* inducing ϕ and has additional structure well-adapted to ϕ. This is the context of the *train track technology* of Bestvina, Feighn and Handel [**7, 4, 6**].

Their work results in a decomposition theory for elements of $\mathrm{Out}(F)$ that is closely analogous to (but more complicated than) the Nielsen-Thurston theory for surface automorphisms [**20**]. The finer features of the topological normal forms that they obtain are adapted to the problems that they wished to solve in each of their papers: the Scott conjecture in [**7**] and the Tits alternative in the series of papers [**4, 6, 5**]. The problem that we solve in this book, that of determining the Dehn functions of all free-by-cyclic groups, requires a further refinement of the train-track technology. Specifically, we must adapt our topological representatives so as to make tractable the problem of determining the isoperimetric properties of the mapping torus of the homotopy equivalence $f : Y_\phi \to Y_\phi$ realizing an iterate of ϕ.

Recall that an automorphism ϕ of a finitely generated free group F is called *positive* if there is a basis $a_1, \ldots, a_n$ for F such that the reduced word representing each $\phi(a_i) \in F$ contains no inverses a_j^{-1}. On the rose (1-vertex graph) with directed edges labelled a_i, one has a natural representative for any automorphism of F. The key feature of positive automorphisms is the fact that the positive iterates of this representative restrict to injections on each edge of the graph. Such maps are the prototypes for train-track representatives.

This discussion suggests a strategy that one might follow in order to prove one Main Theorem: first, one should prove it in the case of positive automorphisms, relying on the simplifications afforded by the positivity hypothesis to confront the web of large-scale cancellation phenomena that must be understood if one is to have any chance of proving the theorem in general. Then, in the general case, one should attempt to follow the architecture of the proof in the positive case, using a suitably refined train-track description of the automorphism in place of the positivity assumption. We shall implement the two stages of this plan in Parts 1 and 3 of this monograph, respectively. Ultimately, this strategy works. However, in Part 3, in order to bring our plan to fruition we have to deal with myriad additional complexities arising from intricate cancellations that do not arise in the positive case.

Roughly speaking, these additional complexities correspond to the fact that most free group automorphisms do not have train track representatives, only *relative* train track representatives. In Part 2 of this monograph, we refine the theory of *improved relative train track maps* due Bestvina, Feighn and Handel [**4**], so as to tease-out features that allow us to adapt the crucial arguments from Part 1. A vital ingredient in this approach is the identification of basic units that will play the role in the general case that single edges (letters) played in the positive case. To this end, we develop a theory of *beads*, whose claim to the role is clinched by the *Beaded Decomposition Theorem 3.2.1.* This theorem is the main objective of Part 2. Indeed we have gone to considerable lengths to distill the entire contribution of Part 2 to Part 3 into this single statement and the important technical refinement of it described in Addendum 2.0.1. We have done so in order that the reader who is willing to accept it as an article of faith may proceed directly from Part 1 to Part 3.

The introduction to each part of the book contains a more detailed explanation of its contents.

Acknowledgements. The first author's work was supported in part by Research Fellowships from the EPSRC of Great Britain and by a Royal Society Wolfson Research Merit Award. Much of this work was undertaken whilst he was a Professor at Imperial College London, from which he was granted two terms of sabbatical leave. The second author was supported in part by a Junior Research Fellowship at Merton College, Oxford, by a Taussky-Todd Instructorship and a Senior Research Fellowship at the California Institute of Technology and by NSF Grant DMS-0504251. We thank these organisations for their support. We also thank the anonymous referee for his careful reading and helpful comments.

Part 1

Positive Automorphisms

An automorphism ϕ of a finitely generated free group F is called *positive* if there is a basis $a_1, \ldots, a_n$ for F such that the reduced word representing each $\phi(a_i) \in F$ contains no inverses a_j^{-1}. Part 1 of this work is dedicated entirely to proving the following special case of the Main Theorem.

THEOREM C. *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.*

This part of the book is organised as follows. In Section 1.1 we recall some basic definitions associated to Dehn functions. In Sections 1.2 and 1.3 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 Theorem C based on the geometry of the *time flow of corridors*. In Section 1.4 we state a sharper version of Theorem C adapted to this strategy and reduce to the study of automorphisms with stability properties that regulate the evolution of corridors. In Section 1.5 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 1.5 and 1.6 reduce us to the nub of the difficulties that one faces in trying to prove Theorem C, 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 1.7. The three main ingredients in this strategy are the elaborate global cancellation arguments in Section 1.8, the machinery of *teams* developed in Section 1.9, and the *bonus scheme* developed in Section 1.10 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 Theorem C. A glossary of constants is included for the reader's convenience.

1.1. Van Kampen Diagrams

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

1.1.1. 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

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

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

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

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 [**26**], detailed proofs of which were given by several authors, states that if a Dehn function is subquadratic, then it is linear — see [**15**, III.H] for a discussion, proof and references.

See [**12**] 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.

1.1.2. Van Kampen diagrams. According to van Kampen's lemma (see [**27**], [**30**] or [**12**]) 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 $\text{Area}(w)$ is the minimum number of 2-cells among all van Kampen diagrams for w. If a van Kampen diagram Δ for w has $\text{Area}(w)$ 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 1.1.1. *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 $f(n+m) \le f(n) + f(m)$ for all $n, m \in \mathbb{N}$. For example, given $r \ge 1$, $k > 0$, the function $n \mapsto kn^r$ is sub-additive.

LEMMA 1.1.2. *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 $\Delta \le f(|\partial\Delta|)$ *for every least-area disc diagram Δ over $\mathcal{P}$, then the Dehn function of $\mathcal{P}$ is $\le f(n)$.*

1.1.3. Presenting $F \rtimes \mathbb{Z}$. We shall establish the quadratic bound required for the Theorem C by examining 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 $a_1, \dots, a_m$ 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

$$\mathcal{P} \cong \langle a_1, \dots, a_m, t \mid t^{-1}a_1tu_1^{-1}, \dots, t^{-1}a_mtu_m^{-1} \rangle. \tag{1.1.1}$$

Throughout Part 1, we shall work exclusively with this presentation.

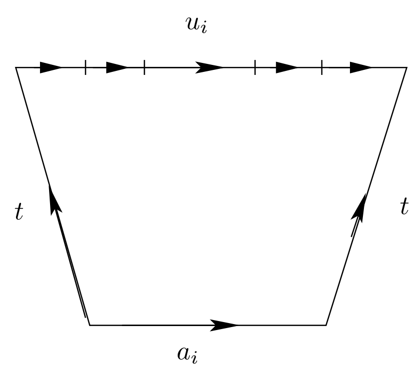

FIGURE 1. A 2-cell in a van Kampen diagram for $F \rtimes_\phi \mathbb{Z}$.

1.1.4. 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 edge 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.)

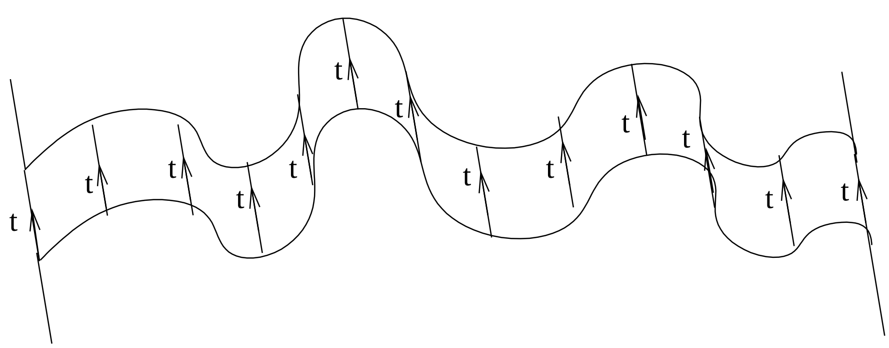

FIGURE 2. A t-corridor

See [**13**] 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 with respect to this notion of time.

Definition 1.1.3 (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.

1.1.5. 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 1.1.2, in order to prove the Theorem C 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 1.1.4. *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. □

1.1.6. 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 [**10**]. 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 1.1.5. *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".

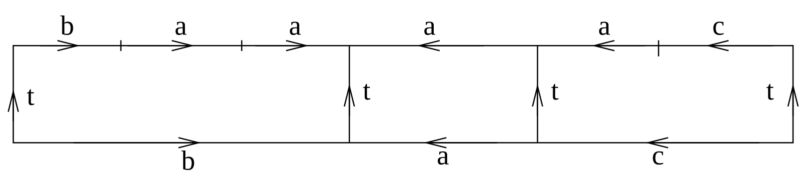

FIGURE 3. An unfolded corridor

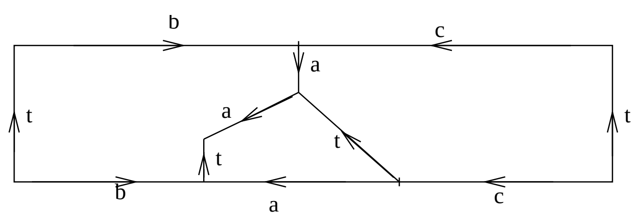

FIGURE 4. The corresponding folded corridor.

1.1.7. 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.

1.2. 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.

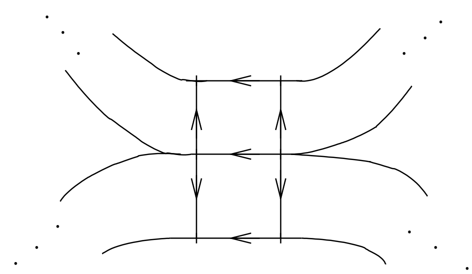

FIGURE 5. Corridors cannot meet this way in a least-area diagram

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

LEMMA 1.2.1. *If $S \neq S'$, then $\bot(S) \cap \bot(S')$ 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 $\bot(S) \cap \bot(S')$ contains no edges. To see that it cannot contain more than one vertex, follow the proof of Proposition 1.2.3(1). $\square$

DEFINITION 1.2.2. 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 L be the maximum of the lengths of the words u_i in our fixed presentation $\mathcal{P}$ of $F \rtimes_\phi \mathbb{Z}$.

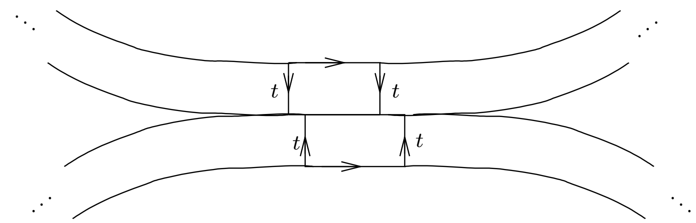

FIGURE 6. A 'singularity'

PROPOSITION 1.2.3 (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 $\mathcal{P}$.*
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 LB.*

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).

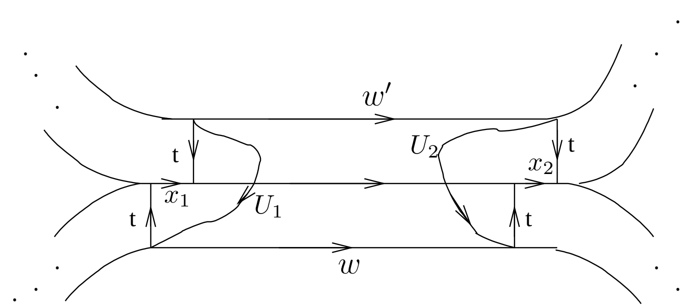

FIGURE 7. The proof of Proposition 1.2.3

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 L. 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 $\perp(S)$ (resp. $\perp(S')$) 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

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

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

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 $|x| \leq L$. (It suffices to take $B_0 = LL_{inv}$, where L_{inv} is the maximum of the lengths of the reduced words $\phi^{-1}(a_i)$.) By definition, $\text{Area}(\Delta_1)+\text{Area}(\Delta_2) \leq 2B_0$, 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 LB.

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 1.2.4 (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 $|I| < B$.*

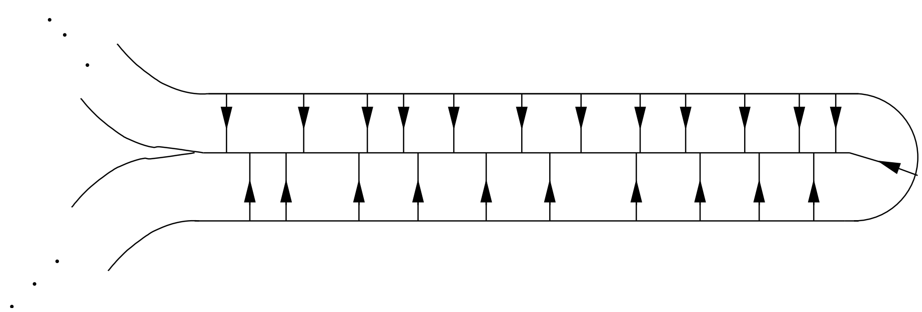

FIGURE 8. Bounded Cancellation Lemma

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

The above lemma is a reformulation of the Bounded Cancellation Lemma from [**21**], which Cooper attributes to Thurston.

REMARK 1.2.5. *'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 1.2.3 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.

1.3. 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 1.3.1 (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 $\varepsilon < \varepsilon''$.

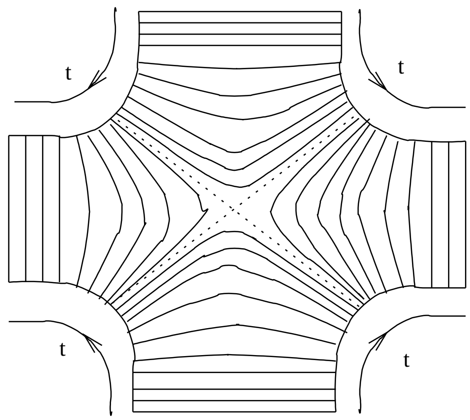

FIGURE 9. Schematic depiction of a singularity

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 1.3.2. 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 1.3.3 (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'.*

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

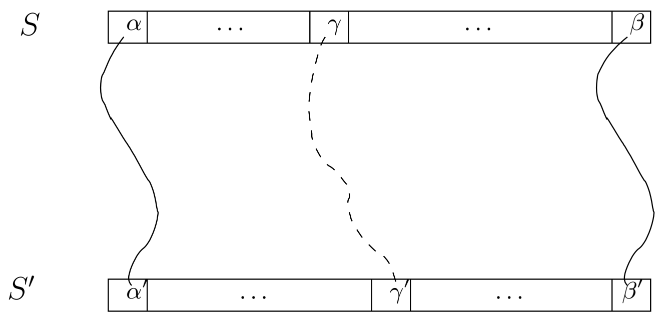

FIGURE 10. The 'loop' picture

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 1.3.2. 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.)

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

1.4. 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, for the remainder of Part 1, 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 Theorem C is an immediate consequence of:

THEOREM 1.4.1. *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 1.4.2. 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.

1.4.1. 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 exponential letters.

EXAMPLE 1.4.3. 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 1.4.4. 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.

[2]Bestvina *et al.* [**4**] use the terminology *non-exponentially-growing* strata

1.4.2. 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 1.4.5. *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.*
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 $k = 1, \ldots, r$ 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 1.4.4.) 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 1.4.6. 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.

1.5. 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 1.5.1 (Preferred futures). For each element $x \in \{a_1, \ldots, a_n\}$ of the basis, 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.

1.5.1. Left-fast, constant letters, etc. In this paragraph, we divide the letters $x \in \{a_1^{\pm 1}, \ldots, a_m^{\pm 1}\}$ into classes according to the growth of the words $\phi^j(x), j = 1, 2, \ldots$, 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 $d \in \{1, \ldots, m-1\}$ 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 1.5.2. 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. [**4**, 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 1.5.3. *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.

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

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 1.8.

Definition 1.5.4 (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 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 $\perp(S)$ 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 1.5.5. *No pair of old edges can 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 $\perp(S)$ 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 $x^{-1}u^{-1}Wux$. □

Corollary 1.5.6. *An edge labelled by a parabolic letter x can only be consumed by an edge labelled y with* $\mathrm{Supp}(x)$ *strictly contained in* $\mathrm{Supp}(y)$.

Remark 1.5.7. A non-constant letter can only be (eventually) consumed from the left (resp. right) by a right-fast (resp. left-fast) letter.

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

REMARK 1.5.8. 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 1.5.9. *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.*

1.6. 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 $\perp(S_0)$.

1.6.1. 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.

1.6.2. 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 1.2.3.

LEMMA 1.6.1. $|S_0(1)| \leq |\partial\Delta|$ and $|S_0(2)| \leq 2B\,|\partial\Delta|$.

LEMMA 1.6.2. $|S_0(3a)| \leq B\,|\partial\Delta|$.

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 1.3.3 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 1.4.1 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 1.3.1.

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 $\mu(S_0)$.

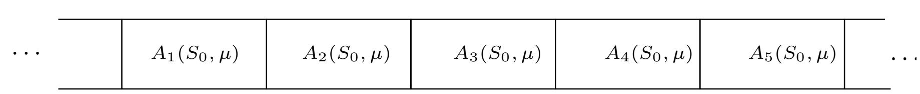

FIGURE 11. The second decomposition of S_0

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 1.6.1 by the equality in the following lemma, which is a tautology. The inequality in this lemma is a restatement of Lemmas 1.6.1 and 1.6.2.

LEMMA 1.6.3.

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

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

LEMMA 1.6.4.

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

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 1.5.9, 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 1.5.3, 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 1.6.5.

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

1.6.4. 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.

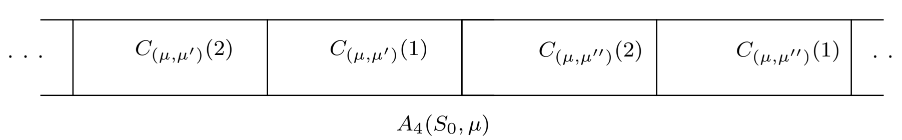

FIGURE 12. $C_{(\mu,\mu')}(1)$ and $C_{(\mu,\mu')}(2)$.

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.

1.6.5. Adjacent Colours. In Corollary 1.3.4 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 $\perp(S)$.

Lemma 1.6.6.

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

Proof. We shall express this proof in the language of the forest $\mathcal{F}$ introduced in Remark 1.3.2. 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. ☐

1.6.6. 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 Theorem C to the study of blocks of constant letters. In this section we achieve the last step of this reduction.

Lemma 1.6.7. *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 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 1.5.6. 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 1.4.5(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).

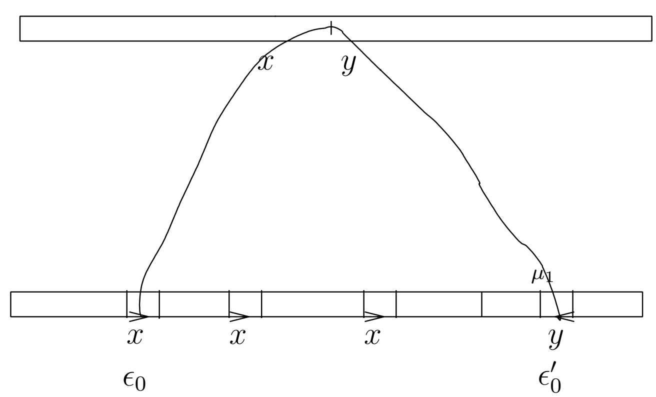

FIGURE 13. The edge labelled ϵ_0' will eventually consume ϵ_0.

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 1.5.4).

[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

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 1.5.4), could only cancel with a new edge (Lemma 1.5.5) in the future of I. And since we have arranged that there be no new edges labelled x, the 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 $B+|\phi(y)|\le 2B$ 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 $|I|\le 4B^2$.

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

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 $|I| \le 2mB^2$. □

COROLLARY 1.6.8.

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

PROOF. This follows immediately from Lemmas 1.6.6 and 1.6.7. □

1.7. 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 Theorem C, 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 Theorem C — see Section 1.11 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 $C_{(\mu,\mu')}(2)$.

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

Our strategy is motivated by the following considerations. Believing Theorem 1.4.1 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 Theorem C will follow. The chosen currencies are apposite because, as we shall see in Section 1.8, 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 sufficient motivation to sustain them through the many technicalities needed to establish the estimates in subsequent sections.

In the following proposition, L 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 1.8.26, and C_1 is the upper bound on the lengths of the intervals $C_{(\mu,\mu')}(1)$ from Lemma 1.6.7, T_0 comes from the Two Colour Lemma 1.8.4 and C_4 comes from Lemma 1.9.4. The constant λ_0 is defined above Definition 1.8.22, and B is the Bounded Cancellation constant from Lemma 1.2.4.

The Constant K_1 is defined to be

$$2C_1+6\lambda_0+2B(5T_0+6T_1+2)+2LC_4(6T_1+8T_0+3)+(B+3)(3T_1+2T_0)L+5L+2.$$

PROPOSITION 1.7.1.

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

1.7.1. 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 1.9 and 1.10. These lemmas are phrased in the language associated to *teams*, the precise definition of which will also be given in Section 1.9. 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 1.9.8). There are less than $2\,|\partial\Delta|$ teams (Lemma 1.9.10).

Each pair (μ,μ') with $C_{(\mu,\mu')}(2)$ non-empty is either a member or a virtual member of a team (Lemma 1.9.10). There are *short* teams (Definition 1.9.6) and long teams, of which some are *distinguished* (Lemma 1.9.29). There are four types of *genesis* of a team, (G1), (G2), (G3) and (G4) (see Subsection 1.9.2). Teams of genesis (G3) have associated to them a pincer $\Pi_{\mathcal{T}}$ (Definition 1.9.12) 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 1.9.29). 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 $\text{down}_2(\mathcal{T})$ are disjoint for different teams, but we shall explain how to account for the amount of 'double-counting' that can occur (see Lemma 1.9.29).

Associated to every team one has the time $t_1(\mathcal{T})$ at which the reaper starts consuming the team (see Subsection 1.9.1). Teams of 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 1.9.13 *et seq.*).

In Section 1.10 we describe a *bonus scheme* that assigns a set of extra edges, $\text{bonus}(\mathcal{T})$ to each team. These bonuses are assigned so as to ensure that $|\text{bonus}(\mathcal{T})| + \|\mathcal{T}\|$ 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 1.7.1.

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 1.6.6 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 1.6.7 that

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

Lemma 1.9.29 accounts for the set of distinguished long teams $\mathrm{D}\Lambda$:

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

For all other teams $\mathcal{T}$ we rely on Lemma 1.10.2 which states

$$\sum_{(\mu,\mu')\in\mathcal{T}\ \text{or}\ (\mu,\mu')\in_v\mathcal{T}} |C_{(\mu,\mu')}(2)| \le \,\|\,\mathcal{T}\,\|\, + |\mathrm{bonus}(\mathcal{T})| + B. \tag{1.7.1}$$

We next consider the *genesis* of teams. All teams of genesis (G4) are short (Lemma 1.9.7). And by Definition 1.9.6 for the short teams $\mathcal{T}\in\Sigma$ we have

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

Lemma 1.9.20 tells us that for teams of genesis (G1) and (G2) we have

$$\|\,\mathcal{T}\,\|\, \le 2LC_4|\mathrm{down}_1(\mathcal{T})| + |\partial^{\mathcal{T}}|,$$

whilst for teams of genesis (G3) we have

$$\|\,\mathcal{T}\,\|\, \le 2LC_4\big(|\mathrm{down}_1(\mathcal{T})| + |Q(\mathcal{T})|\big) + T_0\big(|\chi_P(\mathcal{T})|+1\big) + |\partial^{\mathcal{T}}| + \lambda_0.$$

Let $\mathcal{G}_3$ denote the set of teams of genesis (G3) with $Q(\mathcal{T})$ non-empty. In Definition 1.9.25 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 1.9.24 we prove that

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

and in Corollary 1.9.31 we prove that

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

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

Section 10 is dedicated to the proof of Proposition 1.10.13, which states

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

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

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

where K_1 is

$$2C_1+6\lambda_0+2B(5T_0+6T_1+2)+2LC_4(6T_1+8T_0+3)+(B+3)(3T_1+2T_0)L+5L+2.$$

Thus the proposition is proved. □

REMARK 1.7.2. 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.

By a precisely analogous argument, we also have:

PROPOSITION 1.7.3.

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

where K_1 is the constant defined prior to Proposition 1.7.1.

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 1.8.26).

1.8.1. The Buffer Lemma.

LEMMA 1.8.1. *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 1.4.5) 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 1.8.2. *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.*

1.8.2. The Two Colour Lemma.

DEFINITION 1.8.3. 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 $\mu(S), \mu'(S) \subseteq \perp(S)$ that are adjacent in corridors of van Kampen diagrams, and the following lemma remains valid in that context.

PROPOSITION 1.8.4 (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 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 1.5.3 tells us that the length of the first subword is less than C_0, and the proof of Lemma 1.6.7 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 1.5.3 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 1.8.5. *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 ${T_0}'$.*

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

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 1.4.5), the length of this segment grows at a constant rate ($< L$) 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 LT_0 at time N, and hence is cancelled entirely before time $T_0(L+1)$. □

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.

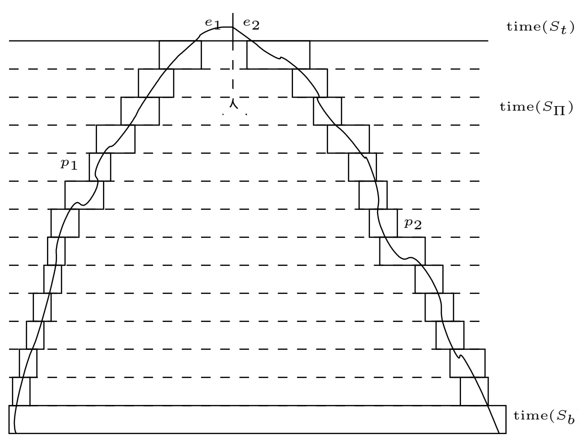

FIGURE 14. A pincer.

DEFINITION 1.8.6. 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

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

PROPOSITION 1.8.7 (Unnested Pincer Lemma). *There exists a constant $\hat{T}_1$, depending only on ϕ, such that for any pincer Π*

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

Fix a pincer Π and assume $\mathrm{Life}(\Pi) \neq 0$. The idea of the proof of Proposition 1.8.7 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 $\mathrm{time}(S_b) + \hat{T}_1$, 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 $\mathrm{time}(S_\Pi)$.

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

DEFINITION 1.8.8. 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 $S = S^1, S^2, \ldots, S^p$ 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 $j = 1, \ldots, r-1$ are consumed entirely), *or else* all of the non-constant letters in $\nu_j(S^p)$, for $j = 1, \ldots, r-1$, 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 $\nu_0(S^p) \ldots \nu_r(S^p) \subset \bot(S^p)$ (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 $\nu_0(S^p) \ldots \nu_r(S^p)$. 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 1.8.9. Let $[\nu_0(S), \ldots, \nu_r(S)]$ be a p-implosive array.

(1) Any implosive subarray of $[\nu_0(S), \ldots, \nu_r(S)]$ 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 $[\nu_i(S), \ldots, \nu_j(S)]$ 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 1.8.10. *The ordered list of colours along each corridor before* time(S_Π) *in a pincer* Π *must contain an implosive array.*

PROOF. At the top of the pincer there is cancellation between non-constant edges. Lemma 1.8.1 tells us that before time(S_Π) 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. Furthermore, the array $[\nu_0(S_{t'}), \ldots, \mu(S_{t'}), \ldots, \nu_r(S_{t'})]$ 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 $p \leq \hat{T}_1$, 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 1.8.7 is an immediate consequence of the following result, which will be proved in (1.8.18).

PROPOSITION 1.8.11 (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 $p \leq \hat{T}_1$.*

The first restriction to note concerning implosive arrays is this:

LEMMA 1.8.12. *If* $[\nu_0(S), \dots, \nu_r(S)]$ *is implosive of Type I, then* $r \leq B$. *If it is implosive of Type II, then* $r < 2B$.

PROOF. In Type I arrays, the interval $\nu_1(S^p) \dots \nu_{r-1}(S^p) \subset \perp(S^p)$ 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 1.8.13. In the light of Lemma 1.8.12, an obvious finiteness argument would provide the bound required for Lemma 1.8.11 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 $[\nu_0(S_b), \dots, \nu_r(S_b)]$.

Previously (Subsection 1.6.3) we partitioned each colour $\nu_j(S_b)$ into five intervals $A_1(S_b, \nu_j), \dots, A_5(S_b, \nu_j)$ 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].

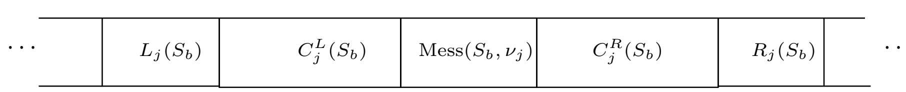

FIGURE 15. The decomposition of the colour ν_j

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

[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.

its right end up to and including the edge one to the left of $\mathrm{Mess}(S_b, \nu_0)$. 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 $\mathrm{Mess}(S_b, \nu_r)$. 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 1.8.14. The length of $A(S) = [\nu_0(S), \ldots, \nu_r(S)]$, written $\| A(S) \|$, is the number of edges in the interval $\mathcal{L}(\nu_0, S) \ldots \mathcal{R}(\nu_r, S) \subset \bot(S)$. (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 $j = 1, \ldots, r-1$ essentially vanishes in S^p.

By definition, no preferred future of any edge in $\mathrm{Mess}(\nu_j, S_b)$ is cancelled before S^p. Hence these intervals do not shrink in length before that time, and as in the proof of Lemma 1.8.12 we can use the Bounded Cancellation Lemma to bound the sum of their lengths:

LEMMA 1.8.15. *After excluding the edges of the enduring block, the sum of the lengths of the intervals* $\mathrm{Mess}(\nu_j, S_b)$ *is at most* $2B$.

Combining this estimate with the bounds from Lemmas 1.5.3 and 1.6.7, we deduce that for $j = 1, \ldots, r-1$

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

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

(1) *If* $[\nu_0(S^{T_0}), \dots, \nu_r(S^{T_0})]$ *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 1.2.4. The (crude) bound in (2) is an immediate consequence of Lemma 1.8.15 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 1.8.7.

Lemma 1.8.17. *If* $A(S^0) = [\nu_0(S^0), \dots, \nu_r(S^0)]$ *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 the array* $A(S^{T_0})$ *contains a very short implosive subarray* $[\nu_k(S^{T_0}), \dots, \nu_l(S^{T_0})]$.

Proof. Assume $p > 2T_0$ and that $[\nu_0(S^0), \dots, \nu_r(S^0)]$ is not short. We claim that there is a block of at least $B + 1$ constant letters in the interval determined by the array $\mathcal{L}(\nu_0, S^{T_0}) \dots \mathcal{L}(\nu_r, S^{T_0})$. 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 1.8.1 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 $k \leq i - 2$.

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 1.8.4. 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 $p \leq 2T_0$, contrary to our hypothesis.

Thus we have proved that the edge which ultimately consumes e_l is coloured ν_k where $k \leq i-2$. 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 $\nu_k \searrow \nu_i$.

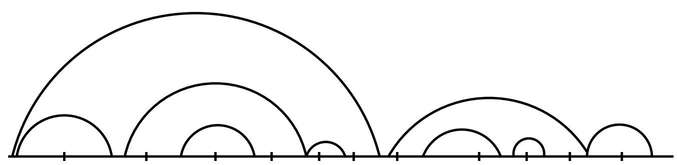

FIGURE 16. The nesting associated to $\searrow$

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. □

1.8.18. *Proof of Regular Implosions (Prop.1.8.11):* Given the bound in Lemma 1.8.16(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 1.8.9(1)). So in the light of Lemmas 1.8.17 and 1.8.16(1), it suffices to let $\hat{T}_1 = \max\{2T_0, \tau\}$. □

1.8.4. Super-Buffers. In this subsection we prove an important cancellation lemma based on Proposition 1.8.7, this lemma involves the following constant.

Definition 1.8.19. We fix an integer T_1' such that one gets repetitions in all T_1'-long subsequences of 5-tuples of reduced words

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

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 $|u_3^k| \le 4B + 1$. That is, for some $t_1 \le t_2 \le T_1'$ 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 1.8.20. Assume $T_1' \ge \hat{T}_1$.

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 $V = V_1V_2\cdots V_k$, where each V_i is either positive or negative; we think of these subwords as having colours $\mu_1, \dots \mu_k$. 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 $t^{-k}Vt^k\phi^k(V)^{-1}$. Thus the following lemma is a comment on the form of such diagrams.

Proposition 1.8.21. *Let $V = V_1V_2V_3$ be a concatenation 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 $k \ge 0$ 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 1.8.7), 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 $1 \le i \le T_1'$, 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 1.5.3 and 1.6.7 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 1.5.3, 1.6.7 tell is that $|R^i|, |L^i| \le C_0 + C_1 + 2B + 1$, 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 1.8.19: 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. □

1.8.5. 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 1.8.7), 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 1.8.17.

Definition 1.8.22. 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 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.

Remark 1.8.23. A nested pincer cannot be both left-loaded and right-loaded (cf. Remark 1.8.9(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 1.8.24. 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 1.8.25. *If the pincer* Π_1 *is nested in* Π_0 *then* $\mathrm{time}(S_t(\Pi_1)) < \mathrm{time}(S_{\Pi_0})$.

[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 $\top(S)$.

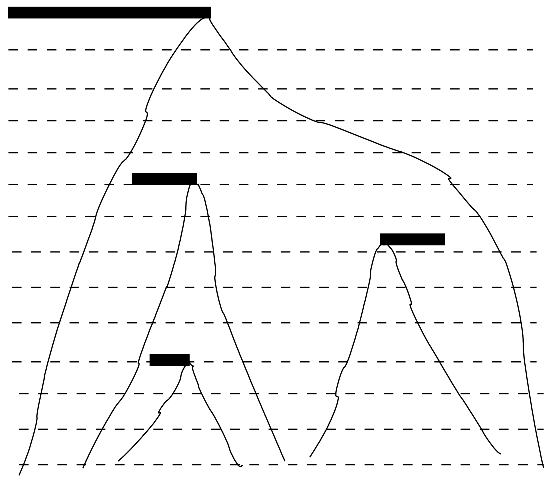

FIGURE 17. A depiction of nesting

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 1.8.1. □

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

THEOREM 1.8.26 (Pincer Lemma). *For any pincer* Π

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

PROOF. The heart of our proof of Proposition 1.8.7 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 $\{\Pi_j\}$, $j = 1, \ldots, J$ be the set of such, indexed in order of appearance from left to right.

For $j = 1, \ldots, J-1$, 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 $p_2(\Pi)$.

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

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 $S^0, S^1, \ldots$

We focus on a single Σ_j, and write its colours in order as $\nu_1, \ldots, \nu_r$. We analyse how the colours in Σ_j come to vanish. The first important observation is that $2 \leq i \leq r-1$, 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 1.8.7, 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 1.8.21. 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 Lemma applies. We would like to apply Proposition 1.8.21 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 1.8.17). 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. □

1.9. Teams and their Associates

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

1.9.1. 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 1.2.4). 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 1.4.5). 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 μ and μ' (as defined in 1.3.2). 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 1.9.1. 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 $\| \hat{\mathcal{T}} \|$.

In a little while we shall define *teams* to be pre-teams satisfying a certain maximality condition (see Definition 1.9.6).

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

REMARK 1.9.2. 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 1.7.1 we saw that it would be desirable if (whatever our final definition of *team* and bonus may be) the following inequality (1.7.1) should hold for all teams:

$$\sum_{(\mu,\mu')\in\mathcal{T} \text{ or } (\mu,\mu')\in_v\mathcal{T}} |C_{(\mu,\mu')}(2)| \leq \parallel \mathcal{T} \parallel + |\text{bonus}(\mathcal{T})| + B. \tag{1.9.1}$$

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 1.9.3. *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 $\hat{\mathfrak{T}}(\mu, \mu')$. □

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 $\text{time}(S_0)$ 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 (1.9.1) in this setting that we are obliged to instigate the bonus scheme described in Section 1.10.

1.9.2. 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 $\hat{\mathfrak{T}}(\mu, \mu')$.

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).

(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).

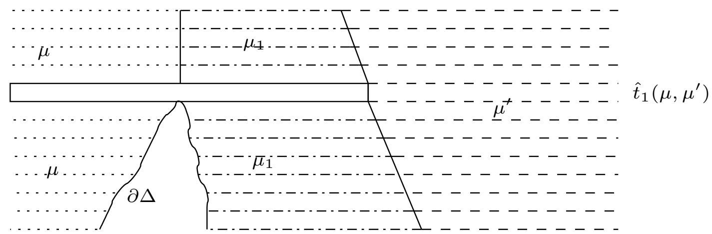

FIGURE 18. A team of genesis (G1)

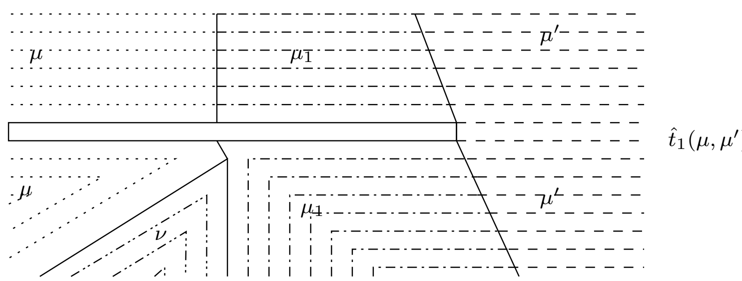

FIGURE 19. A team of genesis (G2)

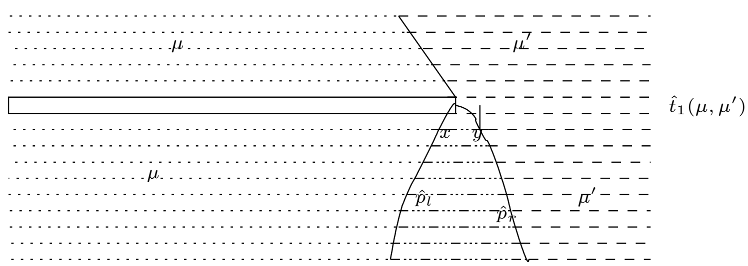

FIGURE 20. A team of genesis (G3)

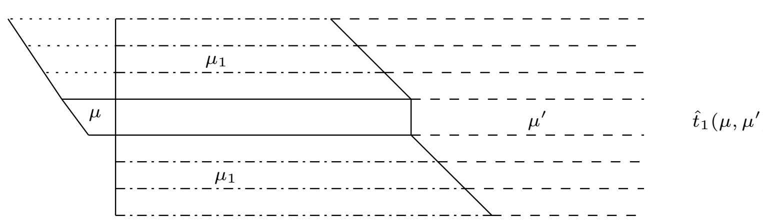

FIGURE 21. A team of genesis (G4)

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

LEMMA 1.9.4. *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 $\perp(S)$ 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, $|I| \leq |\alpha| + 2LC_4$.

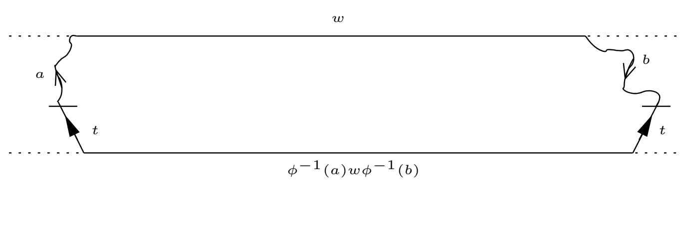

FIGURE 22. The proof of Lemma 1.9.4

PROOF. See Figure 22. Follow the path from the left end of I to $\perp(S)$. This passes through a (possibly empty) path a^{-1}, followed by an edge labelled t^{-1}, where the length of a is less than L (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 L. The path along $\perp(S)$ 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 $L.L_{inv}$. 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. □

REMARK 1.9.5. It is convenient to assume that $LC_4 < \lambda_0$. (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 1.9.6. 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 1.9.7. *Teams of genesis (G4) are short.*

PROOF. Lemma 1.9.4 implies that $\mathfrak{T}$ is in the immediate future of an interval of length at most C_4. And we have decreed (Remark 1.9.5) that $LC_4 < \lambda_0$. ☐

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 1.9.1) we excluded pairs (μ,μ') with $|C_{(\mu,\mu')}(2)| \leq 2B$. The following definitions remove these difficulties.

DEFINITION 1.9.8 (Virtual team members). If a pre-team $\hat{\mathcal{T}}(\mu,\mu')$ of type (G4) is disqualified under the terms of Definition 1.9.6 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 1.9.9. If (μ,μ') is such that $1 \leq |C_{(\mu,\mu')}(2)| \leq 2B$ 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 1.9.10. *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 $|\mathcal{Z}| < 2\,|\partial\Delta|$. □

1.9.3. 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 1.9.11. 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 1.9.12 (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 1.9.13 (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 1.9.14. 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 1.9.13 but do not become $p_l(\mathcal{T})$.

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

Lemma 1.9.15.

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

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

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

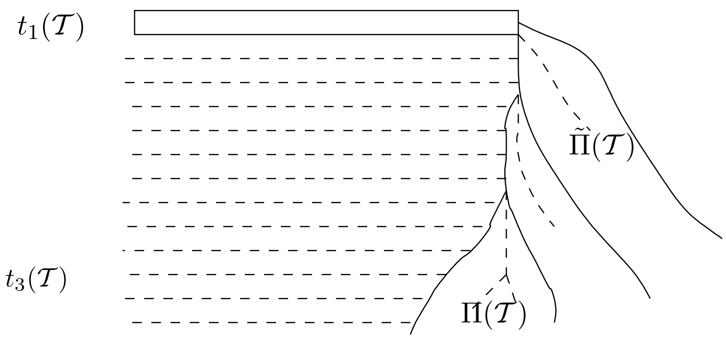

FIGURE 23. The cascade of pincers.

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

1.9.4. The length of teams.

DEFINITION 1.9.16. Define $\mathrm{down}_1(\mathcal{T}) \subset \partial\Delta$ 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 $\mathrm{down}_1(\mathcal{T})$.

DEFINITION 1.9.17. 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 1.9.18.

(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 $\mathcal{T}_2$, $\mathrm{down}_1(\mathcal{T}_1)$ and $\mathrm{down}_1(\mathcal{T}_2)$ are disjoint.

DEFINITION 1.9.19. 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 1.9.20.

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

$$\| \mathcal{T} \| \leq 2LC_4 \, |\mathrm{down}_1(\mathcal{T})| + |\partial^{\mathcal{T}}|.$$

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

$$\| \mathcal{T} \| \leq 2LC_4 \, |\mathrm{down}_1(\mathcal{T})| + |\partial^{\mathcal{T}}| + 2LC_4 \, |Q(\mathcal{T})| + 2LC_4 T_0 \big(|\chi_P(\mathcal{T})| + 1 \big) + \lambda_0.$$

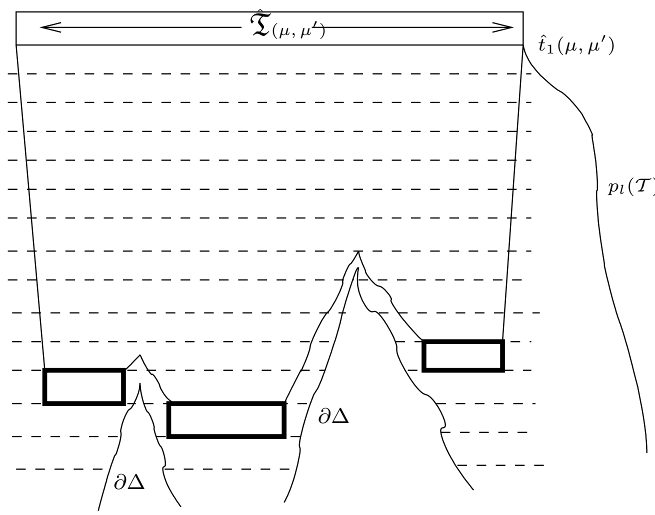

FIGURE 24. Bounding the size of a team in terms of $|\mathrm{down}_1|$ and $|p_l|$

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 1.9.4, 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 $2LC_4k_t$, 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 LC_4 constant letters may be added. This follows from Lemma 1.9.4. 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 $2LC_4$) 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 $2LC_4k_{t+1}$ 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 $2LC_4k_{t+1}$. 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 $\text{down}_1(\mathcal{T})$ 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 $2LC_4$ when moving forward one unit of time (unless added to by $\partial^{\mathcal{T}}$).

