Herman Goulet-Ouellet
University of Coimbra, CMUC, Department of Mathematics, and University of Porto, CMUP, Department of Mathematics
E-mail address: hgouletouellet@student.uc.pt

Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 21–19

SUFFIX-CONNECTED LANGUAGES

HERMAN GOULET-OUELLET

Abstract: Inspired by a series of papers initiated in 2015 by Berthé et al., we introduce a new condition called suffix-connectedness. We show that the groups generated by the return sets of a uniformly recurrent suffix-connected language lie in a single conjugacy class of subgroups of the free group. Moreover, the rank of the subgroups in this conjugacy class only depends on the number of connected components in the extension graph of the empty word. We also show how to explicitly compute a representative of this conjugacy class using the first order Rauzy graph. Finally, we provide an example of suffix-connected, uniformly recurrent language that contains infinitely many disconnected words.

Keywords: Tree sets, Extension graphs, Return words, Rauzy graphs, Stallings algorithm, Free groups.
Math. Subject Classification (2010): 68Q45, 68R15.

1. Introduction

In [4], Berthé et al. introduced the notion of extension graph and used it to study the subgroups generated by the return sets in uniformly recurrent languages. One result achieved in that paper, dubbed the *Return Theorem*, states that if L is a uniformly recurrent language on the alphabet A such that all the extension graphs of L are trees, then the return sets of L are all bases of the free group on A [4, Theorem 4.5]. Moreover, they also show that part of this result holds under weaker assumptions: if we merely assume that the extension graphs of L are connected, then the return sets of L all generate the free group on A [4, Theorem 4.7]. The aim of this paper is to give a weaker condition under which a similar conclusion still holds. To do this, we introduce *suffix extension graphs*, a notion generalizing the extension graphs of [4]. This allows us to define a new condition called *suffix-connectedness*. Our main result is the following:

Received June 7, 2021.
The author is grateful for the financial support provided by the Centre for Mathematics of the University of Coimbra (UIDB/00324/2020, funded by the Portuguese Government through FCT/MCTES), the Centre for Mathematics of the University of Porto (UIDB/00144/2020, funded by the Portuguese Government through FCT/MCTES), as well as a PhD grant from FCT/MCTES (PD/BD/150350/2019). Special thanks go to Jorge Almeida and Alfredo Costa for many helpful discussions and comments which greatly improved this paper.

Theorem 1.1. *Let L be a suffix-connected uniformly recurrent language on an alphabet A. Then the subgroups generated by the return sets of L all lie in the same conjugacy class and their rank is $n - c + 1$, where $n = \mathrm{Card}(A)$ and c is the number of connected components of the extension graph of the empty word.*

Our proof is constructive, in the sense that we can also deduce a way to explicitly compute a representative for this conjugacy class. Moreover, the proof of Theorem 1.1 has two notable consequences that we wish to highlight now. The first one is a characterization of suffix-connected, uniformly recurrent languages whose return sets generate the full free group.

Corollary 1.2. *Let L be a suffix-connected and uniformly recurrent language on the alphabet A. Then the following statements are equivalent:*

(1) *All the return sets of L generate the free group on A.*
(2) *Some return set of L generates a group of rank* $\mathrm{Card}(A)$.
(3) *The extension graph of the empty word is connected.*

The next corollary is a special case of our main result. It involves *neutrality*, which is a combinatorial condition also introduced in [4] (we will recall the definition in Section 8). A *connected set* is a language in which the extension graphs of non-empty words are connected, while a *tree set* is a language in which the extension graph of the empty word is a forest, and all other extension graphs are trees. These conventions differ slightly from [4], but are in line with other papers such as [3, 7]. The term *dendric* has also been used to refer to tree sets, for instance in [8]. A subset of the free group is called *free* if it forms a basis of the subgroup it generates.

Corollary 1.3. *Let L be a uniformly recurrent language on the alphabet A. If L is connected and neutral, then the following conditions are equivalent:*

(1) *Some return set of L is a free subset of the free group on A.*
(2) *All return sets of L are free subsets of the free group on A.*
(3) *L is a tree set.*

Since connectedness implies suffix-connectedness, the assumptions of the Return Theorem place us in the scope of both Corollary 1.2 and 1.3. It follows that the Return Theorem is a direct consequence of the above corollaries.

In order to further motivate this new *suffix-connectedness* condition, we give an example of a uniformly recurrent language which is suffix-connected but contains infinitely many disconnected elements. This language is defined by a primitive substitution. More precisely, we will show the following:

while also having infinitely many disconnected elements:

$$\begin{array}{l} 0 \mapsto 100 \\ 1 \mapsto 032 \\ 2 \mapsto 232 \\ 3 \mapsto 03 \end{array}; \qquad \begin{array}{l} 0 \mapsto 01 \\ 1 \mapsto 2 \\ 2 \mapsto 3 \\ 3 \mapsto 02 \end{array}; \qquad \begin{array}{l} 0 \mapsto 12 \\ 1 \mapsto 2 \\ 2 \mapsto 01 \end{array}.$$

In [8], Dolce and Perrin introduced the notion of an *eventually dendric language*, which requires all but finitely many extension graphs to be trees. This suggests the analogous notion of an *eventually suffix-connected language*, in which all but finitely many words are suffix-connected.

Question 10.2. Can we find a generalization of Theorem 1.1 for eventually suffix-connected languages?

Finally, Dolce and Perrin also showed that the class of eventually dendric languages is closed under two operations, namely conjugacy and complete bifix decoding [8]. We wonder if analogous results hold for suffix-connected languages.

Question 10.3. Is the class of suffix-connected languages closed under complete bifix decoding or conjugacy?

References

[1] J. Almeida and A. Costa, *A geometric interpretation of the Schützenberger group of a minimal subshift*, Ark. Mat. **54** (2016), no. 2, 243–275.
[2] S. Balchin and D. Rust, *Computations for symbolic substitutions*, J. Integer Seq. **20** (2017).
[3] V. Berthé, C. De Felice, V. Delecroix, F. Dolce, J. Leroy, D. Perrin, C. Reutenauer, and G. Rindone, *Specular sets*, Theoret. Comput. Sci. **684** (2017), 3–28.
[4] V. Berthé, C. De Felice, F. Dolce, J. Leroy, D. Perrin, C. Reutenauer, and G. Rindone, *Acyclic, connected and tree sets*, Monatsh. Math. **176** (2015), no. 4, 521–550.
[5] J. A. Bondy and U. S. R. Murty, *Graph theory with applications*, Macmillan London, 1976.
[6] The Sage Developers, *SageMath, the Sage Mathematics Software System (Version 9.2)*, 2020, `https://www.sagemath.org`.
[7] F. Dolce and D. Perrin, *Neutral and tree sets of arbitrary characteristic*, Theoret. Comput. Sci. **658** (2017), 159–174.
[8] ———, *Eventually dendric shift spaces*, Ergodic Theory Dynam. Systems (2020), 1–26.
[9] F. Durand, *A characterization of substitutive sequences using return words*, Discrete Math. **179** (1998), no. 1-3, 89–101.
[10] N. Pytheas Fogg, V. Berthé, S. Ferenczi, C. Mauduit, and A. Siegel (eds.), *Substitutions in dynamics, arithmetics and combinatorics*, Springer Berlin Heidelberg, 2002.
[11] I. Kapovich and A. Myasnikov, *Stallings foldings and subgroups of free groups*, J. Algebra **248** (2002), no. 2, 608–668.
[12] B. Steinberg, *Fundamental groups, inverse Schützenberger automata, and monoid presentations*, Comm. Algebra **28** (2000), no. 11, 5235–5253.

latter boils down to a few computations:

$$\begin{aligned}|\varphi^k(0010)| - |x_k| &= |\varphi^k(001)| + |\varphi^k(0)| - |\varphi^k(2)| + 1\\ &> |\varphi^k(001)| + 1 = d_k;\\ |\varphi^k(2000)| - |x_k| &= |\varphi^k(2000)| - |\varphi^k(2)| + 1\\ &= |\varphi^k(000)| + 1\\ &> |\varphi^k(001)| + 1 = d_k.\end{aligned}$$

Note that σ_k maps $\{0001, 1001\}$ onto the natural embedding of $\mathbf{L}(w_k)$. Since $\mathbf{E}_{d_0,d_0}(y_0)$ is connected, it suffices to show that σ_k defines a graph morphism. Take $u \in \mathbf{L}_{d_0}(y_0)$ and $v \in \mathbf{R}_{d_0}(y_0)$, and suppose that $uy_0v \in L$. Then, it follows that $\varphi^k(uy_0v) \in L$. Since $\sigma_k(u)$ is a suffix of u and $x_k\sigma_k(v)$ is a prefix of v, we conclude that $\sigma_k(u)y_k\sigma_k(v) = \sigma_k(u)\varphi^k(y_0)x_k\sigma_k(v)$ is a factor of $\varphi^k(uy_0v)$. Therefore, it must also be in L, and $\sigma_k\colon \mathbf{E}_{d_0,d_0}(y_0) \to \mathbf{E}_{d_k,d_k}(y_k)$ is a graph morphism. ∎

With some extra work, we were able to show that the map σ_k defined in the previous proof is in fact a graph isomorphism. To prove this, we made use of the following observation, which is a consequence of the Cayley-Hamilton theorem: for any word $x \in A^*$, the sequence $(|\varphi^k(x)|)_{k\in\mathbb{N}}$ follows the linear recurrence determined by the characteristic polynomial of φ. This is a general observation which holds for any substitution, and we believe it could be useful for establishing suffix-connectedness in harder cases.

10. Conclusion

Let us end this paper by suggesting a few ideas for future reasearch.

Firstly, we feel that the proof presented in Section 9, on account of its ad-hoc and technical nature, is somewhat unsatisfactory. We hope it could be improved.