The summands $2LC_4\,|Q(\mathcal{T})|$ and $2LC_4T_0\big(|\chi_P(\mathcal{T})|+1\big)$ 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. □

1.9.5. 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 1.9.13 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 1.8.24.

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

Lemma 1.9.21. *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 1.9.19). 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 1.9.22. *For all $\mathcal{T}\in\mathcal{G}_3$,*

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

Lemma 1.9.23. *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 1.9.21 means that $\Pi_{\mathcal{T}_1}$ is actually nested in $\mathcal{T}_2$ in the sense of Definition 1.8.24. Thus $\chi(\Pi_{\mathcal{T}_1})\cap\chi(\Pi_{\mathcal{T}_2})=\emptyset$ (by Definition 1.8.24). □

COROLLARY 1.9.24. $\sum_{\mathcal{T}\in\mathcal{G}_3} t_3(\mathcal{T}) - t_2(\mathcal{T}) \leq 3T_1\,|\partial\Delta|$.

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 1.9.25. For $\mathcal{T} \in \mathcal{G}_3$ we define $\text{down}_2(\mathcal{T})$ 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

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

(see Corollary 1.9.31).

At this stage our task of bounding $\|\,\mathcal{T}\,\|$ would be complete if the the sets $\text{down}_2(\mathcal{T})$ 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 $\text{down}_2(\mathcal{T}) \neq \emptyset$.

We shall need the following obvious separation property.

LEMMA 1.9.26. *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 $\text{down}_2(\mathcal{T}) \neq \emptyset$ 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 1.9.27. $\mathcal{T}_1 \in \mathcal{G}_3'$ 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 1.9.28. *If there is a team from $\mathcal{G}_3'$ below $\mathcal{T} \in \mathcal{G}_3'$, then $t_1(\mathcal{T}) \geq \mathrm{time}(S_0) \geq t_2(\mathcal{T})$.*

PROOF. The first thing to note is that if $\mathrm{time}(S_0)$ 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 $\mathrm{time}(S_0)$ 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 $\mathrm{time}(S_0) < t_2(\mathcal{T})$ then we are in the non-degenerate situation of Definition 1.9.13 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 1.9.26). 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 $\mathrm{time}(S_0)$, they must also do so at all times up to $t_1(\mathcal{T})$. This contradiction implies that in fact $\mathrm{time}(S_0) \geq t_2(\mathcal{T})$.

Consider Figure 17. Suppose that $\mathcal{T}' \in \mathcal{G}_3'$ is below $\mathcal{T}$. The proof of Lemma 1.9.21 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 $t_1(\mathcal{T}) \geq \mathrm{time}(S_0)$. $\square$

The following is the main result of this section.

LEMMA 1.9.29. *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.

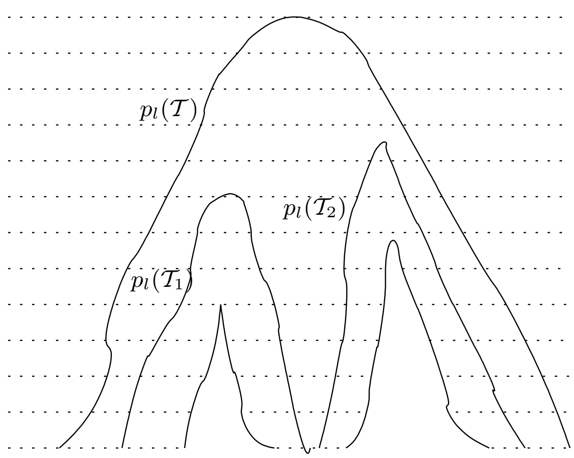

FIGURE 25. A depiction of double-counting

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 1.9.26). 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 1.9.26. 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 $\mathrm{time}(S_0) > t_1(\mathcal{T}_1)$, so Lemma 1.9.28 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 1.8.26) and Proposition 1.8.7 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 1.9.13 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 1.9.28) 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 $t_1(\mathcal{T}_1)$ 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 1.9.28 we have $t_2(\mathcal{T}_0) \leq \mathrm{time}(S_0) \leq t_1(\mathcal{T}_1)$. By combining these inequalities with Lemmas 1.9.22 and

[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 $\Pi_{\mathcal{T}_1}$.

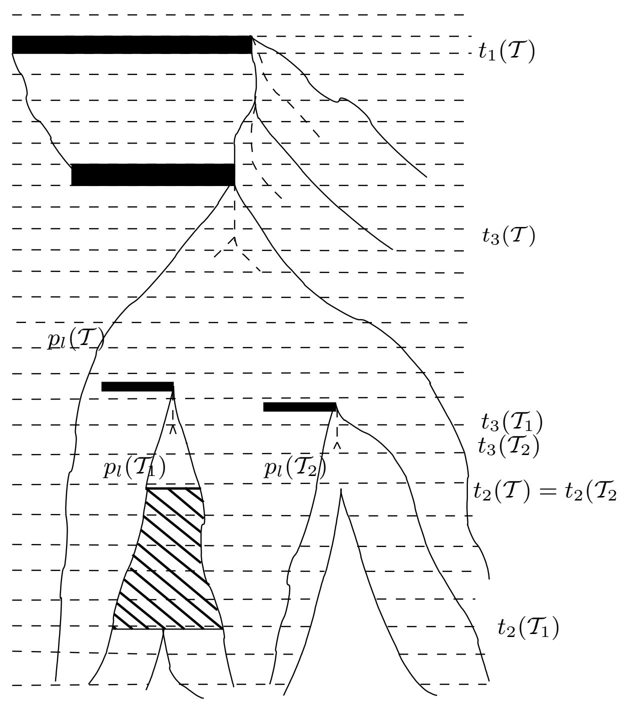

FIGURE 26. Finding the colours $\chi_c(\mathcal{T}_i)$

1.9.15 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 1.9.30. *Summing over the set of teams $\mathcal{T} \in \mathcal{G}_3'$ that are not distinguished, we get*

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

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 1.9.29, we have

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

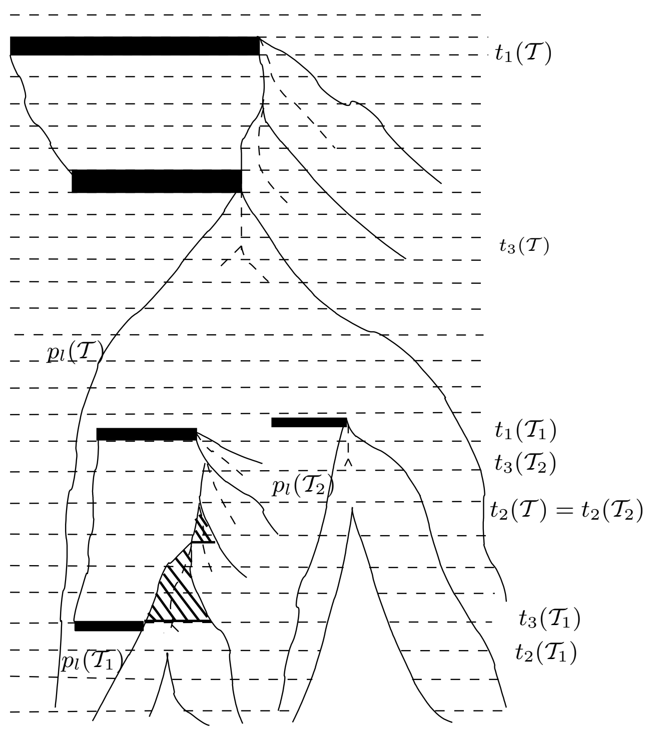

FIGURE 27. Finding the colours $\chi_\delta(\mathcal{T}_i)$

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 1.9.26. Therefore, the set of edges that $\mathcal{T}$ double-counts with any team of lesser depth is exactly $\text{down}_2(\mathcal{T}) \cap \text{down}_2(\mathcal{T}_0)$.

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. $\square$

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

COROLLARY 1.9.31.

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

PROOF. The sets of colours $\chi_c(\mathcal{T})$ and $\chi_\delta(\mathcal{T})$ are disjoint. And the union of the sets $\text{down}_2(\mathcal{T})$ 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 1.9.10). $\square$

1.10. 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 $t_1(\mathcal{T}) \geq \text{time}(S_0)$, 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 1.9.3. We indicated following this lemma how this inequality might fail in the case where $t_1(\mathcal{T}) < \text{time}(S_0)$. 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 $t_1(\mathcal{T}) < \text{time}(S_0)$.

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

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

Lemma 1.10.2. *For any team* $\mathcal{T}$,

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

Proof. If $t_1(\mathcal{T}) \geq \text{time}(S_0)$, this follows immediately from Lemma 1.9.3. If $t_1(\mathcal{T}) < \text{time}(S_0)$ then at each step in time between $t_1(\mathcal{T})$ and 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 1.9.6.)

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

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

LEMMA 1.10.3. *The sum of the lengths of the bonuses provided to all teams by terrors is less than* $2L\,|\partial\Delta|$.

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 L. There is at most one terror for each possible adjacency of colours and hence the total contributions of all terrors is less than $2L\,|\partial\Delta|$. □

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).

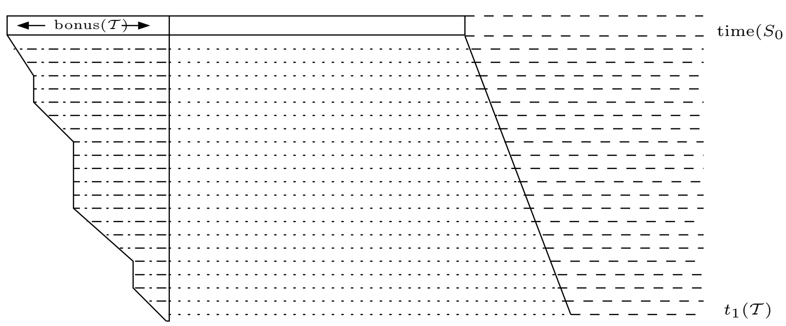

FIGURE 28. The generic situation below time(S_0).

DEFINITION 1.10.4 (Rascals' Pincers). We fix a team $\mathcal{T}$ with $t_1(\mathcal{T}) <$ 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 1.10.5. *The total of all bonuses provided to all teams by rascals e with* $\tau_1(e) \leq \text{time}(S_0)$ *is less than* $(3T_1 + 2T_0 + 1)L\,|\partial\Delta|$.

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 $\tau_1(e) - T_0 \geq \text{time}(S_{\Pi_e})$ (in the degenerate case discussed in the footnote, $\text{time}(S_{\Pi_e})$ 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 $\tau_1(e) \leq \text{time}(S_0)$, 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 $\tau_1(e) \leq \text{time}(S_0)$ 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 bounded by the number of possible adjacencies of colours, the remaining terms contribute at most $(T_1 + T_0)2\,|\partial\Delta|$. Thus

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

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

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

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 1.10.5 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 1.10.6. 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 1.10.7. 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 1.10.5.

Lemma 1.10.8. *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 1.10.9. *The sum of the bonuses provided to all teams by left-biased rascals is less than $(2L + 6LT_1 + 4LT_0 + 2\lambda_0 + 6BT_1 + 4BT_0)\,|\partial\Delta|$.*

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 time(S_{Π_e}) (cf. Definition 1.9.13). Therefore the following is an immediate consequence of the Pincer Lemma.

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

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 bonus($\mathcal{T}$). Then the bonus provided by e is at most $L(\tau_1(e) - \tau_0'(e)) + \lambda_0$. 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 1.9.29 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 1.9.29), and proceeding as in the proof of Lemma 1.9.29 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

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

Therefore, summing over the set of rascals which are not distinguished we get (cf Corollary 1.9.30)

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

And summing over the same set of rascals, we get

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

Therefore, for undistinguished rascals, we have

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

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

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

as required. □

[16] One extends the paths p_e and p_e' of Definition 1.10.4 back in time to $\partial\Delta$ so as to define the order defining depth.

LEMMA 1.10.10. *The sum* $\sum \text{life}(\mathfrak{B})$ *over those* $\mathfrak{B}$ *that are right-biased but do not satisfy conditions (i) or (ii) of Lemma 1.10.8 is at most* $(3T_1B + 2T_0B)\,|\partial\Delta|$.

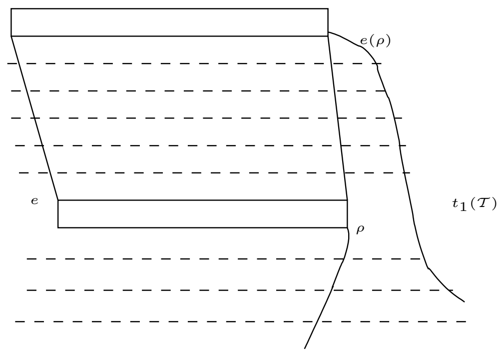

FIGURE 29. A depiction of a right-biased team.

PROOF. Once again, as in Lemmas 1.10.5 and 1.10.9, 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

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

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 1.10.8 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

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

as required. □

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

COROLLARY 1.10.11. *The sum of the effective volumes of all blocks that are right-biased but do not satisfy conditions (i) and (ii) of Lemma 1.10.8 is at most* $(3LT_1B + 2LT_0B)\,|\partial\Delta|$.

LEMMA 1.10.12. *The sum of all blocks that satisfy condition (iv) of Lemma 1.10.8 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).

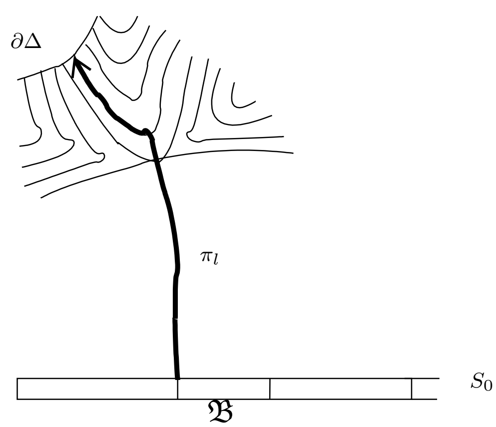

FIGURE 30. The path π_l hits a singularity.

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.

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

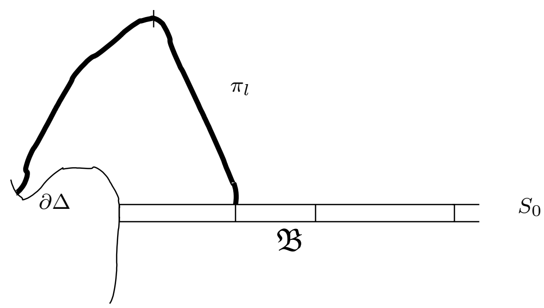

FIGURE 31. The path π_l in cancelled from outside of the future of S_0.

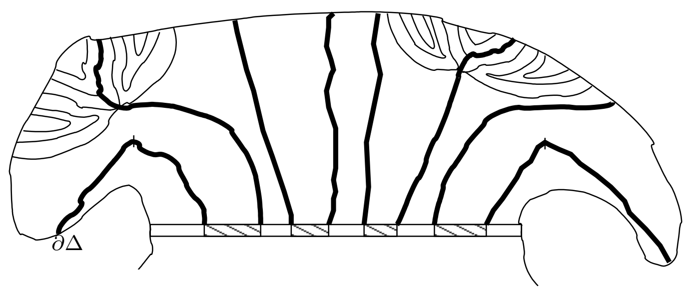

FIGURE 32. The schematic partition of Δ by the paths π_l and π_r.

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 1.2.4). 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:

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

as required. □

Summarising the results of this section we have

LEMMA 1.10.13. *Summing over all teams that are not short, we have*

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

1.11. The Proof of Theorem C

Pulling all of the previous results together, define

$$K_1 = 2C_1+6\lambda_0+2B(5T_0+6T_1+2)+2LC_4(6T_1+8T_0+3)+(B+3)(3T_1+2T_0)L+5L+2,$$

and

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

THEOREM 1.11.1. *$|S_0| \leq K\,|\partial\Delta|$.*

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 1.6.4, Proposition 1.7.1, Lemma 1.6.3, Proposition 1.7.3 and Lemma 1.6.4, respectively,

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

Summing completes the proof of Theorem 1.11.1. □

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

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

which proves the Main Theorem for positive automorphisms, i.e. Theorem C.

1.12. Glossary of Constants

B – the Bounded Cancellation constant (Lemmas 1.2.4 and 1.2.3).

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 1.6.4.

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 1.6.7. Note that one can take $C_1 = 2mB^2$.

L – the maximum of the lengths of the images $\phi(a_i)$ of the basis elements a_i, i.e. the maximum length of $u_1,\ldots,u_m$ in the presentation $\mathcal{P}$ (see equation 1.1.1).

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

T_0 – the constant from the 2-Colour Lemma (Lemma 1.8.4). 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 1.8.7.

T_1' – the constant from Definition 1.8.19. 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 1.8.26.

$C_4 := LL_{inv}$

$\lambda_0 := \max\{2B(T_0+1)+1, LC_4\}$

Finally, K_1 is defined to be

$$2C_1+6\lambda_0+2B(5T_0+6T_1+2)+2LC_4(6T_1+8T_0+3)+(B+3)(3T_1+2T_0)L+5L+2,$$

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

Part 2

Train Tracks and the Beaded Decomposition

Part 2 of this work is dedicated to the construction and analysis of a refined topological representative for a suitable iterate of an arbitrary automorphism of a finitely generated free group. In Part 3 we shall use these representatives to extend the results obtained in Part 1 to the general setting. Our results rely in a fundamental way on the theory of *improved relative train tracks* developed by Bestvina, Feighn and Handel in [**4**].

The properties of the topological representative $f : G \to G$ constructed in [**4**] allow one to control the manner in which a path σ evolves as one looks at its iterated images under f, and one might naively suppose that this is the key issue that one must overcome in translating the proof of our Main Theorem from the positive case (Part 1) to the general case (Part 3). However, upon closer inspection one discovers this is actually only a fraction of the story because when a corridor evolves in the time flow on a van Kampen diagram, the interaction of the forward iterates of the individual edges is such that the basic *splitting* of paths established in [**4**] may get broken. It is to overcome this difficulty that we need the notion of *hard splitting*.

Definition (See Definition 2.2.1). *We say that a decomposition of an edge-path into sub edge-paths $\rho = \rho_1\rho_2$ is a* hard k-splitting *if for* any *choice of tightening of $f^k(\rho) = f^k(\rho_1)f^k(\rho_2)$ there is no cancellation between the image of ρ_1 and the image of ρ_2.*

A decomposition that is a hard k-splitting for all $k \geq 1$ is called a hard splitting. *If $\rho_1 \cdot \rho_2$ is a hard splitting, we write $\rho_1 \odot \rho_2$.*

In the analysis of van Kampen diagrams that forms the core of the proof of the Main Theorem, the class of “broken” paths that one must understand are the residues of the images of a single edge that survive repeated cancellation during the corridor flow. In the language of the topological representative $f : G \to G$, this amounts to understanding *monochromatic paths*, as defined below. Every edge-path ρ in G admits a unique maximal splitting into edge-paths (Lemma 2.2.6); our main task here in Part 2 is to understand the nature of the factors in this splitting and the behaviour of certain larger units into which they naturally accrete when ρ is monochromatic.

To this end, we identify a small number of basic units into which the iterated images of monochromatic paths split; the key feature of this splitting is that it is robust enough to withstand the difficulties caused by cancellation in van Kampen diagrams. The basic units are defined so as to ensure that they enjoy those features of individual edges that proved important in the positive case (see Part 1). We call the units *beads*. The vocabulary of beads is as follows.

Let $f : G \to G$ be a topological representative and let $f_\#(\sigma)$ denote the tightening rel endpoints of the image of an edge-path σ. Following [**7**], if $f_\#(\tau) = \tau$ we call τ a *Nielsen* path. A path ρ in G is called a *growing exceptional path* (GEP) if either ρ or $\bar{\rho}$ is of the form $E_i\bar{\tau}^k\bar{E}_j$ where τ is a

Nielsen path, $k \geq 1$, E_i and E_j are parabolic edges, $f(E_i) = E_i \odot \tau^m$, $f(E_j) = E_j \odot \tau^n$, and $n > m > 0$. If it is ρ (resp. $\overline{\rho}$) that is of this form, then proper initial (resp. terminal) sub edge-paths of ρ are called ΨEPs (*pseudo-exceptional paths*). .

GEPs and ΨEPs are key objects of study for us in Parts 2 and 2. They admit no nontrivial hard splitting, but there is no global bound on their length. Therefore, they must be included as basic units in the Beaded Decomposition Theorem below. Also, there is no uniform bound on the number of iterates required to cancel a GEP or ΨEP when it occurs as a sub-path of the label on a corridor. This leads to considerable technical difficulties in Part 3.

Let $f : G \to G$ be an improved relative train track map and $d, J \geq 1$ integers. Then *d-monochromatic* paths in G are defined by a simple recursion: edges in G are d-monochromatic and if ρ is a d-monochromatic path then every sub edge-path of $f^d_\#(\rho)$ is d-monochromatic.[17] A (J, f)-*atom* is a d-monochromatic edge-path of length at most J that admits no non-vacuous hard splitting into edge-paths.

An edge-path ρ is (J, f)-*beaded* if it admits a hard splitting $\rho = \rho_1 \odot \cdots \odot \rho_k$ where each ρ_i is a GEP, a ΨEP, a (J, f)-atom, or an indivisible Nielsen path of length at most J (where GEPs, ΨEPs and Nielsen paths are defined with respect to the map f).

The following is the most important output of Part 2.

Beaded Decomposition Theorem. For every $\phi \in \mathrm{Out}(F_n)$, there exist positive integers k, d and J such that ϕ^k has an improved relative train-track representative $f : G \to G$ with the property that every d-monochromatic path in G is (J, f)-beaded.

In fact, we do not prove the Beaded Decomposition Theorem *per se*. Instead, we prove a more general statement about futures of arbitrary paths under repeated iteration and cancellation (Theorem 2.3.5). We also need the following:

Addendum 2.0.1. *If f is replaced by an iterate $f_1 = f^l_\#$, then the Beaded Decomposition Theorem is true for f_1 with the same constant as for f.*

This sharpening of the Beaded Decomposition Theorem will prove vital in Part 3: often, we will need to replace f by an iterate, but the iterate we choose will depend on J, so Addendum 2.0.1 is needed to avoid circularity. Related to this point, there are a number of complications concerning how one should interpret beads; these are addressed in Section 2.5.

As is clear from the preceding discussion, our main motivation for developing the Beaded Decomposition is its application in Part 3. The import of Part 2 in Part 3 has been deliberately distilled into this single statement and

[17] See Subsection 2.1.2 for a precise definition of the map $f^d_\#$.

Addendum 2.0.1 so that a reader who is willing to accept these as articles of faith can proceed directly from Part 1 to Part 3.

We expect that our particular refinement of the train-track technology may prove useful in other contexts. This expectation stems from the general point that the development of refined topological representatives leads to insights into purely algebraic questions about free-group automorphisms. See [**14**] for a concrete illustration of this.[18]

2.1. Improved Relative Train Track Maps

In this section we collect and refine those elements of the train-track technology that we shall need. Most of the material here is drawn directly from [**7**] and [**4**].

The philosophy behind train tracks is to find an *efficient* topological representative for an outer automorphism of F. Precisely what it means for a graph map to be *efficient* is spelled out in this section.

2.1.1. Edge-paths and tightening. Let G be a graph. Following [**4**], we try to reserve the term *path* for a map $\sigma : [0,1] \to G$ that is either constant or an immersion (i.e. *tight*). The reverse path $t \mapsto \sigma(1-t)$ will be denoted $\overline{\sigma}$. We conflate the map σ with its monotone reparameterisations (and even its image, when this does not cause confusion). Given an arbitrary continuous map $\rho : [0,1] \to G$, we denote by $[\rho]$ the unique (tight) *path* homotopic rel endpoints to ρ. In keeping with the notation of the previous section, given $f : G \to G$ and a path σ in G, we write $f_{\#}(\sigma)$ to denote $[f(\sigma)]$. We are primarily concerned with *edge-paths*, i.e. those paths σ for which $\sigma(0)$ and $\sigma(1)$ are vertices.

We consider only maps $f : G \to G$ that send vertices to vertices and edges to edge-paths (not necessarily to single edges). If there is an isomorphism $F \cong \pi_1 G$ such that f induces $\mathcal{O} \in \mathrm{Out}(F)$, then one says that f *represents* $\mathcal{O}$.

2.1.2. Replacing f by an Iterate. In order to obtain good topological representatives of outer automorphisms, one has to replace the given map by a large iterate. It is important to be clear what one means by *iterate* in this context, since we wish to consider only topological representatives whose restriction to each edge is an immersion and this property is not inherited by (naive) powers of the map.

Thus we deem the phrase[19] *replacing f by an iterate*, to mean that for fixed $k \in \mathbb{N}$, we pass from consideration of $f : G \to G$ to consideration of the map

[18] [**14**] contains results about the growth of words under iterated automorphisms. A previous version of Part 2 of this book contained an incorrect version of these results. We thank Gilbert Levitt for bringing this error to our attention.

[19] and obvious variations on it

$f^k_\# : G \to G$ that sends each edge E in G to the tight edge-path $f^k_\#(E)$ that is homotopic rel endpoints to $f^k(E)$.

2.1.3. (Improved) Relative train tracks. We now describe the properties of Improved Relative Train Track maps, as constructed in [**7**] and [**4**].

Splittings, Turns and Strata. Suppose that $\sigma = \sigma_1\sigma_2$ is a decomposition of a path into nontrivial subpaths (we do not assume that σ_1 and σ_2 are edge-paths, even if σ is). We say that $\sigma = \sigma_1\sigma_2$ is a *k-splitting* if

$$f^k_\#(\sigma) = f^k_\#(\sigma_1)f^k_\#(\sigma_2)$$

is a decomposition into sub-paths (i.e. for *some* choice of tightening, there is no folding between the f^k-images of σ_1 and σ_2 when $f^k(\sigma_1\sigma_2)$ is tightened). If $\sigma = \sigma_1\sigma_2$ is a k-splitting for all $k > 0$ then it is called a *splitting*[20] and we write $\sigma = \sigma_1 \cdot \sigma_2$. If one of σ_1 or σ_2 is the empty path, the splitting is said to be *vacuous.*

A *turn* in G is an unordered pair of half-edges originating at a common vertex. A turn is *non-degenerate* if it is defined by distinct half-edges, and is *degenerate* otherwise. The map $f : G \to G$ induces a self-map Df on the set of oriented edges of G by sending an oriented edge to the first oriented edge in its f-image. Df induces a map Tf on the set of turns in G.

A turn is *illegal* with respect to $f : G \to G$ if its image under some iterate of Tf is degenerate; a turn is *legal* if it is not illegal.

Associated to f is a *filtration* of G,

$$\emptyset = G_0 \subset G_1 \subset \cdots \subset G_\omega = G,$$

consisting of f-invariant subgraphs of G. We call the sets $H_r := \overline{G_r \smallsetminus G_{r-1}}$ *strata*. To each stratum H_r is associated M_r, the *transition matrix for* H_r; the $(i,j)^{\text{th}}$ entry of M_r is the number of times the f-image of the j^{th} edge crosses the i^{th} edge in either direction. By choosing a filtration carefully one may ensure that for each r the matrix M_r is either the zero matrix or is irreducible. If M_r is the zero matrix, then we say that H_r is a *zero stratum.* Otherwise, M_r has an associated Perron-Frobenius eigenvalue $\lambda_r \geq 1$, see [**36**]. If $\lambda_r > 1$ then we say that H_r is an *exponential stratum*; if $\lambda_r = 1$ then we say that H_r is a *parabolic stratum*[21]. The edges in strata inherit these adjectives, e.g. "exponential edge".

A turn is defined to be *in H_r if both* half-edges lie in the stratum H_r. A turn is a *mixed turn in* (G_r, G_{r-1}) if one edge is in H_r and the other is in G_{r-1}. A path with no illegal turns in H_r is said to be *r-legal.* We may emphasize that certain turns are in H_r by calling them *r-(il)legal turns.*

[20] In the next section, we introduce a stronger notion of *hard* splittings.

[21] Bestvina *et al.* use the terminology *exponentially-growing* and *non-exponentially-growing* for our exponential and parabolic. This difference in terminology explains the names of the items in Theorem 2.1.8 below.

DEFINITION 2.1.1. [**7**, Section 5, p.38] We say that $f : G \to G$ is a *relative train track map* if the following conditions hold for every exponential stratum H_r:

(RTT-i) Df maps the set of oriented edges in H_r to itself; in particular all mixed turns in (G_r, G_{r-1}) are legal.
(RTT-ii) If α is a nontrivial path in G_{r-1} with endpoints in $H_r \cap G_{r-1}$, then $f_\#(\alpha)$ is a nontrivial path with endpoints in $H_r \cap G_{r-1}$.
(RTT-iii) For each legal path β in H_r, $f(\beta)$ is a path that does not contain any illegal turns in H_r.

The following lemma is "the most important consequence of being a relative train track map" [**4**, p.530]; it follows immediately from Definition 2.1.1.

LEMMA 2.1.2. [**7**, Lemma 5.8, p.39] *Suppose that $f : G \to G$ is a relative train track map, that H_r is an exponential stratum and that $\sigma = a_1 b_1 a_2 \dots b_l$ is the decomposition of an r-legal path σ into subpaths a_j in H_r and b_j in G_{r-1}. (Allow for the possibility that a_1 or b_l is trivial, but assume the other subpaths are nontrivial.) Then $f_\#(\sigma) = f(a_1) f_\#(b_1) f(a_2) \dots f_\#(b_l)$ and is r-legal.*

DEFINITION 2.1.3. Suppose that $f : G \to G$ is a topological representative, that the parabolic stratum H_i consists of a single edge E_i and that $f(E_i) = E_i u_i$ for some path u_i in G_{i-1}. We say that the paths of the form $E_i \gamma \bar{E}_i$, $E_i \gamma$ and $\gamma \bar{E}_i$, where γ is in G_{i-1}, are *basic paths of height i*.

LEMMA 2.1.4. [**4**, Lemma 4.1.4, p.555] *Suppose that $f : G \to G$ and E_i are as in Definition 2.1.3. Suppose further that σ is a path or circuit in G_i that intersects H_i nontrivially and that the endpoints of σ are not contained in the interior of E_i. Then σ has a splitting each of whose pieces is either a basic path of height i or is contained in G_{i-1}.*

DEFINITION 2.1.5. A *Nielsen path* is a nontrivial path σ such that $f^k_\#(\sigma) = \sigma$ for some $k \geq 1$.

Nielsen paths are called *periodic Nielsen paths* in [**4**], but Theorem 2.1.8 below allows us to choose an f so that any periodic Nielsen path has period 1 (which is to say that $f_\#(\sigma) = \sigma$), and we shall assume that f satisfies the properties outlined in Theorem 2.1.8. Thus we can assume that $k = 1$ in the above definition. A Nielsen path is called *indivisible* if it cannot be split as a concatenation of two non-trivial Nielsen paths.

DEFINITION 2.1.6 (cf. 5.1.3, p. 561 [**4**]). Suppose that H_i is a single edge E_i and that $f(E_i) = E_i \tau^l$ for some closed Nielsen path τ in G_{i-1} and some $l > 0$. The *exceptional paths of height i* are those paths of the form $E_i \tau^k \bar{E}_j$ or $E_i \bar{\tau}^k \bar{E}_j$ where $k \geq 0$, $j \leq i$, H_j is a single edge E_j and $f(E_j) = E_j \tau^m$ for some $m > 0$.

REMARK 2.1.7. In [**4**] the authors mistakenly say that τ is an indivisible Nielsen path, rather than a primitive Nielsen path (not a proper power). We omit the modifier entirely.

In Definition 2.1.6, the paths do not have a preferred orientation. Thus it is important to note that the paths of the form $E_j\tau^k\bar{E}_i$ and $E_j\bar{\tau}^k\bar{E}_i$ with E_i, E_j and τ as above are also exceptional paths of height i.

2.1.4. The Theorem of Bestvina, Feighn and Handel. A matrix is *aperiodic* if it has a power in which every entry is positive. The map f is *eg-aperiodic* if every exponential stratum has an aperiodic transition matrix.

Theorem 5.1.5 in [**4**] is the main structural theorem for improved relative train track maps. We shall use it continually in what follows, often without explicit mention. We therefore record those parts of it which we need. A map f which satisfies the statements of Theorem 2.1.8 is called an *improved relative train track map.*

THEOREM 2.1.8. *(cf. Theorem 5.1.5, p.562,* [**4**]*) For every outer automorphism $\mathcal{O} \in \mathrm{Out}(F)$ there is an eg-aperiodic relative train track map $f : G \to G$ with filtration $\emptyset = G_0 \subset G_1 \subset \cdots \subset G_\omega = G$ such that f represents an iterate of $\mathcal{O}$, and f has the following properties.*

- *Every periodic Nielsen path has period one.*
- *For every vertex $v \in G$, $f(v)$ is a fixed point. If v is an endpoint of an edge in a parabolic stratum then v is a fixed point. If v is the endpoint of an edge in an exponential stratum H_i and if v is also contained in a noncontractible component of G_{i-1}, then v is a fixed point.*
- *H_i is a zero stratum if and only if it is the union of the contractible components of G_i.*
- *If H_i is a* zero stratum*, then*
  - z-(i) *H_{i+1} is an exponential stratum.*
  - z-(ii) *$f|H_i$ is an immersion.*
- *If H_i is a* parabolic stratum*, then*
  - ne-(i) *H_i is a single edge E_i.*
  - ne-(ii) *$f(E_i)$ splits as $E_i \cdot u_i$ for some closed path u_i in G_{i-1} whose basepoint is fixed by f.*
  - ne-(iii) *If σ is a basic path of height i that does not split as a concatenation of two basic paths of height i or as a concatenation of a basic path of height i with a path contained in G_{i-1}, then either:* (i) *for some k, the path $f^k_\#(\sigma)$ splits into pieces, one of which equals E_i or $\bar{E}_i$; or* (ii) *u_i is a Nielsen path and, for some k, the path $f^k_\#(\sigma)$ is an exceptional path of height i.*
- *If H_i is an* exponential stratum *then*

eg-(i) *There is at most one indivisible Nielsen path ρ_i in G_i that intersects H_i nontrivially. The initial edges of ρ_i and $\bar{\rho}_i$ are distinct (possibly partial) edges in H_i.*

Suppose that $f : G \to G$ is an improved relative train track map representing some iterate ϕ^k of $\phi \in \mathrm{Out}(F_n)$, and that ρ is a Nielsen path in G_r that intersects H_r nontrivially, and suppose that ρ is not an edge-path. Then subdividing the edges containing the endpoints of ρ at the endpoints, gives a new graph G', and the map $f' : G' \to G'$ induced by f is an improved relative train track map representing ϕ^k. To ease notation, it is convenient to assume that this subdivision has been performed. Under this assumption, all Nielsen paths will be edge-paths, and all of the paths which we consider in the remainder of Part 2 will also be edge-paths.

CONVENTION 2.1.9. Since all Nielsen paths in the remainder of Part 2 will be edge-paths, we will use the phrase *'indivisible Nielsen path'* to mean a Nielsen edge-path which cannot be decomposed nontrivially as a concatenation of two non-trivial Nielsen *edge*-paths. In particular, a single edge fixed pointwise by f will be considered to be an indivisible Nielsen path.

For the remainder of this article, we will concentrate on an improved relative train track map $f : G \to G$ and repeatedly pass to iterates $f_\#^k$ in order to better control its cancellation properties.

Recall the following from [**4**, Section 4.2, pp.558-559].

DEFINITION 2.1.10. If $f : G \to G$ is a relative train track map and H_r is an exponential stratum, then define P_r to be the set of paths ρ in G_r that are such that:

(i) For each $k \geq 1$ the path $f_\#^k(\rho)$ contains exactly one illegal turn in H_r.
(ii) For each $k \geq 1$ the initial and terminal (possibly partial) edges of $f_\#^k(\rho)$ are contained in H_r.
(iii) The number of H_r-edges in $f_\#^k(\rho)$ is bounded independently of k.

LEMMA 2.1.11. [**4**, Lemma 4.2.5, p.558] *P_r is a finite $f_\#$-invariant set.*

LEMMA 2.1.12. [**4**, Lemma 4.2.6, p.559] *Suppose that $f : G \to G$ is a relative train track map, that H_r is an exponential stratum, that σ is a path or circuit in G_r and that, for each $k \geq 0$, the path $f_\#^k(\sigma)$ has the same finite number of illegal turns in H_r. Then σ can be split into subpaths that are either r-legal or elements of P_r.*

DEFINITION 2.1.13. If ρ is a path and r is the least integer such that ρ is in G_r then we say that *ρ has weight r*.

If ρ has weight r and H_r is exponential, we will say that ρ is an *exponential path.* We define *parabolic paths* similarly.

LEMMA 2.1.14. *Suppose that σ is an edge-path and that, for some $k \geq 1$, $f_\#^k(\sigma)$ is a Nielsen path. Then $f_\#(\sigma)$ is a Nielsen path.*

PROOF. Suppose that the endpoints of σ are u_1 and v_1 and that the endpoints of $f_\#^k(\sigma)$ are u_2 and v_2. For each vertex $v \in G$, $f(v)$ is fixed by f, so $f(u_1) = u_2$ and $f(v_1) = v_2$. If $f_\#(\sigma) \neq f_\#^k(\sigma)$ then we have two edge-paths with the same endpoints which eventually get mapped to the same path. Thus there is some nontrivial circuit which is killed by f, contradicting the fact that f is a homotopy equivalence. Therefore $f_\#(\sigma) = f_\#^k(\sigma)$ and so is a Nielsen path. □

Always, L will denote the maximum of the lengths of the paths $f(E)$, for E an edge in G.

Later, we will pass to further iterates of f in order to find a particularly nice form.

An analysis of the results in this section allows us to see that there are three kinds of indivisible Nielsen paths. The first are those which are single edges; the second are certain exceptional paths; and the third lie in the set P_r. We will use this trichotomy frequently without mention. The first two cases are where the path is parabolic-weight, the third where it is exponential-weight. It is not possible for Nielsen path to have weight r where H_r is a zero stratum.

OBSERVATION 1. *Let ρ be an indivisible Nielsen path of exponential weight r. Then the first and last edges in ρ are contained in H_r.*

Because periodic Nielsen paths have period 1, the set of Nielsen paths does not change when f is replaced by a further iterate of itself. We will use this fact often.

LEMMA 2.1.15. *Suppose E is an edge such that $|f_\#^j(E)|$ grows linearly with j. Then $f(E) = E \cdot \tau^k$, where τ is a Nielsen path that is not a proper power. The edge-path τ decomposes into indivisible Nielsen paths (each of which is itself an edge-path, by Convention 2.1.9).*

PROOF. The fact that $f(E) = E \cdot \tau^k$, where τ is a Nielsen path follows from conditions ne-(ii) and ne-(iii) of Theorem 2.1.8. [22] □

LEMMA 2.1.16. *Let τ be a Nielsen path and τ_0 a proper initial (or terminal) sub edge-path of τ. No image $f_\#^k(\tau_0)$ contains τ as a sub edge-path.*

PROOF. It is sufficient to prove the lemma for indivisible Nielsen paths, as the result for arbitrary Nielsen paths then follows immediately.

If τ is an indivisible Nielsen path and τ_0 is a proper non-trivial subpath of τ then τ cannot be a single edge. Therefore, either τ is either an indivisible Nielsen path of exponential weight, or an exceptional path.

[22] If Theorem 2.1.8, ne-(iii) held with the Nielsen path τ in the definition of exceptional paths being indivisible, we could also insist that τ be indivisible here.

In case τ is an indivisible Nielsen path of exponential weight, suppose the weight is r. By Lemma 2.1.12 τ contains a single illegal turn in H_r. Suppose that τ_0 does not contain this illegal turn. Then τ_0 is r-legal, and so no iterate of τ_0 contains an illegal turn in H_r. Therefore no iterate of τ_0 can contain τ as a subpath.

Suppose then that τ_0 *does* contain the r-illegal turn in τ. Then, being a proper subpath of τ, the path on one side of the illegal turn in τ_0 and its (tightened) iterates is strictly smaller than the corresponding path in τ. Once again τ cannot be contained as a subpath of any iterate of τ_0.

Finally, suppose τ is an exceptional path. Then $\tau = E_i \rho^k \bar{E}_j$ where ρ is a Nielsen path and E_i and E_j are of weight greater than ρ. Any proper sub edge-path τ_0 of τ contains at most one edge of weight greater than ρ. T he same is true for any iterate of τ_0, and once again no iterate of τ_0 contains τ as a sub-path. □

2.2. Hard Splittings

In this section we introduce a new concept for improved relative train tracks: *hard splittings*. This plays an important role in the subsequent sections of Part 2, and also in Part 3.

Recall that a decomposition of a path $\sigma = \sigma_1 \sigma_2$ is a k-splitting if $f^k_\#(\sigma) = f^k_\#(\sigma_1) f^k_\#(\sigma_2)$; which means that, for *some* choice of tightening, the images of σ_1 and σ_2 do not interact with each other. This leads to the concept of *splittings*. We need a more restrictive notion, where the decomposition is preserved for *every* choice of tightening. For this purpose, we make the following

Definition 2.2.1. [Hard splittings] We say that a k-splitting $\rho = \rho_1 \rho_2$ is a *hard* k-splitting if for *any* choice of tightening of $f^k(\rho) = f^k(\rho_1) f^k(\rho_2)$ there is no cancellation between the image of ρ_1 and the image of ρ_2.

A decomposition which is a hard k-splitting for all $k \geq 1$ is called a *hard* splitting. If $\rho_1 \cdot \rho_2$ is a hard splitting, we write $\rho_1 \odot \rho_2$.

An edge-path is *hard-indivisible* (or *h-indivisible*) if it admits no non-vacuous hard splitting into edge-paths.

Remark 2.2.2. If one works in the universal cover, then $\tilde{\sigma}_1 \tilde{\sigma_2}$ is a k-hard splitting if and only if, inside $\tilde{f}^k(\tilde{\sigma}_1 \tilde{\sigma}_2))$, the intersection $\tilde{f}^k(\tilde{\sigma}_1) \cap \tilde{f}^k(\tilde{\sigma}_2)$ is a single point.

Remark 2.2.3. In the above definition, we allow the possibility that one of the paths in the hard splitting is empty. This is to allow various later statements to be made more concisely.

For example, the phrase 'ρ admits a hard splitting immediately on either side of σ of ρ' (for a path ρ and a sub edge-path σ) allows the possibility that σ is an initial or terminal sub-path of ρ.

EXAMPLE 2.2.4. Suppose that G is the graph with a single vertex and edges E_1, E_2 and E_3. Suppose that $f(E_1) = E_1$, $f(E_2) = E_2E_1$ and $f(E_3) = E_3\bar{E}_1\bar{E}_2$. Then f is an improved relative train track. And $E_3E_2 \cdot \bar{E}_1$ is a 1-splitting, since

$$f(E_3E_2\bar{E}_1) = E_3\bar{E}_1\bar{E}_2E_2E_1\bar{E}_1,$$

which tightens to $E_3\bar{E}_1 = f_\#(E_3E_2)f_\#(\bar{E}_1)$. In fact this is a splitting. However, there is a choice of tightening which first cancels the final $E_1\bar{E}_1$ and then the subpath $\bar{E}_2E_2$. Therefore the splitting $E_3E_2 \cdot \bar{E}_1$ is not a hard 1-splitting.

The following lemma describes the main utility of hard splittings, and the example above shows that it is not true in general for splittings.

LEMMA 2.2.5. *Suppose that $\sigma_1 \odot \sigma_2$ is a hard splitting, that ρ_1 is a terminal subsegment of σ_1, and that ρ_2 is an initial subpath of σ_2. Then $\rho_1 \odot \rho_2$ is a hard splitting.*

PROOF. If there were any cancellation between images of ρ_1 and ρ_2 then there would be a possible tightening between the images of σ_1 and σ_2. □

The following two lemmas will also be crucial for our applications of hard splittings in Part 3.

LEMMA 2.2.6. *Every edge-path admits a unique maximal hard splitting into edge-paths.*

PROOF. This follows by an obvious induction on length from the observation that if $\rho = \rho_1\rho_2\rho_3$, where the ρ_i are edge-paths, and if $\rho = \rho_1 \odot \rho_2\rho_3$ and $\rho = \rho_1\rho_2 \odot \rho_3$ then $\rho = \rho_1 \odot \rho_2 \odot \rho_3$. □

LEMMA 2.2.7. *If $\rho = \rho_1 \odot \rho_2$ and σ_1 and σ_2 are, respectively, terminal and initial subpaths of $f_\#^k(\rho_1)$ and $f_\#^k(\rho_2)$ for some $k \geq 0$ then $\sigma_1\sigma_2 = \sigma_1 \odot \sigma_2$.*

PROOF. For all $i \geq 1$, the untightened path $f^i(\sigma_1)$ is a terminal subpath of the untightened path $f^i(f_\#^k(\rho_1))$, while $f^i(\sigma_2)$ is an initial subpath of $f^i(f_\#^k(\rho_2))$.

The hardness of the splitting $\rho = \rho_1 \odot \rho_2$ ensures that no matter how one tightens $f^{k+i}(\rho_1)f^{k+i}(\rho_2)$ there will be no cancellation between $f^{k+i}(\rho_1)$ and $f^{k+i}(\rho_2)$. In particular, one is free to tighten to obtain $f^i(f_\#^k(\rho_1))f^i(f_\#^k(\rho_2))$ first, and then tighten $f^i(\sigma_1)f^i(\sigma_2)$, and there can be no cancellation between them. (It may happen that when one goes to tighten $f^{k+i}(\rho_1)$ completely, the whole of $f^i(\sigma_1)$ is cancelled, but this does not affect the assertion of the lemma.) □

The purpose of the remainder of this section is to sharpen results from the previous section to cover hard splittings [23].

The following lemma is clear.

[23] Bestvina *et al.* make no explicit mention of the distinction between splittings and hard splittings, however condition (3) of Proposition 5.4.3 on p.581 (see Lemma 2.2.10

LEMMA 2.2.8 (cf. Lemma 4.1.1, p.554 [**4**]). *If* $\sigma = \sigma_1 \odot \sigma_2$ *is a hard splitting, and* $\sigma_1 = \sigma_1' \odot \sigma_2'$ *is a hard splitting then* $\sigma = \sigma_1' \odot \sigma_2' \odot \sigma_2$ *is a hard splitting. The analogous result with the roles of* σ_1 *and* σ_2 *reversed also holds.*

REMARK 2.2.9. The possible existence of an edge-path σ_2 so that $f_\#(\sigma_2)$ is a single vertex means that $\sigma_1\sigma_2 = \sigma_1 \odot \sigma_2$ and $\sigma_2\sigma_3 = \sigma_2 \odot \sigma_3$ need *not* imply that $\sigma_1\sigma_2\sigma_3 = \sigma_1 \odot \sigma_2 \odot \sigma_3$.

Indeed if σ_2 is an edge-path so that $f_\#(\sigma_2)$ is a vertex then $f_\#(\sigma_1)$ and $f_\#(\sigma_3)$ come together in a tightening of $f(\sigma_1\sigma_2\sigma_3)$, possibly cancelling.

In contrast, if $f_\#(\sigma_2)$ (and hence each $f_\#^k(\sigma_2)$) contains an edge, then the hardness of the two splittings ensures that in any tightening $f_\#(\sigma_1\sigma_2\sigma_3) = f_\#(\sigma_1)f_\#(\sigma_2)f_\#(\sigma_3)$, that is $\sigma_1\sigma_2\sigma_3 = \sigma_1 \odot \sigma_2 \odot \sigma_3$.

The following strengthening of Theorem 2.1.8 ne-(ii) is a restatement of (a weak form of) [**4**, Proposition 5.4.3.(3), p.581].

LEMMA 2.2.10. *Suppose* f *is an improved relative train track map and* E *is a parabolic edge with* $f(E) = Eu$*. For any initial subpath* w *of* u*,* $E \cdot w$ *is a splitting.*

COROLLARY 2.2.11. *Suppose* f *is an improved relative train track map,* E *is a parabolic edge and* $f(E) = Eu$*. Then* $f(E) = E \odot u$*.*

The following lemma is straightforward to prove.

LEMMA 2.2.12. *Suppose* H_i *is a parabolic stratum and* σ *is a path in* G_i *that intersects* H_i *nontrivially, and that the endpoints of* σ *are not contained in the interior of* E_i*. Then* σ *admits a hard splitting, each of whose pieces is either a basic path of height* i *or is contained in* G_{i-1}*.*

LEMMA 2.2.13. *If* σ *is a basic path of height* i *that does not admit a hard splitting as a concatenation of two basic paths of height* i *or as a concatenation of a basic path of height* i *with a path of weight less than* i*, then either; (i) for some* k*, the path* $f_\#^k(\sigma)$ *admits a hard splitting into pieces, one of which is* E_i *or* $\bar{E}_i$*; or (ii)* $f(E_i) = E_i \odot u_i$*, where* u_i *is a Nielsen path and, for some* k*, the path* $f_\#^k(\sigma)$ *is an exceptional path of height* i*.*

PROOF. Follows from the proof of [**4**, Lemma 5.5.1, pp.585–590]. □

LEMMA 2.2.14 (cf. Lemma 2.1.12 above). *Suppose that* $f : G \to G$ *is a relative train track map, that* H_r *is an exponentially-growing stratum, that* σ *is a path or circuit in* G_r*, and that each* $f_\#^k(\sigma)$ *has the same finite number of illegal turns in* H_r*. Then* σ *can be decomposed as* $\sigma = \rho_1 \odot \ldots \odot \rho_k$*, where each* ρ_i *is either (i) an element of* P_r*; (ii) an* r*-legal path which starts and ends with edges in* H_r*; or (iii) of weight at most* $r - 1$*.*

below) indicates that they are aware of the distinction and that the term 'splitting' has the same meaning for them as it does here.

PROOF. Consider the splitting of σ given by Lemma 2.1.12. The pieces of this splitting are either (i) elements of P_r, or (ii) r-legal paths. By Definition 2.1.1 RTT-(i), any r-legal path admits a hard splitting into r-legal paths which start and end with edges in H_r, and paths of weight at most $r-1$. The turn at the end of a Nielsen path in the splitting of σ is either a mixed turn (with the edge from H_r coming from the Nielsen path and the other edge being of weight at most $r-1$) or a legal turn in H_r. In either case, σ admits a hard splitting at the vertex of this turn. □

The next result follows from a consideration of the form of indivisible Nielsen paths, noting Definition 2.1.1 and Lemma 2.2.14.

LEMMA 2.2.15. *Any Nielsen path admits a hard splitting into indivisible Nielsen paths.*

REMARK 2.2.16. If $\rho = \rho_1 \odot \rho_2$ is a hard splitting for the map f then it is a hard splitting for $f_\#^k$ for any $k \geq 1$.

We record a piece of terminology which will be important in Part 3.

DEFINITION 2.2.17. A sub edge-path ρ of a path χ is *displayed* if there is a hard splitting of χ immediately on either side of ρ.

2.3. A Small Reduction

In this section we clarify a couple of issues about monochromatic paths, and state Theorem 2.3.2, which immediately implies the Beaded Decomposition Theorem.

Our strategy for proving the Beaded Decomposition Theorem is as follows: given an automorphism $\phi \in \mathrm{Aut}(F_n)$, we start with an improved relative train track representative $f : G \to G$ for some iterate ϕ^k of ϕ, as obtained from the conclusion of Theorem 2.1.8. We analyse the evolution of monochromatic paths, and eventually pass to an iterate of f in which we can prove the Beaded Decomposition Theorem. However, it is crucial to note that monochromatic paths for f are not necessarily monochromatic paths for $f_\#^k$ when $k > 1$. See Section 2.5 for further discussion about some of these issues.

These concerns lead to the following definition, where we are concentrating on a fixed IRTT $f : G \to G$, and so omit mention of f from our notation.

DEFINITION 2.3.1. For a positive integer d, we define *d-monochromatic paths* by recursion: edges in G are d-monochromatic and if ρ is a d-monochromatic path then every sub edge-path of $f_\#^d(\rho)$ is d-monochromatic.

Note that if d' is a multiple of d then every d'-monochromatic path is d-monochromatic but not *vice versa*. Thus if we replace f by an iterate then, for fixed n, the set of n-monochromatic paths may get smaller. The content of the Beaded Decomposition Theorem is that one need only pass to a bounded iterate in order to ensure that all monochromatic paths admit a beaded decomposition. In particular, the Beaded Decomposition Theorem is an immediate consequence of the following theorem.

THEOREM 2.3.2 (Monochromatic paths are beaded). *Let $f : G \to G$ be an improved relative train track map. There exist constants d and J, depending only on f, so that every d-monochromatic path in G is (J, f)-beaded.*

DEFINITION 2.3.3 (Nibbled Futures). Let ρ be a (tight) edge-path. The 0-*step nibbled future of* ρ is ρ.

For $k \geq 1$, a k-*step nibbled future of* ρ is a sub edge-path of $f_\#(\sigma)$, where σ is a $(k-1)$-step nibbled future of ρ. A *nibbled future of* ρ is a k-step nibbled future for some $k \geq 0$.

For $k \geq 0$, the k-*step entire future of* ρ is $f_\#^k(\rho)$.

REMARK 2.3.4. Nibbled futures are not assumed to be non-empty. If a path is empty, any statement we claim about the existence of hard splittings should be interpreted to hold vacuously. The 1-monochromatic paths are precisely the nibbled futures of single edges.

The notion of nibbled futures is central to Parts 2 and 3 of this book. Usually, when proving things about monochromatic paths, we are actually proving things about the nibbled futures of paths of bounded length. In this spirit, rather than just proving Theorem 2.3.2, we prove the following more general theorem about the iterated futures of arbitrary paths. We expect this theorem to have applications beyond those presented in this work.

THEOREM 2.3.5. *If $f : G \to G$ is an improved relative train track map, then there exists an integer d with the following property: for each positive integer n, there exists $J > 0$ so that for every edge-path ρ with $|\rho| \leq n$ and every positive integer k,* every *kd-step nibbled future of ρ is (J, f)-beaded.*

REMARK 2.3.6. It is clear that Theorem 2.3.2 follows immediately from Theorem 2.3.5. Therefore, in order to prove the Beaded Decomposition Theorem, it suffices to prove Theorem 2.3.5.

REMARK 2.3.7. We posted a version of Part 2 of this book on the *ArXiv* in July 2005. In December 2006, Feighn and Handel posted [**24**], in which they develop a powerful refinement of the train track technology. If one employs their *completely split* train track representatives, one can prove the Beaded Decomposition Theorem with considerably greater ease than we do here. One can also streamline significant parts of the proof of Theorem 2.3.5. However, we feel that the effort that this would save the reader is offset by the extra machinery that they would be required to accept or absorb. On this basis, we decided to retain our original proof.

2.4. Nibbled Futures

NOTATION 2.4.1. *Throughout this section and the rest of Part 2, $f : G \to G$ is an improved relative train track map.*

Let L be the maximum of the lengths of the paths $f(E)$ where E ranges over the edges of G.

Monochromatic paths arise as *nibbled futures* in the sense defined below. Thus in order to prove Theorem 2.3.2 we must understand how *nibbled futures* evolve. The results in this section reduce this challenge to the task of understanding the nibbled futures of GEPs.

THEOREM 2.4.2 (First Decomposition Theorem). *For any $n \geq 1$ there exists an integer $V = V(n, f)$ such that if ρ is an edge-path of length at most n then any nibbled future of ρ admits a hard splitting into edge-paths, each of which is either the nibbled future of a* GEP *or else has length at most V.*

The remainder of this section is dedicated to proving Theorem 2.4.2. We begin by examining the entire future of a path of fixed length (Lemma 2.4.4) and then refine the argument to deal with nibbling. In the proof of the first of these lemmas we require the following observation.

REMARK 2.4.3. Suppose that ρ is a tight path of weight r. Since f is an improved relative train track map, the number of r-illegal turns in $f^l_\#(\rho)$ is a non-increasing function of l, bounded below by 0.

LEMMA 2.4.4. *There is a function $D : \mathbb{N} \to \mathbb{N}$, depending only on f, such that, for any $r \in \{1, \dots, \omega\}$, if ρ is a path of weight r, and $|\rho| \leq n$, then for any $i \geq D(n)$ the edge-path $f^i_\#(\rho)$ admits a hard splitting into edge-paths, each of which is either*

(1) a single edge of weight r;
(2) an indivisible Nielsen path of weight r;
(3) a GEP *of weight r; or*
(4) a path of weight at most $r - 1$.

PROOF. If H_r is a zero stratum, then $f_\#(\rho)$ has weight at most $r - 1$, and $D(n) = 1$ will suffice for any n.

If H_r is a parabolic stratum, then ρ admits a hard splitting into pieces which are either basic of height r or of weight at most $r - 1$ (Lemma 2.2.12). Thus it is sufficient to consider the case where ρ is a basic path of weight r and $|\rho| \leq n$. By at most 2 applications of Lemma 2.2.13, we see that there exists a k such that $f^k_\#(\rho)$ admits a hard splitting into pieces which are either (i) single edges of weight r, (ii) exceptional paths of height r, or (iii) of weight at most $r - 1$. By taking the maximum of such k over all basic paths of height r which are of length at most n, we find an integer k_0 so that we have the desired hard splitting of $f^{k_0}_\#(\rho)$ for all basic paths of height r of length at most n. Any of the exceptional paths in these splittings which are not GEPs have bounded length and are either indivisible Nielsen paths or are decreasing in length. A crude bound on the length of the exceptional paths which are not GEPs is $L^{k_0}n$ where L is the maximum length of $f(E)$ over all edges $E \in G$. Thus, those exceptional paths which are decreasing in length will become GEPs within less than $L^{k_0}n$ iterations. Therefore, replacing k_0 by $k_0 + L^{k_0}n$, we may assume all exceptional paths in the hard splitting are GEPs.

Finally, suppose that H_r is an exponential stratum. As noted in Remark 2.4.3, the number of r-illegal turns in $f^l_\#(\rho)$ is a non-increasing function of l bounded below by 0. Therefore, there is some j so that the number of r-illegal turns in $f^{j'}_\#(\rho)$ is the same for all $j' \geq j$. By Lemma 2.2.14, $f^j_\#(\rho)$ admits a hard splitting into pieces which are either (i) elements of P_r, (ii) single edges in H_r, or (iii) paths of weight at most $r-1$. To finish the proof of the lemma it remains to note that if $\sigma \in P_r$ then $f_\#(\sigma)$ is a Nielsen path by Lemma 2.1.14.

Therefore, the required constant for H_r may be taken to be the maximum of $j+1$ over all the paths of weight r of length at most n.

To find $D(n)$ we need merely take the maximum of the constants found above over all of the strata H_r of G. □

In the extension of the above proof to cover nibbled futures, we shall need the following straightforward adaptation of Lemma 2.1.16.

LEMMA 2.4.5. *Let τ be a Nielsen path and τ_0 a proper initial (or terminal) sub-path of τ. No nibbled future of τ_0 contains τ as a sub-path.*

PROPOSITION 2.4.6. *There exists a function $D' : \mathbb{N} \to \mathbb{N}$, depending only on f, so that for any $r \in \{1, \ldots, \omega\}$, if ρ is a path of weight r and $|\rho| \leq n$, then for any $i \geq D'(n)$ any i-step nibbled future of ρ admits a hard splitting into edge-paths, each of which is either*

(1) a single edge of weight r;
(2) a nibbled future of a weight r indivisible Nielsen path;
(3) a nibbled future of a weight r GEP; or
(4) a path of weight at most $r-1$.

Moreover, in Case (3), the GEP lies in the j-step nibbled future of ρ for some $j \leq i$.

REMARK 2.4.7. Each of the conditions (1) – (4) stated above is stable in the following sense: once an edge in a k-step nibbled future is contained in a path satisfying one of these conditions, then any future of this edge in any further nibbled future will also lie in such a path (possibly the future will go from case (1) to case (4), but otherwise which case it falls into is also stable). Thus we can split the proof of Proposition 2.4.6 into a number of cases, deal with the cases separately by finding some constant which suffices, and finally take a maximum to find $D'(n)$. An entirely similar remark applies to a number of subsequent proofs, in particular Theorem 2.8.1.

Remark 2.4.8. Since the statement of Proposition 2.4.6 involves all paths ρ such that $|\rho| \leq n$, if the function $D'(n)$ is chosen to be the smallest function satisfying the conclusion then it is nondecreasing. We will assume that the function D' we use is indeed monotonic.

Proof (Proposition 2.4.6). Let $\rho_0 = \rho$ and for $j > 0$ let ρ_j be a sub edge-path of $f_\#(\rho_{j-1})$.

If H_r is a zero stratum, then $f_\#(\rho)$ has weight at most $r-1$ and it suffices to take $D'(n) = 1$.

Suppose that H_r is an exponential stratum. By Lemma 2.4.4, the $D(n)$-step entire future of ρ admits a hard splitting of the desired form. We consider how nibbling can affect this splitting. As we move forwards through the nibbled future of ρ, cancellation of H_r-edges can occur only at r-illegal turns and at the ends, where the nibbling occurs.

Remark 2.4.3 implies that we can trace the r-illegal turns forwards through the successive nibbled futures of ρ (whilst the r-illegal continues to exist). We compare the r-illegal turns in ρ_k to those in $f_\#^k(\rho)$, the entire future of ρ. We say that the nibbling *first cancels an r-illegal turn at time k* if the collection of r-illegal turns in ρ_{k-1} is the same as the collection in $f_\#^{k-1}(\rho)$, but the collection in ρ_k is *not* the same as that of $f_\#^k(\rho)$. The first observation we make is that if, at time k, the nibbling has not yet cancelled any r-illegal turn then the sequence of H_r-edges in ρ_k is a subsequence of the H_r-edges in $f_\#^k(\rho)$. Therefore, any splitting of the desired type for $f_\#^k(\rho)$ is inherited by ρ_k.

Since there is a splitting of the $D(n)$-step entire future of ρ of the desired form, either there is a splitting of $\rho_{D(n)}$, or else $\rho_{D(n)}$ has fewer r-illegal turns than $f_\#^{D(n)}(\rho)$, and hence than ρ. However, $|\rho_{D(n)}| \leq n.L^{D(n)}$. We apply the above argument to $\rho_{D(n)}$, going forwards a further $D(nL^{D(n)})$ steps into the future. Since the number of illegal turns in H_r in ρ was at most $n-1$, we will eventually find a splitting of the required form within an amount of time bounded by a function of n (this function depends only on f, as required). Denoting this function by D_0, we have that any $D_0(n)$-step nibbled future of any path of exponential weight whose length is at most n admits a hard splitting of the desired form.

Now suppose that H_r is a parabolic stratum. By Lemma 2.2.12, ρ admits a hard splitting into basic edge-paths. Therefore we may assume (by reversing the orientation of ρ if necessary) that $\rho = E_r\sigma$ or $\rho = E_r\sigma\overline{E_r}$ where E_r is the unique edge in H_r and σ is in G_{r-1}. For the nibbled future of ρ to have weight r, the nibbling must occur only on one side (since the only edges of weight r in any future of ρ occur on the ends). We assume that all nibbling occurs from the right. Once again, the $D(n)$-step entire future of ρ admits a hard splitting of the desired form. If $\rho = E_r\sigma\overline{E_r}$ then the $D(n)$-step nibbled future of ρ either admits a hard splitting of the required form, or is of the form $E_r\sigma_1$, where σ_1

is in G_{r-1}. Hence we may assume that $\rho = E_r\sigma$. Suppose that $f(E_r) = E_r u_r$, and that u_r has weight $s < r$.

Consider first the possibility that σ has weight $q > s$ (but less than r by hypothesis). We claim that after a bounded amount of time the nibbled future of ρ admits a splitting into one piece of the form $E_r\sigma'$ where the weight of σ' is strictly less than q, and other pieces which are all of the form required by the statement of the proposition. Then, by induction on weight, we may suppose that we have a splitting into one piece of the form $E_r\sigma''$ where the weight of σ'' is at most s and all of the other pieces have the form required by the proposition.

So, suppose that σ has weight $q > s$. There are three cases to consider. If the weight of σ is that of a zero stratum, then it immediately drops in weight and the claim is proved.

Now suppose that H_q is an exponential stratum. The future of E_r cannot cancel any edges of weight q or higher in the future of σ, so the edges of weight q in the nibbled future of ρ are exactly the same as the edges of weight q in the corresponding nibbled future of σ (recall we are assuming that nibbling only occurs from the right). This $D_0(|\sigma|)$-step nibbled future of σ admits a hard splitting into edge-paths which are either[24] single edges of weight q, the nibbled future of an indivisible Nielsen path of weight q, or of weight at most $q-1$. Let σ_2 be the subpath of the $D_0(|\sigma|)$-step nibbled future of ρ which starts at the right endpoint of E_r up to but not including the first edge of weight q.[25] Then, since mixed turns are legal, the $D_0(n)$-step nibbled future of ρ admits a hard splitting into edge-paths, the leftmost of which is $E_r\sigma_2$.

Suppose now that H_q is a parabolic stratum. It is easy to see that ρ admits a hard splitting into edge-paths, the leftmost of which is either $E_r\sigma_2$ or $E_r\sigma_2\overline{E_q}$, where σ_2 has weight at most $q-1$. Thus we may suppose that ρ itself has this form. Again, either the $D(n)$-step nibbled future of ρ admits a hard splitting of the required form, or the $D(n)$-step nibbled future of ρ has the form $E_r\sigma_3$, where σ_3 has weight at most $q-1$. The arguments in the previous two paragraphs include the possibility that a GEP of weight r occurs as a factor of the hard splitting of the $D(n)$-step nibbled future of ρ. Thus we may assume that in some nibbled future of ρ there will necessarily be a hard splitting on each side of the edge of weight r. (Recall by Remark 2.2.3 that this includes the case that this edge is an initial or terminal subsegment.)

As noted above, by induction we have now proved that going forwards into the nibbled future an amount of time bounded by a function of n, we may assume that ρ has the form $E_r\sigma_4$, where σ_4 has weight at most s (thus σ_4 is the path σ'' from the claim above). Suppose that σ_4 has weight less than s. Then $f_\#(E_r\sigma_4) = E_r \odot \sigma_5$, where σ_5 has weight less than r. This is a splitting

---

[24] GEPs have parabolic weight

[25] In case the nibbled future of σ is empty, this is the entire path.

of the required form which is inherited by an nibbled future. Therefore, we are left with the case that the weight of σ_4 is exactly s.

We now consider what kind of stratum H_s is. Suppose that H_s is parabolic. There are only two ways in which cancellation between weight s edges in the nibbled future of ρ can occur (see Lemma 1.5.5): they might be cancelled by edges whose immediate past is the edge of weight r on the left end of the previous nibbled future; alternatively, they can be nibbled from the right. The $D(n)$-step entire future of ρ admits a hard splitting as $E_r \odot \sigma_6$, where σ_6 has weight at most $r-1$. There is no way that nibbling can affect this splitting.

Finally, suppose that H_s is an exponential stratum. We follow a similar argument to the case when H_r was an exponential stratum. Either the $D(n)$-step nibbled future of ρ admits a hard splitting of the desired kind (which means $\rho_{D(n)} = E_r \odot \sigma_7$ where σ_7 has weight at most $r-1$), or there are fewer s-illegal turns in the future of σ_4 in $\rho_{D(n)}$ than there are s-illegal turns in σ_4. We then apply the same argument to the nibbled future of $\rho_{D(n)}$ until eventually we achieve a hard splitting of the required form.

The last sentence in the statement of Proposition 2.4.6 follows immediately, since in the proof we have only consider paths which arise in the nibbled futures of ρ. This completes the proof of Proposition 2.4.6. □

We are now in a position to prove Theorem 2.4.2. For this we require the following definition.

Definition 2.4.9. Suppose that H_r is a stratum, and $E \in H_r$. An *r-seed* is a non-empty subpath ρ of $f(E)$ which is maximal subject to lying in G_{r-1}.

If the stratum H_r is not relevant, we just refer to *seeds*.

Note that seeds are edge-paths and that the set of all seeds is finite. Also, if H_r is an exponential stratum and $E \in H_r$ then the seeds in $f(E)$ are the sub-paths b_i from Definition 2.1.2.

The following is an immediate consequence of Lemma 2.2.14 and RTT-(i) of Definition 2.1.1.

Lemma 2.4.10. *If $E \in H_r$ is an exponential edge and ρ is an r-seed in $f(E)$ then $f(E) = \sigma_1 \odot \rho \odot \sigma_2$ where σ_1 and σ_2 are r-legal paths which start and finish with edges in H_r.*

Proof (Theorem 2.4.2). Suppose that ρ is a path of length n and that ρ_k is a k-step nibbled future of ρ. Denote by $\rho_0 = \rho, \rho_1, \ldots, \rho_{k-1}$ the intermediate nibbled futures of ρ used in order to define ρ_k.

We begin by constructing a van Kampen diagram[26] Δ_k which encodes the ρ_i, proceeding by induction on k. For $k = 1$ the diagram Δ_1 has a single

[26]in fact, just a stack of corridors. Of course, van Kampen diagrams are not required for this proof, but we find them a convenient way of encoding choices of tightening and nibbling.

(folded) corridor with the bottom labelled by ρ and the path ρ_1 a subpath of the top of this corridor. Suppose that we have associated a van Kampen diagram Δ_{k-1} to ρ_{k-1}, with a unique corridor at each time $t = 0, \ldots, k-2$, such that ρ_{k-1} is a subpath of the top of the latest (folded) corridor. Then we attach a new folded corridor to Δ_{k-1} whose bottom is labelled by ρ_{k-1}. The path ρ_k is, by definition, a subpath of the top of this new latest corridor. By convention, we consider ρ_i to occur at time i.

Choose an arbitrary edge ε in ρ_k on the (folded) top of the latest corridor in Δ_k. We will prove that there is a path σ containing ε in ρ_k so that ρ_k admits a hard splitting immediately on either side of σ and so that σ is either suitably *short* or a nibbled future of a GEP. The purpose of this proof is to find a suitable notion of *short*.

Consider the embedded 'family forest' $\mathcal{F}$ for Δ_k, tracing the histories of edges lying on the folded tops of corridors (see Remark 1.3.2). Let p be the path in $\mathcal{F}$ which follows the history of ε. We denote by $p(i)$ the edge which intersects p and lies on the bottom of the corridor at time i. The edges $p(i)$ form the *past* of ε. We will sometimes denote the edge ε by $p(k)$. It will be an analysis of the times at which the weight of $p(i)$ decreases that forms the core of the proof of the theorem.

The weights of the edges $p(0), p(1), \ldots, p(k)$ form a non-increasing sequence. Suppose this sequence is $W = \{w_0, \ldots, w_k\}$. A *drop* in W is a time t such that $w_{t-1} > w_t$. At such times, the edge $p(t)$ is contained in a (folded) seed in the bottom of a corridor of Δ_k.

We will show that either successive drops occur rapidly, or else we reach a situation wherein each time a drop occurs we lose no essential information by restricting our attention to a small subpath of ρ_i.

To make this localisation argument precise, we define *incidents*, which fall into two types.

An *incident of Type A* is a time t which (i) is a drop; and (ii) is such that there is a hard splitting of ρ_t immediately on either side of the folded seed containing $p(t)$.

An *incident of Type B* is a time t such that $p(t-1)$ lies in an indivisible Nielsen path with a hard splitting of ρ_{t-1} immediately on either side, but $p(t)$ does not; except that we do not consider this to be an incident if some ρ_i, for $i \leq t-1$ admits a hard splitting $\rho_i = \sigma_1 \odot \sigma_2 \odot \sigma_3$ with $p(i) \subseteq \sigma_2$ and σ_2 a GEP. In case of an incident of Type B, necessarily $p(t)$ lies in the nibbled future of a Nielsen path on one end of ρ_t with a hard splitting of ρ_t immediately on the other side.

Define the time t_1 to be the last time at which there is an incident (of Type A or Type B). If there are no incidents, let $t_1 = 0$. If this incident is of Type A, the edge $p(t_1)$ lies in a folded seed, call it π, and there is a hard splitting of ρ_{t_1} immediately on either side of π. If the incident is of Type B, the edge

$p(t_1)$ lies in the 1-step nibbled future of a Nielsen path, call this nibbled future π also. In case $t_1 = 0$, let $\pi = \rho$. We will see that there is a bound, α say, on the length of π which depends only on f and n, and not on the choice of π, or the choice of nibbled future. The bound α will be defined solely in terms of Type B incidents. We postpone the proof of the existence of the bound α while we examine the consequences of its existence.

The purpose of isolating the path π is that it is a path of controlled length and the hard splitting [27] of ρ_{t_1} immediately on either side of π means that we need only consider the nibbled future or π. Suppose that π has weight r.

Claim 1: There exists a constant $\beta = \beta(n, \alpha, f)$ so that one of the following must occur:

(i) for some $t_1 \leq i < k$, the edge $p(i)$ lies in a GEP in $f_\#(\rho_{i-1})$ with a hard splitting immediately on either side;
(ii) case (i) does not occur; $k - t_1 > \beta$; and at some time $i \leq t_1 + \beta$, the edge $p(i)$ lies in an indivisible Nielsen path τ in $f_\#(\rho_{i-1})$ with a hard splitting immediately on either side;
(iii) $k - t_1 \leq \beta$; or
(iv) there is a hard splitting of ρ_k immediately on either side of ε.

This claim implies the theorem, modulo the bound on α, as we shall now explain. In case (i), for all $j \geq i$, the edge $p(j)$ lies in the nibbled future of a GEP, so in particular this is true for $\varepsilon = p(k)$. If case (ii) arises then the definition of t_1 implies that for $j \geq i$, the edge $p(j)$ always lies in a path labelled τ with a hard splitting immediately on either side, for otherwise there would be a subsequent incident. Also, the length of this Nielsen path is at most αL^β. If case (iii) arises, then the nibbled future of π at time k has length at most αL^β.

To prove the claim, we define two sequences of numbers $V_\omega, V_{\omega-1}, \ldots, V_1$ and $V'_\omega, V'_{\omega-1}, \ldots, V'_1$, depending on n and f, as follows (where $D'(n)$ is the function from Proposition 2.4.6):

$$\begin{aligned} V_\omega &:= D'(\alpha), \\ V'_\omega &:= V_\omega + \alpha L^{V_\omega}. \end{aligned}$$

For $\omega > i \geq 1$, supposing V'_{i+1} to be defined,

$$V_i := V'_{i+1} + D'(\alpha L^{V'_{i+1}}).$$

Also, supposing V_i to be defined, we define

$$V'_i := V_i + \alpha L^{V_i}.$$

The constants V'_i and V_i are defined so that Proposition 2.4.6 may be applied successively to paths which satisfy Case (4) of the statement of that result.

[27] this splitting is vacuous in case $t_1 = 0$ and at various other points during this proof which we do not explicitly mention

The key point is that at time t_1 we have control over the length of the path π, and so may apply Proposition 2.4.6 to find a hard splitting of any i-step nibbled future of ρ so long as $i \geq D'(|\pi|)$. If we consider the $D'(|\pi|)$-step nibbled future, then we also have control of each of the h-indivisible paths in this hard splitting, and thus we may apply Proposition 2.4.6 again. Note that the paths which satisfy Cases (1)–(3) of Proposition 2.4.6 also satisfy the requirements of Theorem 2.4.2 (given the as yet unproved bound α), so we have to deal with the paths satisfying Case (4). We deal with these by successive applications of Proposition 2.4.6, considering at each weight the paths satisfying Case (4). The constants V_i and V_i' are tuned to allow this induction on weight to occur. Note that since D' is a nondecreasing function (Remark 2.4.8), we have $V_{i+1} < V_{i+1}' < V_i$ for each i.

Consider the situation at time $t_1 + V_r$ (recall that r is the weight of π). Possibly $k \leq t_1 + V_r$, which is covered by case (iii) of our claim, so long as $\beta > V_r$. Therefore, suppose that $k > t_1 + V_r$.

According to Proposition 2.4.6, and the definition of t_1, at time $t_1 + V_r$ the V_r-step nibbled future of π which exists in $\rho_{t_1+V_r}$ admits a hard splitting into edge-paths, each of which is either:

(1) a single edge of weight r;
(2) a nibbled future of a weight r indivisible Nielsen path;
(3) a nibbled future of a weight r GEP; or
(4) a path of weight at most $r-1$.

We need to augment possibility (3) by recalling that Proposition 2.4.6 also shows that the GEP referred to lies in the j-step nibbled future of π for some $j \leq V_r$.

We analyse what happens when the edge $p(t_1 + V_r)$ lies in each of these four types of path.

Case (1): In the first case, by the definition of t_1, there will be a hard splitting of ρ_k immediately on either side of ε, since in this case if there is a drop in W after $t_1 + V_r$ then there is an incident of Type A, contrary to hypothesis.

Case (3): If $p(t_1 + V_r)$ lies in a path of the third type then we are in case (i) of our claim, and hence content.

The fourth type of path will lead us to an inductive argument on the weight of the path under consideration. But first we consider the nibbled futures of Nielsen paths.

Case (2): Suppose that in $\rho_{t_1+V_r}$ the edge $p(t_1 + V_r)$ lies in the nibbled future of a Nielsen path of weight r, with a hard splitting of $\rho_{t_1+V_r}$ immediately on either side. Suppose that this nibbled future is π_r. If π_r is actually a Nielsen path then we lie in case (ii) of our claim. Thus suppose that π_r is not a Nielsen path. It has length at most αL^{V_r}, and within time αL^{V_r} any nibbled future of π_r admits a hard splitting into edge-paths of types (1), (3) and (4) from the above list. The required bound on length is straightforward, since the length of π is at most α and we are considering a sub-path of a V_r-step nibbled future of π (recall that L is the maximum length of paths $L(E)$ for edges E in G).

To see that any nibbled future of π_r admits a splitting of the required form within time αL^{V_r}, consider the three types of indivisible Nielsen paths. If τ is a Nielsen path which is a single edge fixed pointwise by f, then any nibbled future of τ is either a single edge or empty.

Suppose that τ is an indivisible Nielsen path of weight r and H_r is exponential, and suppose that τ' is a proper subpath of τ. Then there is some iterated image $f^l_\#(\tau')$ of τ' which is r-legal. By Proposition 2.4.6 any $D'(\alpha)$-step nibbled future of τ' is r-legal. Since τ has length at most α, so does τ'. Therefore, if $i \geq D'(\alpha)$ then any i-step nibbled future of τ' admits a hard splitting into paths of the required form. Since $V_r > V_\omega = D'(\alpha)$, it is clear that within time $L^{V_r} > D'(\alpha)$, the nibbled futures of τ' admit a hard splitting of the required form.

Finally suppose that $E_i\tau^k\overline{E_j}$ is an indivisible Nielsen path of parabolic weight, with $k \geq 0$. Thus τ is a Nielsen path of weight less than r, and E_i, E_j are edges such that $f(E_i) = E_i \odot \tau^m$, $f(E_j) = E_j \odot \tau^m$. A 1-step nibbled future of $E_i\tau^k\overline{E_j}$ has one of three forms: (I) $E_i\tau^{k_1}\tau'$, where τ' is a proper sub edge-path of τ; (II) $\tau'\tau^{k_2}\tau''$ where τ' and τ'' are proper sub edge-paths of τ; or (III) $\tau'\tau^{k_3}\overline{E_j}$, where τ' is a proper sub edge-path of τ. Note that cases (I) and (III) are not symmetric because we assume that $k \geq 0$ (and hence $k_1, k_2, k_3 \geq 0$ also).

Case 2(I): In this case, $E_i\tau^{k_1}\tau'$ admits a hard splitting into E_i and $\tau^{k_1}\tau'$, which is of the required sort.

Case 2(II): In this case the path already had weight less than r.

Case 2(III): Suppose we are in case (III), and that μ, the αL^{V_r}-step nibbled future of $\tau'\tau^{k_3}\overline{E_j}$ has a copy of $\overline{E_j}$. Lemma 2.4.5 assures us that no nibbled future of τ' can contain τ as a subpath, and therefore there is a splitting of μ immediately on the right of $\overline{E_j}$, and we are done. If there is no copy of $\overline{E_j}$ in μ, we are also done, since this nibbled future must have weight less than r.

Case (4): Having dealt with cases (1), (2) and (3), we may now suppose that at time $t_1 + V_r + \alpha L^{V_r} = t_1 + V'_r$ the edge $p(t_1 + V'_r)$ lies in an edge-path of weight at most $r - 1$ with a hard splitting of $\rho_{t_1+V'_r}$ immediately on either side.[28] Denote this path by π'_r, chosen to be in the future of π. Note that π'_r has length at most $\alpha L^{V'_r}$.

By Proposition 2.4.6 again, either $k < t_1 + V_{r-1}$ or at time $t_1 + V_{r-1}$ the nibbled future of π'_r admits a hard splitting into edge-paths each of which is either:

[28] Note that again it is possible that $k < t_1 + V'_r$, in which case we are in case (iii) of our claim. We suppose therefore that this is not the case.

(1) a single edge of weight $r-1$;
(2) a nibbled future of a weight $r-1$ indivisible Nielsen path;
(3) a nibbled future of a weight $r-1$ GEP; or
(4) a path of weight at most $r-2$.

We continue in this manner. We may conceivably fall into case (4) each time until $t_1 + V_1$ when it is not possible to fall into a path of weight at most $1-1$! Thus at some stage we must fall into one of the first three cases. This completes the proof of Claim 1.

The existence of α. We must find a bound, in terms of n and f, on the length of indivisible Nielsen paths that arise in the nibbled future of ρ with a hard splitting immediately on either side.[29] To this end, suppose that ε' is an edge which lies in an indivisible Nielsen path τ in a k'-step nibbled future of ρ, and that there is a hard splitting immediately on either side of τ. We again denote the i-step nibbled future of ρ by ρ_i for $0 \leq i \leq k'$.

As above, we associate a diagram $\Delta_{k'}$ to $\rho_{k'}$.[30] Denote by q the path in the family forest of $\Delta_{k'}$ which follows the past of ε'. Let $q(i)$ be the edge in ρ_i which intersects q. Let the sequence of weights of the edges $q(i)$ be $W' = \{w'_0, \ldots, w'_{k'}\}$.

Define incidents of Type A and B for W' in exactly the same way as for W, and let t_2 be the time of the last incident of Type A for W'. If there is no incident of Type A for W' let $t_2 = 0$. Let κ be the folded seed containing $q(t_2)$; in case $t_2 = 0$ let $\kappa = \rho$. Define $\theta = \max\{n, L\}$ and note that $|\kappa| \leq \theta$. The path τ must lie in the nibbled future of κ, so it suffices to consider the nibbled future of κ. Suppose that κ has weight r'.

We deal with the nibbled future of κ in the same way as we dealt with that of π. Let $\kappa_0 = \kappa, \kappa_1, \ldots$ be the nibbled futures of κ.

Claim 2: There exists a constant $\beta' = \beta'(n, f)$ so that one of the following must occur:

(i) for some $t_2 < i < k'$, the edge $q(i)$ lies in a GEP in $f_\#(\kappa_{i-1})$ that has a hard splitting immediately on either side;
(ii) not in case (i), *and* at some time $i \leq k'$ the edge $q(i)$ lies in an indivisible Nielsen path τ_0 in $f_\#(\kappa_{i-1})$ so that $|\tau_0| \leq \theta L^{\beta'}$ and immediately on either side of τ_0 there is a hard splitting, *and* there are no incidents of Type B after time i;
(iii) $k' - t_2 \leq \beta'$; or
(iv) there is a hard splitting of $\kappa_{k'}$ immediately on either side of ε'.

[29]Recall that the definition of Type B incidents excluded the case of Nielsen paths which lie in the nibbled future of a GEP with a hard splitting immediately on either side.

[30]If we are considering Nielsen paths arising in the past of ε above, then we can assume $k' \leq k$ and that $\Delta_{k'}$ is a subdiagram of Δ_k in the obvious way.

Let us prove that this claim implies the existence of α and hence completes the proof of the theorem. By definition, α is required to be an upper bound on the length of an arbitrary Nielsen path τ involved in a Type B incident. We assume this incident occurs at time k' and use Claim 2 to analyse what happens.

Case (i) of Claim 2 is irrelevant in this regard. If case (ii) occurs, the futures of τ_0 are unchanging up to time k', so $\tau = \tau_0$ and we have our required bound. In case (iii) the length of τ is at most $\theta L^{\beta'}$, and in case (iv) τ is a single edge. It suffices to let $\alpha = \theta L^{\beta'}$.

It remains to prove Claim 2. The proof of Claim 2 follows that of Claim 1 almost verbatim, with θ in place of α and κ in place of ρ, etc., *except* that the third sentence in Case (2) of the proof becomes invalid because Type B incidents after time $t_2 + V_r$ may occur.

In this setting, suppose π_r (which occurs at time $t_2 + V_r$) is a Nielsen path, but that we are not in case (ii) of Claim 2, and there is a subsequent Type B incident at time j, say. The length of π_r is at most θL^{V_r}. The Nielsen path at time $j-1$ has the same length as the one at time $t_2 + V_r$. We go forward to time j, where the future of π_r is no longer a Nielsen path, and continue the proof of Case (2) from the fourth sentence of the proof.

Otherwise, the proof of Claim 2 is the same as that of Claim 1 (the above modification is required at each weight, but at most *once* for each weight). The only way in which the length bounds change is in the replacement of θ by α (including in the definitions of V_i and V_i'). This finally completes the proof of Theorem 2.4.2. □

2.5. Passing to an Iterate of f

It is important to be able to replace f by an iterate $f_0 = f_\#^k$, for $k \geq 1$. However, when doing this, it is important to be able to retain control over certain constants (since which iterate we choose will depend on some of these constants). In this section we describe what happens to various definitions when we replace f by an iterate. Suppose that $k \geq 1$, and consider the relationship between f and $f_0 = f_\#^k$.

First, for any integer $j \geq 1$, the set of kj-monochromatic paths for f is the same as the set of j-monochromatic paths for f_0. Therefore, once Theorem 2.3.2 is proved, we will pass to an iterate so that r-monochromatic becomes 1-monochromatic. However, the story is not quite as simple as that.

It is not hard to see that if $\sigma \odot \nu$ is a hard splitting for f, then it is also a hard splitting for f_0.

When f is replaced by f_0, the set of GEPs is unchanged, as are the sets of ΨEPs and indivisible Nielsen paths. Also, the set of indivisible Nielsen paths which occur as sub-paths of $f(E)$ for some linear edge E remains unchanged.