Question 10.1. Is there a more systematic approach to show that a given language is suffix-connected?

In particular, it could be interesting to study other examples of suffix-connected languages defined by primitive substitutions, and see how much of Section 9 can be recycled. According to our computations, the languages defined by the following primitive substitutions are likely to be suffix-connected

Theorem 1.4. *The language of the primitive substitution*

$$\varphi\colon \begin{array}{ccc} 0 & \mapsto & 0001\\ 1 & \mapsto & 02\\ 2 & \mapsto & 001\end{array}$$

is suffix-connected and contains infinitely many disconnected words.

We will also see that, in the language of this substitution, the extension graph $\mathbf{E}(\varepsilon)$ is connected. Therefore, as a result of Corollary 1.2, all the return sets in this language generate the full free group of rank 3. However, further computations reveal that the language of φ has return sets of cardinality 3 and 4, which means that some but not all of them are free subsets of the free group.

This paper is structured as follows. In Section 2, we introduce suffix extension graphs and suffix-connectedness, while also recalling some relevant definitions in more details. Section 3 reviews some basic material about the groups generated by labeled digraphs. Section 4 is devoted to Rauzy graphs. Section 5 presents a technical result that makes up the core of the proof of our main result. In Section 6, we examine the relationship between Rauzy graphs and return sets. In Section 7, we put everything together and give the proof of Theorem 1.1. Section 8 discusses the proof of the two corollaries above. Finally, Section 9 is devoted to our suffix-connected example.

2. Suffix-connectedness

In this paper, L denotes a language on a finite alphabet A of cardinality n, and $F(A)$ denotes the free group on A. We will always suppose that L is recurrent and that $A \subseteq L$. We recall that a language L is *recurrent* if it is closed under taking factors, and if for every two words $u, v \in L$, there exists a non-empty word w such that $uwv \in L$. The *left extensions* and *right extensions* of order k of $w \in L$, are:

$$\mathbf{L}_k(w) = \{u \in L \cap A^k : uw \in L\}, \qquad \mathbf{R}_k(w) = \{v \in L \cap A^k : wv \in L\}.$$

The *extension graph of order* (k, l) of $w \in L$ is a bipartite graph over the disjoint union of $\mathbf{L}_k(w)$ and $\mathbf{R}_l(w)$ (the union of disjoint copies of $\mathbf{L}_k(w)$ and $\mathbf{R}_l(w)$). In this graph, there is an edge between $u \in \mathbf{L}_k(w)$ and $v \in \mathbf{R}_l(w)$ if $uwv \in L$. We denote this graph by $\mathbf{E}_{k,l}(w)$. Note that all extension graphs are simple and undirected. We abbreviate $\mathbf{R}_1$, $\mathbf{L}_1$ and $\mathbf{E}_{1,1}$ respectively by $\mathbf{R}$, $\mathbf{L}$ and $\mathbf{E}$. In the absence of further clarifications, the term *extension graph of* w refers to $\mathbf{E}(w)$. A word is *connected* if its extension graph is connected, and it

is called *disconnected* otherwise. A language is *connected* if all its non-empty words are connected, and it is *disconnected* otherwise.

A word $w \in L$ is called *left special* if $\mathrm{Card}(\mathbf{L}(w)) > 1$. Similarly, w is called *right special* if $\mathrm{Card}(\mathbf{R}(w)) > 1$. By a *bispecial word*, we mean a word which is both left and right special.

Remark 2.1. If w is not bispecial, then $\mathbf{E}(w)$ is a star graph, and in particular a tree. Hence, only bispecial factors can be disconnected.

Given a word $w \in A^*$ and $0 \leq i < |w|$, we denote by $w(i)$ the i-th letter of w. In particular, the first letter of w is $w(0)$. Given $0 \leq i \leq j \leq |w|$, we denote $w[i:j]$ the factor of w defined by:

$$w[i:j] = w(i)w(i+1)\dots w(j-1)$$

Note that $|w[i:j]| = j - i$, $w[i:i]$ is the empty word and $w[0:|w|] = w$. Let $u \in A^*$ with $|u| = k$. We say that an index j is an *occurrence* of u in w if $w[j:j+k] = u$. We also define the *tail* and the *init* of a non-empty word w by putting:

$$\mathrm{tail}(w) = w[1:|w|], \quad \mathrm{init}(w) = w[0:|w|-1].$$

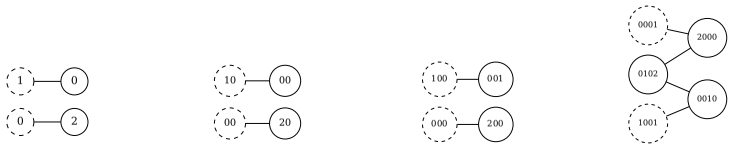
We view tail and init as maps $A^+ \to A^*$. With this, we are now ready to introduce suffix extension graphs.

Definition 2.2. For $w \in L$ and $1 \leq d \leq |w|+1$, the *depth d suffix extension graph* of w is the extension graph $\mathbf{E}_{d,d}(\mathrm{tail}^{d-1}(w))$.

The set $\mathbf{L}(w)$ naturally embeds in the suffix-extension graphs of w. Indeed, let u be the prefix of length $d-1$ of w, which means that u satisfies $w = u\,\mathrm{tail}^{d-1}(w)$. Then $a \mapsto au$ is an injective map $\mathbf{L}(w) \to \mathbf{L}_d(\mathrm{tail}^{d-1}(w))$, with the latter set being viewed as a subset of $\mathbf{E}_{d,d}(\mathrm{tail}^{d-1}(w))$. We call this the *natural embedding* of $\mathbf{L}(w)$ in the depth d suffix extension graph.

Definition 2.3. A word w is called *suffix-connected* if the natural embedding of $\mathbf{L}(w)$ in $\mathbf{E}_{d,d}(\mathrm{tail}^{d-1}(w))$ lies in one connected component, for some $1 \leq d \leq |w|+1$. A language is called suffix-connected if all its non-empty words are suffix-connected.

We note that this definition is sensitive to both increases and decreases in the depth parameter. That is, for a given word w, it may happen that some of the natural embeddings $\mathbf{L}(w)$ lie in a single connected component, while others do not. The next example is a good illustration of this behaviour. It features



$\mathbf{E}_{1,1}(001000100010)$ $\mathbf{E}_{2,2}(01000100010)$ $\mathbf{E}_{3,3}(1000100010)$ $\mathbf{E}_{4,4}(000100010)$

FIGURE 9. Suffix extension graphs of $w_0 = 001000100010$ at depth up to 4. The dashed vertices represent the natural embeddings of $\mathbf{L}(w_0)$.

Hence, the result follows. Similarly, if k is odd, $\varphi^k(2) = x_k 1$, $\varphi^{k+1}(2) = x_{k+1}2$, and

$$x_{k+1}2 = \varphi(x_k 1) = \varphi(x_k)02.$$

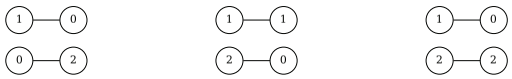
(2) We proceed by induction on k. The basis, $k = 0$, is obvious. Let us assume $\varphi^k(0) = x_k 0 t_k$, for some $t_k \in A^*$. Hence,

$$\varphi^{k+1}(0) = \varphi(x_k 0 t_k) = \varphi(x_k)0001\varphi(t_k) = \begin{cases} x_{k+1}01\varphi(t_k) & \text{if } k \text{ even;} \\ x_{k+1}001\varphi(t_k) & \text{if } k \text{ odd.} \end{cases}$$ ■

By the recursive definition of w_k and a straightforward inductive argument involving Part (1) of the lemma, we have $\varphi^k(y_0)x_k = y_k$ for all $k \in \mathbb{N}$. Moreover, note that x_k is a prefix of both $\varphi^k(2000)$ and $\varphi^k(0010)$, the former by definition and the latter by part (2) of the lemma. Since 2000 and 0010 are right extensions of y_0 (see Figure 9), it follows that $x_k^{-1}\varphi^k(2000)$ and $x_k^{-1}\varphi^k(0010)$ are right extensions of y_k. With these observations in mind, we are ready to conclude the proof of Theorem 1.4. We do this by establishing the following claim.

Claim. For $k \geq 1$, there is a path in $\mathbf{E}_{d_k,d_k}(y_k)$ between $1\varphi^k(001)$ and $2\varphi^k(001)$.

Proof of the claim: Consider the map $\sigma_k \colon \mathbf{E}_{d_0,d_0}(y_0) \to \mathbf{E}_{d_k,d_k}(y_k)$ defined as follows: an element $u \in \mathbf{L}_{d_0}(y_0) = \{0001, 0102, 1001\}$ is mapped to the suffix of length d_k of $\varphi^k(u)$, and an element $v \in \mathbf{R}_{d_0}(y_0) = \{2000, 0010\}$ is mapped to the prefix of length d_k of $x_k^{-1}\varphi^k(v)$. We first need to show that this map is well-defined. This amounts to show that $|\varphi^k(u)| \geq d_k$ for all $u \in \mathbf{L}(y_0)$, and $|\varphi^k(v)| - |x_k| \geq d_k$ for all $v \in \mathbf{R}(y_0)$. The former condition is obvious, and the



$\mathbf{E}(w_0)$ $\mathbf{E}(w_k)$, $k \geq 1$ odd $\mathbf{E}(w_k)$, $k \geq 1$ even

FIGURE 8. Extension graphs of the disconnected words of L.