With the definition as given, the set of (J, f_0)-atoms may be smaller than the set of (J, f)-atoms. This is because an atom is required to be 1-monochromatic. However, we continue to consider the set of (J, f)-atoms even when we pass to f_0, and we also consider paths to be *beaded* if they are (J, f)-beaded.

Since we are quantifying over a smaller set of paths the constant $V(n, f_0)$ in Theorem 2.4.2 is assumed, without loss of generality, to be $V(n, f)$. This is an important point, because the constant V is used to find the appropriate J when proving Theorem 2.3.2. When passing from f to f_0, we need this J to remain unchanged, for the appropriate iterate k which we eventually choose depends crucially upon J (See Addendum 2.0.1).

It is also clear that if $m \leq n$ then without loss of generality we may assume that $V(m, f) \leq V(n, f)$. Once again, this is because we are considering a smaller set of paths when defining $V(m, f)$.

We now want to replace f by a fixed iterate in order to control some of the cancellation within monochromatic paths. The following lemma is particularly useful in the proof of Proposition 2.6.9 below, and also for Theorem 2.8.1. In particular, it will be used to find the value of d in the Beaded Decomposition Theorem. Lemma 2.5.1 allows us to tune the improved relative train track map in order to exclude some troublesome cancellation phenomena that can otherwise occur in nibbled futures.

LEMMA 2.5.1. *There exists $k_1 \geq 1$ so that $f_1 = f_\#^{k_1}$ satisfies the following. Suppose that E is an exponential edge of weight r and that σ is an indivisible Nielsen path of weight r (if it exists, σ is unique up to a change of orientation). Then*

(1) $|f_1(E)| > |\sigma|$.
(2) Moreover, if σ is an indivisible Nielsen path of exponential weight r and σ_0 is a proper subedge-path of σ, then $(f_1)_\#(\sigma_0)$ is r-legal.
(3) If σ_0 is a proper initial sub edge-path of σ then $(f_1)_\#(\sigma_0)$ admits a hard splitting, $f(E) \odot \xi$, where E is the edge on the left end of σ.
(4) Finally, if σ_1 is a proper terminal sub edge-path of σ then $(f_1)_\#(\sigma_1) = \xi' \odot f(E')$ where E' is the edge on the right end of σ.

Now suppose that σ is an indivisible Nielsen path of parabolic weight r and that σ is a sub edge-path of $f(E_1)$ for some linear edge E_1. The path σ is either of the form $E\eta^{m_\sigma}\overline{E'}$ or of the form $E\overline{\eta}^{m_\sigma}\overline{E'}$, for some linear edges E and E'. Then

(1) If σ_0 is a proper initial sub edge-path of σ then

$$(f_1)_\#(\sigma_0) = E \odot \eta \odot \cdots \odot \eta \odot \xi'',$$

where there are more than m_σ copies of η visible in this splitting.

(2) If σ_1 is a proper terminal sub edge-path of σ then

$$(f_1)_\#(\sigma_1) = \xi' \odot \overline{\eta} \odot \cdots \odot \overline{\eta} \odot \overline{E'},$$

where there are more than m_σ copies of $\overline{\eta}$ visible in this splitting;

PROOF. First suppose that H_r is an exponential stratum, that σ is an indivisible Nielsen path of weight r, and that E is an edge of weight r. Since $|f_\#^j(E)|$ grows exponentially with j, and $|f_\#^j(\sigma)|$ is constant, there is certainly some d_0 so that $|f_\#^d(E)| > |\sigma|$ for all $d \geq d_0$.

There is a single r-illegal turn in σ, and if σ_0 is a proper sub edge-path of σ. By Lemma 2.1.16, no future of σ_0 can contain σ as a subpath. The number of r-illegal turns in iterates of σ_0 must stabilise, so by Lemma 2.1.12 there is an iterate of σ_0 which is r-legal. Since there are only finitely many paths σ_0, we can choose an iterate of f which works for all such σ_0.

Suppose now that σ_0 is a proper initial sub edge-path of σ, and that E is the edge on the left end of σ. It is not hard to see that every (entire) future of σ_0 has E on its left end. We have found an iterate of f so that $f_\#^{d'}(\sigma_0)$ is r-legal. It now follows immediately that

$$f_\#^{d'+1}(\sigma_0) = f(E) \odot \xi,$$

for some path ξ. The case when σ_1 is a proper terminal sub edge-path of σ is identical.

Now suppose that H_r is a parabolic stratum and that σ is an indivisible Nielsen path of weight r of the form in the statement of the lemma. The claims about sub-paths of σ follow from the hard splittings $f(E) = E \odot u_E$ and $f(E') = E' \odot u_{E'}$, and from the fact that m_σ is bounded because σ is a subpath of some $f(E_1)$.

As in Remark 2.4.7, we can treat each of the cases separately, and finally take a maximum. □

2.6. The Nibbled Futures of GEPs

In this section f is an improved relative train track map, although we do not suppose yet that we have replaced f by an iterate so that Lemma 2.5.1 holds with $k_1 = 1$.

The entire future of a GEP is a GEP but a nibbled future need not be and Theorem 2.4.2 tells us that we need to analyse these nibbled futures. This analysis will lead us to define *proto-*ΨEP*s*. In Proposition 2.6.9, we establish a normal form for proto-ΨEPs which proves that proto-ΨEPs are in fact the ΨEPs which appear in the Beaded Decomposition Theorem.

To this end, suppose that

$$\zeta = E_i \overline{\tau}^n \overline{E_j}$$

is a GEP, where τ is a Nielsen path, $f(E_i) = E_i \odot \tau^{m_i}$ and $f(E_j) = E_j \odot \tau^{m_j}$. As in Definition 2.1.6, we consider $E_i \overline{\tau}^n \overline{E_j}$ to be unoriented, but here we do not suppose that $j \leq i$. However, we suppose $n > 0$ and thus, since $E_i \overline{\tau}^n \overline{E_j}$ is a GEP, $m_j > m_i > 0$.

The analysis of GEPs of the form $E_j \tau^n \overline{E_i}$ is entirely similar to that of GEPs of the form $E_i \bar{\tau}^n \overline{E_j}$ except that one must reverse all left-right orientations. Therefore, we ignore this case until Definition 2.6.2 below (and often afterwards also!).

We fix a sequence of nibbled futures $\zeta = \rho_{-l}, \ldots, \rho_0, \rho_1, \ldots, \rho_k, \ldots$ of ζ, where ρ_0 is the first nibbled future which is not the entire future. Since the entire future of a GEP is a GEP, we restrict our attention to the nibbled futures of ρ_0.

There are three cases to consider, depending on the type of sub-path on either end of ρ_0.

(1) $\rho_0 = \bar{\tau}_0 \bar{\tau}^m \overline{E_j}$;
(2) $\rho_0 = \bar{\tau}_0 \bar{\tau}^m \bar{\tau}_1$.
(3) $\rho_0 = E_i \bar{\tau}^m \bar{\tau}_1$;

where τ_0 is a (possibly empty) initial sub edge-path of τ, and τ_1 is a (possibly empty) terminal sub edge-path of τ.

In case (1) ρ_0 admits a hard splitting

$$\rho_0 = \overline{\tau}_0 \odot \overline{\tau} \odot \cdots \odot \overline{\tau} \odot \overline{E}_j.$$

Since τ_0 is a sub edge-path of $f(E_i)$, it has length less than L and its nibbled futures admit hard splittings as in Theorem 2.4.2 into nibbled futures of GEPs and paths of length at most $V(L, f)$. These GEPs will necessarily be of strictly lower weight than ρ_0, since $\overline{\tau}_0$ is. Thus, case (1) is easily dealt with by an induction on weight, supposing that we have a nice splitting of the nibbled futures of lower weight GEPs; this is made precise in Proposition 2.6.10. Case (2) is entirely similar.

Case (3) is by far the most troublesome of the three, and it is this case which leads to the definition of *proto*-ΨEP*s* in Definition 2.6.2 below. Henceforth assume $\rho_0 = E_i \bar{\tau}^m \bar{\tau}_1$.

Each of the nibbled futures of ρ_0 (up to the moment of death, Subsection 2.6.1) has a nibbled future of $\overline{\tau}_1$ on the right. If the latter becomes empty at some point, the nibbled future of ρ_0 at this time has the form $E_i \overline{\tau}^{n'} \overline{\tau}_2$, where τ_2 is a proper (but possibly empty) sub edge-path of τ. We restart our analysis at this moment. Hence we make the following

Working Assumption 2.6.1. *We make the following two assumptions on the k-step nibbled futures considered:*

(1) $\rho_0 = E_1 \overline{\tau}^m \overline{\tau}_1$*;*
(2) all nibbling of ρ_k *occurs on the right; and*
(3) the k-step nibbled future $\overline{\tau}_{1,k}$ *of* $\overline{\tau}_1$ *inherited from* ρ_k *is non-empty.*

We will deal with the case $m - km_i < 0$ later, in particular with the value of k for which $m - (k-1)m_i \geq 0$ but $m - km_i < 0$. For now suppose that $m - km_i \geq 0$.

In this case, the path ρ_k has the form

$$\rho_k = E_i \overline{\tau}^{m-km_i} \overline{\tau}_{1,k}.$$

There are (possibly empty) Nielsen edge-paths ι and ν, and an indivisible Nielsen edge-path σ so that

$$\tau = \iota \odot \sigma \odot \nu \text{ and } \tau_1 = \sigma_1 \odot \nu, \tag{2.6.1}$$

where σ_1 is a proper terminal sub edge-path of σ. Now, as in Working Assumption 2.6.1, there is no loss of generality in supposing that

$$\rho_k = E_i \overline{\tau}^{m-km_i} \overline{\nu} \overline{\sigma}_{1,k},$$

where $\overline{\sigma}_{1,k}$ is the nibbled future of $\overline{\sigma}_1$ inherited from ρ_k, and that $\overline{\sigma}_{1,k}$ is non-empty.

Since $|\sigma_1| < L$, by Theorem 2.4.2 the path $\sigma_{1,k}$ admits a hard splitting into edge-paths each of which is either the nibbled future of a GEP, or of length at most $V(L, f)$; we take the (unique) maximal hard splitting of $\sigma_{1,k}$ into edge-paths.

Let $s = \lfloor m/m_i \rfloor + 1$. In ρ_s (but not before) there may be some interaction between the future of E_i and $\overline{\sigma}_{1,s}$. We denote by $\gamma_{\sigma_1}^{k,m}$ the concatenation of those factors in the hard splitting of $\overline{\sigma}_{1,k}$ which contain edges any part of whose future is eventually cancelled by some edge in the future of E_i under any choice of nibbled futures of ρ_k (not just the ρ_{k+t} chosen earlier) and any choice of tightening. Below we will analyse more carefully the structure of the paths $\overline{\sigma}_{1,k}$ and $\gamma_{\sigma_1}^{k,m}$.

We now have $\overline{\sigma}_{1,k} = \gamma_{\sigma_1}^{k,m} \odot \sigma_{1,k}^{\bullet}$. From (2.6.1), we also have

$$\rho_k = E_i \overline{\tau}^{m-km_i} \overline{\nu} \gamma_{\sigma_1}^{k,m} \odot \sigma_{1,k}^{\bullet}. \tag{2.6.2}$$

DEFINITION 2.6.2 (Proto-ΨEPs). Suppose that τ is a Nielsen edge-path, E_i a linear edge such that $f(E_i) = E_i \odot \tau^{m_i}$ and τ_1 a proper terminal sub edge-path of τ such that $\tau_1 = \sigma_1 \odot \nu$ as in (2.6.1). Let $k, m \geq 0$ be such that $m - km_i \geq 0$ and let $\gamma_{\sigma_1}^{k,m}$ be as in (2.6.2). A path π is called a *proto*-ΨEP if either π of $\overline{\pi}$ is of the form

$$E_i \overline{\tau}^{m-km_i} \overline{\nu} \gamma_{\sigma_1}^{k,m}.$$

REMARKS 2.6.3.

(1) The definition of proto-ΨEPs is intended to capture those paths which remain when a GEP is partially cancelled, leaving a path which may shrink in size of its own accord.
(2) By definition, a proto-ΨEP admits no non-vacuous hard splitting into edge-paths.

We now introduce two distinguished kinds of proto-ΨEPs.

DEFINITION 2.6.4. Suppose that

$$\pi = E_i \overline{\tau}^{m-km_i} \overline{\nu} \gamma_{\sigma_1}^{k,m},$$

is a proto-ΨEP as in Definition 2.6.2.

The path π is a *transient* proto-ΨEP if $k = 0$.

The path π is a *stable* proto ΨEP if $\gamma_{\sigma_1}^{k,m}$ is a single edge.

LEMMA 2.6.5. *A transient proto-*ΨEP *is a* ΨEP.

PROOF. With the notation of Definition 2.6.2, in this case $\gamma_{\sigma_1}^{0,m}$ is visibly a sub-path of $\bar{\tau}$, and the proto-ΨEP is visibly a sub-path of a GEP. □

LEMMA 2.6.6. *A stable proto-*ΨEP *is a* ΨEP.

PROOF. Since $\bar{\sigma}$ is a Nielsen path, if α is a nibbled future of $\bar{\sigma}$ where all the nibbling has occurred on the right, then the first edge in α is the same as the first edge in $\bar{\sigma}$.

On the other hand, $\gamma_{\sigma_1}^{k,m}$ is a nibbled future of $\bar{\sigma}$ where all the nibbling has occurred on the right. Therefore, if $\gamma_{\sigma_1}^{k,m}$ is a single edge then it must be a sub-path of $\bar{\sigma}$. It follows immediately that any stable proto-ΨEP must be a ΨEP. □

REMARK 2.6.7. We will prove in Proposition 2.6.9 that after replacing f by a suitable iterate all proto-ΨEPs are either transient or stable, and hence are ΨEPs.

2.6.1. The Death of a proto-ΨEP. Suppose that $\pi = E_i \overline{\tau}^{m-km_i} \overline{\nu} \gamma_{\sigma_1}^{k,m}$ is a proto-ΨEP with nibbled futures satisfying Assumption 2.6.1. Let $q = \lfloor \frac{m-km_i}{m_i} \rfloor + 1$, and consider, π_{q-1}, a $(q-1)$-step nibbled future of π. As before, we assume that the $(q-1)$-step nibbled future of $\gamma_{\sigma_0}^{k,m}$ inherited from a π_{q-1} is not empty and that the edge labelled E_i on the very left is not nibbled.

In π_q, the edge E_i has consumed all of the copies of $\overline{\tau}$ and begins to interact with the future of $\overline{\nu}\gamma_{\sigma_1}^{k,m}$. Also, the future of π at time q need not contain a ΨEP. Hence we refer to the time q as the *death of the* ΨEP. Recall that $\tau = \iota \odot \sigma \odot \nu$ and that $\gamma_{\sigma_1}^{k,m}$ is a k-step nibbled future of $\overline{\sigma}_1$, where σ_1 is a proper subpath of σ. Let $p = m - (k+q-1)m_i$, so that $0 \leq p < m_i$.

The path π_{q-1} has the form

$$\pi_{q-1} = E_i \overline{\tau}^p \overline{\nu} \gamma_{\sigma_1}^{k+q-1,m}.$$

Suppose that π_q is a 1-step nibbled future of π_{q-1}. In other words, π_q is a subpath of $f_\#(\pi_{q-1})$. Consider what happens when $f(\pi_{q-1})$ is tightened to form $f_\#(\pi_{q-1})$ (with any choice of tightening). The p copies of $\overline{\tau}$ (possibly in various stages of tightening) will be consumed by E_i, leaving $\overline{\nu} \odot f(\gamma_{\sigma_1}^{k+q-1,m})$ to interact with at least one remaining copy of $\tau = \iota \odot \sigma \odot \nu$. The paths ν and $\overline{\nu}$ will cancel with each other[31].

[31]The hard splittings imply that this cancellation must occur under *any* choice of tightening.

Lemma 2.4.5 states that $\gamma_{\sigma_1}^{k,m}$ cannot contain σ as a subpath. Therefore, once ν and $\overline{\nu}$ have cancelled, not all of $\overline{\sigma}$ will cancel with $f(\gamma_{\sigma_1}^{k+q-1,m})$. A consequence of this discussion (and the fact that $f(E_i) = E_i \odot \tau^{m_i})$ is the following

LEMMA 2.6.8. *Suppose that* $\pi = E_i\overline{\tau}^{m-km_i}\overline{\nu}\gamma_{\sigma_0}^{k,m}$ *is a proto-*ΨEP*, and let* $q = \lfloor \frac{m-km_i}{m_i} \rfloor + 1$. *Suppose that* π_{q-1} *is a* $(q-1)$*-step nibbled future of* π *satisfying Assumption 2.6.1. If* π_q *is an immediate nibbled future of* π_{q-1} *and* π_q *contains* E_i *then* π_q *admits a hard splitting*

$$\pi_q = E_i \odot \lambda.$$

We now analyse the interaction between $f(\gamma_{\sigma_1}^{k+q-1,m})$ and σ more closely. As usual, there are two cases to consider, depending on whether σ has exponential or parabolic weight[32].

In the following proposition, f_1 is the iterate of f from Lemma 2.5.1 and we are using the definitions as explained in Section 2.5. Also, we assume that proto-ΨEPs are defined using f_1, not f.

PROPOSITION 2.6.9. *Every proto-*ΨEP *for* f_1 *is either transient or stable. In particular, every proto-*ΨEP *for* f_1 *is a* ΨEP*.*

PROOF. Let $\pi = E_i\overline{\tau}^{m-km_i}\overline{\nu}\gamma_{\sigma_1}^{k,m}$ be a proto-ΨEP for f_1.

Lemma 2.6.5 implies that if $k = 0$ then π is a ΨEP. Consider Working Assumption 2.6.1. If Assumption 2.6.1.(2) fails to hold at any point, then we can restart our analysis, and in particular we have a transient proto-ΨEP at this moment. Thus we may suppose that π is an initial sub-path of a k-step nibbled future of a GEP, where $k \geq 1$ and we may further suppose that π satisfies Assumption 2.6.1.(2). We prove that in this case π is a stable proto-ΨEP.

First suppose that σ has exponential weight, r say. If $\overline{\sigma}_0$ is a proper initial sub edge-path of $\overline{\sigma}$ then Lemma 2.5.1 asserts that

$$(f_1)_\#(\sigma_0) = f(E) \odot \xi,$$

and $|f(E)| > |\sigma|$. Note also that $f(E) = E \odot \xi''$ for some path ξ''.

Now, at the death of the proto-ΨEP, the nibbled future of $\gamma_{\sigma_0}^{k,m}$ interacts with a copy of E_i, and in particular with a copy of $f(\sigma)$ (in some stage of tightening). Now the above hard splitting, and the fact that σ is not r-legal whilst $f(E)$ is, shows that $\gamma_{\sigma_1}^{k,m}$ must be a single edge (namely E).

Suppose now that σ has parabolic weight. Since σ has proper sub edge-paths, it is not a single edge and so σ or $\overline{\sigma}$ has the form $E\eta^{m_\sigma}\overline{E'}$. The hard

[32]Recall that there are three kinds of indivisible Nielsen paths: constant edges, parabolic weight and exponential weight. If σ has nontrivial proper sub edge-paths, then it is certainly not a single edge, constant or not.

splittings guaranteed by Lemma 2.5.1 now imply that $\gamma_{\sigma_1}^{k,m}$ is a single edge in this case also.

Therefore, every proto-ΨEP for f_1 is transient or stable, proving the first assertion of the proposition. The second assertion follows from the first assertion, and Lemmas 2.6.5 and 2.6.6. □

Finally, we can prove the main result of this section. In the following, L_1 is the maximum length of $f_1(E)$ over all edges E of G.

The following statement assumes the conventions of Section 2.5.

PROPOSITION 2.6.10. *Under iteration of the map f_1 constructed in Lemma 2.5.1, any nibbled future of a* GEP *admits a hard splitting into edge-paths, each of which is either a* GEP*, a* ΨEP*, or of length at most $V(2L_1, f)$.*

PROOF. Suppose that $E_i\overline{\tau}^n\overline{E_j}$ is a GEP of weight r. We may suppose by induction that any nibbled future of any GEP of weight less than r admits a hard splitting of the required form (the base case $r = 1$ is vacuous, since there cannot be a GEP of weight 1).

Suppose that ρ is a nibbled future of $E_i\tau^n\overline{E_j}$. If ρ is the entire future, it is a GEP and there is nothing to prove. Otherwise, as in the analysis at the beginning of this section, we consider the first time when a nibbled future is not the entire future. Let the nibbled future be ρ_0. In cases (1) and (2) from that analysis, ρ_0 admits a hard splitting into edge-paths, each of which is either (i) $\overline{E}_i$; (ii) $\overline{\tau}$; or (iii) a proper sub edge-path of τ. In each of these cases, Theorem 2.4.2 asserts that there is a hard splitting of ρ into edge-paths, each of which is either of length at most $V(L, f)$ or is the nibbled future of a GEP. Any nibbled future of a GEP which occurs in this splitting is necessarily of weight strictly less than r, and so admits a hard splitting of the required form by induction.

Suppose then that ρ_0 satisfies Case (3), the third of the cases articulated at the beginning of this section. In this case, ρ_0 is a transient proto-ΨEP. Also, any time that Assumption 2.6.1.(2) is not satisfied, the nibbled future of ρ_0 is a transient proto-ΨEP. Thus, we may assume that Assumption 2.6.1 is satisfied. If $m - km_i \geq 0$ then we have

$$\rho = E_i\overline{\tau}^{m-km_i}\overline{\nu}\gamma_{\sigma_1}^{k,m} \odot \sigma_{1,k}^{\bullet}.$$

The first path in this splitting is a stable ΨEP by Proposition 2.6.9. Once again, Theorem 2.4.2 and the inductive hypothesis yield a hard splitting of $\sigma_{1,k}^{\bullet}$ of the required form.

Finally, suppose that Case (3) pertains and $m - km_i < 0$. Let $q = \lfloor \frac{m-km_i}{m_i} \rfloor + 1$ (the significance of this moment – "the death of the ΨEP" – was explained at the beginning of this subsection). By the definition of a ΨEP (Definition 2.6.2), the q-step nibbled future of ρ_0 admits a hard splitting as

$$E_i\overline{\tau}^{m-qm_i}\overline{\nu}\gamma_{\sigma_1}^{q,m} \odot \sigma_{1,q}^{\bullet}.$$

By Lemma 2.6.8, the immediate future of $E_i\overline{\tau}^{m-qm_i}\overline{\nu}\gamma_{\sigma_1}^{q,m}$ admits a hard splitting as $E_i \odot \xi$. Since $\gamma_{\sigma_1}^{r,m}$ is a single edge, we have a bound of $2L_1$ on the length of ξ. Any nibbled future of $E_i \odot \xi$ now admits a hard splitting into edge-paths, each of which is either a GEP, a ΨEP or of length at most $V(2L_1, f)$, by induction on weight and Theorem 2.4.2. □

We highlight one consequence of Proposition 2.6.10:

COROLLARY 2.6.11. *Suppose that $\rho = E_i\overline{\tau}^{m-km_i}\overline{\nu}\gamma$ is a* ΨEP. *Any immediate nibbled future of ρ (with all nibbling on the right) has one of the following two forms:*

(1) $\rho' \odot \sigma$, where ρ' is a ΨEP *and σ admits a hard splitting into atoms; or*

(2) $E_i \odot \sigma$, where σ admits a hard splitting into atoms.

In particular, this is true of $f_\#(\rho)$.

There are entirely analogous statements in case ρ is a ΨEP *where $\overline{\rho}$ has the above form and all nibbling occurs on the left.*

2.7. Proof of the Beaded Decomposition Theorem

In this section, we finally prove Theorem 2.3.5. As noted in Remark 2.3.6, this immediately implies the Beaded Decomposition Theorem.

PROOF (THEOREM 2.3.5). Take $d = k_1$, the constant from Lemma 2.5.1. Let L_1 be the maximum length of $f_\#^{k_1}(E)$ for any edge $E \in G$, let $s = \max\{2L_1, n\}$, and let $J = V(s, f)$, where V is the constant from Theorem 2.4.2.

Suppose that ρ is a path so that $|\rho| \leq n$, and let ρ' be a kd-step nibbled future of ρ for some positive integer k. Then ρ' is a k-step nibbled future of ρ with respect to $f_1 = f_\#^{k_1}$. By Proposition 2.6.9, every proto-ΨEP for f_1 is a ΨEP.

By Theorem 2.4.2, ρ' admits a hard splitting into edge-paths, each of which is either the nibbled future of a GEP or else has length at most $V(n, f)$. By Proposition 2.6.10, if we replace f by f_1 then any nibbled future of a GEP admits a hard splitting into edge-paths, each of which is either a GEP, a ΨEP or else has length at most $V(2L_1, f)$. By Lemma 2.2.8, the splitting of the nibbled future of a GEP is inherited by ρ.

We have shown that ρ is (J, f)-beaded, as required. □

PROOF (ADDENDUM 2.3.5). We have already remarked that, for a fixed m, the constant $V(m, f)$ from Theorem 2.4.2 remains unchanged when f is replaced by an iterate.

As in Section 2.5, we retain the notion of (J, f)-beaded with the original f when passing to an iterate of f

Therefore, when f is replaced by an iterate, Theorem 2.3.5 remains true with the same constant J. This immediately implies that the same is true of the Beaded Decomposition, which is what we were required to prove. □

2.8. Refinements of the Beaded Decomposition Theorem

The Beaded Decomposition Theorem is the main result of Part 2. In this section, we provide a few further refinements that will be required for future applications.

Throughout this section we suppose that f has been replaced with f_1 from Lemma 2.5.1, whilst maintaining the conventions for definitions from Section 2.5. When we refer to f we mean this iterate f_1. With this in mind, a *monochromatic path* is a 1-monochromatic path for f. Similarly, armed with Theorem 2.3.2, we refer to (J, f)-beads, simply as *beads*, and a path which is (J, f)-beaded will be referred to simply as *beaded*. The constant L now refers to the maximum length $|f(E)|$ for edges $E \in G$ *with the new* f.

In the following theorem, the *past* of an edge is defined with respect to an arbitrary choice of tightening.

THEOREM 2.8.1. *There exists a constant D_1, depending only on f, with the following properties. Suppose $i \geq D_1$, that χ is a monochromatic path and that ε is an edge in $f^i_\#(\chi)$ of weight r whose past in χ is also of weight r. Then ε is contained in an edge-path ρ so that $f^i_\#(\chi)$ has a hard splitting immediately on either side of ρ and ρ is one of the following:*

(1) a Nielsen path;
(2) a GEP*;*
(3) a ΨEP*; or*
(4) a single edge.

PROOF. Let χ be a monochromatic path. For any $k \geq 0$, denote $f^k_\#(\chi)$ by χ_k. In a sense, we prove the theorem 'backwards', by fixing an edge ε_0 of weight r in $\chi_0 = \chi$ and considering its futures in the paths χ_k, $k \geq 1$. The purpose of this proof is to find a constant D_1 so that if ε is any edge of weight r in χ_i with past ε_0, and if $i \geq D_1$ then we can find a path ρ around ε satisfying one of the conditions of the statement of the theorem.

Fix $\varepsilon_0 \in \chi_0$. By Theorem 2.3.2, there is an edge-path π containing ε_0 so that χ admits a hard splitting immediately on either side of π and π either (I) is a GEP; (II) has length at most J; or (III) is a ΨEP. In the light of Remark 2.4.7, it suffices to establish the existence of a suitable D_1 in each case. To consider the futures of ε_0 in the futures $f^k_\#(\chi)$ of χ, it suffices to consider the futures of ε_0 within the (entire) futures of π. Therefore, for $k \geq 0$, let $\pi_k = f^k_\#(\pi)$. Suppose that we have chosen, for each k, an edge ε_k in π_k such that: (i) ε_k lies in the future of ε_0; (ii) ε_k has the same weight as ε_0; and (iii) ε_k is in the future of ε_{k-1} for all $k \geq 1$.

Case (I): π is a GEP. In this case, the path π_k is a GEP for all k, any future of ε_0 lies in π_k, and there is a hard splitting of χ_k immediately on either side of π_k. Therefore, the conclusion of the theorem holds in this case with $D_1 = 1$.

Case (II): $|\pi| \leq J$. Denote the weight of π by s. Necessarily $s \geq r$. By Lemma 2.4.4 the path $\pi_{D(J)}$ admits a hard splitting into edge-paths, each of which is either

(1) a single edge of weight s;
(2) an indivisible Nielsen path of weight s;
(3) a GEP of weight s; or
(4) a path of weight at most $s-1$.

We consider which of these types of edge-paths our chosen edge $\varepsilon_{D(J)}$ lies in. In case (1) there is a hard splitting of $\pi_{D(J)}$ immediately on either side of the edge $\varepsilon_{D(J)}$, so for all $i \geq D(J)$ there is a hard splitting of π_i immediately on either side of ε_i, since ε_i and $\varepsilon_{D(J)}$ both have the same weight as ε_0. For cases (2) and (3), $\varepsilon_{D(J)}$ lies in an indivisible Nielsen path or GEP with a hard splitting of $\pi_{D(J)}$ immediately on either side, so for all $i \geq D(J)$ any future of ε_0 in π_i, and in particular ε_i, lies in an indivisible Nielsen path of GEP immediately on either side of which there is a hard splitting of π_i.

Finally, suppose we are in case (4) and not in any of cases (1)–(3). Then $\varepsilon_{D(J)}$ lies in an edge-path $\tilde{\rho}$ with a hard splitting of $\pi_{D(J)}$ immediately on either side, and that $\tilde{\rho}$ is not a single edge, an indivisible Nielsen path, or a GEP[33]. We need only consider the future of $\tilde{\rho}$. For $k \geq 0$, let $\rho_{D(J)+k} = f_\#^k(\tilde{\rho})$ be the future of $\tilde{\rho}$ in $\pi_{D(J)+k}$. Now, $|\tilde{\rho}| \leq JL^{D(J)}$ so by Lemma 2.4.4 the edge-path $\rho_{D(J)+D(JL^{D(J)})}$ admits a hard splitting into edges paths, each of which is either

(1) a single edge of weight $s-1$;
(2) an indivisible Nielsen path of weight $s-1$;
(3) a GEP of weight $s-1$; or
(4) a path of weight at most $s-2$.

We proceed in this manner. If we ever fall into one of the first three cases, we are done. Otherwise, after $s-r+1$ iterations of this argument, the fourth case describes a path of weight strictly less than r. Since the weight of each ε_i is r, it cannot lie in such a path, and one of the first three cases must hold. Thus we have found the required bound D_1 in the case that $|\pi| \leq J$.

Case (III): π is a ΨEP.

Let $\pi = E_i \overline{\tau}^{m-km_i} \overline{\nu} \gamma_{\sigma_1}^{k,m}$ as in Definition 2.6.2. We consider where in the path π the edge ε_0 lies. First of all, suppose that ε_0 is the unique copy of E_i. Since ε_0 is parabolic, it has a unique weight s future at each moment in time. Let $q = \lfloor \frac{m-km_i}{m_i} \rfloor + 1$, the moment of death. For $1 \leq p \leq q-1$, the edge ε_p

[33] In this case necessarily $s \leq r-1$

is the leftmost edge in a ΨEP and there is a hard splitting of π_p immediately on either side of this ΨEP. For $p \geq q$, Lemma 2.6.8 ensures that there is a hard splitting of π_p immediately on either side of ε_p. Therefore in this case the conclusion of the theorem holds with $D_1 = 1$.

Now suppose that the edge ε_0 lies in one of the copies of $\overline{\tau}$ in π, or in the visible copy of $\overline{\nu}$. Then any future of ε_0 lies in a copy of τ or ν respectively, which lies in a ΨEP with a hard splitting immediately on either side, until this copy of $\overline{\tau}$ or $\overline{\nu}$ is consumed by E_i. Again, the conclusion of the theorem holds with $D_1 = 1$.

Finally, suppose that ε_0 lies in $\gamma_{\sigma_1}^{k,m}$. For ease of notation, for the remainder of the proof γ will denote $\gamma_{\sigma_1}^{k,m}$. By Proposition 2.6.9 γ is a single edge. Until the q-step nibbled future of π, any future of γ of the same weight is either γ or will have a splitting of π immediately on either side.

Since σ is an indivisible Nielsen path, and γ is a single edge, γ is the leftmost edge of $\overline{\sigma}$. Therefore $[\sigma\gamma]$ is a proper sub edge-path of σ.

Suppose that σ has exponential weight (this weight is r). By Lemma 2.5.1 and the above remark, $f_\#(\sigma\gamma)$ is r-legal. Therefore, any future of γ which has weight r will have, at time q and every time afterwards, a hard splitting immediately on either side.

Suppose now that σ has parabolic weight r. Since $[\sigma E]$ is a proper sub edge-path of σ, and since there is a single edge of weight r in $f(E)$ and this is cancelled, it is impossible for γ to have a future of weight r after time q. □

Recall that the number of strata for the map $f : G \to G$ is ω. Recall also the definition of *displayed* from Definition 2.2.17

Lemma 2.8.2. *Let χ be a monochromatic path. Then the number of displayed* ΨEP*s in χ of length more than J is less than 2ω.*

Proof. Suppose that χ is a monochromatic path, and that ρ is a subpath of χ, with a hard splitting immediately on either side, such that ρ is a ΨEP, and $|\rho| > J$. Then, tracing through the past of χ, the past of ρ must have come into existence because of nibbling on one end of the past of χ. Suppose this nibbling was from the left. Then all edges to the left of ρ in χ have weight strictly less than that of ρ, since it must have come from a proper subpath of an indivisible Nielsen path in the nibbled future of the GEP which became ρ. Also, any ΨEP to the left of ρ must have arisen due to nibbling from the left. Therefore, there are at most ω ΨEPs of length more than J which came about due to nibbling from the left. The same is true for ΨEPs which arose through nibbling from the right. □

Lemma 2.8.3. *Let D_1 be the constant from Theorem 2.8.1, and let $f_2 = (f_1)_\#^{D_1}$. If ρ is an atom, then either $(f_2)_\#^{\omega}(\rho)$ is a beaded path all of whose beads are Nielsen paths and* GEP*s, or else there is some displayed edge $\varepsilon \subseteq (f_2)_\#^{\omega}(\rho)$*

so that all edges in $(f_2)^{\omega}_{\#}(\rho)$ *whose weight is greater than that of* ε *lie in Nielsen paths and* GEP*s.*

PROOF. Suppose that ρ is an atom of weight r. If H_r is a zero stratum and $(f_2)_{\#}(\rho)$ has weight s then H_s is not a zero stratum. Thus, by going forwards one step in time if necessary, we suppose that H_r is not a zero stratum, so $(f_2)_{\#}(\rho)$ has weight r.

By Theorem 2.8.1, all edge of weight r in $(f_2)_{\#}(\rho)$ are either displayed or lie in Nielsen paths or GEPs (since we are considering the entire future of an atom, ΨEPs do not arise here). If all edge of weight r in $(f_2)_{\#}(\rho)$ lie in Nielsen paths or GEPs then we consider the atoms in $(f_2)_{\#}(\rho)$ of weight less than r (this hard splitting exists since ρ and hence $(f_2)_{\#}(\rho)$ are monochromatic paths). We now consider the immediate future of these atoms in $(f_2)^2_{\#}(\rho)$, etc. It is now clear that the statement of the lemma is true. □

Finally, we record an immediate consequence of the Beaded Decomposition Theorem and Proposition 2.6.10:

THEOREM 2.8.4. *Suppose that* σ *is a beaded path. Any nibbled future of* σ *is also beaded.*

Part 3

The General Case

In Part 3, we bring together the techniques developed in Parts 1 and 2 to prove the main result of this book.

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

In Part 1 we proved the Main Theorem in the case of *positive* automorphisms. That proof proceeded via an analysis of van Kampen diagrams in the universal cover of the mapping torus $R \times [0,1]/\langle (x,0) \sim (f(x),1)\rangle$, where R is a 1-vertex graph with fundamental group F and f is the obvious homotopy equivalence with $f_* = \phi$.

Such f are the prototypes for the *improved relative train track maps* of Bestvina, Feighn and Handel [**4**]. In Part 2 we refined the train track technology in pursuit of topological representatives of arbitrary automorphisms that share with the prototypes f features that proved crucial in Part 1. We identified *beads* as the basic units of an edge-path that play the role in the general setting that single edges (letters) played in the case of positive automorphisms. The claim of beads to this role was underscored by the Beaded Decomposition Theorem.

With these technical innovations in hand, we now set about the task of adapting the arguments of Part 1 to the general case, following the proof from Part 1 as closely as possible and providing the (often fierce) technical details needed to translate each step into the more general context provided by the topological representatives constructed in Part 2. We shall not repeat the proofs of technical lemmas from Part 1 when the adaptation is obvious. Nor shall we repeat our account of the intuition underlying our overall strategy of proof and intermediate strategies at key stages.

Unfortunately, the adaptation to the general case is not entirely smooth. Thus at times we are obliged to break from the narrative that parallels Part 1 in order to deal with phenomena that do not arise in the case of positive automorphisms — Section 3.7, for example. But we as far as possible we have organised matters so that, having taken account of the new phenomena, we can return to the main narrative with the new phenomena controlled and packaged into concise terminology. Thus, with considerable technical exertions in our wake, we are able to arrange matters so that the final stages of the proof of our Main Theorem consist only of references to the corresponding sections of Part 1 with a brief explanation of what changes, if any, must be made in the general setting.

We have already noted that, from the analysis of improved relative train tracks in Part 2, it emerged that beads are the correct analogue for the role played by 'letters' in the positive case. An important manifestation of this is that the Main Theorem can be reduced to a statement concerning the existence of a linear bound (in terms of $|\partial\Delta|$) on the number of beads along the bottom of any corridor in a van Kampen diagram Δ in the universal cover of the mapping tori that we consider. In contrast to the positive case, however, the existence of such a bound does not immediately imply the Main Theorem, because there is no global bound on the length of a bead.

Nevertheless, proving a bound on the number of beads is by far the bulk of our work, occupying Sections 3.6–3.11, which closely follow Sections 1.6–1.10 (with different numbering and modified structure). In Section 3.12 we explain how the bound on the number of beads, together with the ideas from the *Bonus Scheme* in Section 3.11, finally gives the Main Theorem.

In Section 3.13 we explain how to deduce estimates on the geometry of van Kampen diagrams for *all* mapping tori of free group automorphisms from the specially-crafted ones that we work with during our main proof. The key estimate – the linear bound on the length of t-corridors – when reformulated algebraically, yields the *Bracketing Theorem* stated in the introduction.

In Section 3.14 we explain how our proof of the Main Theorem allows one to reprove the main result of [**19**].

We suggest that readers approach Part 3 as follows. First, they must be familiar with the structure of the argument in Part 1 and the vocabulary of beads in Part 2. This will enable them to skim smoothly through Sections 3.1–3.4 of the current paper. Next, they can gain an accurate overview of the proof of the Man Theorem reading the introduction to each of Sections 3.1–3.12 together with the titles of their subsections (and the introductions to subsections when they exist). There is then no alternative but to delve into the details of the proof.

Section 3.13 can be read independently. The argument in Section 3.14 is easy to understand in outline, but the proof appeals to detailed results from Sections 3.6, 3.10 and 3.11.

3.1. The Structure of Diagrams

Associated to any finite group-presentation $\Gamma = \langle \mathcal{A} \mid \mathcal{R} \rangle$ one has the standard combinatorial 2-complex $K(\mathcal{A} : \mathcal{R})$ with fundamental group Γ and directed edges labelled by the $a \in \mathcal{A}$. There is a 1-1 correspondence between words in the letters $\mathcal{A}^{\pm 1}$ and combinatorial loops in the 1-skeleton of $K(\mathcal{A} : \mathcal{R})$. Words such that $w = 1$ in Γ correspond to loops that are null-homotopic. Van Kampen's Lemma explains the connection[34] between free equalities demonstrating the membership $w \in \langle\langle \mathcal{R} \rangle\rangle$ and combinatorial null-homotopies for the corresponding loops.

Such a null-homotopy is given by a van Kampen diagram over $\langle \mathcal{A} \mid \mathcal{R} \rangle$, which is a 1-connected, combinatorial planar 2-complex Δ in $\mathbb{R}^2$ with a basepoint; each oriented edge is labelled by a generator $a_i^{\pm 1}$ with $a_i \in \mathcal{A}$ and the boundary label on each face is some $r_j^{\pm 1}$ with $r_j \in \mathcal{R}$ (read from a suitable

[34] For a complete account of the equivalences in this subsection, see [**12**].

basepoint). There is a unique label-preserving map from the 1-skeleton of Δ to the 1-skeleton of the standard 2-complex $K(\mathcal{A} : \mathcal{R})$, and this extends to a combinatorial map $\Delta \to K(\mathcal{A} : \mathcal{R})$.

Van Kampen's Lemma implies that the number of faces in a least-area van Kampen diagram with boundary label w is the least number N of factors among free equalities $w = \prod_{j=1}^{N} u_j r_j u_j^{-1}$. Thus the *Dehn function* of $\langle \mathcal{A} \mid \mathcal{R} \rangle$ can be defined to be the minimal function $\delta(n)$ such that every null-homotopic edge-loop of length at most n in $K(\mathcal{A} : \mathcal{R})$ is the restriction to $\partial\Delta$ of a combinatorial map $\Delta \to K(\mathcal{A} : \mathcal{R})$ where Δ is a 1-connected, planar combinatorial 2-complex. When described in this manner, it is natural to call the Dehn function the *combinatorial isoperimetric function* of $K(\mathcal{A} : \mathcal{R})$; the combinatorial isoperimetric function of an arbitrary compact combinatorial 2-complex is defined in the same way.

There is a standard diagrammatic argument for showing that the Dehn functions of quasi-isometric groups are $\simeq$ equivalent — see [**1**]. In that argument, it is unimportant that the complexes considered have only one vertex. Thus if K is any compact combinatorial 2-complex with fundamental group Γ, then the combinatorial isoperimetric function of K is $\simeq$ equivalent to the Dehn function of Γ. We shall exploit the freedom stemming from this equivalence. Specifically, we shall prove the Main Theorem by establishing a quadratic upper bound on the combinatorial isoperimetric function of a carefully-crafted 2-complex M with fundamental group $F \rtimes_{\phi^r} \mathbb{Z}$, where $r > 0$. In other words, we identify a constant $C > 0$ such that every null-homotopic combinatorial loop of length at most n in $M^{(1)}$ is the boundary of a combinatorial map to M from a 1-connected planar 2-complex with at most Cn^2 2-cells. In fact, we prove something more refined than this (see Section 3.3 below).

REMARK 3.1.1. Note that we are free to pass from $F \rtimes_{\phi} \mathbb{Z}$ to the finite-index subgroup $F \rtimes_{\phi^r} \mathbb{Z}$ because the $\simeq$ class of the Dehn function of a group is an invariant of commensurability.

Henceforth we shall use the term *van Kampen diagram* to refer to the domain of a combinatorial map to M from a 1-connected planar 2-complex, with oriented edges *labelled* by letters representing the oriented edges of the target. (Note that this agrees with the standard terminology in the special case $M = K(\mathcal{A} : \mathcal{R})$.) Such a diagram is said to be *least-area* if it has the least number of 2-cells among all diagrams with the same boundary label.

3.1.1. The Mapping Torus. Let G be a compact graph and let $f : G \to G$ be a continuous map that sends each edge e_i of G to an immersed edge-path $u_i = \varepsilon_1 \dots \varepsilon_m$ in G. We attach to each vertex $v \in G$ a new edge t_v joining v to $f(v)$. We then attach one 2-cell to this augmented graph for each edge e_i; the 2-cell is attached along the edge path $t_v^{-1} e_i t_{v'} u_i^{-1}$, where v and v' are the initial and terminal vertices of e_i and where the inverse is taken in the path groupoid (i.e. u_i^{-1} is u_i traversed backwards). The resulting 2-complex is the *mapping torus* of f, which we shall denote $M(f)$.

In this part of the book we are primarily concerned with van Kampen diagrams over $M(f)$, where f is a homotopy equivalence representing a given free-group automorphism ϕ. In this case $\pi_1(M(f)) \cong \pi_1(G) \rtimes_\phi \mathbb{Z}$. The 1-cells in such a diagram Δ_0 are either labelled by some t_u or by an edge $e \in G$. We will refer to all of the edges t_u as *t-edges* and, when it does not cause confusion, denote them simply by t. For the other edges in Δ_0, it is necessary to distinguish between the edge and its *label* in G.

NOTATION 3.1.2 (Labels $\check{\rho}$). *If an edge ε in a van Kampen diagram over $M(f)$ is labelled by an edge in G, then we write $\check{\varepsilon}$ to denote that label. More generally, if an edge-path ρ in such a diagram contains no t-edges, we write $\check{\rho}$ to denote the path in G that labels ρ.*

3.1.2. Time, folded t-corridors, singularities and bounded cancellation. Assume we are in the setting of the previous paragraph. A *t-corridor* (more simply, *corridor*) is then defined exactly as in Section 1.1.4, and we have the corresponding notion of *time* (which may be thought of as a map to $\mathbb{R}$ that is constant on non-t edges, integer-valued on vertices, and sends the endpoints of each t-edge to integers that differ by 1). As in Subsections 1.1.5 and 1.1.6, we see that each least-area diagram is the union of its corridors, and we may assume that the tops of all corridors are *folded*. (In Subsection 3.2.1 we shall specify how this folding is to be done, but for the results in this subsection it is not necessary to prescribe it.)

We write $\bot(S)$ and $\top(S)$ to denote the *top* and *bottom* of a (folded) corridor, respectively. *Singularities* are defined exactly as in Part 1.

We restrict our attention to least-area disc diagrams. The argument used to prove Lemma 1.2.1 applies *verbatim* in the present setting to prove:

LEMMA 3.1.3. *If S and S' are distinct corridors in a least-area diagram, then $\bot(S) \cap \bot(S')$ consists of at most one point.*

Let L be the maximum length of $f(E)$ for E an edge in G. As in Proposition 1.2.3 we have

PROPOSITION 3.1.4 (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 any least-area diagram over $M(f)$.*
3. *If Δ is a least-area diagram over $M(f)$, then there are less than $2|\partial\Delta|$ non-degenerate singularities in Δ, and each has length at most LB.*

PROOF. Except for one minor difficulty, the proof from Part 1 translates directly to the current setting. The minor difficulty is that in the current context the map f is a homotopy equivalence rather than a group automorphism, and f^{-1} is not defined as a topological map. Thus, given a path ρ, we need a canonical path σ in G such that $f_\#(\sigma) = \rho$, where $f_\#$ is tightening rel endpoints.

Consider $\widetilde{M(f)}$, the universal cover of $M(f)$. Its 1-skeleton consists of a collection of trees (copies of the universal cover of G) joined by t-edges. Consider a lift to $\widetilde{M(f)}$ of the unique edge-path $\tau_0\rho\tau_1^{-1}$ such that the τ_i are t-edges. Both endpoints of this lift lie in one of the trees $T \cong \tilde{G}$; define $\widetilde{\sigma}$ to be the unique injective path which joins them in T, and define σ to be the image of $\widetilde{\sigma}$ in $M(f)$. ☐

As in Lemma 1.2.4, the above result yields as a special case (cf. [**21**] and [**4**, Lemma 2.3.1, pp.527–528]):

Lemma 3.1.5 (Bounded Cancellation Lemma). *There is a constant B, depending only on f, so that if I is an interval consisting of $|I|$ edges on the bottom of a (folded) corridor S in a least-area diagram over $M(f)$, and every edge of I dies in S, then $|I| < B$.*

3.1.3. Past, Future and Colour in Diagrams. These concepts, for edges and 2-cells in van Kampen diagrams Δ, are defined exactly as in Section 1.3. The *immediate past* (or *ancestor*) of an edge at the top of a corridor in any diagram is the unique edge at the bottom of the corridor that lies in the same 2-cell; the *entire past* of an edge is defined by taking the transitive closure of the relation "is the immediate past of". The past of a 2-cell is defined similarly. The *future* of an edge e_0 is the set of edges that have e_0 in their past. The future of 2-cells is defined similarly. The evolution of edges is described by a graph $\mathcal{F}$ whose vertices are the 1-cells e of Δ, which has an edge connecting each e to its immediate ancestor. Note that $\mathcal{F}$ is a forest. Its connected components define *colours* in Δ; each edge not labelled t is assigned a unique colour, as is each 2-cell. Note that colours are in bijection with a subset of the edges of the boundary of the diagram. The union of the 2-cells in a corridor S that have colour μ will be denoted $\mu(S)$.

As in Part 1, simple separation arguments yield the following observations.

Lemma 3.1.6. *Each $\mu(S)$ is connected and intersects each of $\top(S)$ and $\bot(S)$ in an interval.*

Lemma 3.1.7 (cf. Lemma 1.5.9). *Let $\varepsilon_1, \varepsilon_2$ and ε_3 be three (not necessarily adjacent) edges that appear in order of increasing subscript 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.*

Again following Part 1, given a diagram Δ we define $\mathcal{Z}$ to be the set of pairs (μ, μ') such that the coloured regions $\mu(S)$ and $\mu'(S)$ are adjacent in some corridor S. The proof of Lemma 1.6.6 establishes:

LEMMA 3.1.8.

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

3.2. Adapting Diagrams to the Beaded Decomposition

We refer the reader to Part 2 for the definitions and results which we require here about improved relative train track maps, nibbled futures, monochromatic paths, hard splittings and the language of *beads* — including (J, f)-atoms, GEPs and ΨEPs and what it means for a path to be (J, f)-beaded. We shall proceed under the assumption that the reader is familiar with each of these terms, and work axiomatically with the following outputs from Part 2.

THEOREM 3.2.1 (Beaded Decomposition Theorem, Part 2). *For every $\phi \in \mathrm{Out}(F_r)$, there exist positive integers k, r and J such that ϕ^k has an improved relative train-track representative $f_0 : G \to G$ with the property that every $(f_0)^r_\#$-monochromatic path in G is (J, f_0)-beaded.*

We remind the reader that beads are either monochromatic paths (in case they are atoms) or else GEPs or ΨEPs (which may be monochromatic, but do not have to be). Thus, by the above theorem and Proposition 2.6.10, any nibbled future of a (J, f_0)-bead is (J, f_0)-beaded. Any hard splitting of an edge-path is inherited by its (nibbled) futures, by definition. And if one refines a hard splitting by decomposing the factors in a hard splitting, the result is again a hard splitting (Lemma 2.2.8). Thus we have:

COROLLARY 3.2.2 (Theorem 2.8.4). *Let $f = (f_0)^r_\#$ be as in the Beaded Decomposition Theorem above. If an edge-path σ in G is (J, f_0)-beaded, then any f-nibbled future of σ is (J, f_0)-beaded. In particular, $f_\#(\sigma)$ is also (J, f_0)-beaded.*

REMARK 3.2.3. An important point to recall from Part 2 is that the decomposition of an edge-path into (J, f_0)-beads is canonical.

The value of the constant J in the Beaded Decomposition Theorem will be of no importance in what follows, so we drop it from the terminology. Similarly, we will fix the map f_0. Once we have passed to the power $f = (f_0)^r_\#$, the above results remain true when f is replaced by an iterate. Therefore, we refer simply to "beads" and "beaded paths".

3.2.1. Refolding corridors according to the Beaded Decomposition.

Henceforth[35]*, we consider only diagrams over the mapping torus of* $M(f)$*, where* f *is an iterate of* $(f_0)^r_{\#}$ *as in the Beaded Decomposition Theorem. In Section 3.4, we will fix the map* f *once and for all.*

We return to the matter of how best to fold the tops of corridors in least area diagrams over $M(f)$. Given an arbitrary least-area diagram, we refold the tops of corridors in order of increasing time. The process begins with edges at the minimal time on the boundary of the diagram, where there is no folding to be done provided the boundary label is reduced.

Focussing on a particular corridor S, our folding up to time(S) defines the histories of all edges up to this time and hence assigns colours to the edges on $\perp(S)$, decomposing it as a concatenation of monochromatic paths, one for each of the colours $\mu(S)$. Theorem 3.2.1 decomposes each of these labels as a hard splitting of beads σ_i. The hardness of the splitting means that after tightening the $f(\sigma_i)$, their concatenation will be a tightening of $f_{\#}(\check{\mu}(S))$. We insist that the first step in the tightening of the naive top of S, is that determined by the tightening of labels just described: i.e. we first tighten beads *within* colours, each according to a left-to-right convention (which labels inherit from the orientation of the corridors within the diagram). Then, as a second step, we tighten (again with a left-to-right convention) the concatenation of the tightened images of the colours. A diagram which is folded according to these conventions will be called *well-folded*.

The key point of this convention is that the hard splitting of the label on each colour is carried into the future — of course the futures of the original beads may split into a concatenation of several beads, and some beads at the ends of each colour may be cancelled by interaction with neighbouring colours, but *each bead (more precisely*[36]*, bead-labelled arc) in the beaded decomposition of each coloured interval on* $\top(S)$ *is contained into the future of a unique bead-labelled arc of the same colour on* $\perp(S)$*.* Thus $\top(S)$ is a concatenation of beads, each with a definite colour, where neighbouring beads are separated by a hard splitting if they are of the same colour but perhaps not if they are of a different colour. (It also becomes sensible to discuss the future of a bead in a [well-folded] diagram.)

We henceforth suppose (usually without comment) that our diagram has been refolded according to this convention.

Definition 3.2.4. [cf. Definition 3.6.2] The *bead length* of $[S]_\beta$, of a corridor S in a well-folded diagram is the number of beads along $\perp(S)$.

[35]There are exceptions to this in Theorem 3.3.1, Section 3.13 and Section 3.14

[36]we shall generally drop this cumbersome distinction in the sequel

REMARK 3.2.5. It is important to note that the decomposition of $\perp(S)$ and $\top(S)$ into coloured intervals is *not* a hard splitting in general. Indeed it is the analysis of the cancellation between these intervals as one flows S forwards in time that forms the meat of this part of the book.

3.2.2. Abstract Futures of Beads. Given an edge-path ρ in G, expressed as a concatenation of monochromatic edge-paths $\rho = \rho_1 \dots \rho_m$, consider the van Kampen diagram $\Delta(l, \rho)$ with boundary label equal to $t^{-l}\rho t^l \overline{f^l_\#(\rho)}$; this is a simple stack of corridors. The above convention dictates how we should fold the corridors of Δ and determines the future at each time up to l for each bead in the beaded decompositions of the ρ_i.

We define the (full) *abstract future of a bead in* ρ to be (the label on) its future in $\Delta(l, \rho)$.

3.3. Linear Bounds on the Length of Corridors

In any least-area diagram, each corridor has at least two edges on the boundary, namely its t-edges. The *length* of a corridor S is defined to be the number of 2-cells that it contains. The area of a least-area diagram is the sum of the lengths of its corridors, and therefore our Main Theorem is an immediate consequence of:

THEOREM 3.3.1. *Let ϕ be an automorphism of a finitely generated free group and let f be a topological representative for a positive power of ϕ. There is a constant K, depending only on f, so that each corridor in a least-area diagram Δ over $M(f)$ has length at most $K\,|\partial\Delta|$.*

Note that the Main Theorem actually depends only on establishing Theorem 3.3.1 for a single topological representative f^k of a suitable power of our given free group automorphism ϕ; in the next section we shall articulate what that suitable power is. The bulk of this part of the book will then be devoted to proving the existence of the constant K for this particular f^k. (In Section 3.13 we shall deduce Theorem 3.3.1 from this special case.)

Having restricted attention to a particular f^k, we may further restrict our attention to diagrams that are well-folded in the sense of Subsection 3.2.1, since refolding the corridors of an arbitrary a diagram does not change the configuration of corridors or their length. In a well-folded diagram, the top of each corridor S is a concatenation of beads, and the vast majority of our work (up to and including Section 3.11) goes into proving the following result.

THEOREM 3.3.2. *If f and k are as above, then there is a constant K_1 such that all corridors S in well-folded, least-area diagrams Δ over $M(f^k_\#)$, have bead length $[S]_\beta \leq K_1\ |\partial\Delta|$.*

The linear bound on the length of S that we require for Theorem 3.3.1 does not follow directly from this estimate because there is no uniform bound on the length of certain beads, namely GEPs and ΨEPs. However, we shall see in Section 3.12 that the ideas developed in Part 1 to implement the Bonus Scheme adapt to the current setting to provide the following estimate:

PROPOSITION 3.3.3. *There are constants J and K_2, depending only on f, such that the beads β on $\perp(S)$ of length greater than J satisfy*

$$\sum_{\beta} |\beta| \leq K_2 \, |\partial\Delta|.$$

The constant J in the above statement is the one from Theorem 3.2.1.

3.4. Replacing f by a Suitable Iterate

In order to establish the bound on the length of corridors required to prove Theorem 3.3.1, we must analyse how corridors grow as they flow into the future and assess what cancellation can take place to inhibit this growth. This is much more difficult than in Part 1 because now we must cope with the cancellation that takes place within colours. But in common with our approach in Part 1, we can appeal to Remark 3.1.1 repeatedly in order to replace our topological representative f by some iterate of f that affords a more stable situation in which cancellation phenomena are more amenable to analysis.

In the present setting, we have to be a little careful about specifying what we mean by "an iterate", because we wish to consider only topological representatives whose restriction to each edge is an immersion, and this property is not inherited by powers of the map. To avoid this problem, we deem the phrase[37] *replacing f by an iterate*, to mean that for fixed $k \in \mathbb{N}$, we pass from consideration of $f : G \to G$ to consideration of the map $f_{\#}^k : G \to G$ that sends each edge E in G to the tight edge-path $f_{\#}^k(E)$ that is homotopic rel endpoints to $f^k(E)$.

When we replace f by $f_{\#}^k$, we leave behind the mapping torus $M(f)$ and consider instead $M(f_{\#}^k)$, which although homotopic to a k-sheeted covering of $M(f)$ is distinct from it.

A corridor in a van Kampen diagram over $M(f_{\#}^k)$ can be divided into a stack of k corridors in order to yield a van Kampen diagram over $M(f)$. This observation will play little role in our arguments, but it highlights one reason for hoping to simplify diagrams by passing to an iterate of f: the van Kampen diagrams over $M(f_{\#}^k)$ are a proper subset (after subdivision[38] of Δ) of the diagrams over $M(f)$; in the diagrams of this subset, corridors flow unhindered for at least k steps in time.

[37] and obvious variations on it

[38] the obvious subdivision of a diagram Δ is called the *k-refinement*

3.4.1. Finding the desired iterate. We have already passed to a large iterate in order to obtain the Beaded Decomposition Theorem. In the present subsection we pass to further iterates in order to control the behaviour of the images of beads.

Before settling on a specific f for the remainder of the paper, we must remove an irritating ambiguity concerning the ordering of strata in the filtration associated to the train track structure. This is required in order to render the choices in Section 3.5 coherent.

DEFINITION 3.4.1. Suppose that $f : G \to G$ is an improved relative train track map, and that H_i, H_j are strata for f. We say that H_i and H_j are *interchangeable* if one can reorder the strata, so that one still has an improved relative train track structure, but the order of H_i and H_j is reversed.

If H_i and H_j are interchangeable, and $i > j$, then no iterate of any edge in H_i crosses an edge in H_j (and neither do the iterates of any edges occurring in the iterated images of edges in H_i).

CONVENTION 3.4.2. *We suppose that for any improved relative train track map that we consider, if H_i and H_j are interchangeable strata so that H_i is an exponential stratum and H_j is a parabolic stratum then $i > j$.*

We further assume that if $H_i = \{E_i\}$ and $H_j = \{E_j\}$ are interchangeable parabolic strata and $n \mapsto |f^n(E_i)|$ grows exponentially while $n \mapsto |f^n(E_j)|$ grows polynomially, then $i > j$. And if both these functions grow polynomially, then the degree of polynomial growth of the former is at least as great as the latter.

In the following lemma, ω is the number of strata in the train track structure for f. Also recall that an edge ε in a path σ is said to be *displayed* if there is a hard splitting $\sigma = \sigma_1 \odot \varepsilon \odot \sigma_2$. The definition of a *displayed sub edge-path* is entirely analogous, and will be used later.

LEMMA 3.4.3. *One can replace f by an iterate to ensure that if ρ is any atom then either the beads of $f_\#^\omega(\rho)$ are Nielsen paths and* GEP*s only, or else there is a displayed edge ϵ in $f_\#^\omega(\rho)$ so that*

(1) ε is of highest weight amongst all displayed edges in all $f_\#^k(\rho)$, for $k \geq 1$, and
(2) the growth of $n \mapsto |f_\#^n(\epsilon)|$ is at least as large as that of any displayed edge in any $f_\#^k(\rho)$.

PROOF. Lemma 2.8.3 contains all but statement (2), whose validity is assured by Convention 3.4.2. □

Our next two results capture the *end stability* that Proposition 1.4.5 provided in the case of positive automorphisms. This is the first stage in our analysis at which we encounter an awkward point that does not arise in Part

1, namely there may exist beads (more specifically atoms) ρ such that $f_\#(\rho)$ is a single vertex.

DEFINITION 3.4.4. A *vanishing bead (atom)* ρ is one with $f_\#(\rho)$ a single vertex.

LEMMA 3.4.5. *There exists a constant k_0, depending only on f so that the map $f_0 = f_\#^{k_0}$ satisfies the following properties. Let ρ be a non-vanishing bead, let $i \in \{1, \dots, \omega\}$, and let σ_i be the leftmost bead in $(f_0)_\#(\rho)$ of weight at least i.*

(1) If σ_i is not a GEP *or a* ΨEP *then the leftmost bead of weight at least i in $(f_0)^j_\#(\rho)$ is the same for all $j \geq 1$. Furthermore, in this case σ_i is a single (displayed) edge or a Nielsen bead.*

(2) If σ_i is a GEP *or a* ΨEP *then the leftmost bead of weight at least i in $(f_0)^j_\#(\rho)$ is contained in the (abstract) future of σ_i for all $j \geq 1$.*

PROOF. If σ is a bead then all iterated images of σ are beaded paths, and a simple finiteness argument shows that there is a bound on the number of beads which are not GEPs or ΨEPs. $\square$

An entirely similar argument applies to rightmost beads, of course. In order to deal with the different types of beads, we also need the following variant.

LEMMA 3.4.6. *There exists a constant k_1, depending only on f, so that the map $f_1 = f_\#^{k_1}$ satisfies the following properties. Let ρ be a non-vanishing bead and let σ be the leftmost bead in $(f_1)^j_\#(\rho)$ which is not a Nielsen bead.*

(1) If σ is not a GEP *or a* ΨEP *then for all $j \geq 1$ the leftmost bead in $(f_1)^j_\#(\rho)$ which is not a Nielsen bead is σ. Furthermore, in this case σ is a (displayed) edge.*

(2) If σ is a GEP *or a* ΨEP *then for all $j \geq 1$ the leftmost bead in $(f_1)_\#(\rho)$ which is not a Nielsen bead is in the future of σ.*

We are finally in a position to articulate all of the properties that we want to arrange for f by replacing it with an iterate.

PROPOSITION 3.4.7. *There is a constant D_2 that depends only on f, so that if we replace f by $f_\#^{D_2}$ then,*

(1) the conclusion of Lemma 2.5.1 holds with $k_1 = 1$: in particular, if ε is an exponential edge of weight i, then $f(\varepsilon)$ is longer than the unique indivisible Nielsen path of weight i (if it exists);

(2) the conclusion of Theorem 2.8.1 holds with $D_1 = 1$;

(3) the conclusion of Lemma 3.4.3 holds;

(4) the conclusions of Lemmas 3.4.5 and 3.4.6 hold; and

(5) if ρ is a bead then $f_\#(\rho)$ contains at least three displayed copies of any exponential edge that is displayed in any $f^j_\#(\rho)$, $j \geq 1$. Moreover,

the leftmost (and rightmost) such displayed edge ε is contained in a displayed path of the form $f(\varepsilon)$.

Power Decree: *For the remainder of the paper, we will assume that $f : G \to G$ is an improved relative train track map that satisfies the properties in Proposition 3.4.7. We shall also operate under Convention 3.4.2.*

Let L be the maximal length of $f(E)$, for edges $E \in G$.

3.5. Preferred Futures of Beads

The reader who is comparing our progress to Part 1 will find that we are now in the position that we were at the start of Section 1.5. Thus we now want to define the preferred future of a bead ρ (in three senses[39]) and then begin a study of fast beads.

Unfortunately, the definition of the preferred future of a bead in a diagram is much more cumbersome than the analogue in Part 1.

3.5.1. Abstract Preferred Futures and Growth. First we note that if beads (or more generally edge paths in G) are ever going to vanish in the sense of Definition 3.4.4, then they do so immediately.

LEMMA 3.5.1. *If σ is an edge path in G and $f_{\#}^k(\sigma)$ is a vertex for some $k \geq 1$, then $f_{\#}(\sigma)$ is already a vertex.*

PROOF. For all vertices $v \in G$, $f(v)$ is a fixed point of f. Therefore, the endpoints of $f_{\#}^j(\sigma)$ are the same for all $j \geq 1$. If $f_{\#}^k(\sigma)$ is a point, then the endpoints of $f_{\#}^k(\sigma)$ are equal, hence the tight path $f_{\#}(\sigma)$ is a loop. Since f is a homotopy equivalence, this loop must be trivial. □

DEFINITION 3.5.2 (Abstract preferred futures). The (immediate) *preferred future* of a non-vanishing bead σ is a particular bead in the beaded decomposition of $f_{\#}(\sigma)$, as defined below. The *k-step preferred future* is then defined by an obvious recursion.

(1) If σ is a GEP then $f_{\#}(\sigma)$ is also a GEP, and we define the preferred future of σ to be $f_{\#}(\sigma)$.
(2) If σ is a ΨEP then either σ or $\overline{\sigma}$ has the form $\sigma = E\overline{\tau}^k\nu\gamma$. If it is σ, then by Corollary 2.6.11, $f_{\#}(\sigma)$ is either of the form $\sigma' \odot \xi$, where σ' is a ΨEP (which has the same weight as σ), or else of the form $E \odot \xi$, where E has the same weight as σ and is the unique highest weight edge in $f_{\#}(\sigma)$. In the first case, the preferred future of σ is σ'. In the second case, the preferred future of σ is E. The preferred future of a ΨEP σ where $\overline{\sigma}$ has the above form is defined in an entirely analogous way.
(3) If σ is a Nielsen path then the preferred future of σ is $f_{\#}(\sigma) = \sigma$.

[39]in $f_{\#}(\rho)$, in a diagram, and in a concatenation of beaded paths

(4) Finally, we consider a non-vanishing atom σ.
(a) If the beaded decomposition of $f_\#(\sigma)$ consists entirely of Nielsen paths and GEPs, then we fix a highest weight GEP to be the preferred future of σ; otherwise, we fix a highest weight Nielsen path.
(b) If not, then let ε be the edge described in Lemma 3.4.3, fix a displayed occurrence of ε in $f_\#(\sigma)$ (in case ε is exponential, choose a displayed occurrence that is neither leftmost nor rightmost[40]) and define this to be the preferred future of ε.

REMARK 3.5.3. Suppose that ε is an edge in G, considered as a bead, and suppose that ε is not contained in a zero-stratum. Then ε has a preferred future, which is an edge contained in the same stratum as ε. We always assume that the preferred future of ε is a (fixed) occurrence of ε in $f_\#(\varepsilon)$ which satisfies the requirements of the above definition. This situation is very close in spirit to the definition of preferred future in Part 1.

We now divide the beads into classes according to the growth of the paths $f_\#^k(\sigma)$, $k = 1, 2, \ldots$. Specifically, we define left-fast and left-slow beads in accordance with Subsection 1.5.1.

DEFINITION 3.5.4 (Left-fast beads). GEPs and Nielsen paths are *left-slow*.
Suppose that α is an atom or a ΨEP. Then α is *left-fast* if the distance between the left end of $f_\#^k(\alpha)$ and the left end of the preferred future of α in $f_\#^k(\alpha)$ grows at least quadratically with k, and *left-slow* otherwise.

Note that if a ΨEP σ is left-fast then it is $\overline{\sigma}$ which it is of the form $E\overline{\tau}^k\nu\gamma$.

REMARK 3.5.5. We only care that fast growth be super-linear, but it happens that this is the same as being at least quadratic (cf. [**14**]).

The concepts of *right-fast* and *right-slow* beads are entirely analogous.

3.5.2. Preferred future in diagrams. In this subsection we define the notion of 'preferred futures' within van Kampen diagrams. We also define 'biting' and 'consumption', which are the analogues in this paper of 'consumption' from Section 1.5.

The folding convention of Subsection 3.1.2 expresses $\perp(S)$ as the concatenation of coloured paths $\mu(S)$, each labelled by a monochromatic path in G. The Beaded Decomposition Theorem gives us a hard splitting into beads

$$\check{\mu(S)} = \check{\beta}_1 \odot \check{\beta}_2 \odot \cdots \odot \check{\beta}_{m_\mu},$$

and it is convenient to refer to the sub-paths $\beta_i \subseteq \perp(S)$ carrying the labels $\check{\beta}_i$ as beads, as we did in Subsection 3.1.2.

---

[40]this exists by Proposition 3.4.7

If $\mu_1, \ldots, \mu_k$ are the colours appearing in S, in order, then the label on $\top(S)$ is obtained by tightening

$$f_\#(\mu_1 \check{(S)}) \cdots f_\#(\mu_k \check{(S)}).$$

The path $f_\#(\mu_1 \check{(S)}) \cdots f_\#(\mu_k \check{(S)})$ is called the *semi-naive future* of S.

We have adopted a left-to-right convention to remove any ambiguity in how one tightens the semi-naive future to obtain the label of $\top(S)$.

We previously defined the (immediate) future of a bead $\beta \subset \bot(S)$ to consist of those edges of $\top(S)$ whose immediate past lies in β. Since it is integral to what we shall do now, we re-emphasize:

LEMMA 3.5.6. *The immediate future of a bead $\beta \subset \bot(S)$ is a (possibly empty) interval equipped with a hard-splitting into beads.*

If ρ is the immediate future of β, then ρ is also an interval in the semi-naive future of S, and hence its label $\check{\rho}$ is a specific sub-path of $f_\#(\check{\beta})$. [Note that one has more than the path $\check{\rho}$ here, one also has its position within $f_\#(\check{\beta})$; thus, for example, we would distinguish between the two visible copies of $\check{\rho}$ in $f_\#(\check{\beta}) = \check{\rho}\sigma\check{\rho}$.]

DEFINITION 3.5.7 (Preferred and tenuous futures in Δ). Consider a bead $\beta \subset \mu(S) \subset \bot(S)$ in Δ whose immediate future $\rho \subset \top(S)$ determines the subpath $\check{\rho}_0$ of $\check{\beta}$ in G.

If the (abstract) preferred future $\check{\beta}_+$ of $\check{\beta}$, as defined in Definition 3.5.2, is entirely contained in $\check{\rho}_0$, then the corresponding sub-path β_+ of ρ is the *preferred future* of β.

If $\check{\rho}_0$ does not contain $\check{\beta}_+$, then β does not have a preferred future. In this situation we say that the future of β is *tenuous*.

REMARK 3.5.8. Note that, if it exists, the preferred future of a bead $\beta \subset \mu(S)$ is a bead in the beaded decomposition of both ρ and the μ-coloured interval of $\top(S)$.

Also, if a bead happens to be a single edge ε whose label is not contained in a zero stratum, the preferred future is a single (displayed) edge, with the same label as ε.

DEFINITION 3.5.9 (Biting and consumption). If the future of a bead $\beta \subset \bot(S)$ is tenuous, we say that β is *bitten in S*. If, in the notation of (3.5.7), *no edge* of the preferred future of $\check{\beta}$ appears in $\check{\rho}$, then we say that β is *consumed* in S.

REMARK 3.5.10. The above definition says in particular that any bead whose label is a vanishing atom is consumed.

Let $\beta' \subset \bot(S)$ be a bead whose label is non-vanishing. If β' is bitten in S, there is a specific edge ε in the semi-naive future of S that, during the tightening process, is the first to cancel with an edge ε' in the interval labelled by the preferred future of $\check{\beta}'$. The edge ε is in the immediate future of a bead β, necessarily of a different colour than β'.

Definition 3.5.11. In the above situation, we say that β *bites* β' *from the left* if β lies to the left of β' in S, and that β *bites* β' *from the right* if β lies to the right of β' in S. We say that the edges ε and ε' discussed above *exhibit* the biting.

The above concepts of biting and consumption replace the single, simpler, notion of consumption from Section 1.5: there, since the preferred future was a single edge, if it was bitten it was consumed. In Part 1, a frequently used concept was for an edge to be 'eventually consumed'. In this part of the book, we need the following replacement:

Definition 3.5.12. Suppose that $\rho_1 \subset \mu_1(S)$ and $\rho_2 \subset \mu_2(S)$ are beads in $\perp(S)$. We say that ρ_1 is *eventually bitten* by ρ_2 if there is a corridor S' which contains a preferred future β_1 of ρ_1 and a bead β_2 in the future of ρ_2 so that β_2 bites β_1 in S'.

With these definitions in hand, we have the following, which is an appropriate replacement for 1.5.3

Lemma 3.5.13 (cf. Lemma 1.5.3). *There exists a constant C_0 with the following property: if ρ is a bead such that $f_\#(\rho)$ contains a left-fast displayed edge E and if $UV\rho$ is a (tight) path with $V\rho = V \odot \rho$ and $|V| \geq C_0$ then for all $j \geq 1$ the preferred future of E is not bitten when $f^j(UV\rho)$ is tightened. Moreover, $|f^j_\#(UV\rho)| \to \infty$ as $j \to \infty$.*

Proof. We first prove the result in the special case that $V\rho$ is a nibbled future of a left-fast edge E_1, where ρ is the preferred future of E_1. In other words, we will prove the existence of a constant C_0' so that if $|V| \geq C_0'$ then the statement of the lemma holds for the particular path $UV\rho$. (We will later reduce to this special case.)

Note that V and $V\rho$ are monochromatic paths, and thus admit a beaded decomposition. Suppose first that V does not contain any beads of length greater than J. In this case, the proof is entirely parallel to that of Lemma 1.5.3, where we count using the number of non-vanishing beads rather than the number of edges.

In case V contains long GEPs or long ΨEPs, we note that the cancellation by U on the left, and possibly by one of the edges in the GEP or ΨEP on the right can only decrease the length of a GEP or ΨEP by at most $2B$ at each iteration. Thus it is straightforward to include long GEPs and ΨEPs into the above calculation. We now turn to the general case.

Suppose that V is an arbitrary path so that $V\rho = V \odot \rho$. Then V can shrink of its own accord (it needn't be beaded), and can be cancelled by the future of U. However, there is certainly a constant C_0 so that if $|V| \geq C_0$ then by the time this shrinking of V combined with cancelling by the future of U can have reduced V to the empty path, the future of the edge E has at least C_0' edges to the left of its preferred future. We are then in the special case that we dealt with first. □

The following two lemmas are proved in an entirely similar manner to Lemma 1.5.5. Recall that displayed edges are particular types of beads, and the (abstract) preferred futures of beads were defined in Definition 3.5.2. Recall from Remark 3.5.8 that the preferred future of a displayed edge whose label is not contained in a zero stratum is a single displayed edge.

Lemma 3.5.14. *Let $\chi_1\sigma\chi_2$ be a tight path in G. Suppose that χ_1 and χ_2 are monochromatic and that, for $i = 1, 2$, the edge E_i is displayed in χ_i and that E_i is not in a zero stratum. Suppose that σ is a concatenation of beaded paths. Then the preferred futures of E_1 and E_2 cannot cancel each other in any tightening of $f_\#(\chi_1)f_\#(\sigma)f_\#(\chi_2)$.*

Suppose that S is a corridor in a well-folded diagram, and that $\mu_1(S)$ and $\mu_2(S)$ are non-empty paths in $\perp(S)$, where μ_1 and μ_2 are colours. Suppose further that for $i = 1, 2$ there is a displayed edge ε_i such that $\breve{\varepsilon}_i$ is not contained in a zero stratum. Then the edges in the semi-naive future of S corresponding to the preferred futures of ε_1 and ε_2 do not cancel each other when folding the semi-naive future of $\perp(S)$ to form $\top(S)$.

Lemma 3.5.15. *Let S be a corridor and suppose that ε_1 and ε_2 are edges in $\perp(S)$ whose labels lie in parabolic strata. In the naive future of each ε_i (that is, before even the beads have been tightened), there is a unique edge ε_i' with the same label as ε_i. At no stage during the tightening of $\top(S)$ can ε_1' cancel with ε_2'.*

Corollary 3.5.16. *A displayed edge in any coloured interval $\mu(S)$ which is labelled by a parabolic edge $\breve{E}_i \in H_i$ can only be consumed by an edge whose label is in $\overline{G \smallsetminus G_i}$.*

3.5.3. Abstract paths, futures and biting. In many of the arguments in later sections, we wish to work with concatenations of beaded paths in G rather than sides of corridors in diagrams. This is done as in Subsection 3.2.2 by associating to such a path $\rho = \rho_1 \dots \rho_m$, with the ρ_i beaded, the van Kampen diagram $\Delta(l, \rho)$ with boundary label $t^{-l}\rho t^l \overline{f_\#^l(\rho)}$. But we modify the usual definition of colour by defining the colours on the bottom of the first (earliest) corridor not to be single edges but rather to be intervals labelled ρ_i. We then use the definitions of the previous subsection (biting, preferred future *etc.*) to define the associated concepts *for beads in ρ*.

We emphasize, ρ itself need not beaded; only the ρ_i are. We also emphasize that edges do not have preferred futures, only beads do.

However, some beads are single, displayed edges, and when considered as beads they do have a preferred future.

3.6. Counting Fast Beads

This section is the analogue of Section 1.6; it is here that the proof of the Main Theorem begins in earnest.

Let Δ be a minimal area van Kampen diagram, folded according to the convention of Section 3.1.2, and fix a corridor S_0 in Δ. As explained in Section 3.3, the core of our task is to bound the number of beads in the decomposition of $\perp(S_0)$. In order to do so, we must undertake a detailed study of the preferred futures of these beads.

First we dispense with the case that $\check{\beta}$ is a vanishing atom.

LEMMA 3.6.1. *Suppose that $\mathcal{S}$ is the collection of beads in S_0 which are not vanishing atoms. If $\sum_{\beta \in \mathcal{S}} |\beta| = D$ then $|S_0| \leq B(D+1)$.*

PROOF. This follows immediately from the Bounded Cancellation Lemma. □

Narrowing our focus in the light of this lemma, we define:

DEFINITION 3.6.2 (Bead norm). Given a concatenation $\rho = \rho_1 \dots \rho_m$ of beaded paths, we define the *bead norm* of ρ, denoted $\| \rho \|_\beta$, to be the number of non-vanishing beads in the concatenation. (This is poor notation, since the norm depends on the decomposition into the ρ_i and not just the edge-path ρ. But in the contexts we shall use it, specifically $\perp(S_0)$, it will always be clear which decomposition we are considering.)

REMARK 3.6.3. All beads have length at least 1. Thus bead norm is dominated by length. In particular, estimates concerning Bounded Singularities and Bounded Cancellation remain true when distance is replaced by bead norm; cf. Lemma 3.6.6.

REMARK 3.6.4. An important advantage of bead norm over edge-length is that when one takes the repeated images $f_\#^k(\chi)$ of a monochromatic path, its length can decrease, due to cancellation within beads, whereas bead norm cannot.

In Definition 3.2.4 we defined the bead length $[S]_\beta$ of a corridor S in a well-folded diagram. It is convenient for our future arguments to concentrate on non-vanishing atoms, and hence on bead norm rather than bead length. However, an immediate consequence of the Bounded Cancellation Lemma is the following bi-Lipschitz estimate:

LEMMA 3.6.5. *Suppose S is a corridor in a well-folded corridor. Then*

$$\| S \|_\beta \leq [S]_\beta \leq B \| S \|_\beta .$$

3.6.1. The first decomposition of S_0. [cf. Subsection 1.6.1]

Let β be a bead in S_0 that is not a vanishing atom. As we follow the preferred future of β forwards in time, one of the following events must occur:

1. The last preferred future of β intersects the boundary of Δ nontrivially.
2. The last preferred future of β intersects a singularity nontrivially.
3. The last preferred future of β is bitten in a corridor S.

We remark that, unlike in Part 1, these events are not mutually exclusive; this is because a bead can consist of more than one edge.

We shall bound the bead norm of S_0 by finding a bound on the number of non-vanishing beads in each of the three cases.

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

3a. The preferred future of β is bitten by a bead that is not in the future of S_0.

3b. The preferred future of β is bitten by a bead that is in the future of S_0.

3.6.2. Bounding the easy bits. [cf. Subsection 1.6.2]

Label the non-vanishing beads which fall into the above classes $S_0(1)$, $S_0(2)$, $S_0(3a)$ and $S_0(3b)$, respectively. We shall see, just as in Part 1, that $S_0(3b)$ is by far the most troublesome of these sets.

The following lemma is proved in an entirely similar way to Lemmas 1.6.1 and 1.6.2, using the Bounded Cancellation Lemma and simple counting arguments.

Lemma 3.6.6.

(1) $\| S_0(1) \|_\beta \leq |\partial\Delta|$.
(2) $\| S_0(2) \|_\beta \leq 2B\, |\partial\Delta|$.
(3) $\| S_0(3a) \|_\beta \leq B\, |\partial\Delta|$.

We have thus reduced our task of bounding $\| S_0 \|_\beta$ to bounding the numbers of beads in $S_0(3b)$, i.e. to understanding cancellation *within* the future of S_0. The bound on the number of beads in $S_0(3b)$ is proved in an analogous way to Part 1, and takes up a large part of the remainder of this part of the book (through Section 3.11).

3.6.3. The chromatic decomposition. [cf. Subsection 1.6.3]

Fix a colour μ and consider the interval $\mu(S_0)$ in $\perp(S_0)$ consisting of beads coloured μ.

We shall subdivide $\mu(S_0)$ into five (disjoint but possibly empty) subintervals according to the fates of the preferred futures of the beads.

Let $l_\mu(S_0)$ be the rightmost bead β in $\mu(S_0)$ such that $f_\#(\check{\beta})$ contains a left-fast displayed edge ϵ so that the preferred future of ϵ is eventually bitten from the left from within the future of S_0. Let $A_1(S_0, \mu)$ be the set of beads in $\mu(S_0)$ from the left end up to and including $l_\mu(S_0)$.

Let $A_2(S_0, \mu)$ consist of those beads which are not in $A_1(S_0, \mu)$ but whose preferred futures are bitten from the left from within the future of S_0.

Let $A_3(S_0, \mu)$ denote those beads which do not lie in $A_1(S_0, \mu)$ or $A_2(S_0, \mu)$ and which fall into the set $S_0(1) \cup S_0(2) \cup S_0(3a)$.

All of the beads which are not in $A_1(S_0, \mu)$, $A_2(S_0, \mu)$ or $A_3(S_0, \mu)$ must have their preferred future bitten from the right from within the future of S_0.

Analogous to the definition of $l_\mu(S_0)$, we define a bead $r_\mu(S_0)$: the bead $r_\mu(S_0)$ is the leftmost bead β' so that $f_\#(\check{\beta}')$ contains a right-fast displayed edge whose preferred future is eventually bitten from the right from within the future of S_0.

Let $A_4(S_0, \mu)$ denote those beads which are not in $A_1(S_0, \mu)$, $A_2(S_0, \mu)$ or $A_3(S_0, \mu)$ and which lie strictly to the left of $r_\mu(S_0)$.

Finally, let $A_5(S_0, \mu)$ denote those edges not in $A_1(S_0, \mu)$, $A_2(S_0, \mu)$, $A_3(S_0, \mu)$ or $A_4(S_0, \mu)$ which lie to the right of $r_\mu(S_0)$ (include $r_\mu(S_0)$ in $A_5(S_0, \mu)$ if it has not already been included in one of the earlier sets).

Now Lemma 3.6.6 immediately implies

Lemma 3.6.7.

$$\sum_\mu \| A_3(S_0, \mu) \|_\beta \leq (3B + 1)\, |\partial \Delta|.$$

We also have

Lemma 3.6.8. *Let C_0 be the constant from Lemma 3.5.13 above. Then*

(1) $\| A_1(S_0, \mu) \|_\beta, \| A_5(S_0, \mu) \| \leq C_0$*; and*

(2) $|A_1(S_0, \mu) \smallsetminus l_\mu(S_0)|, |A_5(S_0, \mu) \smallsetminus r_\mu(S_0)| \leq C_0$.

Proof. We prove the bounds only for $A_1(S_0, \mu)$, the proofs for $A_5(S_0, \mu)$ being entirely similar.

The entire future of beads in $A_1(S_0, \mu)$ other than $l_\mu(S_0)$ must be eventually consumed from the left from within the future of S_0; cf. Lemma 1.5.9.

If $\| A_1(S_0, \mu) \|_\beta$ or $|A_1(S_0, \mu) \smallsetminus l_\mu(S_0)|$ were greater than C_0 then we would conclude from Lemma 3.5.13 that no left-fast bead in the immediate future of $l_\mu(S_0)$ could be bitten at any stage from the left from within the future of S_0, contrary to the definition of $l_\mu(S_0)$. □

As we continue to follow the proof from Part 1, our next goal is to reduce the task of bounding the bead norm of S_0 to that of bounding the number of Nielsen beads contained in $A_2(S_0, \mu)$ and $A_4(S_0, \mu)$. We focus exclusively on $A_4(S_0, \mu)$, the arguments for $A_2(S_0, \mu)$ being entirely similar.

In outline, our argument proceeds in analogy with the subsections beginning with Subsection 1.6.4, commencing with the decomposition of $A_4(S_0, \mu)$ into subintervals $C_{(\mu,\mu')}$. But we quickly encounter a new phenomenon that requires an additional section of argument – HNP cancellation – which does not arise in the case of positive automorphisms.

3.6.4. The decomposition of $A_4(S_0, \mu)$ into the $C_{(\mu,\mu')}$. All beads in $A_4(S_0, \mu)$ are eventually bitten from the right from within the future of S_0. For a colour $\mu' \neq \mu$, define a subset $C_{(\mu,\mu')}$ of $A_4(S_0, \mu)$ as follows: given a bead $\sigma \in A_4(S_0, \mu)$, there is a bead σ' in S_0 so that σ is eventually bitten by σ'. If σ' is coloured μ' then $\sigma \in C_{(\mu,\mu')}$.

The sets $C_{(\mu,\mu')}$ form intervals in S_0.

3.7. HNP-Cancellation and Reapers

The results of the previous section reduce the task of bounding $\| S_0 \|_\beta$ to that of establishing a bound on the sum of the bead norms of the monochromatic intervals $C_{(\mu,\mu')}$. In Part 1, the corresponding intervals (also labelled $C_{(\mu,\mu')}$) contained no exponential edges. In the current context, however, there may be exponential edges *trapped* in Nielsen paths, which may themselves be contained in beads of any type. This raises the concern that our attempts to control the length of the $C_{(\mu,\mu')}$ in the manner of Part 1 will be undermined by the *release* of these trapped edges when the Nielsen path is bitten, leading to rapid growth in subsequent nibbled futures of the Nielsen path. Our purpose in this section is to develop tools to control this situation, specifically Lemmas 3.7.22 and 3.7.23.