1 is a right extension of x, it follows from Step 1 that x ends with 00, so by Step 3, $\mathbf{E}(x) \cong \mathbf{E}(\varphi(x)0) = \mathbf{E}(w)$. This implies that x is disconnected, so by induction $x = w_k$ for some $k \in \mathbb{N}$. As x ends with 00, k is odd and $w = \varphi(w_k)0 = w_{k+1}$. ∎

This, combined with the graph isomorphism identified in Step 3, allows us to explicitly compute the extension graphs of all the disconnected words of L. These extension graphs are shown in Figure 8.

Step 5. Now that we know exactly which are the disconnected words of L, it remains to show that these words are suffix-connected. For $k \in \mathbb{N}$, let us write $d_k = |\varphi^k(001)| + 1$ and $y_k = \text{tail}^{d_k-1}(w_k)$. This means that $w_k = \varphi^k(001)y_k$ and the depth d_k suffix extension graph of w_k is precisely $\mathbf{E}_{d_k,d_k}(y_k)$. For the case $k = 0$, we have $d_0 = 4$ and $y_0 = 000100010$. Notably, the depth 4 suffix extension graph of w_0, which is shown in Figure 9, is connected, hence w_0 is suffix-connected at depth 4. We will show that w_k is suffix-connected at depth d_k for all $k \in \mathbb{N}$.

Let us first note that the natural embedding of $\mathbf{L}(w_k)$ in $\mathbf{E}_{d_k,d_k}(y_k)$ is given by right multiplication by $\varphi^k(001)$. Before concluding the proof of Theorem 1.4, we need to establish the following technical lemma, which gives some properties of the words $x_k = \text{init}(\varphi^k(2))$.

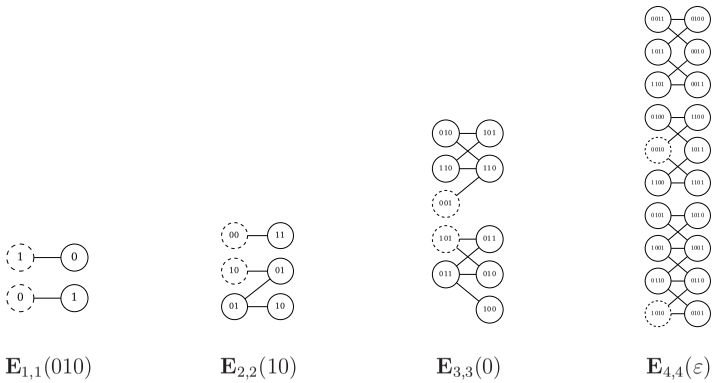
Lemma 9.2. *For all $k \in \mathbb{N}$, the following hold:*

(1) $x_{k+1} = \begin{cases} \varphi(x_k)00 & \text{if } k \text{ is even;} \\ \varphi(x_k)0 & \text{if } k \text{ is odd.} \end{cases}$

(2) x_k0 *is a prefix of* $\varphi^k(0)$.

Proof: **(1)** If k is even, then $\varphi^k(2) = x_k2$ and $\varphi^{k+1}(2) = x_{k+1}1$. It follows that

$$x_{k+1}1 = \varphi(x_k2) = \varphi(x_k)001.$$



$\mathbf{E}_{1,1}(010)$ $\mathbf{E}_{2,2}(10)$ $\mathbf{E}_{3,3}(0)$ $\mathbf{E}_{4,4}(\varepsilon)$

FIGURE 1. Suffix extension graphs of the word 010 in the language of the Thue-Morse substitution. The dashed vertices represent the natural embeddings of $\mathbf{L}(010)$.

a language defined by a primitive substitution, and such languages are well known to be uniformly recurrent (see for instance [10, Proposition 1.2.3]).

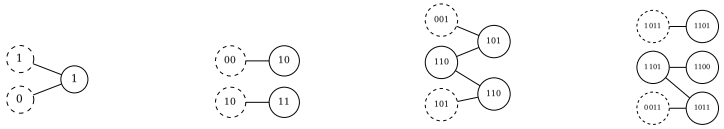
Example 2.4. Let us consider the following binary substitution, known as the *Thue-Morse* substitution:

$$\mu\colon \begin{array}{ccc} 0 & \mapsto & 01 \\ 1 & \mapsto & 10 \end{array}.$$

Let L be the language defined by μ. That is, L is the set of factors of all words of the form $\mu^n(a)$ for $n \in \mathbb{N}$ and $a \in \{0, 1\}$. Figure 1 gives all the suffix extension graphs of the word $010 \in L$, which show that L is not suffix-connected.

On the other hand, Figure 2, gives some extension graphs of $01100 \in L$. These graphs show that the natural embeddings of a given word can alternate between being connected and disconnected as the depth increases.

Replacing tail by init and $\mathbf{L}$ by $\mathbf{R}$ yields the dual notions of *prefix extension graphs* and *prefix-connectedness*. Note that the depth 1 suffix and prefix extension graphs of w both coincide with $\mathbf{E}(w)$, so a connected word or language is both prefix and suffix-connected.

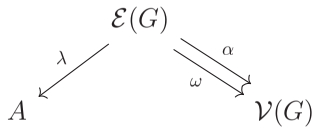


$\mathbf{E}_{1,1}(01100)$ $\mathbf{E}_{2,2}(1100)$ $\mathbf{E}_{3,3}(100)$ $\mathbf{E}_{4,4}(00)$

FIGURE 2. First four suffix extension graphs of the word 01100 in the language of the Thue-Morse substitution. The dashed vertices represent the natural embeddings of $\mathbf{L}(01100)$.

3. Stallings equivalence

Let us start this section by clarifying some basic terminology. A *labeled digraph* over the alphabet A (or, more simply, a *digraph*) is a diagram of sets G of the following form:



One can think of $\mathcal{V}$ as the set of *vertices*, $\mathcal{E}$ as the set of *edges*, and A, the alphabet, as the set of *labels*. The maps α, ω and λ give us respectively the origin, terminus and label of a given edge. For our purposes, we may assume that there are no redundant edges, meaning that $(\alpha, \lambda, \omega)$ are jointly injective. This means in effect that $\mathcal{E}$ may be considered a subset of $\mathcal{V} \times A \times \mathcal{V}$ whenever convenient.

Given an edge $e = (x, a, y)$, we consider its formal inverse $e^{-1} = (y, a^{-1}, x)$. From now on, we use the term *edge* both for elements of $\mathcal{E}(G)$ and for their formal inverses. Two edges are said to be *consecutive* if the last component of the first is equal to the first component of the second. A *path* is a sequence of consecutive edges. We can naturally extend the maps α, ω to paths, and talk about *consecutive paths*. Two consecutive paths can be composed, and any path can be inverted; we write respectively pq and p^{-1}. A self-consecutive path is called a *loop*. As expected, if p, q are consecutive, then so are q^{-1}, p^{-1} and the relation $(pq)^{-1} = q^{-1}p^{-1}$ holds.

The labeling map λ also naturally extends, mapping the set of all paths to the free group $F(A)$. This map satisfies $\lambda(pq) = \lambda(p)\lambda(q)$ and $\lambda(p^{-1}) = \lambda(p)^{-1}$.

For the converse, we proceed by induction on $|w|$. The smallest disconnected word, w_0, provides the basis for the induction. Let us consider a disconnected word $w \in L$ such that $|w| > |w_0| = 12$. Since w_0 is also the longest bispecial factor starting with 001 (see Step 2), we may assume that w starts with 000.

We start by treating the case $w \in A^*00$. By Step 1, we know that:

$$\mathbf{L}(w) = \{1, 2\}, \quad \mathbf{R}(w) = \{0, 1\}.$$

Let us write $w = w'00$. Since $w0 \in L$, it follows that w' cannot end with 0. Let us consider $z_1, z_2, z_3, z_4 \in L$ such that:

$$\varphi(z_1) = s_1 2 w t_1, \quad \varphi(z_2) = s_2 1 w t_2, \quad \varphi(z_3) = s_3 w 0 t_3, \quad \varphi(z_4) = s_4 w 1 t_4.$$

By repeatedly applying Lemma 9.1, we deduce that:

- z_1 has a factor of the form $1x_1$ such that $\varphi(x_1) = w'$.
- z_2 has a factor of the form ax_2 such that $\varphi(x_2) = w'$ and $a \in \{0, 2\}$.
- z_3 has a factor of the form $x_3 0$ such that $\varphi(x_3) = w'$.
- z_4 has a factor of the form $x_4 2$ such that $\varphi(x_4) = w'$.

Since φ is injective, we deduce $x_1 = x_2 = x_3 = x_4 = x$, and x is bispecial. Note that $|x| \leq 2$ would imply $|w| = |\varphi(x)| + 2 \leq 10$, which is a contradiction. Thus, we may assume $|x| \geq 3$. Since 2 is a right extension of x, we deduce by Step 1 that $x \in A^*10$. By Step 4, $\mathbf{E}(x) \cong \mathbf{E}(\varphi(x)00) = \mathbf{E}(w)$; thus, x is disconnected. By induction, $x = w_k$ for some $k \in \mathbb{N}$, and since x ends with 10, k is even. Therefore $w = \varphi(w_k)00 = w_{k+1}$.

The case $w \in A^*10$ is handled in a similar fashion. Let us go quickly over the argument. This time, we have

$$\mathbf{L}(w) = \{1, 2\}, \quad \mathbf{R}(w) = \{0, 2\}.$$

We write $w = w'0$. Take $z_1, z_2, z_3, z_4 \in L$ such that:

$$\varphi(z_1) = s_1 2 w t_1, \quad \varphi(z_2) = s_2 1 w t_2, \quad \varphi(z_3) = s_3 w 0 t_3, \quad \varphi(z_4) = s_4 w 2 t_4.$$

Again, it follows from Lemma 9.1 that:

- z_1 has a factor of the form $1x_1$ such that $\varphi(x_1) = w'$.
- z_2 has a factor of the form ax_2 such that $\varphi(x_2) = w'$ and $a \in \{0, 2\}$.
- z_3 has a factor of the form $x_3 a$ such that $\varphi(x_3) = w'$ and $a \in \{0, 2\}$.
- z_4 has a factor of the form $x_4 1$ such that $\varphi(x_4) = w'$.