We must also deal with a second threat that arises from the phenomenon described in Example 3.7.6; we call this *Half Nielsen Path (HNP-) cancellation.*

Recall that a ΨEP is an edge path ρ in G; it is associated to a GEP and either ρ or $\bar\rho$ is of the form $E\bar\tau^k\bar\nu\gamma$ where E is an edge with $f_\#(E) = E \odot \tau^m$, where τ and ν are Nielsen paths, and $\bar\gamma\nu$ is a terminal segment of τ (and $m, k > 0$). These are the prototypes of the following types of paths.

Definition 3.7.1. Suppose that E is a linear edge with $f_\#(E) = E \odot \tau^m$, where τ is a Nielsen path and $m > 0$. Suppose further that ν is a Nielsen path and γ an edge-path so that $\bar\gamma\nu$ is a terminal segment of τ.

A PEP is a path ρ so that either ρ or $\bar\rho$ has the form $E\bar\tau^k\bar\nu\gamma$ where $k > 0$.

Remark 3.7.2. Every ΨEP is a PEP, but an arbitrary PEP has no GEP associated to it.

It is important to note that in the following definition the PEP being discussed is *not* assumed to be a bead in the decomposition of $\perp(S)$. (Beads along $\perp(S)$ are monochromatic whereas we want to discuss HNP cancellation, as in Definition 3.7.7, in the context of adjacent colours interacting.)

Definition 3.7.3 (HNP cancellation). Let S be a corridor in a well-folded diagram, let ε and ε' be edges in the naive (unfolded) future of $\perp(S)$ that cancel in the passage to $\top(S)$ and assume that ε is to the left of ε'.

Suppose further that the past of ε is e with label $\breve{e} = E$ a linear edge and that ε' is in the future of an edge e_γ whose label is an edge γ.

We call the cancellation of ε and ε' *left HNP-cancellation* and write $\varepsilon \lhd \varepsilon'$ if the interval from e to e_γ in $\bot(S)$ (inclusive) is labelled by a PEP of the form $E\overline{\tau}^k\overline{\nu}\varnothing\gamma$, where τ is a Nielsen path so that $\tau = \xi\nu$, where ξ and ν are Nielsen paths, and $\overline{\varnothing\gamma}$ is a terminal sub edge-path of ξ.

Right HNP-cancellation is defined by reversing the roles of ε and ε' and insisting upon a PEP in $\bot(S)$ of the form $\bar{\gamma}\bar{\varnothing}\nu\tau^k\bar{E}$. It is denoted $\varepsilon \rhd \varepsilon'$.

When we are unconcerned about the distinction between left and right, we refer simply to *HNP-cancellation*.

We extend this definition to concatenations of beaded paths in G by using the obvious stack-of-corridors diagram as in Subsection 3.2.2.

Remark 3.7.4. HNP-cancellation occurs at the 'moment of death' of the PEP; see Section 2.6 for an explanation of the significance of this moment and an analysis of it (in the language of ΨEPs).

Lemma 3.7.5. *Suppose that $E\overline{\tau}^k\overline{\nu}\varnothing\gamma$ is a* PEP *which exhibits an HNP-cancellation, as in Definition 3.7.3. Then $\varnothing$ is empty, so γ is the first edge of $\overline{\xi}$.*

Proof. The assumption that HNP-cancellation occurs means that we can restrict our attention to cancellation when tightening

$$f(E\overline{\tau}^k\overline{\nu}\varnothing\gamma).$$

This can be written as

$$E\tau^m f(\overline{\tau}^k\overline{\nu})f(\varnothing\gamma).$$

The path $\overline{\tau}^k\overline{\nu}\varnothing\gamma$ admits a hard splitting $\overline{\tau} \odot \cdots \odot \overline{\tau} \odot \overline{\nu} \odot \varnothing\gamma$. Therefore, under any choice of tightening, the m copies of τ cancel with the k copies of $f(\overline{\tau})$ (partially tightened), then with $f(\overline{\nu})$; they then begin to interact with $f(\varnothing\gamma)$. Just as in the proof of Proposition 2.6.9, under the assumptions of Lemma 2.5.1, there is only a single edge in $\varnothing\gamma$ whose future can interact with $f(E)$ when tightening. □

We now present the deferred example that explains the need to consider HNP-cancellation. This will also lead us to a further definition — *HNP biting* — that encodes a genuinely troublesome situation where HNP cancellation must be accounted[41] for. Fortunately, many other instances of HNP-cancellation are swept-up by our general cancellation and finiteness arguments, allowing us to avoid a detailed analysis of the possible outcomes.

[41] We usually account for it by excluding it from our definitions. When it cannot be excluded, we often sidestep it, using the notions of 'robust future' and 'robust past' given in Definitions 3.7.12 and 3.7.13 below.

The problem at the heart of the following example did not arise in Part 1 because the natural realisation of a positive automorphism does not map any linear edge across other linear edges.

EXAMPLE 3.7.6. Suppose that u is a Nielsen path, and that E_1 and E_2 are edges so that $f(E_i) = E_i u^k$ for $i = 1, 2$ and some integer $k > 0$. For any integer j, the path $\tau_j = E_1 u^j \overline{E_2}$ is an indivisible Nielsen path.

Suppose that E_3 is an edge so that $f(E_3) = E_3 \tau_j^l$, for some integers j and l (with $l > 0$). For ease of notation, we will assume that $l = 1$.

Consider the path $\rho = E_3 \overline{\tau}_j^r E_2$, for some $r > 0$. Then ρ is a PEP.

In the iterated images $f_\#(\rho)$, the visible copy of E_2 has a unique future labelled E_2, which we will call the 'preferred future' of E_2 for the purposes of this example. After $r+1$ iterations of ρ under $f_\#$ (and any choice of tightening at each stage), the future of E_3 cancels the preferred future of the visible copy of E_2. If we encode the evolution of ρ in a stack diagram as in Subsection 3.2.2 then the cancellation of E_2 is HNP-cancellation.

In the following discussion, we assume that the reader is familiar with Part 1, in particular the vocabulary of teams and reapers.

The phenomenon described in the above example causes problems when the sub-path $\rho_1 = \overline{\tau}_j^r E_2$ of ρ is monochromatic and E_2 is displayed in ρ_1. In this situation, it shows that the most obvious adaptation of Lemma 1.6.7 would be false. It is for this reason that we must exclude HNP-biting in Definition 3.8.7.

Similarly, because Example 3.7.6 renders a naive version of the results of Section 1.8 false, HNP-biting must be excluded from the Two Colour Lemma and the associated results in Section 3.9.

A situation in which we cannot exclude HNP-biting by decree arises in the analysis of teams and in particular the definition of a *reaper* (Subsection 3.7.3). Suppose that ρ labels some interval in the bottom of a corridor, with many copies of $\overline{u}$ to its immediate right. In this case, the edge ε_2 labelled E_2 will consume copies of $\overline{u}$ in the first r units of time, but its future will then be cancelled (assuming no other cancellation occurs from either side, and that there are no singularities, etc.). Since ε_2 was acting as the reaper of a team, we must find a continuing manifestation of it at subsequent times, for otherwise we will lose control over the length of teams (r being arbitrary) and the structure of our main argument will fail. This problem is solved by introducing the *robust future* of ε_2 (Definition 3.7.12), which in this case is an edge labelled E_1 that 'replaces' the preferred future of ε_2 when it is cancelled.

DEFINITION 3.7.7. Suppose that χ_1 and χ_2 are beaded paths in G and $\chi_1\chi_2$ is tight. Suppose that there is a bead $\rho_1 \subset \chi_1$ and a bead $\rho_2 \subset \chi_2$ so that

(1) either ρ_1 is a displayed edge γ in χ_1 which is linear or else ρ_1 is a displayed ΨEP in χ_1 of the form $E\bar{\tau}^k\bar{\nu}\gamma$, where γ is a linear edge;
(2) when tightening $f_\#(\chi_1)f_\#(\chi_2)$ to form $f_\#(\chi_1\chi_2)$, ρ_1 bites ρ_2 and the edge ε' in the exhibiting pair $(\varepsilon', \varepsilon)$ (see Definition 3.5.11) is in the future of γ;
(3) moreover[42], $\varepsilon' \rhd \varepsilon$.

Under these circumstances we say that ρ_2 is *left-HNP-bitten* by ρ_1 and we write $\rho_1 \LEFTcircle \rho_2$. There is an entirely analogous definition of *right-HNP-biting* $\rho_1 \RIGHTcircle \rho_2$, and when we are unconcerned about the direction we will refer simply[43] to *HNP-biting.*

We make the analogous definition for HNP-biting within diagrams.

Definition 3.7.8. Suppose that χ_1 and χ_2 are beaded paths and that ρ_1 is a bead in χ_1. We say that ρ_1 is *eventually HNP-bitten* by χ_2 if ρ_1 is eventually bitten by χ_2 (Definition 3.5.12) and this biting is HNP-biting.

We make the analogous definition within diagrams.

Definition 3.7.9. Suppose that E and E' are edges in G. We say that E and E' are *indistinguishable* if there is a Nielsen path τ and an integer $s > 0$ so that $f(E) = E\tau^s$ and $f(E') = E'\tau^s$.

The edges E_1 and E_2 in Example 3.7.6 are indistinguishable.

3.7.1. Parabolic HNP-cancellation and robust futures. The following is a simple (but key) observation, and has an obvious application to HNP-cancellation of edges of parabolic weight.

Lemma 3.7.10. *Suppose that τ, ν, ν' and σ are Nielsen paths, with σ irreducible and $\tau = \nu'\bar{\sigma}\nu$. Suppose further that γ is the initial edge of σ, and that $f(\gamma) = \gamma \odot \xi^l$ for some Nielsen path ξ. Then σ has the form $\gamma\xi^r\overline{\gamma'}$ where r is some integer and γ' is an edge so that γ and γ' are indistinguishable.*

Moreover, suppose that E is an edge so that $f(E) = E \odot \tau^m$, and let $\rho = E\bar{\tau}^i\bar{\nu}\gamma$ be a PEP *with $0 \le i < m$. Then $f_\#(\rho)$ has the form $E \odot \tau^{m-i-1}\nu'\gamma'\bar{\xi}^j$ where γ and γ' are indistinguishable.*

Proof. The first assertion is an immediate consequence of the structure of indivisible Nielsen paths of parabolic weight, and the second is then obvious (a detailed analysis of the Nielsen paths of parabolic weight is undertaken in Section 2.1). □

Definition 3.7.11. In general, non-displayed edges ε in diagrams do not have preferred futures. But if $\check{\varepsilon}$ has parabolic weight, there is a unique edge of the same weight in $f_\#(\check{\varepsilon})$, and it is natural to define the (immediate) preferred

[42]The PEP implicit in the symbol $\rhd$ is not the ΨEP in (1).

[43]We swap orientation in Definition 3.7.8 so as to emphasize this point immediately.

future of ε to be the corresponding edge in the immediate future of ε. (If ε happens to be displayed, this agrees with our earlier definition.)

In Section 3.9, when proving the Pincer Lemma, we will have to exclude HNP-biting. This will also be the case in the applications of the Pincer Lemma in Sections 3.10 and 3.11. Thus, in following the future of a linear edge γ when HNP-cancellation occurs, we would like to ignore the preferred future (which disappears), and rather follow the future of the indistinguishable edge γ' from Lemma 3.7.10 above. Thus we make the following

DEFINITION 3.7.12 (Robust Futures for Parabolic Edges). Suppose that ε is a (not necessarily displayed) edge in a colour $\mu(S)$, and that $\breve{\varepsilon}$ is contained in a parabolic stratum. If the preferred future of ε is cancelled from the left [resp. right] by HNP-cancellation in $\top(S)$, then Lemma 3.7.10 provides an edge γ' that is indistinguishable from $\breve{\varepsilon}$ and survives in the tightened path $f_\#(E\bar{\tau}^k\bar{\nu}\o\gamma)$ [resp. its reverse] considered in Definition 3.7.3.

We define the *robust future* of an edge $\varepsilon \subseteq \bot(S)$ as follows. If the preferred future of ε survives in $\top(S)$, then the robust future of ε is just the preferred future of ε. If the preferred future is cancelled by HNP-cancellation, then the robust future of ε is the above edge labelled γ', provided this survives in $\top(S)$. Otherwise there is no robust future.

DEFINITION 3.7.13 (Robust Pasts for Linear Edges). Let ε' be an edge of $\top(S)$ and suppose that both it and its immediate past are labelled by linear edges. If ε' is not the robust future of any edge then the robust past of ε' is the past of ε'. But if ε' *is* the (immediate) robust future of ε then the robust past of ε' is ε.

Just as for preferred futures, the notions of robust future and robust past can be extended arbitrarily many steps forwards or backwards in time by iterating the definition.

3.7.2. A setting where we require cancellation lemmas. Consider the following situation. Let $\chi_1\sigma\chi_2$ be a tight path in G with χ_1 and χ_2 monochromatic and σ a path with a preferred decomposition into monochromatic paths (each of which comes equipped with a beaded decomposition). We will analyse the possible interaction between χ_1 and χ_2 in iterates of $\chi_1\sigma\chi_2$ under f (where the tightening follows the convention of Subsection 3.5.3).

As ever, the following lemma remains valid with left/right orientation reversed.

LEMMA 3.7.14. *Suppose that χ_1, χ_2 and σ are as above, and suppose that each non-vanishing bead in χ_2 is eventually bitten by a bead from χ_1 in some iterated image $f_\#^k(\chi_1\sigma\chi_2)$ of $\chi_1\sigma\chi_2$.*

Suppose further that ρ is a bead in χ_2 so that $f_\#(\rho)$ has parabolic weight, and that ρ is eventually left-HNP-bitten by a bead from χ_1 in the evolution of $\chi_1\sigma\chi_2$. Then ρ is the rightmost non-vanishing bead in χ_2.

PROOF. Pass to the iterate $f_\#^{k-1}(\chi_1\sigma\chi_2)$ so that the preferred future of ρ lies in a PEP π, which exhibits the (eventual) HNP-biting of ρ in the tightening to form $f_\#^k(\chi_1\sigma\chi_2)$. Let ρ_1 be the preferred future of ρ in $f_\#^{k-1}(\chi_1\sigma\chi_2)$. Since $f_\#(\rho)$ has parabolic weight, ρ_1 has parabolic weight, and is either a displayed edge or a displayed ΨEP or GEP. We must prove that no bead to the right of ρ_1 is eventually bitten by the future of χ_1.

By Definition 3.7.7 and Lemma 3.7.5 the PEP π has the form $\gamma\bar{\tau}^k\bar{\nu}\varepsilon$, where

(1) γ is an edge so that $f(\gamma) = \gamma \odot \tau^m$;
(2) γ is either a displayed edge in the future of χ_1 in $f_\#^{k-1}(\chi_1\sigma\chi_2)$ or else if the rightmost edge in a displayed ΨEP; and
(3) ε is contained in ρ_1.

Let α be the displayed edge or ΨEP containing γ.

Let ρ_1' be the terminal part of ρ_1 from ε to its right end, and let χ_2' be the terminal part of the future of χ_2 in $f_\#^{k-1}(\chi_1\sigma\chi_2)$, from ε to its right end.

Since ρ_1 is displayed, we have $\chi_2' = \rho_1' \odot \beta$ for some path β.

By Lemma 3.7.10, when tightening to form $f_\#^k(\chi_1\sigma\chi_2)$, the edge ε is replaced by an indistinguishable edge ε' which comes from the future of α. Suppose that δ is that part of $f_\#(\alpha\rho_1')$ from ε' to the right end. Since α is a (linear) edge or a ΨEP, the edge ε' survives in all iterates of α (under any choices of cancellation). Similarly, since ε and ε' are indistinguishable, ε' survives in all iterates of δ (under any choices of tightening). This implies that we have a hard splitting $f_\#(\alpha\chi_2') = f_\#(\alpha\rho_1') \odot f_\#(\beta)$, and the fact that α is displayed implies that no bead in β can be eventually bitten by the future of χ_1, as required. □

In applications of Lemma 3.7.14 (and of Lemmas 3.7.22 and 3.7.23 below), we usually take $\chi_1 = \check{\mu}_1(S)$ and $\chi_2 = \check{\mu}_2(S)$, where μ_1 and μ_2 are colours and S is some corridor, and we will choose σ to be the label of that part of $\bot(S)$ which lies strictly between $\mu_1(S)$ and $\mu_2(S)$.[44] Since the folding conventions of Subsections 3.1.2 and 3.5.3 are compatible, and because of the hardness of our splittings, the interaction between μ_1 and μ_2 in the future of S can be analysed by studying the interaction between the futures of χ_1 and χ_2 in iterated images of $\chi_1\sigma\chi_2$ under f.

3.7.3. Reapers. In Part 1 proving the existence of reapers was straightforward (see Section 1.9). In the current context, however, we have to work harder to prove that a suitable incarnation of a reaper exists, because of the

[44]However, it will also be convenient sometimes to take χ_1 to be a subinterval of $\check{\mu}_1(S)$ consisting of an interval of beads.

phenomena discussed in the preceding subsection. At the heart of our difficulties is the fact that Nielsen atoms need not be single edges.

DEFINITION 3.7.15. A *beaded Nielsen path* in a corridor S is a subinterval $\sigma \subset \bot(S)$ so that $\breve{\sigma}$ is a beaded path all of whose beads are Nielsen paths.

Note that in the above definition we do not assume that σ is a single colour, or even that each bead in $\breve{\sigma}$ is contained in a single colour. Examples of beaded Nielsen paths include that part of a GEP between the extremal edges, and the sub-paths $\overline{\tau}^i$ of a PEP $E\overline{\tau}^k\bar{\nu}\not{o}\gamma$.

Although the beads in a beaded Nielsen path might not be displayed in a path $\mu(\breve{S})$, it is still possible to define the future of a bead in a beaded Nielsen path, and the notions of preferred future and biting still make sense. We will use this observation in the sequel.

The following notion is parallel to that of Definition 1.10.1, which was pivotal in the bonus scheme (cf. Section 3.11 below). Here, it plays a more central role.

DEFINITION 3.7.16 (Swollen present and swollen future). Suppose S is a corridor and that $I \subseteq \bot(S)$ is a beaded Nielsen path in S. The *swollen present* of I is the[45] maximal subinterval $I' \subseteq \bot(S)$ such that (i) $I \subseteq I'$; (ii) I' is a beaded Nielsen path in S; and (iii) the beads of I are beads of I'.

The *left swollen present* of I is that part of the swollen present from the left end up to the right end of I, whilst the *right-swollen present* goes from the left end of I to the right end of the swollen present.

If the actual future of I is a beaded Nielsen path the (immediate) *swollen future* $sw_1(I)$ of I is the swollen present of the (actual) future of I. With a similar qualification, the *swollen future* $sw_k(I)$ at $\text{time}(S) + k$ is defined to be $sw_1(sw_{k-1}(I))$.

With the same qualifications, the left and right swollen futures are defined in the obvious ways.

The first qualification in the above definition is required because it is possible that the immediate future of a beaded Nielsen path is not a beaded Nielsen path. Thus we must be careful only to apply this concept in cases where we know the swollen future to exist.

DEFINITION 3.7.17 (Reapers). Suppose that S is a corridor and $I \subset \bot(S)$ is a beaded Nielsen path in S with nonempty swollen future $sw_1(I)$. Suppose that α is an edge in $\bot(S)$ immediately adjacent to I on the left. We say that α is a *left-reaper for I* if (i) $\breve{\alpha}$ is a linear edge; (ii) $\breve{\alpha}$ bites some of the future of $\breve{I}$ in $f_\#(\breve{\alpha I})$; and (iii) the robust future of α is immediately adjacent to $sw_1(I)$ in $\top(S)$.

[45]Uniqueness is immediate from the observation that if a terminal sub-path σ of a Nielsen path τ is itself Nielsen then σ is a concatenation of beads in τ.

There is an entirely analogous definition of *right-reapers*. As usual, when we are unconcerned about the direction we will refer to *reapers*.

Definition 3.7.18 (Left-edible). Let S be a corridor in a well-folded diagram, and $I \subset \bot(S)$ a beaded Nielsen path. We say that I is *left-edible* if each bead in I is eventually bitten by a bead coloured μ in the future of S, where $\mu(S)$ lies to the left of I.

Right-edible paths are defined with a reversal of the left-right orientation.

In the remainder of this section we work towards proving Propositions 3.7.19 and 3.7.21.

Proposition 3.7.19. *Let S be a corridor in a well-folded diagram and $I \subset \bot(S)$ a left-edible path so that $|I| \geq B + J$. Then the immediate future of I in $\top(S)$ is left-edible.*

The following lemma is straightforward, and allows us to focus our attention on the time when cancellation between colours begins.

Lemma 3.7.20. *Let S be a corridor in a well-folded diagram and let $I \subset \bot(S)$ be a left-edible colour, all of whose beads are eventually bitten by beads coloured μ. Let S^I be the corridor in the future of S so that the first biting of a bead in the left swollen future of I by something coloured μ occurs in S^I. Then the left swollen future of I in $\bot(S^I)$ is left-edible.*

In the following statement B is the Bounded Cancellation Constant from Proposition 3.1.5 and J is the constant from the Beaded Decomposition Theorem 3.2.1. The corridor S^I is as in Lemma 3.7.20 above, and I^λ is the left swollen future of I in S^I.

Proposition 3.7.21. *Suppose that S is a corridor in a well-folded diagram and $I \subset \bot(S)$ is a left-edible path, all of whose beads are eventually bitten by beads coloured μ. Suppose also that $|I| \geq B + J$. Then*

(1) the immediate future of I^λ in $\top(S^I)$ has an associated left reaper α, which is coloured μ; and

(2) for each bead in the immediate future of I^λ, when it is eventually bitten the biting is by the robust future of α.

3.7.4. Two Cancellation Lemmas. The following lemma is useful in the proof of Lemma 3.8.8 below. We record it now because a variation on it (Lemma 3.7.23) is needed in the proof of Proposition 3.7.21.

We revert to the setting described in Subsection 3.7.2.

Lemma 3.7.22. *Assume that in the iterates of $\chi_1\sigma\chi_2$ (i.e. forward-images under $f_\#$) each bead in χ_2 is eventually bitten by a bead in χ_1. Suppose that χ_2 has weight i, where H_i is an exponential stratum, and that all beads of weight i in χ_2 are Nielsen beads. Let ρ be a bead of weight i in χ_2.*

(1) If ρ is not bitten in $f_\#(\chi_1\sigma\chi_2)$ but is eventually bitten in the image $f_\#^k(\chi_1\sigma\chi_2)$ then ρ is entirely consumed in $f_\#^k(\chi_1\sigma\chi_2)$.
(2) If ρ is bitten but not entirely consumed in $f_\#(\chi_1\sigma\chi_2)$ then ρ is the rightmost bead in χ_2.

Proof. There is at most one indivisible Nielsen path of weight i and the lemma is vacuous unless there is exactly one.

Let β be a bead in χ_2 of weight i, and suppose that an edge η in the future of χ_1 is the edge which cancels the rightmost edge in the preferred future of β to exhibit the biting of β by χ_1. Since β is an indivisible Nielsen path, it has edges of weight i on both ends, as does its preferred future, and so η has weight i. Suppose that the past of η in $\chi_1\sigma\chi_2$ has weight i. Then by Theorem 2.8.1 and Assumption 3.4.7, η is either a displayed edge in the future of χ_1, or else is contained in a Nielsen bead. Suppose first that η is contained in a Nielsen bead τ. Since η is to cancel with an edge in β, the path τ must have weight i. Hence $\tau = \bar{\beta}$, and β is entirely consumed when it is bitten.

Suppose then that η is displayed in the future of χ_1. By Assumption 3.4.7.(5) we may assume that the edge η is contained in a displayed path of the form $f(\eta)$. Since $f(\eta)$ is i-legal, and β is not, it is not possible for the illegal turn in β (of weight i) to be cancelled by any iterates of η. However, $|f(\eta)| > |\beta|$, by Assumption 3.4.7(1), so it is not possible for the displayed copy of $f(\eta)$ to be cancelled by the future of β. Therefore, in this case β must be the rightmost bead in χ_2.

Furthermore, suppose that β and η are as above, and the past of η in $\chi_1\sigma\chi_2$ has weight i, and suppose moreover that β is not bitten in $f_\#(\chi_1\sigma\chi_2)$. Then β is bitten by η in some $f_\#^k(\chi_1\sigma\chi_2)$, and $k \geq 2$. Thus we may assume that the immediate past of η is also displayed and is η. By applying Lemma 3.4.5 and noting that the rightmost edge of β must be $\bar{\eta}$, we see that the sub-path between the immediate past of β and the immediate past of η has the form $\cdots\bar{\eta}\omega\eta\cdots$ for some path ω. The path ω must start and finish at he same vertex, and in order for the written copy of $\bar{\eta}$ to cancel with the written copy of η it must be that $f_\#(\omega)$ is a point. However, ω is not a point, because otherwise the past of β and the past of η would already cancel. This contradicts the fact that f is a homotopy equivalence. The same argument shows that if η is contained in a Nielsen bead and β is not bitten in $f_\#(\chi_1\sigma\chi_2)$ then β cannot be bitten by η.

Therefore, if β is bitten by an edge η whose past in $\chi_1\sigma\chi_2$ has weight i then β is close to the left end of χ_2, and is either entirely consumed when bitten or is the rightmost bead in χ_2.

We may now assume that the bead ρ is cancelled by an edge η whose past in χ_2 has weight greater than i. The above arguments show that we may assume that the immediate past of η also has weight greater than i, and by Lemma 3.4.5 we may assume that this past is contained in a displayed edge, a GEP, or a ΨEP. It is easy to see that the immediate past of η cannot have exponential weight and cannot be a GEP. Thus we may assume that the immediate past of η is either the edge on the left end of a ΨEP of the form $\gamma\nu\tau^k\overline{E}$, (and that the edge γ is parabolic) or else is displayed and parabolic.

Lemma 3.4.5 and the above arguments imply that this immediate past of η must be a linear edge, and the above arguments now imply that if ρ is bitten in a corridor it must be entirely consumed. □

The following variant of Lemma 3.7.22 is the one we need in the proof of Proposition 3.7.21. We continue to study $\chi_1\sigma\chi_2$ as in Subsection 3.7.2.

LEMMA 3.7.23. *Suppose that χ_2 is a beaded Nielsen path and each of its beads is eventually bitten by a bead in χ_1 in some iterated image of $\chi_1\sigma\chi_2$ under f.*

Let ρ be a bead in χ_2 which is not bitten in $f_\#(\chi_1\sigma\chi_2)$. If ρ is bitten but not consumed in some iterated image of $\chi_1\sigma\chi_2$ then ρ is the rightmost bead in χ_2.

PROOF. We follow the proof of Lemma 3.7.22 above, with the added wrinkle that there may be parabolic weight Nielsen paths to consider in χ_2. In this case there needn't be a unique Nielsen path of weight i.

Suppose that ρ is as in the statement of the Lemma. If ρ has exponential weight, then the arguments of the proof of Lemma 3.7.22 give the required properties. If ρ has parabolic weight, Lemma 3.5.15 implies that when ρ is bitten by an edge η in the future of χ_1, the immediate past of η has weight greater than that of ρ. Also, this immediate past must be parabolic. Arguing as in the proof of Lemma 3.7.22, one sees that either ρ is entirely consumed when bitten, or else ρ is the rightmost bead in χ_2. □

COROLLARY 3.7.24. *Suppose that I is a beaded Nielsen path in $\perp(S)$ for some corridor S of a well-folded diagram, and suppose that all beads of I are eventually bitten from the left by beads in a single colour μ. Then, with the possible exception of B beads on the left end and one bead on the right (the final one bitten), whenever μ bites a Nielsen bead in the future of I, it consumes it entirely.*

Proof of the Proposition 3.7.19

PROOF. If the immediate future of I in $\top(S)$ were not left-edible, then Corollary 3.7.24 would ensure that no bead in I which is not bitten in S is ever bitten by μ. However, the assumption on the length of I (and the Bounded Cancellation Lemma) ensure that there *are* beads in I not bitten in S. The fact that I *is* left-edible therefore ensures that the future of I in $\top(S)$ is also left-edible. □

Proof of the Proposition 3.7.21

PROOF. Let S' be the corridor containing the immediate past of I^{λ}. Lemma 3.7.23 implies that in $\top(S')$ there is an edge ρ in μ which cancels a whole Nielsen path in the future of I.

Since $|I| \geq B + J$, there is a bead in I not bitten in $\top(S)$. The proof of Lemma 3.7.23 now implies that there is a reaper as in the statement of the proposition. □

3.8. Non-fast and Unbounded Beads

With the technical exertions of the previous section behind us, we are now able to return to the main argument, picking up the flow of Part 1 at Subsection 1.6.6. Thus our next purpose is to reduce the task of bounding the bead norm of the intervals $C_{(\mu,\mu')}$ to that of bounding the lengths of certain long blocks of Nielsen atoms. These blocks are the analogue of the intervals $C_{(\mu,\mu')}(2)$ from Part 1, and will be the building blocks of the *teams* introduced in Section 3.10 (in analogy with Section 1.9).

DEFINITION 3.8.1. Suppose that $\rho = \gamma\nu\tau^k\overline{E_i}$ is a PEP (with $k \geq 0$). We say that ρ is *left-slow* if γ is empty or a concatenation of left-slow beads.

There is an entirely analogous definition of *right-slow* PEPs of the form $\rho = E_i\overline{\tau}^k\overline{\nu\gamma}$.

Often, we will just speak of *slow* PEPs, since a single PEP can only be left-slow or right-slow, but not both.

DEFINITION 3.8.2. Suppose that the bead ρ is such that $f_{\#}(\rho)$ is not a Nielsen bead. Then the function $n \mapsto |f_{\#}^n(\rho)|$ grows at least linearly. In this case, we call ρ an *unbounded bead*.

DEFINITION 3.8.3. A beaded path is called *right-tame* if all of its beads are GEPs, slow ΨEPs, Nielsen paths and atoms which do not have a right-fast displayed edge in their immediate future.

The next lemma follows immediately from the definition.

LEMMA 3.8.4. *$A_4(S_0, \mu)$ is a right-tame path.*

LEMMA 3.8.5. *Suppose that α is a non-vanishing atom which is not right-fast. Then either all of the beads in $f_{\#}(\alpha)$ are Nielsen paths and* GEP*s, or else the preferred future of α is parabolic.*

PROOF. The only modification to Lemma 3.4.3 is the exclusion of exponential edges in the second case, which is valid because such an edge would obviously contradict the fact that α is not right-fast. □

Definition 3.8.6. Suppose that σ is a right-tame path. The *untrapped weight* of σ is the largest j so that $f_{\#}(\sigma)$ contains a bead of weight j which is not Nielsen.

Definition 3.8.7. Suppose that, for some pair $(\mu, \mu') \in \mathcal{Z}$, the untrapped weight of $C_{(\mu,\mu')}$ is j. For each $1 \le i \le j$, define ρ_i to be the leftmost bead in $C_{(\mu,\mu')}$ so that $f_{\#}(\rho_i)$ has an unbounded bead of weight at least i that is not HNP-bitten in the future of S_0.[46]

Let $\mathcal{E}_i$ denote those beads in $C_{(\mu,\mu')}$ from the right end up to and including ρ_i, and let $\mathcal{D}_i = \mathcal{E}_i \smallsetminus \mathcal{E}_{i+1}$.

The following is the analogue of Lemma 1.6.7

Lemma 3.8.8. *For all $1 \le i \le \omega$ there is a constant $C_1(i)$ so that for each of the paths $C_{(\mu,\mu')}$ and decomposition into intervals $\mathcal{D}_i$ as above, we have*

$$\| \mathcal{D}_i \|_{\beta} \le C_1(i).$$

Proof. As far as possible, we try to follow the proof of Lemma 1.6.7. However, due to the phenomena described in Section 3.7, the proof here is somewhat more complicated.

We go forward to the time, t say, which is one step before the moment when μ' first starts to bite the preferred futures. By virtue of Remark 3.6.4, and the definition of $\mathcal{D}_i$, there are at least as many beads in the future of $\mathcal{D}_i$ at time t as there are in S_0. Therefore, it is sufficient to bound the number of beads in the future of $\mathcal{D}_i$ at time t; to ease the notation, we write $\mathcal{D}_i$ for this future, i.e. pretend that $t = \text{time}(S_0)$.

It is possible that there exist beads $\rho \in \mathcal{D}_i$ so that $f_{\#}(\rho)$ has weight greater than i. In such a case, all of the beads in $f_{\#}(\rho)$ of weight greater than i are Nielsen beads.

Consider the highest weight k for which there is a bead ρ in $\mathcal{D}_i$ with $f_{\#}(\rho)$ of weight k, and suppose that $k > i$. Suppose first that ρ has exponential weight. Then by Lemma 3.7.22 either $\mathcal{D}_i$ has bead norm at most B (and length at most $\ell = JB(B+1)$), or else ρ is entirely consumed when it is bitten. In the first case ρ is the leftmost bead in $\mathcal{D}_i$, and also in $C_{(\mu,\mu')}$. A similar argument applies when ρ has parabolic weight.

Thus, excluding cases where $|\mathcal{D}_i| < \ell$, we may treat the Nielsen beads of weight higher than i as indivisible units, which are entirely consumed when bitten. We are therefore in the situation of the proof of Lemma 1.6.7, where the unbounded beads in $\mathcal{C}_i$ grow apart at a linear rate, and so must be cancelled quickly. Otherwise, the proof is entirely parallel to the one from Part 1. ☐

We are trying to reduce the task of bounding the bead norm to that of bounding the size of intervals consisting entirely of Nielsen beads, which are

[46] Note that it is possible that $\rho_i = \rho_{i+1}$ for some i.

each consumed by a reaper. In order to make this reduction, we still have some HNP-biting to deal with. In order to deal with this, we need an analogue of Lemma 1.9.4.

Recall that L is the maximal length of $f(E)$ where E is an edge in G.

PROPOSITION 3.8.9 (cf. Lemma 1.9.4). *There is a constant C_4 depending only on f which satisfies the following properties. If I is an interval on $\top(S)$ labelled by a beaded path all of whose beads are Nielsen atoms, then the path labelling the past of I in $\bot(S)$ is of the form $u\alpha v$ where α is a beaded path all of whose beads are Nielsen atoms and $|u|$ and $|v|$ are less than C_4.*

If the past of I begins (respectively ends) with a point fixed by f, then u (respectively v) is empty.

In particular, $|I| \leq |\alpha| + 2LC_4$.

PROOF. The interval $I \subset \top(S)$ is a beaded path, all of whose beads are Nielsen paths of length at most J. Therefore, along I there are points where I admits a hard splitting and these points occur with a frequency of at least one every J edges. Since these points are vertices, the set of labels of points at which the splitting occurs is finite. Consider the path from $\top(S)$ to $\bot(S)$ starting from one of these vertices. The label of this path is $w\bar{t_i}$ where w is a (possibly empty) path in G of length at most L, and t_i is one of the edges from the mapping torus $M(f)$. (We are about to use a finiteness argument and it will be important that the repetition we infer includes the labels of the points on $\bot(S)$. Thus it is important which of the t-edges this path includes.)

Since the data we record — the label of the vertex on $\top(S)$, the path $w\bar{t_i}$ and the label of the end of this path on $\bot(S)$) — range over a finite set, there is a constant C' such that in the interval within C' vertices of the left end of I there will be repetition of these data. Since the vertices occur at least every J edges, this repetition occurs within $C'J$ of the left end of I.

Once we have found this repetition, we have an interval $\lambda \subset \bot(S)$, an interval $\eta \subset \top(S)$ and a path w_0 of length at most L such that $f_\#(\lambda) = w_0\eta\bar{w}_0$. Therefore, the free homotopy class of $f_\#(\lambda)$ is the same as that of $\eta = f_\#(\eta)$, since η is a beaded path all of whose beads are Nielsen paths. Since f is a homotopy equivalence, the free homotopy class of λ must be the same as that of η.

Suppose that $\eta = p_1 \dots p_m$ where each p_i is an indivisible Nielsen path. Now, λ is tight, so $\lambda = \sigma p_i p_{i+1} \dots p_m p_1 \dots p_{i-1}\bar{\sigma}$, for some path σ. Thus, if '$\sim$' denotes free homotopy,

$$f(\lambda) \sim f_\#(\sigma) p_i \dots p_{i-1} f_\#(\bar{\sigma}),$$

which tightens to

$$w_0 p_1 \dots p_m \bar{w}_0.$$

By the Bounded Cancellation Lemma, tightening the path $f(\lambda)$ as written above reduces the length of $f_\#(\sigma)$ by less than B, and the result has length at most $2L+|\eta|$. This implies that $|f_{\#}(\sigma)| < L+B$. Therefore, $\| \sigma \|$ is bounded, and by a small increase we may also assume that $i = 1$. By considering only one vertex out of every $B(L+B)$, we can find such a path η where there is some p_j in the middle of λ such that the path from the copy of $p_j \subset \top(S)$ to the copy of $p_j \subset \bot(S)$ is a single edge labelled t, for some j.

We have argued that, for some path η of bounded length which lies on the left end of I, the past of η is of the form $u\eta u'$ where $|u|$ and $|u'|$ are bounded, and the paths from the splitting points in $\eta \subset I$ to $\bot(S)$ consist of single edges labelled t.

Consider the analogous situation on the right end of I. We can find a path $\eta' \subset I$ lies at the right end of I such that the past of η' is of the form $v'\eta'v$ where $|v|$ and $|v'|$ are bounded and the paths from the vertices of $\eta' \subset I$ to $\bot(S)$ consist of single edges labelled t.

Consider the paths along $\bot(S)$ and $\top(S)$ from the left end of η to the right end of η'. We have a path $\rho \subset \bot(S)$ with fixed points of f on either end which maps to a Nielsen path $f_{\#}(\rho) \subset I \subset \top(S)$. The same argument as in the proof of Lemma 2.1.14 then shows that $\rho = f_{\#}(\rho)$. Hence ρ is a beaded path, all of whose beads are Nielsen paths, and the paths u and v on either side of ρ are of bounded length as required. This proves the first assertion in the statement of the lemma.

The second assertion follows similarly, and the final assertion follows immediately from the first. ☐

Consider a pair $(\mu, \mu') \in \mathcal{Z}$, and recall the definition of the subintervals $\mathcal{E}_i$ from Definition 3.8.7.

PROPOSITION 3.8.10. *There is a constant C_5, depending only on f so that the following holds. For each $(\mu, \mu') \in \mathcal{Z}$, the interval $C_{(\mu,\mu')} \smallsetminus \mathcal{E}_1$ in $A_4(S_0, \mu)$ has the form uNv where u and v are such that $\| u \|_{\beta}, \| v \|_{\beta} \leq C_5$ and N is a beaded path all of whose beads are Nielsen beads.*

PROOF. By Lemma 3.7.14, for each adjacency of colours (μ, μ') there can only be one bead in $\mu(S)$ which is eventually HNP-bitten by μ'.

The result now follows from Proposition 3.8.9 and the definition of $\mathcal{E}_1$. ☐

DEFINITION 3.8.11. For $(\mu, \mu') \in \mathcal{Z}$, define $C_{(\mu,\mu')}(2) := N$, the beaded Nielsen path from Proposition 3.8.10.

The sum of our arguments to this point has reduced the task of bounding the sum of the bead norms of the intervals $\mu(S_0)$ in S_0 to that of bounding the sum of the lengths of the intervals $C_{(\mu,\mu')}(2)$ for pairs $(\mu, \mu') \in \mathcal{Z}$.

We summarise the results from this section as follows.

PROPOSITION 3.8.12. *There is a constant C_1, depending only on f, so that*

$$\| C_{(\mu,\mu')} \|_{\beta} \leq \| C_{(\mu,\mu')}(2) \|_{\beta} + C_1.$$

REMARK 3.8.13. Since the intervals $C_{(\mu,\mu')}(2)$ consist entirely of Nielsen beads, we have the following obvious relationship between length and bead norm:

$$|C_{(\mu,\mu')}(2)| \leq \| C_{(\mu,\mu')}(2) \|_\beta \leq J|C_{(\mu,\mu')}(2)|.$$

Therefore, in order to finish the bound on bead norm, it is sufficient to bound the total lengths of the intervals $C_{(\mu,\mu')}(2)$.

It is important for the remainder of the paper that the path $C_{(\mu,\mu')}(2)$ is a beaded path that consists entirely of Nielsen atoms. This is a stronger statement than just asserting it is a Nielsen path, since we require a decomposition into beads of uniformly bounded size, each of which is a Nielsen path. This makes the path $C_{(\mu,\mu')}(2)$ very similar to the long blocks of constant letters which played such a prominent role in Part 1

At this point the reader may benefit from consulting Section 1.7, which outlines the strategy for the remainder of the proof of the Main Theorem (the strategy from the positive case still holds here). For the remainder of this part of the book, we will mostly continue without reminding the reader of this strategy.

3.9. The Pleasingly Rapid Disappearance of Colours

We are now at the point in our arguments where we need to formulate and prove the Pincer Lemma, as in Section 1.8. In Part 1 the Pincer Lemma was proved by counting colours which *essentially vanished*, which is to say they came to consist entirely of constant letters. For positive automorphisms, this is a well-defined event and can only occur once for each colour. For general automorphisms, the analogues of constant letters are indivisible Nielsen paths. However, since Nielsen paths can contain non-constant edges, indivisible Nielsen paths are not indivisible in an absolute sense (the terminology refers to the fact that an indivisible Nielsen path cannot be split into two Nielsen paths). Thus, it is possible that a colour can be labelled by a Nielsen path at some time t but not at some later time $t + k$. There are two ways to circumvent this problem. The first is to concentrate on the times when a colour decreases in weight, whilst the second is to focus on the times when a colour becomes Nielsen and seek compensation when a colour subsequently ceases to be Nielsen. We mostly pursue the second idea but there are aspects of the first also.

The version of the Pincer Lemma which we need in this part of the book is Theorem 3.9.27.

The ideas in the proof of the Pincer Lemma here are very similar to those in Part 1 but the execution is somewhat different.

Definition 3.9.1. Suppose that I is a non-empty beaded Nielsen path and that U and V are beaded paths. We say that I is *stably Nielsen* in the path UIV if the future[47] of I in $f_\#(UIV)$ is also a non-empty Nielsen beaded path.

Suppose that μ_1, μ_2 and μ_3 are colours in a well-folded diagram and that the intervals $\mu_1(S), \mu_2(S)$ and $\mu_3(S)$ are non-empty and adjacent in $\perp(S)$. If $\check{\mu_2(S)}$ is a non-empty Nielsen path, then we say that $\mu_2(S)$ is *stably Nielsen* if, in the above sense, $\check{\mu_2(S)}$ is stably Nielsen in $\check{\mu_1(S)}\check{\mu_2(S)}\check{\mu_3(S)})$.

Lemma 3.9.2 (Relative Buffer Lemma). *Let $i \in \{1, \dots, \omega - 1\}$ and let $I \subset \perp(S)$ be an edge-path labelled by edges in G_i. 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 edge in the future of $\mu_1(S)$ with label in $G \smallsetminus G_i$ will ever cancel with an edge in the future of $\mu_2(S)$ with label in $G \smallsetminus G_i$.*

Proof. Given Lemmas 3.4.5 and 3.4.6, the proof of Lemma 1.8.1 applies modulo changes of terminology. □

We now need the following 'two-sided' version of Proposition 3.7.19.

Lemma 3.9.3. *Let μ_1, μ_2, μ_3 and S be as in Definition 3.9.1, and suppose that $\mu_2(S)$ is stably Nielsen. Then for all corridors S' in the future of S, if $\mu_1(S')$ and $\mu_3(S')$ are nonempty then $\mu_2(S')$ is a (possibly empty) Nielsen path.*

Proof. Whilst $\mu_1(S')$ and $\mu_3(S')$ are non-empty, any bead in μ_2 which is bitten must be bitten by a bead coloured either μ_1 or μ_3. Let I_1 be the set of (Nielsen) beads in $\mu_2(S)$ which are eventually bitten by a bead coloured μ_1 (and are bitten whilst $\mu_1(S')$ and $\mu_3(S')$ are non-empty). Define I_2 to be those beads in $\mu_2(S)$ which are bitten by a bead coloured μ_3 (with the same proviso).

Suppose that I_1 and I_2 are non-empty. They form intervals, and I_1 is to the left of I_2.

Proposition 3.7.21, and the fact that $\mu_2(S)$ is stably Nielsen, implies that unless I_1 is immediately consumed there is a left reaper coloured μ_1 associated to I_1, and similarly there is a right reaper coloured μ_3 associated to I_2. The properties of reapers in Definition 3.7.17 imply the result.

In case one or both of I_1 and I_2 are empty (or immediately consumed), there is at most one reaper to consider, but the result follows in the same way. □

Lemma 3.9.4 (Buffer Lemma). *Suppose, for some corridor S in a well-folded diagram, that $I \subset \perp(S)$ is a beaded Nielsen path and that $\mu_1(S)$ and $\mu_2(S)$ lie either side of I, immediately adjacent to it. Suppose further that $\check{I}$ is stably Nielsen in $\check{\mu_1(S)}\check{I}\check{\mu_2(S)}$. Provided that the whole of I does not die in S, no bead in $\mu_1(S)$ can be eventually bitten by a bead coloured μ_2 (and vice versa), unless it is (eventually) HNP-bitten.*

[47] as defined in (3.2.2)

PROOF. Given Lemmas 3.4.5, 3.4.6 and 3.9.3, and the exclusion of HNP-biting, the proof of Lemma 1.8.1 applies. □

The proof of the following lemma follows that of Lemma 1.8.1.

LEMMA 3.9.5 (Weighted Buffer Lemma). *Suppose, for some corridor S in a well-folded diagram, that $I \subset \bot(S)$ is a beaded path consisting of Nielsen beads and beads of weight at most i, and that $\mu_1(S)$ and $\mu_2(S)$ lie either side of I, immediately adjacent to it. Suppose further that the only beads of $f_\#(\mu_1\breve{(S)}\breve{I}\mu_2\breve{(S)})$ that are in the future of I and have weight greater than i are Nielsen beads.*

Then, provided that the whole of I does not die in S, no bead in $\mu_1(S)$ can be eventually bitten by a bead coloured μ_2 (and vice versa), unless it is (eventually) HNP-bitten.

3.9.1. The Two Colour Lemma. Example 3.7.6 can be used to construct examples where the above two results are false if HNP-biting is not excluded. The same is true of the results in this section. This accounts for the caution that the reader will note in Sections 3.10, 3.11 and 3.12, where we are careful to ensure that the Pincer Lemma is applied only to pincers that involve no HNP-biting.

DEFINITION 3.9.6 (Stable f-neutering). Suppose that U and V are beaded paths, that for some k the futures of V in $f_\#^k(UV)$ and $f_\#^{k+1}(UV)$ are Nielsen, but that the future of V in $f_\#^{k-1}(UV)$ contains a non-Nielsen bead.

Denote the futures of U and V in $f_\#^{k-1}(UV)$ by U^{k-1} and V^{k-1}, respectively. Let β be the rightmost non-Nielsen bead in $f_\#(V^{k-1})$. If the biting of β in the tightening of $f_\#(U^{k-1})f_\#(V^{k-1})$ to form $f_\#^k(UV)$ is not HNP-biting then we say that U *stably left f-neuters* V in k steps.

The definition of *stable right f-neutering* is identical with the roles of U and V reversed, and when we are unconcerned about the direction we will refer simply to *stable f-neutering*.

In the light of Proposition 3.7.19, once stably f-neutered, the subsequent futures of V remain beaded Nielsen paths.

PROPOSITION 3.9.7 (Two Colour Lemma, cf. Proposition 1.8.4). *There exists a constant T_0, depending only on f, so that if U and V are beaded paths and U stably f-neuters V then it does so in at most T_0 steps.*

PROOF. Denote the future of U in $f_\#^i(UV)$ by U^i and the future of V by V^i.

As in the proof of Proposition 1.8.4, we will decompose each of the paths V^i into an *unbounded part* and a *bounded part*. The bounded part will be an interval on the right end of V^i whose immediate (abstract) future is a beaded

Nielsen path. The unbounded interval lies on the left end of V^i, and we will bound its length.

This would be a straightforward adaptation of the proof from Part 1 if Proposition 3.8.12 provided a bound of the length of that part of $C_{(\mu,\mu')}$ not contained in $C_{(\mu,\mu')}(2)$. However, the bound in Proposition 3.8.12 is just a bound on bead norm. Thus, we need to deal with the possibility of long GEPs and ΨEPs.

The following enumerated claims will together yield an upper bound on the length of the unbounded part of V^i, which in the course of the proof will be decomposed into V^i_{fast} and V^i_{nc}

Three of the claims concern the existence of a constant k_j that depends only on f; we use the abbreviation $\exists k_j = k_j(f)$.

Claim 1: $\exists k_1 = k_1(f)$ such that any GEP in V^i has length less than k_1.

This follows in a straightforward way from the Buffer Lemma 3.9.4 and the fact that the obvious preferred future of the rightmost edge in any GEP in V^i must eventually cancel with an edge from the future of U^i.

Next we consider long ΨEPs in V^i. Suppose that ρ is a ΨEP in V^i. Then the label on ρ or $\bar\rho$ has the form $E\bar\tau^k\bar\nu\gamma$, where τ is Nielsen path, $f(E) = E \odot \tau^m$ and $\bar\gamma\nu$ is a terminal segment of τ. We consider a number of different cases. First we dismiss a case that follows immediately from Lemma 3.5.13 and from the fact that exponential edges are left-fast:

Claim 2: If $\check\rho = E\bar\tau^k\bar\nu\gamma$ and γ is an exponential edge then the right end of ρ lies within C_0 of the left end of V^i.

Next we consider V^i_{fast}, which is defined to consist of those beads from the left end of V^i up to and including the rightmost bead in V^i whose immediate (abstract) future contains a left-fast bead.

Claim 3: $\exists k_2 = k_2(f)$ such that $|V^i_{\text{fast}}| \leq k_2$.

This follows immediately from Lemma 3.5.13 unless the rightmost bead in V^i_{fast} is a ΨEP. (Note that this rightmost bead is not a GEP, since a GEP does not have a left-fast bead in its immediate abstract future.)

Suppose, then, that the rightmost bead in V^i_{fast} is a ΨEP, say ρ. If $\check\rho = E\bar\tau^k\bar\nu\gamma$, then we are done by Claim 2. So suppose that $\check\rho = \bar\gamma\nu\tau^k\bar E$. Let ε be the edge in ρ whose label is $\bar E$. The preferred future of ε is to be cancelled by an edge in the future of U^i. By an obvious finiteness argument (as in the proof of Proposition 1.8.4), there is a constant p so that the path V^p contains no left-fast beads. This gives a bound on the amount of time before the future of ρ is bitten, and hence a bound on the amount that the future of ρ can shrink before then. Suppose that V^j is the first future of V^i in which the future of ρ has been bitten. Because the preferred future of ε is to be cancelled, Proposition 3.7.21 and the Buffer Lemma 3.9.4 imply that the length of the future in V^j of ρ is bounded above by a constant depending only on f.

The required bound on $|V^i_{\rm fast}|$ is now at hand: Lemma 3.5.13 bounds the length of $V^i_{\rm fast} \smallsetminus \rho$, and the combination of the bound on j and the bound on the length of the future of ρ in V^j gives a bound on the length of ρ. This completes the proof of Claim 3. We remark that the above argument also gives a bound on the amount of time it takes for $V^1_{\rm fast}$ to be entirely consumed.

We now define a set $V^i_{\rm nc}$ as follows: Let $\rho_{\rm nc}$ be the rightmost bead in V^i whose immediate abstract future is not Nielsen. We define $V^i_{\rm nc}$ as follows:

(1) if $\rho_{\rm nc} \in V^i_{\rm fast}$ then $V^i_{\rm nc} = \emptyset$;
(2) if $\rho_{\rm nc}$ is not a ΨEP, then $V^i_{\rm nc}$ consists of those beads from (but not including) the rightmost bead in $V^i_{\rm fast}$ up to and including $\rho_{\rm nc}$;
(3) if $\rho_{\rm nc}$ is a ΨEP with label of the form $\bar\gamma\nu\tau^k\bar E$ or $\rho_{\rm nc}$ is a ΨEP with label of the form $E\bar\tau^k\bar\nu\gamma$ and γ is not a Nielsen path, then $V^i_{\rm nc}$ consists of those beads in V^i from (but not including) the rightmost bead in $V^i_{\rm fast}$ up to and including $\rho_{\rm nc}$;
(4) finally, if $\rho_{\rm nc}$ is a ΨEP with label of the form $E\bar\tau^k n\bar u\gamma$ and γ is either empty or a Nielsen path, then $V^i_{\rm nc}$ consists of that interval from (but not including) the rightmost bead in $V^i_{\rm fast}$ up to and including the leftmost edge in $\rho_{\rm nc}$ (the label of this leftmost edge is E).

Note that in Case 4 the bead $\rho_{\rm nc}$ is certainly not contained in $V^i_{\rm fast}$.

Claim 4: $\exists k_3 = k_3(f)$ such that $|V^i_{\rm nc}| \le k_3$.

The proof of Claim 3 above established an upper bound on the time before all of $V^i_{\rm fast}$ is entirely consumed, and hence also on the time before the future of $V^i_{\rm nc}$ begins to be consumed. We now follow the proof of Lemma 3.8.8, which establishes an upper bound on the time that can elapse before the final non-constant bead in V^i is bitten. We will be done if we can bound this time from below by a positive constant times $|V^i_{\rm nc}|$.

In the current setting, we have non-constant beads in $V^i_{\rm nc}$ that may not be growing apart like those in the proof of Lemma 3.8.8.[48] But there *is* a lower bound on the rate at which the surviving futures of these beads can come together. Hence the length of $V^i_{\rm nc}$ provides a lower bound on the amount of time that must elapse before V^j becomes stably Nielsen, since the future of $V^i_{\rm nc}$ must be entirely consumed before this time. (Note that in Case 4, the preferred future of the edge $\bar E$ in $\rho_{\rm nc}$ must be eventually consumed by the future of U^i.) This proves Claim 4.

The *unbounded part* of V^i is the union of $V^i_{\rm fast}$ and $V^i_{\rm nc}$, whilst the *bounded part* is the remainder of V^i. The sum of the previous four claims bound the length of the unbounded part of V^i by a constant that depends only on f.

There is a similar bound on the number of edges in U^i that have an edge in their future that cancels with an edge in the future of V^i. (Here we need

[48]This is because we are now measuring length rather than bead-norm.

the hypothesis that the path V^k becoming stably Nielsen does not arise from HNP-biting.)

At this stage, we can follow the proof of Proposition 1.8.4 directly. After an amount of time bounded by a constant that depends only on f, either the future of V becomes stably Nielsen or empty, or else there is a repetition of the following data: (i) the unbounded part of V^i plus the leftmost $B+J$ edges of the bounded part; (ii) a terminal segment of U^i containing all of the edges that can ever interact with the future of V. Once we have such a repetition, if the future of V has not become stably Nielsen or vanished then it never will, contrary to hypothesis. □

We need a weighted version of neutering and the two-colour lemma.

Definition 3.9.8 ((f,i)-neutering). Fix $i \in \{1, \dots, \omega\}$ and let U and V be beaded paths. Suppose that for some k the future of V in $f_\#^k(UV)$ has weight less than i, but that the future of V in $f_\#^{k-1}(UV)$ has weight at least i.

Denote the futures of U and V in $f_\#^{k-1}(UV)$ by U_{k-1} and V_{k-1}, respectively. Let β be the rightmost bead in $f_\#(V_{k-1})$ of weight at least i. If the biting of β in the tightening of $f_\#(U_{k-1})f_\#(V_{k-1})$ to form $f_\#^k(UV)$ is not HNP-biting then we say that U (f,i)-*neuters* V in at most k steps.

Proposition 3.9.9 (Weighted Two Colour Lemma). *There exists a constant T_0', depending only on f, so that for any $i \in \{1, \dots, \omega\}$, if U and V are beaded paths and U (f,i)-neuters V then it does so in at most T_0' steps.*

Proof. We decompose the futures of U and V in $f_\#^k(UV)$ as in Lemma 3.9.7.

The proof is similar to that of Lemma 3.9.7, except that when we appeal to the proof of Proposition 3.8.8 we assume that we have a path $\mathcal{E}_j$ with $j \geq i$. Otherwise, the proof of Lemma 3.9.7 above and that of Proposition 1.8.4 can now be followed *mutatis mutandis.* □

By replacing T_0 by T_0' if necessary, we may assume that $T_0 \geq T_0'$. We henceforth make this assumption.

3.9.2. The disappearance of colours: Pincers and implosions.

Definition 3.9.10. Consider a pair of non-constant edges ε_1 and ε_2 which cancel in a corridor S_t of Δ, and suppose that, for $i = 1, 2$, the immediate past of ε_i lies in a bead of some $\mu_i(S_t)$ that is either a unbounded atom, a GEP or a ΨEP. Suppose further that the cancellation of ε_1 and ε_2 is not HNP-cancellation, and that $\mu_1 \neq \mu_2$. Consider the paths p_1, p_2 in $\mathcal{F} \subset \Delta$ tracing the histories of ε_1 and ε_2. Suppose that at time τ_0 the paths p_1 and p_2 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 sub-diagram 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.

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. Define $\widetilde{\chi}(\Pi)$ to be the set of colours $\mu \notin \{\mu_1, \mu_2\}$ such that there is a 2-cell in Π coloured μ. Finally, define

$$\mathrm{life}(\Pi) = \mathrm{time}(S_\Pi) - \mathrm{time}(S_b).$$

See Section 1.8 for illustrative pictures.

Proposition 3.9.11 (Unnested Pincer Lemma, cf. Proposition 1.8.7). *There exists a constant T_1, depending only on f, such that for any pincer Π*

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

In the proof of Proposition 1.8.7 (Regular Implosions) the strategy was to identify a constant T_1 such that over each period of time of length T_1 within a pincer, at least one colour became constant. There are a number of impediments to implementing this strategy in the current situation. The first is that Nielsen paths can consist of edges which are not constant edges, so if a colour *becomes Nielsen* then it may cease to be Nielsen at some stage in the future. In order to overcome this impediment, we make the following

Definition 3.9.12. Suppose that for some colour μ and some corridor S, the path $\check{\mu}(S)$ is stably Nielsen, and let ν_1 and ν_2 be the colours immediately on either side of μ in S. If there is some corridor S' in the future of S in which $\check{\mu}(S')$ is not Nielsen and S' is the earliest such corridor, then we say that μ is *resuscitated* in S'. By Lemma 3.9.3, at least one of ν_1 and ν_2 is not adjacent to μ in S', so either $\nu_1(S')$ or $\nu_2(S')$ is empty. If $\nu_i(S')$ is empty, we say that ν_i *sacrifices itself* for μ.

Remark 3.9.13. A colour can sacrifice itself for at most one colour.

A colour may become stably Nielsen and be resuscitated a number of times, but a different colour must sacrifice itself for each resuscitation.

The concept of 'becoming stably Nielsen' is analogous to that of a colour 'essentially vanishing' in Section 1.8. However, the concept of 'resuscitation' does not have an analogue in Part 1.

Fix a pincer Π and assume that $\mathrm{life}(\Pi) > 1$. The strategy to prove Proposition 3.9.11 is to identify a constant T_1 so that during the life of Π, in each $T_1/2$ steps of time there is a colour that becomes stably Nielsen (perhaps vanishing) In order to obtain the bound in the statement of Proposition 3.9.11, we then count the colours which become stably Nielsen or vanish, and the colours which sacrifice themselves for those that are resuscitated. A colour can therefore be counted twice – once for disappearing (or for the last time it becomes stably Nielsen), and once as a sacrifice – but no colour is counted more than twice. Thus Proposition 3.9.11 is an immediate consequence of the following result whose proof will occupy the remainder of this subsection.

Proposition 3.9.14. *There is a constant T_1, depending only on f, so that for any pincer Π in a minimal area van Kampen diagram over $M(f)$, in any interval of time of length $T_1/2$, at least one colour in $\tilde{\chi}(\Pi)$ becomes stably Nielsen or vanishes.*

Definition 3.9.15 (p-implosive arrays). Let p be a positive integer and S a corridor. A *p-implosive array* of colours in S is an ordered tuple $A(S) = [\nu_0(S), \dots, \nu_r(S)]$, with $r > 1$, such that

(1) each pair of colours $\{\nu_j, \nu_{j+1}\}$ is separated in S only by a stably Nielsen (or empty) path;
(2) in each of the corridors $S = S^1, S^2, \dots, S^p$ in the future of S, no $\nu_j(S^i)$ is empty or a stably Nielsen path, $j = 1, \dots, r-1$;
(3) in S^p, *either* an edge coloured ν_0 from a unbounded atom, a GEP or a ΨEP cancels with an edge coloured ν_r from a unbounded atom, a GEP or a ΨEP (and hence the colours ν_j with $j = 1, \dots, r-1$ are consumed entirely), or *else* each of the colours ν_j $(j = 1, \dots, r-1)$ become stably Nielsen or vanish, while ν_0 and ν_r are not Nielsen in $f_\#(\nu_0(\breve{S}^p) \cdots \nu_r(\breve{S}^p))$ (although they may nevertheless become stably Nielsen or even disappear in S^p because of 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 stably Nielsen path which lies between $\nu_0(S^p)$ and $\nu_r(S^p)$ (if either $\nu_0(S^p)$ begins or $\nu_r(S^p)$ ends with an interval of Nielsen atoms include these in the residual block). Note that the residual block may be empty. The *enduring block* of the array is the set of stably Nielsen paths in $\perp(S)$ that have a future in the residual block.

Note that there may exist some *unnamed colours* between $\nu_j(S)$ and $\nu_{j+1}(S)$; if they exist, these form a stably Nielsen path.

Remark 3.9.16. Let $[\nu_0(S), \dots, \nu_r(S)]$ be a p-implosive array.

(1) Any q-implosive sub-array of $[\nu_0(S), \dots, \nu_r(S)]$ has $q = p$.
(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 come from displayed unbounded atoms, GEPs or ΨEPs, then the sub-array $[\nu_i(S), \dots, \nu_j(S)]$ is p-implosive of Type I.
(3) If u, v and w are beaded edge-paths such that u, v and $f_\#(uwv)$ are Nielsen paths then w is a Nielsen path. 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 then they are consecutive, ν_j, ν_{j+1}.

(4) Likewise, 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 they must be consecutive.

LEMMA 3.9.17. *Let* Π *be a pincer. The ordered list of colours along each corridor before* $\mathrm{time}(S_\Pi)$ *in a pincer* Π *must contain a p-implosive array for some p.*

PROOF. The definition of p-implosive array is designed so that when a colour becomes stably Nielsen (or disappears) in a pincer there is a p-implosive array. See the proof of Lemma 1.8.10 for more details. □

DEFINITION 3.9.18. Suppose that $A(S) = [\nu_0(S), \dots, \nu_r(S)]$ is a p-implosive array. We say that $A(S)$ is an *HNP-implosive array* if either

(1) $A(S)$ is of Type I and in S^p the cancellation between ν_0 and ν_r is HNP-biting, or
(2) $A(S)$ is of Type II and in S^p, for some $0 < i < r$, ν_0 and ν_i are involved in HNP-biting or for some $0 < j < r$, ν_j and ν_r are involved in HNP-biting.

In order to follow the arguments from Part 1, we need to sharpen Lemma 3.9.17: HNP-cancellation can beget p-implosive arrays with p arbitrarily large, and therefore we must argue for the frequent occurrence of p-implosive arrays that are not HNP-implosive. A first step in this direction is given by the following

LEMMA 3.9.19. *Let* Π *be a pincer, and let* μ_1 *and* μ_2 *be the colours associated to the bounding-paths* p_1 *and* p_2 *of* Π*. Then there is no HNP-biting between beads in* μ_1 *and* μ_2 *within* Π*.*

PROOF. Follows from Lemmas 3.7.14 and 3.7.22. □

When we are unconcerned about p in a p-implosive array, we refer merely to an *implosive array.* The first restriction to note concerning implosive arrays is this:

LEMMA 3.9.20. *If* $[\nu_0(S), \dots, \nu_r(S)]$ *is implosive of Type I, then* $r \leq B$*. If it is implosive of Type II, then* $r \leq 2B$*.*

PROOF. In Type I arrays, the interval $\nu_1(S^p) \cdots \nu_{r-1}(S^p) \subset \bot(S^p)$ is to die in S^p, so the bound is an immediate consequence of 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. □

PROOF OF PROPOSITION 3.9.14. We give a suitable formulation of 'short' so that in any corridor S within Π, S contains a short p-implosive array. Proposition 3.9.14 then follows from an obvious finiteness argument.

Let $A(S) = [\nu_0(S), \dots, \nu_r(S)]$ be the implosive array guaranteed to exist by Lemma 3.9.17, and suppose that $p \geq 2T_0$ (if not then a colour becomes stably Nielsen or vanishes within $2T_0$ of time(S)).

We can decompose each of the colours $\nu_j(S)$ in analogy with Part 1, using the decomposition in Section 3.6.3 above.

We fix a constant Λ_1 so that if $\| A(S) \| > \Lambda_1$ then one of the following must occur in S^{T_0}:

(1) there is a block of displayed Nielsen atoms in some $\nu_j(S^{T_0})$ of length at least $J + 4B$,
(2) there is a displayed GEP in some $\nu_j(S^{T_0})$ of length at least $J+4B+2$,
(3) there is a displayed ΨEP in some $\nu_j(S^{T_0})$ of length at least $J + 4B + L + 1$, or
(4) there is an interval of unnamed colours in $A(S)$ (which form a stably Nielsen block) of length at least $J + 4B$ between $\nu_0(S^{T_0})$ and $\nu_r(S^{T_0})$.

In the remainder of the proof, we shall use the term *block* to refer generically to the identified interval in whichever of the above cases we find ourselves. Increasing Λ_1 if necessary, we may assume that the past of the block in S satisfies the relevant condition from (1) – (4) with the bound increased by $2BT_0$.

For such a block I in S^{T_0}, consider the first edge on either side of this block which is not contained in a Nielsen path. These edges may be on one end of a GEP or a ΨEP (including the GEP or ΨEP from condition (2) or (3)), or may be contained in unbounded atoms. Call these edges ε_1 and ε_2.

The Buffer Lemma 3.9.4 implies that either (i) one of ε_1 and ε_2 must be 'stabbed in the back' – we do not exclude the possibility that this stabbing happens by HNP-biting, or (ii) there is HNP-cancellation across the above block.

We first dispose of case (ii). Suppose, for ease of notation, that the edge ε_1 HNP-bites the edge ε_2 across the above block I. Let ε_1 have weight k. Then all edges in I and ε_2 must have weight less than k. Let ε_2' be the first edge to the right of I that has weight at least k. Then the Relative Buffer Lemma 3.9.2 implies that either ε_1 or ε_2' must be stabbed[49] in the back (again, this could be by HNP-biting).

We have argued that some edge must be stabbed in the back. Suppose that this stabbing is of an edge ε in S^{T_0} and that ε has weight k_1. Consider first the possibility that ε is stabbed in the back via HNP-biting. Then this occurs by an edge ε' of weight at least $k_1 + 1$. Now, either this stabbing in the back occurs within T_0 of S^{T_0}, or by the Weighted Two Colour Lemma (3.9.9) there is another block as in (1) – (4) above. This block has higher weight than

[49]Note that if there is no such edge ε_2' in Π then ε_1 must be stabbed in the back, by Lemmas 3.9.2 and 3.9.19.

the previous block, and as above leads to another stabbing in the back. If this stabbing is HNP-biting, pass to a yet higher weight stabbing, and so on.

Eventually (after less than ω iterations of this argument), we get an edge ε stabbed in the back with the stabbing not HNP-biting. Suppose that ε has weight k_2. Suppose for ease of notation that ε is to the left of the long block, and suppose that ε is coloured ν_i. Because of the block of Nielsen atoms to the non-stabbing side of ε, the Two Colour Lemma (Proposition 3.9.7) implies that if the edge ε' which stabs ε in the back is coloured by ν_j then $i - j > 1$; we then write $\nu_j \searrow \nu_i$.

Passing to an innermost pair $\nu_{l_1} \searrow \nu_{l_2}$ between ν_i and ν_j we can see that there are no blocks in S^{T_0} satisfying any of (1) – (4) above, for otherwise there would be a further stabbing, leading to a related pair of colours between our innermost pair, contradicting the innermost nature of this pair.

Once there are no such blocks, we have a bound on the length of the p-implosive array implicit in the relation $\nu_{l_1} \searrow \nu_{l_2}$. An obvious finiteness argument now finishes the proof. □

We have already seen how Proposition 3.9.14 implies Proposition 3.9.11. Just as in Section 1.8, we must now deal with the possibility of 'nested pincers'.

3.9.3. Super-buffers.

Definition 3.9.21. We consider sequences of 5-tuples of tight edge-paths in G.

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

with $|u_{k,1}|$ and $|u_{k,2}|$ at most $C_0 + C_1 + 2B(B+1) + 1$, while $|u_{k,2}|$ and $|u_{k,4}|$ are at most $C_0 + C_1 + J$ and $|u_{k,3}| \leq 4B(B+1) + 1$.[50] We fix an integer T_1' sufficiently large to ensure that for any sequence of length T_1' there will be a repetition, i.e. some $t_1 < t_2 \leq T_1'$ with

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

We also choose $T_1' \geq T_1$.

With appropriate changes of terminology and the results of the previous subsection in hand, the proof of Proposition 1.8.21 yields:

Lemma 3.9.22. *Let $V = V_1V_2V_3$ be a tight concatenation of three beaded paths in G. If the future of V_2 is not stably Nielsen in $f_\#^{T_1'}(V)$ then the future of V_2 is not stably Nielsen in $f_\#^k(V)$ for any $k \geq 0$.*

[50]The purpose of these constants is just as in Definition 1.8.19, with appropriate changes due to Lemmas 3.6.8 and 3.6.1 and Proposition 3.8.12.

3.9.4. Nesting and the Pincer Lemma. Let $\lambda_0 = J + 2B(T_0 + 1) + 1$, which is the obvious analogue of the constant of the same name in Section 1.8. As in Remark 1.9.5, it is convenient to assume that $LC_4 < \lambda_0$, and we increase λ_0 to make this so. (This makes certain statements in Section 3.10 easier, but has no serious affect.)

DEFINITION 3.9.23. 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 $p_1(\Pi_1) \cap \top(S)$ or $p_2(\Pi_1) \cap \top(S)$ is not contained in any stably Nielsen path and this future [51] lies in a beaded path consisting of Nielsen beads and beads of weight strictly less than the weight of the edges containing $p_1(\Pi_1) \cap \top(S)$ and $p_2(\Pi_1) \cap \top(S)$, and that this beaded path has at least λ_0 non-vanishing beads. Then we say that Π_1 is *nested* in Π_0.

REMARK 3.9.24. Besides the obvious translations, the above differs from Definition 1.8.22 in that the path at the top of the pincer may now consist of Nielsen beads and lower weight beads, whereas in Part 1 it consisted entirely of constant letters. This more general setting does not make any of the proofs in this section harder (because of the Weighted Two Colour Lemma), but is needed because of the more complicated definition of the 'cascade of pincers' below (Definition 3.10.17).

DEFINITION 3.9.25. For a pincer Π_0, let $\{\Pi_i\}_{i \in I}$ be the set of all pincers nested in Π_0. Then define

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

The corridor S_t was defined in Definition 3.9.10.

LEMMA 3.9.26 (cf. Lemma 1.8.25). *If the pincer* Π_1 *is nested in* Π_0 *then* $\mathrm{time}(S_t(\Pi_1)) < \mathrm{time}(S_{\Pi_0})$.

PROOF. The existence of the beaded path at the top of the pincer Π_1 makes this an immediate consequence of the Weighted Buffer Lemma 3.9.5. ☐

Define $T_1 = T_1' + 2T_0$. The following theorem is the main result of this section, and is the strict analogue of Theorem 1.8.26. The proof in the current context follows the proof from Part 1 *mutatis mutandis*.

THEOREM 3.9.27 (Pincer Lemma). *For any pincer* Π

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

---

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

3.10. Teams

By virtue of Lemma 3.8.12, Remark 3.8.13 and the results of Section 3.6, we have reduced the task of bounding the bead norm of S_0 to that of bounding the lengths of certain blocks $C_{(\mu,\mu')}(2)$ which consist of Nielsen beads coloured μ all of which are to be eventually bitten by beads coloured μ' in the future of S_0. By Proposition 3.7.21, if such a block has length at least $B+J$, then there is an associated reaper, which consumes Nielsen beads in $C_{(\mu,\mu')}(2)$ at a constant rate (and entirely consumes any bead it bites, up to the final bead). Note that to each pair (μ,μ') there is at most one associated reaper.

This puts us in the situation where we can develop the technology of *teams* as in Section 1.9. However, there are a number of key differences to Part 1: we already had to work hard in Section 3.7 to establish the existence of a *reaper* for $C_{(\mu,\mu')}(2)$, and now we have to work harder to identify the times $\hat{t}_1(\mu,\mu')$ and $t_1(\mathcal{T})$ attached to a pair $(\mu,\mu') \in \mathcal{Z}$ and a team $\mathcal{T}$, using the *robust* past of the reaper instead of the actual past; this is required in order that the Pincer Lemma apply to teams of genesis (G3). It is worth remarking that once we have identified the pincer $\Pi_{\mathcal{T}}$ associated to a team $\mathcal{T}$ of genesis (G3), we revert to an analysis of actual pasts (as in the definition of pincer).

Note that the colour of the edges in the robust future of an edge may not always be the same, contrary to the actual future. In fact, whenever the robust past is not the actual past, the colour changes. This explains a slight difference between Definition 3.10.3 below and Definition 1.9.1.

Consider an interval $C_{(\mu,\mu')}(2)$ so that $|C_{(\mu,\mu')}(2)| > B+J$, and let ϵ^{μ} be the reaper associated to $C_{(\mu,\mu')}(2)$ in Proposition 3.7.21 above. Let t_0 be the time at which ϵ^{μ} first bites a Nielsen bead in $C_{(\mu,\mu')}(2)$, and let β_μ be the rightmost bead in the future of $C_{(\mu,\mu')}(2)$ at this time. Note that β_μ is a Nielsen bead. Let ϵ_μ be the rightmost edge in β_μ.

Remark 3.10.1. Since $|C_{(\mu,\mu')}(2)| > B+J$, and each bead of $C_{(\mu,\mu')}(2)$ is to be bitten by μ', the colour of ϵ^{μ} is μ'.

Lemma 3.10.2. *Suppose that the immediate past of ϵ_μ exists (i.e. that ϵ_μ does not lie on $\partial\Delta$). Then the immediate past of ϵ_μ lies in some bead σ, and σ contains the immediate past of each edge in β_μ.*

The above lemma, applied at each stage in the past, implies that we can follow the past of the edge ϵ_μ and deduce consequences about the past of all edges in β_μ.

We now define a time $\hat{t}_1(\mu,\mu')$ as follows: We go back to the last point in time when (i) the past of ϵ_μ and the robust past of ϵ^{μ} lay in a common corridor; and (ii) ϵ_μ is contained in a beaded Nielsen path whose swollen present is immediately adjacent to the robust past of ϵ^{μ}.

We denote this corridor $S_{\uparrow}$.

DEFINITION 3.10.3. The robust past of ϵ^μ at time $\hat{t}_1(\mu,\mu')$ is called the *reaper*, and is denoted $\hat{\rho}(\mu,\mu')$. The interval $\hat{\mathfrak{T}}(\mu,\mu')$ is the maximal beaded Nielsen path in $\perp(S_\uparrow)$ all of whose beads are eventually bitten by $\hat{\rho}(\mu,\mu')$. The *pre-team* $\hat{\mathcal{T}}(\mu,\mu')$ is defined to be the set of pairs $(\mu_1,\mu_2)\in\mathcal{Z}$ so that (i) the robust past of ϵ^μ is coloured μ_2 at some time between $\hat{t}_1(\mu,\mu')$ and t_0; and (ii) $\hat{\mathfrak{T}}(\mu,\mu')$ contains some edges coloured μ_1. The number of beads in $\hat{\mathfrak{T}}(\mu,\mu')$ is denoted $\parallel\hat{\mathcal{T}}\parallel$.

As in Section 1.9, we will define *teams* to be pre-teams satisfying a certain maximality condition (see Definition 3.10.6 below).

REMARK 3.10.4. Just as in Remark 1.9.2, 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 for which $\nu(S_0)$ is empty.

LEMMA 3.10.5 (cf. Lemma 1.9.3). *If* $\hat{t}_1(\mu,\mu')\geq\text{time}(S_0)$ *then*

$$\sum_{(\mu_1,\mu_2)\in\hat{\mathcal{T}}(\mu,\mu')}|C_{(\mu,\mu')}(2)|\leq\ \parallel\hat{\mathcal{T}}(\mu,\mu')\parallel+B(B+1).$$

PROOF. The extra $B(B+1)$ is to account for the beads consumed before the reaper comes into play. Otherwise the proof is just as in Part 1. □

3.10.1. The Genesis of pre-teams. [cf. Subsection 1.9.2]
We consider the various events that may occur at $\hat{t}_1(\mu,\mu')$ which prevent us pushing the pre-team back one step in time. Recall that $S_\uparrow$ is the corridor at time $\hat{t}_1(\mu,\mu')$ which contains $\hat{\mathcal{T}}(\mu,\mu')$. Suppose that μ_2 is the colour of $\hat{\rho}(\mu,\mu')$.

There are four types of events:

(G1) The immediate past of $C_{(\mu,\mu_2)}(S_\uparrow)$ is separated from the robust past of $\hat{\rho}(\mu,\mu')$ by an intrusion of $\partial\Delta$.
(G2) We are not in Case (G1), but the immediate past of $C_{(\mu,\mu_2)}(S_\uparrow)$ is separated from the robust past of $\hat{\rho}(\mu,\mu')$ because of a singularity.
(G3) The immediate past of $C_{(\mu,\mu_2)}(S_\uparrow)$ is still in the same corridor as the robust past of $\hat{\rho}(\mu,\mu')$, but the swollen present of the immediate past of $C_{(\mu,\mu_2)}(S_\uparrow)$ is not immediately adjacent to the robust past of $\hat{\rho}(\mu,\mu')$.
(G4) We are not in any of the above cases, but the immediate past of the rightmost edge in $C_{(\mu,\mu_2)}(S_\uparrow)$ is not contained in a beaded Nielsen path.

We now make the definition of a team.

DEFINITION 3.10.6 (cf. Definition 1.9.6). All pre-teams $\hat{\mathcal{T}}(\mu,\mu')$ with $\hat{t}_1(\mu,\mu')\geq\text{time}(S_0)$ are defined to be teams, but the qualification criteria for pre-teams with $\hat{t}_1(\mu,\mu')<\text{time}(S_0)$ 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 $(\mu_1, \mu_2) \in \hat{\mathcal{T}}(\mu, \mu')$ but $\hat{t}_1(\mu, \mu') > \hat{t}_1(\mu_1, \mu_2)$ and hence $(\mu, \mu') \notin \hat{\mathcal{T}}(\mu_1, \mu_2)$. To avoid double-counting in our estimates on $\| \mathcal{T} \|$ we disqualify the (intuitively smaller) pre-team $\hat{\mathcal{T}}(\mu_1, \mu_2)$ 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 (just as in Section 1.9).

A team is said to be *short* if $\| \mathcal{T} \| \leq \lambda_0$ or $\sum_{(\mu_1,\mu_2)\in\mathcal{T}} |C_{(\mu_1,\mu_2)}(2)| \leq \lambda_0$. Let Σ denote the set of short teams.

LEMMA 3.10.7 (cf. Lemma 1.9.7). *Teams of genesis (G4) are short.*

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 $\mathcal{T}(\mu, \mu')$ with genesis of type (G4) may have been disqualified, leaving (μ, μ') teamless; second, in our initial discussion of pre-teams we excluded pairs (μ, μ') with $|C_{(\mu,\mu')}(2)| \leq B + J$. The following definitions remove these difficulties.

DEFINITION 3.10.8 (Virtual team members). If a pre-team $\hat{\mathcal{T}}(\mu, \mu')$ of type (G4) is disqualified under the terms of Definition 3.10.6 and the smaller team necessitating disqualification is $\hat{\mathcal{T}}(\mu_1, \mu_2)$, then we define $(\mu, \mu') \in_v \hat{\mathcal{T}}(\mu_1, \mu_2)$ and $\hat{\mathcal{T}}(\mu, \mu') \subset_v \hat{\mathcal{T}}(\mu_1, \mu_2)$. 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 3.10.9. If (μ, μ') is such that $1 \leq |C_{(\mu,\mu')}(2)| \leq B + J$ 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 3.10.10 (cf. Lemma 1.9.10). *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 $|\mathcal{Z}| < 2\,|\partial\Delta|$. $\square$

3.10.2. Pincers associated to teams of genesis (G3). [cf. Subsection 1.9.3]

In this subsection we describe a pincer $\Pi_{\mathcal{T}}$ canonically associated to each team of genesis (G3), as in Subsection 1.9.3. The only real difference between the definitions here and those in Part 1 is the use of robust past and beaded Nielsen paths. Sadly, this variation leads to complications in the cascade of pincers; see Definition 3.10.17 and Remark 3.9.24.

Definition 3.10.11 (cf. Definition 1.9.11). The *narrow past* of a team $\mathcal{T}$ at time t consists of those beaded Nielsen paths whose beads are displayed in their colour and whose future is contained 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).

The following lemma follows from the definition of teams of genesis (G3) in a straightforward manner.

Lemma 3.10.12. *Let $\mathcal{T}$ be a team of genesis (G3). There exist beads $y(\mathcal{T})$ and $y_1(\mathcal{T})$ of different colours, both lying strictly between the immediate past of the swollen present of $\mathcal{T}$ and the robust past of $\hat{\rho}(\mu, \mu')$, so that $y(\mathcal{T})$ is bitten by $y_1(\mathcal{T})$ and this is not HNP-biting.*

Definition 3.10.13 (The Pincer $\tilde{\Pi}_{\mathcal{T}}$). Choose a leftmost pair of beads $y(\mathcal{T}), y_1(\mathcal{T})$ satisfying Lemma 3.10.12, and let $x(\mathcal{T})$ be the leftmost edge in $y(\mathcal{T})$. Let $x_1(\mathcal{T})$ be the edge in $y_1(\mathcal{T})$ which is the past of the edge which cancels with the leftmost edge in the immediate future of $x(\mathcal{T})$.

Define $\tilde{p}_l(\mathcal{T})$ to be the path in the family forest $\mathcal{F}$ that traces the history of $x(\mathcal{T})$ to $\partial\Delta$, and let $\tilde{p}_r(\mathcal{T})$ be the path that traces the history of $x_1(\mathcal{T})$.

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.

Remark 3.10.14. Since the pair $y(\mathcal{T}), y_1(\mathcal{T})$ in Definition 3.10.13 are the leftmost pair satisfying Lemma 3.10.12, any non-vanishing beads which lie between $\mathfrak{T}$ and this pair are involved in HNP-biting and are of lower weight than $y_1(\mathcal{T})$, by the Weighted Buffer Lemma 3.9.5.

Lemma 3.10.15. *The segments of the paths $\tilde{p}_l(\mathcal{T})$ and $\tilde{p}_r(\mathcal{T})$, together with the path joining them along the bottom of the corridor at time $\tilde{t}_2(\mathcal{T})$ form a pincer.*

Proof. Note that when choosing the beads $y(\mathcal{T})$ and $y_1(\mathcal{T})$ we excluded HNP-cancellation. That the paths in the statement of the lemma form a pincer then follows immediately from the definition of pincers. $\square$

We denote the pincer described in Lemma 3.10.15 above by $\tilde{\Pi}_{\mathcal{T}}$.

3.10.3. The cascade of pincers. 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 its corridors. However, when there are only two colours, the situation is more complicated.

Recall that the constant T_0 is as in Proposition 3.9.7, subject to the requirement that $T_0 \geq T_0'$ as in the assumption immediately after Proposition 3.9.9. The pincer S_Π associated to a pincer Π is defined in Definition 3.9.10.

LEMMA 3.10.16. *One of the following must occur:*

(1) $\mathrm{time}(S_{\tilde{\Pi}_\mathcal{T}}) > t_1(\mathcal{T}) - T_0$*;*
(2) the path $\tilde{p}_l(\mathcal{T})$ *and the entire narrow past of* $\mathcal{T}$ *are not in the same corridor at time* $t_1(\mathcal{T}) - T_0$*; or*
(3) at time $t_1(\mathcal{T}) - T_0$ *the path* $\tilde{p}_l(\mathcal{T})$ *and the narrow past of* $\mathcal{T}$ *are separated by a path which does not split as a beaded path whose beads are either Nielsen paths or of weight less than* $\tilde{p}_l(\mathcal{T})$*.*

PROOF. If not, the Weighted Two Colour Lemma (Lemma 3.9.9) would give a contradiction, since there *is* to be interaction between the beads $y(\mathcal{T})$ and $y_1(\mathcal{T})$ at time $t_1(\mathcal{T})$, and this interaction is not HNP-biting. □

We now consider each of the three cases in turn, seeking a definition of times $t_2(\mathcal{T})$ and $t_3(\mathcal{T})$ and (possibly) a pincer $\Pi_\mathcal{T}$. The following definition is entirely analogous to Definition 1.9.13, with the appropriate translations.

DEFINITION 3.10.17 (cf. Definition 1.9.13).

(1) Suppose 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[52] 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) Suppose that Case (1) does not occur and $\mathrm{time}(S_{\tilde{\Pi}_\mathcal{T}}) > t_1(\mathcal{T}) - T_0$. 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) Suppose that neither Case (1) or Case (2) occurs: In this case, Lemma 3.10.16(3) pertains. We pass to the latest time at which there is a path between $\tilde{p}_l(\mathcal{T})$ and the narrow past of $\mathcal{T}$ which has an edge of at least the same weight as $\tilde{p}_l(\mathcal{T})$ at this time and is not contained in a Nielsen path. Choose a pair of beads $y'(\mathcal{T})$, $y_1'(\mathcal{T})$ as in Lemma 3.10.12, as well as edges $x'(\mathcal{T})$, $x_1'(\mathcal{T})$. Let $\tilde{p}_l'(\mathcal{T})$ be the path tracing the history of $x'(\mathcal{T})$. Let $\tilde{p}_r'(\mathcal{T})$ trace the history of the edge $x_1'(\mathcal{T})$ that cancels $x'(\mathcal{T})$. Let $\tilde{t}_2'(\mathcal{T})$ 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 these

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

paths after time $\tilde{t}'_2(\mathcal{T})$ and 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 $\tilde{p}''_l(\mathcal{T})$, *etc.*. We iterate this analysis until we fall into Case (1) or (2), at which point we acquire the desired definitions of $\Pi_{\mathcal{T}}, t_2(\mathcal{T})$ and $t_3(\mathcal{T})$.

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 edges (one at each time) lying on the leftmost of the primed $\tilde{p}_l(\mathcal{T})$ from the top of $\pi_{\mathcal{T}}$ to time $t_1(\mathcal{T})$.