By injectivity of φ, we have $x_1 = x_2 = x_3 = x_4 = x$ and x is bispecial. Moreover, $|x| \leq 2$ would imply $|w| = |\varphi(x)| + 1 \leq 9$, which contradicts our standing assumption that $|w| > 12$. Thus, we may assume $|x| \geq 3$. Since

if $x \in A^*10$, then $00\sigma(b)$ is a prefix of $\varphi(b)$, for all $b \in \mathbf{R}(x)$. In particular, $\sigma(a)y\sigma(b)$ is a factor of $\varphi(axb)$. Thus,

$$axb \in L \implies \varphi(axb) \in L \implies \sigma(a)y\sigma(b) \in L.$$

This shows that $\sigma\colon \mathbf{E}(x) \to \mathbf{E}(y)$ is a graph morphism, and that y is bispecial. Moreover, it follows from Step 1 that σ is bijective on vertices. It remains only to show that σ is onto on edges.

Let us suppose that $\sigma(a)y\sigma(b) \in L$ for $a, b \in A$. We need to show that $axb \in L$. The fact that $\sigma(a)y\sigma(b) \in L$ implies:

$$\exists z \in L,\ \varphi(z) = s\sigma(a)y\sigma(b)t.$$

Note that $y = \varphi(x)0$ or $\varphi(x)00$. In both cases, $\varphi(x)$ is a prefix of y ending with 1 or 2. Moreover, we have $\sigma(a) \in \{1, 2\}$ by Step 1. Thus, in the factorization $s\sigma(a)y\sigma(b)t$ given above, there must be one cutting point at the start of y, and one at the end of $\varphi(x)$. By Lemma 9.1, this implies that z has a factor of the form $cx'd$, where:

(1) $\varphi(x') = \varphi(x)$;
(2) $\sigma(a)$ is a suffix of $\varphi(c)$;
(3) $\varphi(d)$ starts with $0\sigma(b)$ if $x \in A^*00$; or $00\sigma(b)$ if $x \in A^*10$.

Injectivity of φ implies $x = x'$, so $c \in \mathbf{L}(x)$ and $d \in \mathbf{R}(x)$. Using Step 1, we can then deduce from a case-by-case analysis that $a = c$ and $b = d$. This proves that σ is onto on edges, thus finishing the proof. ∎

Step 4. Recall that the longest bispecial factor of L that starts with 001 is also its smallest disconnected element. We will see that *all* disconnected elements of L arise from this word. Consider the sequence of words $\{w_k\}_{k\in\mathbb{N}}$ defined by:

$$w_0 = 001000100010, \qquad w_{k+1} = \begin{cases} \varphi(w_k)00 & \text{if } k \text{ is even;} \\ \varphi(w_k)0 & \text{if } k \text{ is odd.} \end{cases}$$

For the purpose of the proof below, it is useful to notice that w_k ends with 10 if k is even, and with 00 if k is odd. We now prove the following claim.

Claim. A word $w \in L$ is disconnected if and only if $w = w_k$ for some $k \in \mathbb{N}$.

Proof: We already saw that w_0 is disconnected, and one can check via explicit computations that so is w_1. Since w_k starts with 000 whenever $k \geq 1$, it follows from Step 3 that w_k is disconnected for all $k \in \mathbb{N}$.

We write $p\colon x \xrightarrow{u} y$ as a shorthand for $\alpha(p) = x$, $\omega(p) = y$, $\lambda(p) = u$. The set of all labels of loops over a given vertex x forms a subgroup of $F(A)$, which we call the *group of G at x*. Note that under the assumption that G is connected (any two vertices can be joined by a path), all the groups of G lie in the same conjugacy class of subgroups of $F(A)$.

Let $\equiv$ be an equivalence relation on the vertices of a digraph G. Then $\equiv$ can also be seen as an equivalence relation on $\mathcal{E}(G)$,

$$(x, a, y) \equiv (x', b, y') \iff x \equiv x', a = b, y \equiv y'.$$

The *quotient digraph* $G/\equiv$ is then defined by:

$$\mathcal{V}(G/\equiv) = \mathcal{V}(G)/\equiv, \quad \mathcal{E}(G/\equiv) = \mathcal{E}(G)/\equiv,$$

together with the following adjancency and labeling maps:

$$\alpha(x/\equiv) = \alpha(x)/\equiv, \quad \omega(x/\equiv) = \omega(x)/\equiv, \quad \lambda(x/\equiv) = \lambda(x).$$

The definition of $G/\equiv$ can be summarized by the following commutative diagrams:

$$\begin{array}{ccc} A & \xleftarrow{\lambda} & \mathcal{E}(G) \\ & \nwarrow^{\lambda} & \downarrow{\scriptstyle \equiv} \\ & & \mathcal{E}(G/\equiv) \end{array} \qquad \begin{array}{ccc} \mathcal{E}(G) & \xrightarrow{\alpha} & \mathcal{V}(G) \\ \downarrow{\scriptstyle \equiv} & & \downarrow{\scriptstyle \equiv} \\ \mathcal{E}(G/\equiv) & \xrightarrow{\alpha} & \mathcal{V}(G/\equiv) \end{array} \qquad \begin{array}{ccc} \mathcal{E}(G) & \xrightarrow{\omega} & \mathcal{V}(G) \\ \downarrow{\scriptstyle \equiv} & & \downarrow{\scriptstyle \equiv} \\ \mathcal{E}(G/\equiv) & \xrightarrow{\omega} & \mathcal{V}(G/\equiv) \end{array}$$

The natural projection $G \to G/\equiv$ is a *digraph morphism*, meaning that it preserves the maps α, ω, λ. If, conversely, $\phi\colon G \to H$ is a digraph morphism, then the quotient $G/\ker(\phi)$ is isomorphic to $\mathrm{Im}(\phi)$, where $\ker(\phi) = \{(x, y) : \phi(x) = \phi(y)\}$. Note that for a digraph morphism $\phi\colon G \to G'$ to be onto, it needs to be onto on both $\mathcal{V}(G')$ and $\mathcal{E}(G')$. The latter condition can be written as follows:

$$\forall (x, a, y) \in \mathcal{E}(G'), \exists (x', a, y') \in \mathcal{E}(G),\ \phi(x') = x \wedge \phi(y') = y.$$

We say that an equivalence relation $\equiv$ on $\mathcal{V}(G)$ is *group-preserving* if the group of G at x is equal to the group of $G/\equiv$ at $x/\equiv$, for all $x \in \mathcal{V}(G)$. We also call *group-preserving* a digraph morphism whose kernel is a group-preserving relation. Note that the group of G at x is always a subgroup of the group of $G/\equiv$ at $x/\equiv$. Therefore, to prove that $\equiv$ is group-preserving, one only needs to prove the reverse inclusion. Moreover, in the case of a connected digraph, this inclusion needs only to be checked on a single vertex.

The family of group-preserving equivalence relations of a digraph G also has the property of being closed under taking subrelations. Indeed, let us suppose that $\equiv_1$ is group-preserving and consider $\equiv_2 \subseteq \equiv_1$. Then, the canonical surjection of $\equiv_1$ factors through that of $\equiv_2$, giving us the following commutative diagram:

$$\begin{array}{ccc} G & \xrightarrow{\equiv_1} & G/{\equiv_1} \\ {\scriptstyle \equiv_2}\downarrow & \nearrow & \\ G/{\equiv_2} & & \end{array}$$

Let us fix $x \in \mathcal{V}(G)$ and let H, H_1, H_2 be respectively the group of G at x; the group of $G/{\equiv_1}$ at $x/{\equiv_1}$; and the group of $G/{\equiv_2}$ at $x/{\equiv_2}$. Then the diagram above implies $H \leq H_2 \leq H_1$, while the fact that $\equiv_1$ is group-preserving implies $H = H_1$. Thus, $H_2 = H$ and $\equiv_2$ is also group-preserving.

A well-known algorithm due to Stallings implies that a digraph always has a greatest group-preserving equivalence relation. We now proceed to give a description of this equivalence relation, starting with the following definition.

Definition 3.1. The *Stallings equivalence* of G is the least equivalence relation on $\mathcal{V}(G)$ closed under the two following rules:

(F) If (x, y), (u, x'), (y', v) are related, and (x, a, x'), (y, a, y') are edges in G, then (u, v) are related.

(F') If (x, y), (u, x'), (y', v) are related, and (x', b, x), (y', b, y) are edges in G, then (u, v) are related.

We denote the Stallings equivalence by $\equiv_S$.

Note that if two equivalence relations are closed under either rule (F) or (F'), then so is their intersection (this follows immediately from the definitions). Moreover, the total relation $\mathcal{V}(G) \times \mathcal{V}(G)$ is trivially closed under the two rules. Hence, the relation $\equiv_S$ is simply the intersection of all equivalence relations on $\mathcal{V}(G)$ that are closed under (F) and (F').

By a *trivially-labeled path*, we mean a path whose label is the identity element of $F(A)$. The next result relates Stallings equivalence with trivially-labeled paths, and can be seen as a reformulation of Stallings algorithm.

Proposition 3.2. *Let G be a connected digraph. The equivalence $\equiv_S$ is, alternatively,*

(1) *the equivalence relation induced by trivially-labeled paths;*

$\mathbf{E}(0010)$ $\mathbf{E}(00100)$ $\mathbf{E}(00100010)$ $\mathbf{E}(001000100010)$

FIGURE 7. Extension graphs of all bispecial factors of L starting with 001.