Definition 3.10.18 (cf. Definition 1.9.14). Let $\mathcal{T}$ be a long team of genesis (G3). We define $\chi_P(\mathcal{T})$ to be the set of colours containing the paths $\tilde{p}_l(\mathcal{T}), \tilde{p}'_l(\mathcal{T}), \tilde{p}''_l(\mathcal{T}), \ldots$ that arise in Case (3) of Definition 3.10.17 but do not become $p_l(\mathcal{T})$.

Lemma 3.10.19 (cf. Lemma 1.9.15).

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

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

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

3.10.4. The length of teams. This subsection follows Subsection 1.9.4. We consider the lengths of arbitrary teams.

Definition 3.10.20 (cf. Definition 1.9.16). Let $\mathcal{T}$ be a team. Define $\text{down}_1(\mathcal{T}) \subset \partial\Delta$ to consist of those edges e that are labelled by some t_i 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 $p_l(\mathcal{T})$.

Definition 3.10.21 (cf. Definition 1.9.17). Define $\partial^{\mathcal{T}} \subset \partial\Delta$ to be the intersection of the narrow past of $\mathcal{T}$ with $\partial\Delta$.

Lemma 3.10.22 (cf. Lemma 1.9.18).

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

DEFINITION 3.10.23 (cf. Definition 1.9.19). 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 $t_3(\mathcal{T})$, 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 reduces the task of bounding the total length of teams to that of bounding the size of the sets $Q(\mathcal{T})$. Its proof follows that of Lemma 1.9.20.

LEMMA 3.10.24 (cf. Lemma 1.9.20).

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

$$\| \mathcal{T} \| \leq 2LC_4|\mathrm{down}_1(\mathcal{T})| + |\partial^{\mathcal{T}}|.$$

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

$$\| \mathcal{T} \| \leq 2C_4|\mathrm{down}_1(\mathcal{T})| + |\partial^{\mathcal{T}}| + 2LC_4|Q(\mathcal{T})| + 2LC_4T_0(|\chi_P(\mathcal{T})| + 1) + \lambda_0.$$

3.10.5. Bounding the size of $Q(\mathcal{T})$. Let $\mathcal{G}_3$ be the set of long teams of genesis (G3) for which $Q(\mathcal{T})$ is nonempty. Our goal for the remainder of this section is to find a bound for $\sum_{\mathcal{T}\in\mathcal{G}_3} |Q(\mathcal{T})|$.

LEMMA 3.10.25 (cf. Lemma 1.9.22). *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 3.10.26 (cf. Lemma 1.9.23). *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 disjoint or else one is contained in the other. In the latter case, say $\Pi_{\mathcal{T}_1} \subset \Pi_{\mathcal{T}_2}$, the definition of nesting (Definition 3.9.23), and of the pincer associated to a team (Definition 3.10.17) ensure that $\Pi_{\mathcal{T}_1}$ is actually nested in $\Pi_{\mathcal{T}_2}$ (cf. Remark 3.9.24). ☐

COROLLARY 3.10.27 (cf. Corollary 1.9.24).

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

We have now reduced our task for this section to bounding the number of edges in the $Q(\mathcal{T})$ which occur before $t_2(\mathcal{T})$; this is the cardinality of the following set.

DEFINITION 3.10.28 (cf. Definition 1.9.25). For a team $\mathcal{T} \in \mathcal{G}_3$ we define $\mathrm{down}_2(\mathcal{T})$ to be the set of edges in $\partial\Delta$ that lie at the right-hand end of a corridor containing an edge in $Q(\mathcal{T})$ before time $t_2(\mathcal{T})$.

Just as in Part 1, it is not necessarily the case that the sets $\mathrm{down}_2(\mathcal{T})$ are disjoint for distinct teams, and we must deal with the possibility of 'double-counting'.

The left-to-right ordering defined on paths in $\mathcal{F}$ in Section 1.9 is defined in the current context exactly as in Part 1.

Notation: Let $\mathcal{G}_3'$ be the set of teams $\mathcal{T} \in \mathcal{G}_3$ with $\mathrm{down}_2(\mathcal{T}) \neq \emptyset$.

Lemma 3.10.29 (cf. Lemma 1.9.26). *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 to the right of $p_r(\mathcal{T})$, then there is no corridor connecting p to q at any time $t < t_2(\mathcal{T})$.*

Definition 3.10.30 (cf. Definition 1.9.27). $\mathcal{T}_1 \in \mathcal{G}_3'$ is said to be *below* $\mathcal{T}_2 \in \mathcal{G}_3'$ if $p_l(\mathcal{T}_1)$ and $p_r(\mathcal{T}_1)$ both lie between $p_l(\mathcal{T}_2)$ and $p_r(\mathcal{T}_2)$ in the left-to-right ordering.

$\mathcal{T}_1$ is *to the left* of $\mathcal{T}_2$ if both $p_l(\mathcal{T}_1)$ 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 $\mathcal{T}$ 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 those teams in $\mathcal{G}_3'$ at any given depth.

Lemma 3.10.31 (cf. Lemma 1.9.28). *If there is a team from $\mathcal{G}_3'$ below a team $\mathcal{T} \in \mathcal{G}_3'$, then $t_1(\mathcal{T}) \geq \mathrm{time}(S_0) \geq t_2(\mathcal{T})$.*

Proof. The proof from Part 1 works almost verbatim. In particular, the same proof shows that $\mathrm{time}(S_0) \geq t_2(\mathcal{T})$.

To see that $t_1(\mathcal{T}) \geq \mathrm{time}(S_0)$, suppose that $\mathcal{T}'$ is a team below $\mathcal{T}$. Associated to the team $\mathcal{T}'$ we have the beaded Nielsen path $\mathfrak{T}'$, which is to be consumed by some reaper. The definitions of nesting and of the pincer $\Pi_{\mathcal{T}'}$ ensure that this consumption of $\mathfrak{T}'$ must occur before time $t_1(\mathcal{T})$. On the other hand, $\mathfrak{T}$ has a non-empty future or past in S_0. $\square$

With the preceding results in hand, a direct translation of the proof of Lemma 1.9.29 finishes the work of this section:

Lemma 3.10.32 (cf. Lemma 1.9.29). *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*
$$|\mathrm{down}_2(\mathcal{T}_0) \cap \mathrm{down}_2(\mathcal{T})| \leq T_1\Big(1+|\chi_c(\mathcal{T})|\Big) + T_0\Big(|\chi_\delta(\mathcal{T})|+2\Big).$$

Corollary 3.10.33 (cf. Corollary 1.9.30). *Summing over the set of teams $\mathcal{T} \in \mathcal{G}_3'$ that are not distinguished, we get*

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

Summing over the same set of teams again, we finally obtain:

Corollary 3.10.34.

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

3.11. The Bonus Scheme

This section closely follows Section 1.10. We have at last reached a stage where the proofs from Part 1 can be translated without significant modification.

In the previous section we 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 the team is such that $t_1(\mathcal{T}) \ge \mathrm{time}(S_0)$, then no member of the team is virtual and we have the inequality

$$\| \mathcal{T} \| \ge \sum_{(\mu_1,\mu_2)\in\mathcal{T}} |C_{(\mu_1,\mu_2)}| - B(B+1),$$

established in Lemma 3.10.5. This inequality might fail in case $t_1(\mathcal{T}) < \mathrm{time}(S_0)$. The *bonus scheme* assigns additional edges to teams in order to compensate for this failure.

By definition, at time $t_1(\mathcal{T})$ the reaper $\rho = \rho_{\mathcal{T}}$ lies immediately to the right of $\mathfrak{T}$. The beads of $\mathfrak{T}$ not consumed from the right by ρ by $\mathrm{time}(S_0)$ have a preferred future in S_0. This preferred future, if contained in a single colour, lies in $C_{(\mu_1,\mu_2)}(2)$ for some member $(\mu_1, \mu_2) \in \mathcal{T}$. It could also intersect more than one colour [53]. However, not all beads in the $C_{(\mu_1,\mu_2)}(2)$ need arise in this way: some may not have a Nielsen bead as an ancestor at time $t_1(\mathcal{T})$. And if (μ_1, μ_2) is only a virtual member of $\mathcal{T}$, then no bead of $C_{(\mu_1,\mu_2)}(2)$ lies in the future of $\mathfrak{T}$. The *bonus* beads in $C_{(\mu_1,\mu_2)}(2)$ are a certain subset of those that do not have a Nielsen bead as an ancestor at time $t_1(\mathcal{T})$. They are defined as follows.

Definition 3.11.1. 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 defined as in Definition 3.7.16 with respect to the interval $\mathfrak{T}$, which lies at time $t_1(\mathcal{T})$.

[53] Since Nielsen beads have bounded length, and there is a bound on the number of adjacencies of colours, there are relatively few such beads.

Let ϵ be a non-Nielsen bead that lies immediately to the left of the swollen future of $\mathcal{T}$, but whose immediate ancestor is not a right linear edge in this position. If the path from ϵ to the reaper $\rho_{\mathcal{T}}$ of $\mathcal{T}$ is a GEP, then we say that ϵ is a *rascal*. Otherwise, if ϵ provides more Nielsen beads than the reaper consumes, then ϵ is a *terror*.

In both cases, the *bonus provided by* ϵ is the set of beads in the swollen future of $\mathcal{T}$ in S_0 that have ϵ as their most recent ancestor which is not a Nielsen bead, and which are eventually consumed by $\rho_{\mathcal{T}}$.

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

LEMMA 3.11.2 (cf. Lemma 1.10.2). *For any team* $\mathcal{T}$*,*

$$\sum_{(\mu_1,\mu_2)\in\mathcal{T} \ \ or \ \ (\mu_1,\mu_2)\in_v\mathcal{T}} |C_{(\mu_1,\mu_2)}(2)| \leq \| \mathcal{T} \| + |\text{bonus}(\mathcal{T})| + B + J.$$

Note that the GEP which contains a rascal in the above definition is not displayed. We now proceed to bound the total bonus provided to teams by all rascals and terrors. Terrors are straightforward to deal with.

LEMMA 3.11.3 (cf. Lemma 1.10.3). *The sum of the lengths of the bonuses provided to all teams by terrors is less than* $2L\,|\partial\Delta|$*.*

PROOF. Let ϵ by a terror, associated to a team $\mathcal{T}$. Since the region from ϵ to the reaper of $\mathcal{T}$ is not a GEP, ϵ must be right-fast. Therefore, it will be separated from the team to which it is associated after one unit of time. Hence the bonus that ϵ provides is at most L.

That there can be at most one terror per adjacency of colours follows in a straightforward manner from Lemma 3.4.6 and the definition of terror.

Thus the total contribution of all terrors is less than $2L\,|\partial\Delta|$. □

In parallel with Definition 1.10.4, we make the following

DEFINITION 3.11.4. Fix a team $\mathcal{T}$ with $t_1(\mathcal{T}) < \text{time}(S_0)$ and consider the interval of time $[\tau_0(\epsilon), \tau_1(\epsilon)]$, where $\tau_0(\epsilon)$ is the time at which a rascal ϵ appears at the left end of the swollen future of $\mathcal{T}$, and $\tau_1(\epsilon)$ is the time at which the robust future of ϵ is no longer to the immediate left of the future of the swollen future of $\mathcal{T}$.

In the case where the robust future $\hat{\epsilon}$ of ϵ at time $\tau_1(\epsilon)$ is cancelled from the left by an edge e', we define $\tau_2(\epsilon)$ to be the earliest time when the pasts of $\hat{\epsilon}$ and e' are in the same corridor. The path in $\mathcal{F}$ that traces the past of $\hat{\epsilon}$ is denoted p_ϵ and the past following the ancestors of e' from $\tau_2(\epsilon)$ to $\tau_1(\epsilon)$ is denoted p'_ϵ. The pincer[54] formed by p_ϵ, p'_ϵ and the corridor joining them at time $\tau_2(\epsilon)$ is denoted Π_ϵ.

[54]we include the degenerate case here where the "pincer" has no colours other than those of ϵ and e'.

The only essential difference between the above definition and Definition 1.10.4 is the use of the robust future of ϵ rather than the pp-future.

With this definition in hand, the remaining results from Section 1.10 may be translated directly, yielding in particular:

PROPOSITION 3.11.5 (cf. Lemma 1.10.13). *Summing over all teams that are not short, we have*

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

3.12. From Bead Norm to Length

The output of the results up to now is a bound for the bead norm of our corridor S_0. In order to complete the proof of Theorem 3.3.1 in the case of the specified IRTT f (which implies our Main Theorem) we need to turn this into a bound on the length of S_0. For this we need to bound the total length of the GEPs and ΨEPs in S_0 which have length more than J (or indeed any other fixed length). In this section we explain how the techniques of the bonus scheme can be used to establish such a bound.

If a bead ρ in $\mu(S_0)$ has length greater than J, it is either a GEP or a ΨEP. If it is a ΨEP then we may trace its past: at each time, this past is either of length at most J or else is a ΨEP or a GEP. Whilst this past remains a ΨEP, the number of Nielsen paths will decrease with each backwards step in time, so at some point in the past of ρ, it must become a GEP.

Suppose now that ρ is a GEP. The past of a GEP is either a GEP or else has length at most J. Thus, the length of the GEP decreases as we go into the past until eventually it is of length at most J.

There is a strong analogy between teams of genesis (G4) and long GEPs and ΨEPs. On one end of a long bead is a linear edge which consumes the Nielsen beads in the middle. This linear edge can be considered as a reaper. On the other end of a GEP is a linear edge which can be considered as a rascal. The moment when the past of a ΨEP becomes a GEP is analogous to $\tau_1(\epsilon)$ from the bonus scheme, and so a ΨEP in S_0 can be thought of as a team with a rascal ϵ with $\tau_1(\epsilon) \le \mathrm{time}(S_0)$. Similarly, a long GEP in S_0 can be thought of as a team with a rascal ϵ so that $\tau_1(\epsilon) > \mathrm{time}(S_0)$.

We can define the bonus associated to such a rascal exactly as we did in the previous section. Since we are in the setting of genesis type (G4), all of the Nielsen beads in a long GEP or ΨEP are in the bonus. Thus it is enough to bound the total of the bonuses associated to long GEPs and ΨEPs.

The only thing we need to be able to follow the bonus scheme directly is a bound on the number of long GEPs and ΨEPs in S_0.

LEMMA 3.12.1. *The number of beads of length greater than J in S_0 is less then $4\,|\partial\Delta|$.*

Proof. Let ρ be a bead in S_0 of length greater than J, and assign a time $\tau_1(\rho)$ to ρ as described above. If ρ is a GEP then $\tau_1(\rho) > \text{time}(S_0)$, whilst if ρ is a ΨEP then $\tau_1(\rho) \leq \text{time}(S_0)$.

Let ρ' be the past or future of ρ at time $\tau_1(\rho) - 1$. Consider the 'event' at time $\tau_1(\rho)$ which stops the robust future of ρ' being a GEP.

This 'event' is either an intrusion of the boundary, a singularity, or else there is an associated pincer caused by a cancellation from another colour. There are less than $|\partial\Delta|$ events of each of the first two types.

The Buffer Lemma ensures that there is at most one event of the third type for each adjacency of colours. An application of Lemma 3.1.8 completes the proof. □

A bound on the total length of long beads in S_0 now follows exactly as in the bonus scheme from Section 3.11 (the detailed arguments being in Section 1.10).

3.12.1. The end of the main road. In Section 3.3 we discussed how our Main Theorem follows from Theorem 3.3.2 and Proposition 3.3.3. The bound that we just established on the total length of long beads in S_0 proves Proposition 3.3.3. The output of our estimates in the previous sections bounded the bead norm of S_0 by a linear function of $|\partial\Delta|$, and Theorem 3.3.2 follows from this because

$$[S]_\beta \leq B \parallel S \parallel_\beta,$$

(see Lemma 3.6.5).

Thus the proof of the Main Theorem is finally at an end, and the reader can join us in wondering why a statement as simple and engaging as this theorem should require such a complicated proof.

3.13. Corridor Length Functions and Bracketing

In this section we prove Theorem 3.3.1 in full generality and deduce the Bracketing Theorem from it. Our proof of Theorem 3.3.1 proceeds via a discussion of *corridor length functions* for more general semidirect products and mapping tori. Such functions should be regarded as measuring the complexity of van Kampen diagrams in the spirit of isoperimetric and isodiametric functions. We prove the following results (see Subsection 3.13.2 for precise definitions of the terms involved).

Proposition 3.13.1. *Let G_1 and G_2 be compact combinatorial complexes with fundamental group Π, and for $i = 1, 2$ let $f_i : G_i^{(1)} \to G_i^{(1)}$ be an edge-path map of 1-skeleta inducing $\phi \in \text{Out}(\Pi)$. Then the t-corridor length function for the mapping torus $M(f_1)$ is $\simeq$ equivalent to that of $M(f_2)$.*

PROPOSITION 3.13.2. *If* Π *is finitely generated and* $\Gamma = \Pi \rtimes_\phi \mathbb{Z}$ *is finitely presented, then for every positive integer* p*, the corridor length function of* Π *is* $\simeq$ *equivalent to that of* $\Gamma_p = \Pi \rtimes_{\phi^p} \mathbb{Z}$

In the previous section we completed the proof of Theorem 3.3.1 in the case of one particular IRTT representative f of a certain power of an arbitrary free-group automorphism ϕ. The above results complete the proof in the general case. Before turning to the proof of these results, we explain how the Bracketing Theorem stated in the introduction is obtained by applying Theorem 3.3.1 to the most naive topological representation of a free group automorphism ϕ.

3.13.1. The Bracketing Theorem. The terms in the following theorem were defined in the introduction.

Theorem . *There exists a constant* $K = K(\phi, \mathcal{B})$ *such that any word* $w \equiv e_1 \dots e_n$ *that represents the identity in* $F \rtimes_\phi \mathbb{Z}$ *admits a* t*-complete bracketing* $\beta_1, \dots, \beta_m$ *such that the content* c_i *of each* β_i *satisfies* $d_F(1, c_i) \le Kn$*.*

PROOF. We work with the mapping torus M of the obvious realisation of ϕ on the graph with one vertex whose edges are indexed by $\mathcal{B}$. Given a word w, we consider a minimal-area van Kampen diagram over M with boundary label w. We insert a bracket $w_1(w_2)w_3$ if and only if there is a t-corridor whose ends are labelled by the initial and terminal letters of w_2. (One must allow t-corridors of zero length in this description; one would exclude them by making the easy reduction to words that have no proper sub-words that are null-homotopic.)

These brackets are pairwise compatible because distinct t-corridors cannot cross. And because every t-edge in the boundary of a van Kampen diagram is the end of a (perhaps zero-length) corridor, the bracketing is complete. The content of the bracket is the freely reduced form of the label along the top or bottom of the corridor (according to the orientation of the sentinels). In the former case, the length of the corridor bounds the length of this label, and in the latter case one has to multiply the length by at most $L = \max\{|\phi(b)| : b \in \mathcal{B}\}$. □

3.13.2. Corridor length functions. If Π is a group with finite generating set $\mathcal{A}$ and $\phi \in \mathrm{Aut}(\Pi)$ is such that $\Gamma = \Pi \rtimes_\phi \mathbb{Z}$ is finitely presented, then Γ has a finite presentation of the form

$$\big\langle \mathcal{A},\, t \mid \mathcal{R},\, t^{-1}at = \hat{\phi}(a)\ (a \in \mathcal{A})\big\rangle,$$

where t is the generator of the visible $\mathbb{Z}$, the relations $\mathcal{R}$ involve only the letters $\mathcal{A}$, and $\hat{\phi}(a) \in F(\mathcal{A})$ is equal to $\phi(a)$ in Π.

We are concerned with the geometry of t-corridors in van Kampen diagrams over such presentations. Thus we associate to the presentation the *t-corridor length function* $\Lambda : \mathbb{N} \to \mathbb{N}$, which is defined as follows. For each $w \in F(\mathcal{A} \cup \{t\})$

with $w = 1$ in Γ, we choose a van Kampen diagram for w in which the length of the longest t-corridor is as small as possible, and we define $\lambda_t(w)$ to be this length. We then define

$$\Lambda(n) := \max\{\lambda_t(w) \mid w =_\Gamma 1,\ |w| \leq n\}.$$

More generally, since we have a well-defined notion of van Kampen diagram and t-corridor in the setting of mapping tori of *edge-path maps*[55] of combinatorial complexes, we can define the *t-corridor length function* for such a complex.

3.13.3. Invariance under change of topological representative. The scheme of the following proof follows the standard method of showing that features of the geometry of van Kampen diagrams are preserved under quasi-isometry. However, one has to be careful to deal only with fibre-preserving maps in order to retain control over the t-corridor structure.

PROOF OF PROPOSITION 3.13.1. We have a cocompact action of $\Gamma = \Pi \rtimes_\phi \mathbb{Z}$ on the universal cover $X_i = \widetilde{M}(f_i)$ for $i = 1, 2$, where the action of Π leaves invariant the connected components $C_{i,m}$ of the preimage of $G_i \subset M(f_i)$ and the generator t of $\mathbb{Z}$ acts so that $t^r.C_{i,m} = C_{i,m+r}$.

The cocompactness of the actions means that there exist constants δ_1, δ_2 so that every vertex in $C_{i,m}$ is within a distance δ_i of any Π-orbit of vertices in $C_{i,m}$, where distance is measured in the combinatorial metric on the 1-skeleton (unit edge lengths).

We define Γ-equivariant quasi-isometries between the 1-skeleta of the X_i as follows. First we pick base vertices $x_i \in C_{i,0}$ and define $g_1 : \gamma.x_1 \mapsto \gamma.x_2$ and $g_2 : \gamma.x_2 \mapsto \gamma.x_1$. Then, for each vertex $v \in C_{i,m} \smallsetminus \Gamma.x_i$ we choose a closest element $v' \in \Gamma.x_i \cap C_{i,m}$ and define $g_i(v) := g(v')$. Next, we extend to the edges in $C_{i,m}$ by sending each to a shortest edge path connecting the images of its vertices. Finally, we extend g_i to t-edges in X_i so that it sends each such homeomorphically onto the t-edge joining the images of its endpoints.

With the maps g_1, g_2 in hand, we can now push van Kampen diagrams back and forth between X_1 and X_2 as in the standard proof of the qi-invariance of Dehn functions (cf. [**15**], page 143). Thus, given a loop ℓ in the 1-skeleton of X_1, labelled $u_1 t^{\varepsilon_1} u_2 \dots u_l t^{\varepsilon_l}$ we consider the loop $g_1 \circ \ell$ in $X_2^{(1)}$ and fill it with a van Kampen diagram Δ so as minimize the length of the longest t-corridor. We will be done if we can bound $\lambda_t(\ell)$ by a linear function of this length.

Viewing Δ as a map from a cellulated 2-disc to X_2, we compose it with g_2 to obtain a map to X_1. This new map is obtained from Δ by simply changing the labels on the edges: the t-edges are unchanged while the edges labelled by 1-cells in G_2 are now labelled by edge-paths in the 1-skeleton of G_1 whose length is bounded by the constants of the quasi-isometry g_2; the boundary

[55] an *edge-path map* is a cellular map that sends edges to edge-paths

label of the diagram will be $\ell' = v_1 t^{\varepsilon_1} v_2 \dots v_l t^{\varepsilon_l}$, where the v_j are edge-paths of uniformly bounded length and each v_j is contained in the same component C_{1,m_j} as u_j. (This is the point at which we use the fact that we chose our quasi-isometries to respect fibres.) The faces of this diagram can be filled with van Kampen diagrams in X_1; in the case of 2-cells with no t-labels, we use only lifts of 2-cells from G_1; in the case of 2-cells labelled $t^{-1}\rho t\sigma$ we divide them into (short) t-corridors in the obvious manner. The result[56] is a van Kampen diagram for ℓ' in X_1 whose t-corridors are in bijection with those of Δ and whose length is bounded by k times the length of those in Δ, where k is a constant that depends only on our quasi-isometries.

To complete the desired diagram filling our original loop ℓ, we need an annular diagram between ℓ and ℓ' that does not disrupt the structure of t-corridors in Δ'. To this end, we join the vertices of u_j to those of v_j by paths in C_{i,m_j} of minimal length and fill the resulting loop with a diagram mapping to C_{i,m_j}; this gives a diagram Δ'' with holes corresponding to the occurrences of $t^{\pm 1}$ in ℓ. Next, if the arc joining the termini of u_j and v_j is labelled ρ_i, then we insert a t-corridor into the hole associated to $\dots u_j t u_{j+1} \dots$, where the bottom of the t-corridor is labelled ρ_j. (If t is replaced by t^{-1}, the bottom of the corridor is the arc σ_{j+1} joining the initial vertex of u_{j+1} to that of v_{j+1}.) To complete the construction of Δ, one uses 2-cells in $C_{i,m_{j+1}}$ to fill the loop formed by the top of the t-corridor and σ_{j+1}. ∎

COROLLARY 3.13.3. *If Π is finitely generated and $\Gamma = \Pi \rtimes_\phi \mathbb{Z}$ is finitely presented then, up to $\simeq$ equivalence, the t-corridor length function of $\Pi \rtimes_\phi \mathbb{Z}$ depends only on the semidirect product (i.e. although it depends on the form of the finite presentation, it does not depend on the choice of $\mathcal{A}$ and $\hat{\phi}$).*

3.13.4. Passing to Powers. The purpose of this subsection is to prove Proposition 3.13.2.

Let $(\mathcal{A} \cup \{t\})^{\pm 1}$ be as above. Identifying $\Gamma_p = \Pi \rtimes_{\phi^p} \mathbb{Z}$ with the subgroup $\Pi \rtimes p\mathbb{Z}$ of Γ, we take generators $\mathcal{A} \cup \{\tau\}$ where $\tau = t^p$ in Γ. To each word $w \in (\mathcal{A}^{\pm} \cup \{t^{\pm 1}\})^*$ that equals $1 \in \Gamma$ we associate a word w_p in the free group on $\mathcal{A} \cup \{\tau\}$ according to the following scheme. First we draw a path on the integer lattice in $\mathbb{R}^2$ that begins at the origin and proceeds up one space as we read t, down one as we read t^{-1} and moves one space to the right as we read a letter from $\mathcal{A}^{\pm}$. We shall modify w by replacing certain open segments of this path that lie in the vertical intervals $[mp, (m+1)p]$; these segments are of two types, called *bumps* and *steps*.

If both endpoints of the subpath are at height mp and none of its edge are at height $(m+1)p$, then the segment is called an *up-bump*. If the initial endpoint

[56] A familiar problem in this type of argument arises from degeneracies that threaten the planarity of the diagram; such problems are removed by surgery [**30**]. In the current setting these surgeries take place only in the regions between the t-corridors and therefore do not affect our discussion.

is at height mp, the terminus at height $(m+1)p$ and all other vertices are at heights in $(mp,(m+1)p)$, then the segment is called an *up-step*. A *down-bump* and *down-step* are defined similarly.

When we have replaced all steps and bumps from the path defined by w, the horizontal segments of the resulting path will all run at heights divisible by p.

To this end, we write $w = u_1v_1u_2v_2\ldots$ where u_1 is the first non-trivial prefix of w whose exponent sum in t is $0 \mod p$ and v_1 is the (possibly empty) subword before the next $t^{\pm1}$, then u_2 is the first non-trivial prefix of w whose exponent sum in t is $0 \mod p$, and so on. Each u_i labels either a bump or a step.

If u_i labels a bump then we replace it by the reduced word $U_i \in F(A)$ that is equal in Γ to u_i. If $u_i = t^{\varepsilon}u_i'$, $\varepsilon = \pm1$, is a step, then we replace it by the unique reduced word $t^{\varepsilon p}U_i$ with $U_i \in F(A)$ and $t^{\varepsilon}U_i = u_i$ in Γ.

Let $\tilde{w}_p \in (\mathcal{A}^{\pm} \cup \{t^{\pm1}\})^*$ be the word obtained from w by the above process and let $w_p \in (\mathcal{A}^{\pm} \cup \{t^{\pm p}\})^*$ be the word obtained from $\tilde{w}_p$ by (starting from the left) replacing sub-words labelled $t^{\pm p}$ by $\tau^{\pm p}$ and then freely reducing.

As usual, in the following lemma $L = \max\{|\phi(a)| : a \in \mathcal{A}\}$.

LEMMA 3.13.4. *$w = \tilde{w}_p = w_p$ in Γ and $|w_p| \le |\tilde{w}_p| \le L^{p-1}|w|$.*

PROOF. The bound on $|\tilde{w}_p|$ comes from the following observation. For a bump labelled u_i, one can pass from u_i to U_i by deleting all letters $t^{\pm1}$ from u_i and replacing each occurrence of $a \in \mathcal{A}$ in u_i, say $u_i = \alpha a \beta$, by the freely reduced word in $F(A)$ representing $\phi^r(a)$, where $-r$ is the exponent sum of t in α. Similarly, if a step is labelled $u_i = t^{\varepsilon}u_i'$, then U_i is obtained by deleting all t from u_i' and replacing each occurrence of $a \in \mathcal{A}$ in u_i, say $u_i' = \alpha a \beta$, by the freely reduced word in $F(A)$ representing $\phi^{\varepsilon(p-r)}(a)$, where εr is the exponent sum of t in α. □

The replacement scheme described in the preceding proof corresponds to the construction of a singular-disc diagram $A(w)$ exhibiting the equality $w = \tilde{w}_p$ in Γ. Specifically, for each bump or step, one draws the vertical line joining each vertex to the height where it will be pushed, one labels it by the appropriate power of t, and then one fills-in the resulting line of rectangles with 2-cells whose boundary labels have the form $t^{-1}at\phi^{-1}(a)$. (Starting from this specific planar embedding one will in general have to flip some of the components of the interior in order to get an embedded diagram $A(w)$ with boundary cycle $\tilde{w}_p w_p^{-1}$.)

LEMMA 3.13.5. *$A(w)$ is a union of t-corridors; each has at most one of its ends on the boundary arc labelled $\tilde{w}_p$, and the length of a t-corridor in $A(w)$ is at most $L^{p-1}\max|u_i|$, where the u_i are the sub-words of w labelling bumps and steps.*

Proof. The diagram $A(w)$ consists of a string of disc diagrams, one for each bump or step. A t-corridor in a disc corresponding to a bump labelled u_i has both of its ends on the arc labelled u_i, while a t-corridor in a disc corresponding to a step labelled tu_i' may have one end on the corresponding arc labelled t^p in $\tilde{w}_p$ and one on the arc labelled u_i' or (if the change in height along u_i' is not monotone) both ends on the arc labelled u_i'. In all cases, the label on the bottom side of the corridor is a concatenation of less than $|u_i|$ words of the form $\phi^r(a)$ with $a \in \mathcal{A}$ and $|r| \leq p-1$. $\square$

Proof of Proposition 3.13.2. As we discussed immediately before subsection 3.4.1, the set of diagrams for Γ_p is, after p-refinement, a subset of the diagrams over Γ, and hence the corridor length function of the latter $\preceq$-dominates that of the former. (There are some constants to take account of here, such as a factor of p in length coming from the p-refinement, and an L^{p-1} needed to estimate the area of a t-corridor in terms of the corresponding τ-corridor, but these are trivial matters.) Thus the true content of the proposition is that the corridor length function of Γ is $\preceq$-bounded above by that of the Γ_p.

For each freely-reduced word $W \in (\mathcal{A}^{\pm} \cup \{t^{\pm p}\})^*$ that is null-homotopic in Γ_p we fix a van Kampen diagram $\Delta(W)$ whose τ-corridors have length at most $\Lambda(|W|)$. Then, for each freely-reduced $w \in (\mathcal{A}^{\pm} \cup \{t^{\pm 1}\})^*$ that is null-homotopic in Γ we define a van Kampen diagram $\Delta_p(w)$ as follows. First, we replace $\Delta(w_p)$ by its p-refinement (which has boundary label $\tilde{w}_p$). We then attach to this the singular-disc diagram $A(w)$ along the portion of its boundary labelled $\tilde{w}_p$.

We claim that the length of each t-corridor in $\Delta_p(w)$ is at most

$$L^{p-1}\,(2 + \Lambda(L^{p-1}|w|)).$$

It follows from Lemma 3.13.5 that each of the t-corridors in $\Delta_p(w)$ is either contained in the annular diagram $A(w)$, or else is a layer in the p-refinement of a τ-corridor from $\Delta(w_p)$, possibly augmented on each end by a t-corridor in $A(w)$. (The fact that there are no t-corridors in $A(w)$ with both ends on the boundary arc labelled $\tilde{w}_p$ is crucial here.)

The length of a t-corridor in $A(w)$ is at most $L^{p-1}|v|$. The length of a τ-corridor from $\Delta(w_p)$ is at most $\Lambda(|w_p|) \leq \Lambda(L^{p-1}|w|)$, and the length of each layer in its refinement is therefore at most $L^{p-1}\,\Lambda(L^{p-1}|w|)$. $\square$

3.14. On a Result of Brinkmann

The following theorem is the main result in [**19**]. It plays a vital role in the first proof that the conjugacy problem is solvable for free-by-cyclic groups [**8**] (our Corollary B).

Theorem 3.14.1. [**19**, Theorem 0.1] *Let $\phi : F \to F$ be an automorphism of a finitely generated free group. Then there exists a constant $K \geq 1$ such* *that for any pair of exponents* N, i *satisfying* $0 \le i \le N$*, the following two statements hold:*

(1) If w *is a cyclic word in* F*, then*

$$\| \phi^i(w) \| \le K\Big(\| w \| + \| \phi^N(w) \| \Big),$$

where $\| w \|$ *is the length of the cyclic reduction of* w *with respect to some word metric on* F*.*

(2) If w *is a word in* F*, then*

$$|\phi^i(w)| \le K\Big(|w| + |\phi^N(w)|\Big),$$

where $|w|$ *is the word length of* w*.*

The purpose of this section is to explain how to extract Theorem 3.14.1 from our proof of the Main Theorem. We regard words and cyclic words in F_n as, respectively, based and unbased loops in the graph R with one vertex and n edges; the assertions of Theorem 3.14.1 are then statements about how the lengths of the tightened images of such loops grow when one applies the obvious topological realisation $\overline{\phi}$ of ϕ. As in the previous subsection, these assertions will follow if we can establish the corresponding bounds with $\overline{\phi} : R \to R$ replaced by a topological (IRTT) representative $f : G \to G$ of a power of ϕ satisfying Assumption 3.4.7.

Remark 3.14.2. The proof given below shows that the constant K of Theorem 3.3.1 suffices for Theorem 3.14.1. Brinkmann [**19**] states that (his constant) K can be computed effectively, but we do not see how to prove this. Indeed, given his approach (and ours), this assertion would seem to require an effective construction of an improved relative train track representative for ϕ, and a proof that such a construction exists does not seem to be available at the moment.

The following lemma allows a proof of the assertions in (1) and (2) to be undertaken simultaneously.

Lemma 3.14.3. *If* σ *is a nontrivial loop in* G*, then for some* $j \ge 1$*, the loop* $f_\#^j(\sigma)$ *admits a splitting at a vertex.*

Proof. According to [**4**, Lemma 4.1.2, p.554], σ admits a splitting $\sigma = \sigma_1$, where σ_1 is a path, but we argue further to arrange for this splitting to be at a vertex.

We divide the argument into a number of cases, depending on the largest i so that the stratum H_i contains an edge of σ_1. If this H_i is a zero stratum, $f_\#(\sigma_1) \subset G_{i-1}$ and an obvious induction applies. If H_i parabolic, then we apply [**4**, Lemma 4.1.4] to the circuit σ to obtain a splitting into paths, at least one of which is a basic path, and so has a vertex at one end. If H_i is an exponential stratum, then there is a positive integer K so that the number of i-illegal turns in $f_\#^k(\sigma_1)$ is the same for all $k \geq K$. In this case, since all Nielsen paths of exponential weight are edge-paths and all periodic paths are Nielsen, [**4**, Lemma 4.2.6] implies that $f_\#^K(\sigma_1)$ admits a splitting into sub-paths which are either r-legal or pre-Nielsen paths. If all sub-paths of $f_\#^K(\sigma_1)$ are pre-Nielsen paths, then $f_\#^{K+1}(\sigma_1)$ is a Nielsen path, and we ensured in Section 2.1 that all Nielsen paths are edge-paths.

Suppose, then, that $f_\#^K(\sigma_1)$ contains an r-legal path ρ of weight r in its splitting. Then an iterate $f_\#^i(\rho)$ of ρ contains a displayed edge ε of weight r, and the path $f_\#^{K+i}(\sigma_1)$ splits immediately on either side of ε. Since σ has weight i, the splitting of $f_\#^{K+i}(\sigma_1)$ induces a splitting of $f_\#^{K+i}(\sigma)$ at a vertex, as required. $\square$

In order to prove the statements (1) and (2), we analyze the van Kampen diagram Δ over the mapping torus of $f : G \to G$ that has boundary label $t^{-k}\sigma t^k f_\#^k(\sigma)^{-1}$. This is a simple stack of corridors as consider in Subsection 3.2.2.

In the restricted setting of stack diagrams, many of the difficulties that had to be overcome in the proof of Main Theorem do not arise (there are no singularities, for example), but there remain difficulties that one does not encounter in the context of positive automorphisms.

The number of edges in $\partial\Delta$ not labelled t is the quantity that determines the upper bound we seek, $n := |\sigma| + |f^N(\sigma)|)$. We must bound the length of each corridor in Δ linearly in terms of n. Theorem 3.3.1 provides a bound in terms of $|\partial\Delta|$, so we must argue is that in the context of stack diagrams, one can dispose of the contribution of the t-edges to this bound. In order to do so, we make an exhaustive list of those places in the proof of Theorem 3.3.1 where t-edges were accounted for, and we explain why, in each case, they are not required in the setting of simple stack diagrams.

(1) The t-edges contributed to the bound on the size of $S_0(2)$ and $S_0(3a)$ in Section 3.6, but these sets do not arise in stack diagrams.

(2) The t-edges were required in determining the sets $\mathrm{down}_1(\mathcal{T})$ used to bound the lengths of teams (see Definition 3.10.20). But $\mathrm{down}_1(\mathcal{T})$ was used only to bound the lengths of those teams whose narrow past had several components at some time in the past, and this cannot happen in a stack diagram.

(3) The t-edges entered the definition of $\mathrm{down}_2(\mathcal{T})$, which was used to bound the number of edges in $Q(\mathcal{T})$ before time $t_2(\mathcal{T})$ (see Definition 3.10.28). But there are no such edges in a stack of corridors, so we do not have to worry about double-counting, and an improved bound on the lengths of teams can be derived directly from the Pincer Lemma, noting that there are less than $2|\partial\Delta|$ adjacencies of colours.

(4) In the bonus scheme, the set ∂^e is used to bound the size of the interval of time $[\tau_0(e), \tau_2(e)]$, but in a stack of corridors it is clear that $\tau_0(e) = \tau_2(e)$, so the edges ∂^e are not required.

(5) Likewise, when bounding the size of the bonuses provided by rascals, we do not need to use the edges $\mathrm{down}_2(e)$ if our diagram is simply a stack of corridors

(6) A final use of t-edges is hidden in our references to Part 1 in the implementation of the Bonus scheme, specifically the bound on the sum of the lengths of blocks satisfying condition (iv) of the 'tautologous tetrad'. This is unnecessary in stack diagrams because there are no singularities and no edges that are cancelled by edges from outside the future of S_0, so the paths π_l and π_r travel forwards in time until they hit the boundary and $\sum |\mathrm{bdy}(\mathfrak{B})| < n$ bounds the size of the sum of all such blocks. $\square$

Bibliography

[1] J. Alonso, Inégalités isopérimétriques et quasi-isométries, *C. R. Acad. Sci. Paris*, **311** (1990), 761-764.

[2] 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.

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

[4] 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.

[5] M. Bestvina, M. Feighn and M. Handel, Solvable subgroups of $\mathrm{Out}(F_n)$ are virtually abelian, *Geom. Ded.* **104** (2004), 71–96.

[6] M. Bestvina, M. Feighn and M. Handel, The Tits alternative for $\mathrm{Out}(F_n)$, II: A Kolchin type theorem, *Ann. of Math. (2)*, **161** (2005), 1–59.

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

[8] O. Bogopolski, A. Martino, O. Maslakova and E. Ventura, Free-by-cyclic groups have solvable conjugacy problem, *Bull. London Math. Soc* **38** (2006), 787–794.

[9] N. Brady and M.R. Bridson, On the absence of biautomaticity in certain graphs of abelian groups, preprint.

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

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

[12] 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.

[13] 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.

[14] M.R. Bridson and D. Groves, The growth of conjugacy classes under free-group automorphisms, in preparation.

[15] M.R. Bridson and A. Haefliger, *Metric spaces of non-positive curvature*, Springer-Verlag, Berlin, 1999.

[16] M.R. Bridson and L. Reeves, On the absence of automaticity in certain free-by-cyclic groups, in preparation.

[17] M.R. Bridson and K. Vogtmann, *Automorphism groups of free, surface, and free-abelian groups*, in "Problems on mapping class groups and related topics", Proc. Sympos. Pure Math. **74**, B. Farb (ed.), Amer. Math. Soc., Providence, RI, 2006. pp. 301–316.

[18] P. Brinkmann, Hyperbolic automorphisms of free groups, *GAFA*, **10** (2000), 1071–1089.

[19] P. Brinkmann, Dynamics of free group automorphisms, preprint.

[20] A.J. Casson and S. Bleiler, Automorphisms of surfaces after Nielsen and Thurston, *LMS Student Texts* **9**, Cambridge Unversity Press, Cambridge, 1988.

[21] D. Cooper, Automorphisms of free groups have finitely generated fixed point sets, *J. Algebra*, **111** (1987), 453–456.

[22] 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.

[23] M. Feighn and M. Handel, Mapping tori of free group automorphisms are coherent, *Ann. of Math. (2)*, **149** (1999), 1061–1077.

[24] M. Feighn and M. Handel, The Recognition Theorem for $\mathrm{Out}(F_n)$, preprint.

[25] S.M. Gersten, The automorphism group of a free group is not a CAT(0) group, *Proc. Amer. Math. Soc.*, **121** (1994), 999–1002.

[26] M. Gromov, Hyperbolic groups, in *Essays in group theory* (S.M. Gersten, ed.), Springer Verlag, MSRI Publ. **8** (1987), 75-263.

[27] E.R. van Kampen, On some lemmas in the theory of groups, *Amer. J. Math.*, **55** (1933), 268–273.

[28] B. Leeb, 3-Manifolds with(out) metrics of nonpositive curvature, *Invent. Math.* **122** (1995), 277–289.

[29] M. Lustig, Structure and conjugacy for automorphisms of free groups I,II, *MPI-Preprint series* (2000) **241** and (2001) **4**.

[30] R.C. Lyndon and P.E. Schupp, Combinatorial group theory, Springer-Verlag, Berlin, 1977.

[31] N. Macura, Quadratic isoperimetric inequality for mapping tori of polynomially growing automorphisms of free groups, *GAFA*, **10** (2000), 874–901.

[32] A.Yu. Ol'shanskii and M.V. Sapir, Groups with small Dehn functions and bipartite chord diagrams, *Geom. Funct. Anal.* **16** (2006), 1324–1376..

[33] P. Papasoglu, On the asymptotic cone of groups satisfying a quadratic isoperimetric inequality, *J. Diff. Geom.* **44** (1996), 789–806.

[34] S. Schleimer, Polynomial time word problems, preprint.

[35] Z. Sela, The Nielsen-Thurston classification and automorphisms of a free group I, *Duke Math. J.*, **84** (1996), 379–397.

[36] E. Seneta, Non-negative matrices and Markov chains, Springer-Verlag, New York, 1981.

Index