All in all, we exhausted all cases and found four bispecial factors:

$$0010, 00100, 00100010, 001000100010.$$

This proves the claim. ∎

We give the extension graphs of these four bispecial factors in Figure 7. The longest among these, which has length 12, is the only one which is disconnected. We also saw that all the bispecial factors of L of length at most 4 are connected, and it is not hard from there to complete the picture and show that 001000100010 is both the *longest bispecial factor starting with 001* and the *smallest disconnected factor* of L. This can be done by explicit computations for the only three missing bispecial factors of length at most 12, which are 00010, 000100 and 000100010.

Step 3. Next, we give conditions ensuring stability of some extension graphs under $x \mapsto \varphi(x)0$ or $x \mapsto \varphi(x)00$.

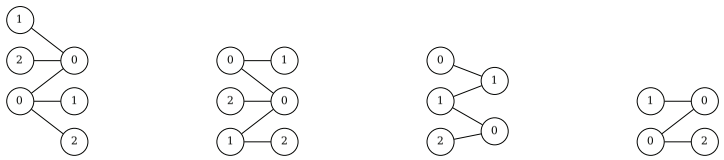
Claim. Let x be a bispecial factor of L starting with 000. Then:

$$\mathbf{E}(x) \cong \begin{cases} \mathbf{E}(\varphi(x)0) & \text{if } x \in A^*00 \\ \mathbf{E}(\varphi(x)00) & \text{if } x \in A^*10. \end{cases}$$

Proof of the claim: Let us put:

$$y = \begin{cases} \varphi(x)0 & \text{if } x \in A^*00 \\ \varphi(x)00 & \text{if } x \in A^*10. \end{cases}$$

Let σ be the permutation of $A = \{0, 1, 2\}$ fixing 0 and exchanging 1, 2. We readily deduce from Step 1 that $\sigma(a)$ is a suffix of $\varphi(a)$, for all $a \in \mathbf{L}(x)$. Similarly, if $x \in A^*00$, then $0\sigma(b)$ is a prefix of $\varphi(b)$, for all $b \in \mathbf{R}(x)$; and



$\mathbf{E}(\varepsilon)$ $\mathbf{E}(0)$ $\mathbf{E}(00)$ $\mathbf{E}(0010)$

FIGURE 6. Extension graphs of all the bispecial words of length at most 4 in L.

and $\mathbf{L}(y) \supseteq \{0, 2\}$. By Step 1, it follows that $|y| < 3$, which leaves us with only thirteen possibilities. Further accounting for the fact that $2 \in \mathbf{L}(y)$ and $\varphi(y) \neq \varepsilon$, we narrow it down to only two possibilities, namely $y = 0$ and $y = 00$. Trying out both values, we obtain either:

$$u = 001\varphi(0)0 = 00100010; \text{ or}$$
$$u = 001\varphi(00)0 = 001000100010.$$

A direct computation shows that both of those words are bispecial.

Finally, we treat the case $u = 001x00$. By Step 1, $u0 = 001x000 \in L$. Since $0000 \notin L$, it follows that x cannot end with 0. Moreover, we also have $0u, 1u, u1 \in L$, so there exist $z_1, z_2, z_3 \in L$ such that:

$$\varphi(z_1) = s_1 0001x00t_1, \quad \varphi(z_2) = s_2 1001x00t_2, \quad \varphi(z_3) = s_3 001x001t_3.$$

Again, since $0000 \notin L$, s_1 cannot end with 0. Recalling that cutting points are located exactly after the occurrences of 1 or 2, we apply Lemma 9.1 to conclude that there exist: a factor of z_1 of the form $0y_1$ such that $\varphi(y_1) = x$; a factor of z_2 of the form $2y_2$ such that $\varphi(y_2) = x$; and a factor of z_3 of the form $y_3 2$ such that $\varphi(y_3) = x$. Since φ is injective, $y_1 = y_2 = y_3 = y$, and $\mathbf{L}(y) \supseteq \{0, 2\}$, $2 \in \mathbf{R}(y)$. By Step 1, we conclude that $|y| < 3$, which again leaves us with thirteen possible values for y. Accounting for the fact that $2 \in \mathbf{R}(y)$ and $2 \in \mathbf{L}(y)$ narrows this to only two possibilities: $y = \varepsilon$ and $y = 0$. Testing both possibilities, we find that $y = 0$ does not yield a bispecial factor, leaving us with only one bispecial factor for that case:

$$u = 001\varphi(\varepsilon)00 = 00100.$$

FIGURE 3. The rules defining Stallings equivalence. The arrows represent edges, the thick lines represent existing relations, and the dashed lines represent the relations deduced from each rule.

(2) *the greatest group-preserving equivalence of* G.

For the proof of this result, the following definition will be useful: given an equivalence relation $\equiv$ on $\mathcal{V}(G)$, an $\equiv$*-path* in G is a sequence of edges $p = (e_1, \ldots, e_k)$ satisfying $\alpha(e_{i+1}) \equiv \omega(e_i)$. The notions of *label* and *length* extend in a straightforward way to $\equiv$-paths. We also use the notation $p\colon x \xrightarrow{u} y$ for $\equiv$-paths, to mean $\alpha(p) = x$, $\omega(p) = y$ and $\lambda(p) = u$. Finally, we adopt the convention that an $\equiv$-path of length 0 is a pair $x \equiv y$.

Proof of Proposition 3.2: **(1)** Let us denote by $\sim$ the relation induced by trivially-labeled paths and by $\approx$ the relation induced by trivially-labeled $\equiv_S$-paths. Clearly $\sim$ is contained in $\approx$. Let us show that $\approx$ is contained in $\equiv_S$.

We proceed by induction on the length of the trivially-labeled $\equiv_S$-path. Note that by definition, an $\equiv_S$-path of length 0 is nothing but a pair $x \equiv y$, so the basis of the induction is trivial. Let us suppose that there is a trivially-labeled $\equiv_S$-path $p\colon x \to y$ of length $k \geq 1$. Write $p = (e_1, \ldots, e_k)$. Since p is trivially-labeled, k is even and there must exist i such that $\lambda(e_i) = a^{-1}$ and $\lambda(e_{i+1}) = a$, where a is either a letter, or the inverse of a letter. Write $e_i = (u', a^{-1}, u)$ and $e_{i+1} = (v, a, v')$, where $u \equiv_S v$. If $a \in A$, then we may use rule (F) to conclude $u' \equiv_S v'$. Otherwise, one uses rule (F') to obtain the same conclusion. It follows that $p' = (e_1, \ldots, e_{i-1}, e_{i+2}, \ldots, e_k)$ is also a trivially-labeled $\equiv_S$-path between x and y. Since p' has length $k - 2 < k$, we conclude by induction that $x \equiv_S y$.

We finish the proof of (1) by showing that $\equiv_S$ is contained in $\sim$. By definition of $\equiv_S$, it suffices to show that $\sim$ is closed under the rules (F) and (F'). Suppose that $u \sim x'$, $x \sim y$, $y' \sim v$, and that there are two edges $e = (x, a, x')$ and $f = (y, a, y')$. Consider trivially-labeled paths $p_1\colon u \to x'$, $q\colon x \to y$ and $p_2\colon y' \to v$. Then, the composition $p_1 e^{-1} q f p_2$ is a trivially-labeled path in G

between u and v. Thus, $u \sim v$, which proves $\sim$ is closed under (F). The proof for (F') is similar.

(2) We first show that $\equiv_S$ is a group-preserving equivalence relation, and then we show it is the greatest. Let us fix any path $p\colon x/\equiv_S \to y/\equiv_S$ in the quotient $G/\equiv_S$. We say that an $\equiv_S$-path q in G *lifts* p if $q\colon x' \to y'$ with $x \equiv_S x'$, $y \equiv_S y'$ and $\lambda(p) = \lambda(q)$. Note that any path in the quotient $G/\equiv_S$ admits such a lift in G. If $q = (e_0, \dots, e_k)$ lifts p, we put $D(q) = \{0 \leq i < k : \omega(e_i) \neq \alpha(e_{i+1})\}$. Clearly, q is a path if and only if $D(q)$ is empty. Assume $j = \max(D(q))$, and consider a trivially-labeled path $r\colon \omega(e_j) \to \alpha(e_{j+1})$ in G, which we know exists by Part (1). Let $q = q_1q_2$ be the factorization of q where $|q_1| = j + 1$. Then, $q' = q_1rq_2$ is an $\equiv_S$-path between x' and y' satisfying $\lambda(q') = \lambda(q)$ and $D(q') = D(q) \setminus \{j\}$. Thus, we may assume that q is a path. Composing on both ends with trivially-labeled paths $x \to x'$ and $y' \to y$, we get a lift of p which is a path between x and y in G. This result applied to loops shows that $\equiv_S$ is group-preserving.

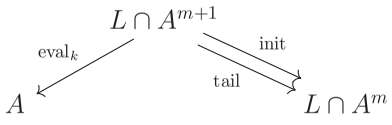
Finally, let us suppose that $\equiv$ is another group-preserving congruence, and let $x \equiv y$. Choose any path $p\colon x \to y$. Then $p/\equiv$ is a loop over $y/\equiv$ in $G/\equiv$. Since $\equiv$ is group-preserving, there is a loop q over y with $\lambda(q) = \lambda(p/\equiv) = \lambda(p)$. It follows that pq^{-1} is a trivially-labeled path between x and y, so $x \equiv_S y$. ∎

From now on, we will use the three equivalent descriptions of $\equiv_S$ interchangeably.

4. Rauzy graphs

Recall that we defined the two maps init and tail by $\mathrm{init}(x) = x[0 : |x| - 1]$ and $\mathrm{tail}(x) = x[1 : |x|]$. For $k \in \mathbb{N}$, let us also define the map eval_k by $\mathrm{eval}_k(x) = x(k)$. Note that init and tail are defined on A^+, while eval_k is defined on $A^{>k}$.

Definition 4.1. Let L be a recurrent language on A and $m, k \in \mathbb{N}$ with $k \leq m$. The *k-labeled Rauzy graph* of level m of L is the digraph $G_{m,k}$ defined by the diagram:



Special cases of these labeled Rauzy graphs have appeared in the litterature, including in [4] with $k = m$, and in [1] with $m = 2k$.

Step 2. We now know that all long enough bispecial factors must start with either 000 or 001, and end with either 00 or 10. We restrict the possibilities even further by proving the following claim:

Claim. The only four bispecial factors of L starting with 001 are:

$$0010, \quad 00100, \quad 00100010, \quad 001000100010.$$

The proof of this claim makes use of the concept of *cutting points*, by which we mean the following: in a word of the form $\varphi(z)$, a cutting point is an index $0 \leq j \leq |\varphi(z)| - 1$ such that $j = |\varphi(z_1)|$, for some prefix z_1 of z. We observe that in the specific case of φ, the cutting points are located exactly after the occurrences of the letters 1 and 2. The following elementary lemma will be useful.

Lemma 9.1. *Let $\varphi(z) = u_1 \dots u_n$ be a factorization such that $|u_1 \dots u_k|$ is a cutting point for all $1 \leq k < n$. Then there is a factorization $z = z_1 \dots z_n$ such that $\varphi(z_i) = u_i$ for all $1 \leq i \leq n$.*

Proof: By assumption, $u_1 \dots u_{j-1}, u_1 \dots u_j \in \varphi(A^*)$ for all $1 \leq j \leq n$. Since $\varphi(A)$ is a prefix code, we have $u_j \in \varphi(A^*)$ and we may write $u_j = \varphi(z_j)$ for some z_j. Then $\varphi(z) = \varphi(z_1 \dots z_n)$ and as φ is injective, $z = z_1 \dots z_n$ as required. ∎

Proof of the claim: Let us start by noting that the only bispecial factors of L with length at most 4 are:

$$\varepsilon, \quad 0, \quad 00, \quad 0010.$$

This can be proven simply by inspecting the sets $L \cap A^k$ for $2 \leq k \leq 6$. The extension graphs of these four words can be found in Figure 6. From now on, we work only with bispecial factors of length at least 5.

Let us suppose that u is bispecial, $|u| \geq 5$ and 001 is a prefix of u. We distinguish two cases: $u = 001x10$ and $u = 001x00$.

We start by the case $u = 001x10$. Let $x' = x1$. By Step 1, we know that $0u, 1u \in L$. Thus, there exist $z_1, z_2 \in L$ such that:

$$\varphi(z_1) = s_1 0001 x' 0 t_1, \qquad \varphi(z_2) = s_2 1001 x' 0 t_2.$$

Since $0000 \notin L$, it follows that s_1 ends with either 1 or 2. Therefore $|s_1|$ is a cutting point in $\varphi(z_1)$. Similarly, there is a cutting point in $\varphi(z_1)$ at the end of x'. It follows from Lemma 9.1 that z_1 has a factor of the form $0y_1$ such that $\varphi(y_1) = x'$. With similar arguments, we conclude that z_2 has a factor of the form $2y_2$ such that $\varphi(y_2) = x'$. Since φ is injective, we find $y_1 = y_2 = y$,

Some of these steps involve the computation of the sets $L \cap A^k$ for several values of k, some of them quite large. We will omit the details of these computations and provide only the results. These computations can be checked either by hand (e.g. with the algorithm described in [2, Section 3.2]), or perhaps more appropriately using SageMath [6]. At the time of writing, a SageMath web interface can be accessed at the address `https://sagecell.sagemath.org`. To compute the set $L \cap A^k$, simply evaluate the following line of code in the web interface:

`WordMorphism({0:[0,0,0,1],1:[0,2],2:[0,0,1]}).language(`k`)`.

Step 1. We prove the following claim.

Claim. Let x be a right special factor of L of length at least 2. Then one of the two following alternatives hold:

(1) x ends with 00 and $\mathbf{R}(x) = \{0, 1\}$.
(2) x ends with 10 and $\mathbf{R}(x) = \{0, 2\}$.

Dually, let y be a left special factor of L of length at least 3. Then one of the two following alternatives hold:

(1) y starts with 000 and $\mathbf{L}(y) = \{1, 2\}$.
(2) y starts with 001 and $\mathbf{L}(y) = \{0, 1\}$.

Proof of the claim: Direct computations reveal that:

$$L \cap A^3 = \{000, 001, 010, 020, 100, 102, 200\}.$$

Hence, the only two right special factors in $L \cap A^2$ are 00 and 10, and they satisfy respectively:

$$\mathbf{R}(00) = \{0, 1\}, \qquad \mathbf{R}(10) = \{0, 2\}.$$

Since the sets $\mathbf{R}(x)$ are weakly increasing under taking suffixes, the first part follows.

Similarly, we find:

$$L \cap A^4 = \{0001, 0010, 0100, 0102, 0200, 1000, 1001, 1020, 2000\}.$$

Therefore $L \cap A^3$ contains only two left special factors, 000 and 001, satisfying respecitvely:

$$\mathbf{L}(000) = \{1, 2\}, \qquad \mathbf{L}(001) = \{0, 1\}.$$

Since the sets $\mathbf{L}(x)$ are weakly increasing under taking prefixes, the second part follows as well. ∎

The maps init, tail and eval_k used to define the Rauzy graphs are jointly injective, and moreover the following diagrams commute:

$$\begin{array}{ccc} A^{\geq 2} & \xrightarrow{\text{init}} & A^+ \\ \downarrow{\scriptstyle \text{tail}} & & \downarrow{\scriptstyle \text{tail}} \\ A^+ & \xrightarrow{\text{init}} & A^* \end{array} \qquad \begin{array}{ccc} A^{>k+1} & \xrightarrow{\text{init}} & A^{>k} \\ & {\scriptstyle \text{eval}_k}\searrow & \downarrow{\scriptstyle \text{eval}_k} \\ & & A \end{array} \qquad \begin{array}{ccc} A^{>k+1} & \xrightarrow{\text{tail}} & A^{>k} \\ & {\scriptstyle \text{eval}_{k+1}}\searrow & \downarrow{\scriptstyle \text{eval}_k} \\ & & A \end{array}$$

Therefore, init and tail also define onto digraph morphisms for $m \geq 1$:

$$\begin{aligned} &\text{init} \colon G_{m,k} \to G_{m-1,k} && (0 \leq k \leq m-1) \\ &\text{tail} \colon G_{m,k} \to G_{m-1,k-1} && (1 \leq k \leq m). \end{aligned}$$

These morphisms will allow us to relate the groups defined the Rauzy graphs. In the next definition, we introduce a convenient notation for these groups.

Definition 4.2. Let (u, v) be such that $uv \in L$, $|u| = k$ and $|u| + |v| = m$. We denote by $H_{u,v}$ the group of $G_{m,k}$ at uv. We call $H_{u,v}$ a *Rauzy group* of L.

The fact that tail and init define digraph morphisms immediately implies that:

$$H_{u,v} \leq H_{\text{tail}(u),v}, \quad H_{u,v} \leq H_{u,\text{init}(v)}.$$

We further note that $H_{ua,v} = a^{-1} H_{u,av} a$. Since we are assuming that L is recurrent, the Rauzy graphs are connected and it follows that $H_{u,v}$ and $H_{u',v'}$ lie in the same conjugacy class whenever $|uv| = |u'v'|$.

5. Paths in suffix extension graphs

For this section, it is useful to introduce a local version of suffix-connectedness. We do this in the next definition.

Definition 5.1. Let $m, e \in \mathbb{N}$ with $1 \leq e \leq m + 1$. We say that L is (m, e)-*suffix-connected* if for every $w \in L \cap A^m$, there exists $1 \leq d \leq e$ such that the natural embedding of $\mathbf{L}(w)$ in $\mathbf{E}_{d,d}(\text{tail}^{d-1}(w))$ lies in a single connected component.

Remark 5.2. This local version of suffix-connectedness has the following feature: suppose that $1 \leq e \leq e' \leq m + 1$ and that L is (m, e)-suffix-connected; then L is also (m, e')-suffix-connected. In particular, if we suppose that L is suffix-connected, then it must be $(m, m + 1)$-suffix-connected for all $m \geq 1$.

The main result of this section is the following proposition, which is the main ingredient in the proof of Theorem 1.1:

Proposition 5.3. *Assume that L is recurrent and $(m-1,e)$-suffix-connected, where $m \geq 1$ and $1 \leq e \leq m$. Then,* $\ker(\mathrm{tail})$ *is a group-preserving equivalence relation of $G_{m,k}$ whenever $e \leq k \leq m$.*

The proof relies on the following lemma:

Lemma 5.4. *Let L be a recurrent language on A, $m \in \mathbb{N}$, $0 \leq k \leq m$, $d \geq 1$ and $x \in L \cap A^{m+d}$. Then there exists a path p_x in $G_{m,k}$ such that:*

$$p_x \colon \mathrm{init}^d(x) \xrightarrow{x[k:k+d]} \mathrm{tail}^d(x)$$

Proof: We proceed by induction on d. If $d = 1$, then x itself is an edge in $G_{m,k}$ providing the required path.

For the induction step, we assume that $d > 1$. Let $x' = \mathrm{init}(x)$ and $x'' = \mathrm{tail}^{d-1}(x)$. Note that $|x''| = |x| - d + 1 = m + 1$, so x'' is an edge in $G_{m,k}$, which we see as a path of length 1. Moreover, the induction hypothesis gives us a path p' such that

$$p' \colon \mathrm{init}^{d-1}(x') \xrightarrow{x'[k:k+d-1]} \mathrm{tail}^{d-1}(x')$$

Recalling that init and tail commute, we find that:

$$\mathrm{tail}^{d-1}(x') = \mathrm{tail}^{d-1} \circ \mathrm{init}(x) = \mathrm{init} \circ \mathrm{tail}^{d-1}(x) = \mathrm{init}(x'').$$

Hence, p' and x'' are consecutive, and we may form the composition $p = p'x''$. Note that p is a path between $\mathrm{init}^d(x)$ and $\mathrm{tail}^d(x)$, as required. Moreover, the label of this path is given by:

$$x'[k : k+d-1]x''(k) = x[k : k+d-1]x(k+d-1) = x[k : k+d],$$

and this concludes the proof. ∎

We are now ready to prove the proposition above.

Proof of Proposition 5.3: Let us fix a pair of vertices identified by the digraph morphism $\mathrm{tail} : G_{m,k} \to G_{m-1,k-1}$, that is to say two words $ax, bx \in L \cap A^m$ where $x \in L \cap A^{m-1}$ and $a, b \in A$. We want to show $ax \equiv_S bx$, which amounts to find a trivially-labeled path in $G_{m,k}$ between ax and bx.

By assumption, there exists $d \leq e$ such that the natural embedding of $\mathbf{L}(x)$ in $\mathbf{E}_{d,d}(\mathrm{tail}^{d-1}(x))$ lies in one connected component. Let us write $y = \mathrm{tail}^{d-1}(x)$, and let u, v be the natural embeddings of $a, b \in \mathbf{L}(x)$ inside $\mathbf{E}_{d,d}(y)$. In other words, u and v satisfy $ax = uy$ and $bx = vy$. Let us consider a path in $\mathbf{E}_{d,d}(y)$

$\mathrm{rank}(K_{u,v})$ is equal to the right hand side. Since (2) holds exactly when $\mathrm{Card}(\mathcal{R}_{u,v}) = \mathrm{rank}(K_{u,v})$ for all such u, v, the result follows. ∎

9. Suffix-connected example

This section is devoted to the proof of Theorem 1.4. We consider the following substitution on the alphabet $A = \{0, 1, 2\}$:

$$\varphi \colon \begin{array}{rcl} 0 & \mapsto & 0001 \\ 1 & \mapsto & 02 \\ 2 & \mapsto & 001 \end{array} .$$

Note that φ is primitive (since for every $a, b \in A$, a occurs in $\varphi^3(b)$), and that $\varphi(A)$ is a *prefix code* (no word in $\varphi(A)$ is a proper prefix of another). In particular, this implies that φ is injective, a fact that will be used several times. We recall that the *language* defined by φ is the subset of all words $w \in A^+$ such that w is a factor of $\varphi^n(a)$ for some $a \in A$ and $n \in \mathbb{N}$. For the current section, L denotes the language of φ. As we already mentioned, it is well known that the language of a primitive substitution is uniformly recurrent. We will show that L is suffix-connected, and deduce that all the return sets of L generate the full free group $F(A)$.

The proof, being a bit lengthy, is organized in 5 steps. Let us give a quick outline of each step:

(1) We show that every right special factor of length at least 2 either ends with 00 and satisfies $\mathbf{R}(x) = \{0, 1\}$; or ends with 10 and satisfies $\mathbf{R}(x) = \{0, 2\}$. Similarly, we show that every left special factor of length at least 3 either starts with 000 and satisfies $\mathbf{L}(x) = \{1, 2\}$; or starts with 001 and satisfies $\mathbf{L}(x) = \{0, 1\}$.

(2) We show that L contains only 4 bispecial factors starting with 001 and we compute them.

(3) We show that if x is a bispecial factor that starts with 000, then

$$\mathbf{E}(x) \cong \begin{cases} \mathbf{E}(\varphi(x)0) & \text{if } x \text{ ends with } 00, \\ \mathbf{E}(\varphi(x)00) & \text{if } x \text{ ends with } 10. \end{cases}$$

(4) We define inductively a sequence of words $(w_k)_{k\in\mathbb{N}}$ of increasing lengths, and we show that the disconnected elements of L are precisely the members of that sequence.

(5) We define a sequence of integers $(d_k)_{k\in\mathbb{N}}$ such that $\mathbf{L}(w_k)$ embeds in one connected component of $\mathbf{E}_{d_k,d_k}(\mathrm{tail}^{d_k-1}(w_k))$.

Before proving Corollary 1.3, we need some preliminary material. A word $w \in L$ is called *neutral* if:

$$1 - \chi(\mathbf{E}(w)) = 0,$$

where $\chi(\mathbf{E}(w))$, the characteristic of $\mathbf{E}(w)$, is the difference between the number of vertices and edges in $\mathbf{E}(w)$. A *neutral language* is a language in which all non-empty words are neutral. The next result, quoted from [7, Corollary 5.4], will be useful to prove Corollary 1.3.

Lemma 8.1. *If L is recurrent and neutral, then for all u, v with $uv \in L$,*

$$\mathrm{Card}(\mathcal{R}_{u,v}) = \mathrm{Card}(A) - \chi(\mathbf{E}(\varepsilon)) + 1.$$

With this lemma in mind, let us recall the statement of Corollary 1.3: if L is uniformly recurrent, connected and neutral, then the following statements are equivalent:

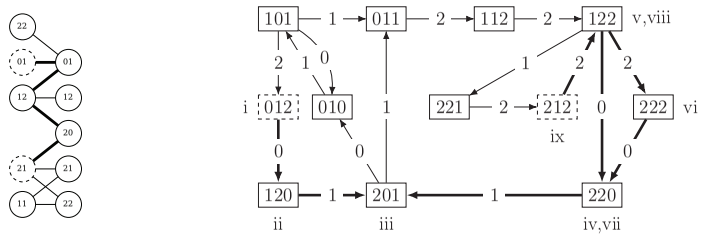
(1) Some return set of L is a free subset of the free group $F(A)$.
(2) All return sets of L are free subsets of the free group $F(A)$.
(3) L is a tree set.

Proof of Corollary 1.3: We first recall the following fact, which is a straightforward consequence of the well-known Hopfian property of $F(A)$: a finite subset of $F(A)$ is free if and only its cardinality agrees with the rank of the subgroup it generates. Moreover, note that under our assumptions, the return sets of L all have the same cardinality (by Lemma 8.1), as well as the same rank (by Theorem 1.1). Therefore if *some* return set is free, then *all* return sets are free, that is to say (1) and (2) are equivalent.

To prove the equivalence of (2) and (3), we use the following fact about graphs: a simple graph G is a forest if and only if it has exactly $\chi(G)$ connected components [5, Exercise 2.1.7 (b)]. On the one hand, this implies that in a connected neutral language, the extension graph of any non-empty word must be a tree (since neutrality implies $\chi(\mathbf{E}(w)) = 1$). Thus, a neutral connected language is a tree set if and only if $\mathbf{E}(\varepsilon)$ is a forest, if and only if $\chi(\mathbf{E}(\varepsilon)) = c$, where c denotes the number of connected components of $\mathbf{E}(\varepsilon)$. This is also equivalent to the following equality:

$$\mathrm{Card}(A) - \chi(\mathbf{E}(\varepsilon)) + 1 = \mathrm{Card}(A) - c + 1.$$

Let us fix u, v with $uv \in L$. Lemma 8.1 implies that $\mathrm{Card}(\mathcal{R}_{u,v})$ is equal to the left-hand side of the previous equation, while Theorem 1.1 implies that



$\mathbf{E}_{2,2}(\mathrm{tail}(12))$ $G_{3,3}$

FIGURE 4. A trivially-labeled path between a pair of words in $\ker(\mathrm{tail})$ induced by a path in a depth 2 suffix extension graph. The Roman numerals indicate the order in which the vertices are visited in the Rauzy graph. This takes place in the language defined by the primitive substitution $0 \mapsto 12, 1 \mapsto 2, 2 \mapsto 01$.

joining u and v. Since $\mathbf{E}_{d,d}(y)$ is bipartite, this path must have the following form:

$$u = s_0, t_0, s_1, t_1, \ldots, t_{j-1}, s_j = v,$$

where $s_i \in \mathbf{L}_d(y)$, $t_i \in \mathbf{R}_d(y)$. The fact that this forms a path in $\mathbf{E}_{d,d}(y)$ means that, for each $0 \leq i < j$, we have:

$$s_i y t_i, s_{i+1} y t_i \in L.$$

Let us put $w_i = s_i y t_i$ and $z_i = s_{i+1} y t_i$. By Lemma 5.4, there exist paths:

$$p_i \colon \mathrm{init}^d(w_i) \xrightarrow{w_i[k:k+d]} \mathrm{tail}^d(w_i) \qquad q_i \colon \mathrm{init}^d(z_i) \xrightarrow{z_i[k:k+d]} \mathrm{tail}^d(z_i)$$

We notice that $\mathrm{init}^d(w_i) = s_i y$, $\mathrm{tail}^d(w_i) = y t_i = \mathrm{tail}^d(z_i)$, $\mathrm{init}^d(z_i) = s_{i+1} y$. Therefore, p_i, q_i^{-1} are consecutive and their composition is a path $s_i y \to s_{i+1} y$. Moreover, since $k \geq e \geq d$, it follows that

$$w_i[k : k+d] = (y t_i)[k-d : k] = z_i[k : k+d].$$

Therefore, $p_i q_i^{-1}$ is trivially-labeled. Composing these paths for $i = 0, \ldots, j-1$ gives us a trivially-labeled path between $ax = uy = s_0 y$ and $s_j y = vy = bx$. ∎

By combining Proposition 5.3 with Remark 5.2, it then follows that for each $m \geq 1$, the map tail defines a group-preserving morphism:

$$\mathrm{tail} \colon G_{m,m} \to G_{m-1,m-1}.$$

But clearly, the class of all group-preserving morphisms is closed under composition. Therefore, in a suffix-connected language, the following is a group-preserving morphism for all $m \geq 1$:

$$\text{tail}^{m-1} \colon G_{m,m} \to G_{1,1}.$$

We immediately deduce the following:

Corollary 5.5. *Let L be a suffix-connected recurrent language and $u \in L$ with $u \neq \varepsilon$. Then $H_{u,\varepsilon} = H_{b,\varepsilon}$, where b is the last letter of u.*

Let us highlight another particular case of this result. The condition of being $(m,1)$-suffix-connected is equivalent to being m-connected, meaning that $\mathbf{E}(w)$ is connected for all $w \in L \cap A^m$, which in turn is equivalent to the dual condition of being $(m,1)$-prefix-connected. Combining Proposition 5.3 with its dual for the special case $e = 1$, we obtain the following result, which is reminescent of [4, Proposition 4.2]:

Corollary 5.6. *If L is a $(m-1)$-connected recurrent language, where $m \geq 1$, then:*

1. *For $0 \leq k \leq m-1$,* ker(init) *is a group-preserving equivalence relation of $G_{m,k}$;*
2. *For $1 \leq k \leq m$,* ker(tail) *is a group-preserving equivalence relation of $G_{m,k}$.*

6. Return sets

Let us recall that the *return set to (u,v) in L* is the set of all words $r \in L$ such that $urv \in L$, urv starts and ends with uv, and contains exactly two occurrences of uv. We denote this set by $\mathcal{R}_{u,v}$. For basic properties of return sets, see [9].

Definition 6.1. Let (u,v) be such that $uv \in L$. The subgroup of $F(A)$ generated by $\mathcal{R}_{u,v}$ is denoted by $K_{u,v}$. We call this a *return group* of L.

Our main result for this section relates the return groups with the Rauzy groups.

Proposition 6.2. *Let L be a recurrent language and let u, v be such that $uv \in L$.*

1. *$K_{u,v} \leq H_{u,v}$.*
2. *If $\mathcal{R}_{u,v}$ is finite and s is one of its longest elements, then $H_{u,sv} \leq K_{u,v}$.*

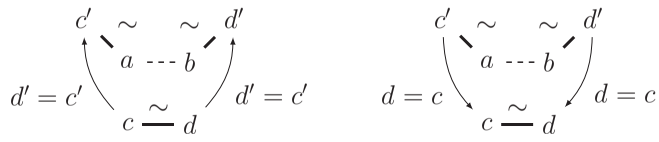


FIGURE 5. The rules (F) and (F') as they appear in the proof of Lemma 7.2 when showing that $\equiv_S \subseteq \sim$.

It only remains to show that $G_{1,1}/\equiv_S$ has n edges. We do this by showing that the labeling map $\lambda \colon G_{1,1}/\equiv_S \to A$ is a bijection. Fix a letter $a \in A$. Since L is recurrent, there exists $b \in A$ with $ba \in L$. Hence, there is at least one edge labeled a in $G_{1,1}$, and therefore also in $G_{1,1}/\equiv_S$. Hence, $\lambda \colon G_{1,1}/\equiv_S \to A$ is surjective. Now suppose that $G_{1,1}$ has two edges e, f labeled a. As noted before, $\lambda = \omega$ in $G_{1,1}$, so $\omega(e) = a = \omega(f)$. Applying rule (F'), we conclude that $\alpha(e) \equiv_S \alpha(f)$. In particular, $e/\equiv_S = f/\equiv_S$, which proves that the labeling map $\lambda \colon G_{1,1}/\equiv_S \to A$ is injective. ∎

8. Proof of the corollaries

Let us start this section by recalling the statement of Corollary 1.2: if L is suffix-connected and uniformly recurrent, then the following statements are equivalent:

1. All the return sets of L generate the full free group $F(A)$.
2. Some return set of L generates a group of rank $\text{Card}(A)$.
3. The extension graph of the empty word is connected.

Proof of Corollary 1.2: **(1) implies (2).** Trivial.

(2) implies (3). By Theorem 1.1, all return groups of L have rank $n - c + 1$ where c is the number of connected components of $\mathbf{E}(\varepsilon)$ and $n = \text{Card}(A)$. Under the assumption (2), we therefore have $n = n - c + 1$ and $c = 1$.

(3) implies (1). If $\mathbf{E}(\varepsilon)$ is connected, then by Corollary 5.6, there is a group-preserving morphism $G_{1,1} \to G_{0,0}$. But note that $G_{0,0}$ has a single vertex with loops labeled by the letters of A. Thus, the group generated by $G_{0,0}$ is equal to the full free group $F(A)$, and so are all the groups of the level 1 Rauzy graph $G_{1,1}$. But recall that, for a suffix-connected language, all the return groups lie in the conjugacy class generated by the level 1 Rauzy groups (see Lemma 7.1), and so the result follows. ∎

Lemma 7.2. *Let L be a recurrent language. Then, the groups of the Rauzy graph $G_{1,1}$ have rank $n - c + 1$, where $n = \mathrm{Card}(A)$ and c is the number of connected components of $\mathbf{E}(\varepsilon)$.*

Proof: A well-known consequence of Stallings algorithm is that the rank of any group generated by a connected digraph G is

$$\mathrm{Card}(\mathcal{E}(G/\equiv_S)) - \mathrm{Card}(\mathcal{V}(G/\equiv_S)) + 1$$

(see [11, Lemma 8.2]). Thus, we need only to show that the quotient $G_{1,1}/\equiv_S$ has c vertices and n edges.

Let us start by showing that $G_{1,1}/\equiv_S$ has c vertices. By definition, we have $\mathcal{V}(G_{1,1}) = A = \mathbf{L}(\varepsilon)$. Let $\sim$ be the relation defined as follow: for $a, b \in A$, we have $a \sim b$ exactly when, viewed as elements of $\mathbf{L}(\varepsilon)$, a and b lie in the same connected component of $\mathbf{E}(\varepsilon)$. Note that the relation $\sim$ has precisely c classes because every connected component of $\mathbf{E}(\varepsilon)$ contains at least one vertex in $\mathbf{L}(\varepsilon)$. Therefore, it suffices to show that $\sim = \equiv_S$.

We now prove the inclusion $\sim \subseteq \equiv_S$. Since $\mathbf{E}(\varepsilon)$ is bipartite, any path in $\mathbf{E}(\varepsilon)$ between elements of $\mathbf{L}(\varepsilon)$ has even length, and thus it suffices to argue for elements related by paths of length 2. Let us assume that $a, b \in \mathbf{L}(\varepsilon) = A$ are related by a path of length 2 inside $\mathbf{E}(\varepsilon)$. By definition of $\mathbf{E}(\varepsilon)$, the existence of such a path means that there is some $c \in A$ such that $ac, bc \in L$. But recall that $\mathcal{E}(G_{1,1}) = L \cap A^2$, so we may view $e = ac$ and $f = bc$ as edges in $G_{1,1}$, both of which have label c. The path (e, f^{-1}) is then a trivially-labeled path between a and b, so $a \equiv_S b$ as required.

Let us prove the inclusion $\equiv_S \subseteq \sim$. By definition of $\equiv_S$, we only need to show that $\sim$ is closed under the two rules (F) and (F'). We argue for each separately. Let us fix $a, b, c, c', d, d' \in A$ such that $a \sim c'$, $b \sim d'$, $c \sim d$.

(F) We assume that there are edges $e\colon c \to c'$, $f\colon d \to d'$ such that $\lambda(e) = \lambda(f)$. Note that the maps tail and eval_1 agree on A^2, so by definition $\omega = \lambda$ in the Rauzy graph $G_{1,1}$. Thus, under our current assumptions, $c' = d'$. Hence, $a \sim c' \sim b$ and $a \sim b$ by transitivity.

(F') We assume that there are edges $e\colon c' \to c$, $f\colon d' \to d$ such that $\lambda(e) = \lambda(f)$. Since $\omega = \lambda$ in $G_{1,1}$, we deduce that $d = c$. By definition, $\mathcal{E}(G_{1,1}) = L \cap A^2$, $e = c'c \in L$ and $f = d'd = d'c \in L$. In particular, there is an edge in $\mathbf{E}(\varepsilon)$ joining c' and c, and another one joining d' and c. Hence, $a \sim c' \sim d' \sim b$ and $a \sim b$ by transitivity.

The following lemma recalls several properties of Rauzy graphs that will be relevant. By a *positive path*, we mean a path which consists only of edges in $\mathcal{E}(G)$ or, equivalently, which contains no formal inverses.

Lemma 6.3. *Let L be a recurrent language and let u, v be such that $uv \in L$, $|u| = k$ and $|u| + |v| = m$.*

(1) *Any element $w \in L$ is the label of a positive path in $G_{m,k}$.*
(2) *Any label w of a positive path in $G_{m,k}$ of length at most $m + 1$ is in L.*
(3) *If $p\colon x \to uv$ is a positive path in $G_{m,k}$, then $\lambda(p)$ is suffix-comparable with u. Moreover, there is at least one such path satisfying $\lambda(p) = u$.*
(4) *If $q\colon uv \to y$ is a positive path in $G_{m,k}$, then $\lambda(q)$ is prefix-comparable with v. Moreover, there is at least one such path satisfying $\lambda(q) = v$.*

All of these properties follow from the definition of $G_{m,k}$ in a straightforward manner. Parts (1) and (2) are standard and can be found for instance in [4, Section 4.1]. Parts (3) and (4) are analogous to [1, Lemma 4.5].

We are now ready to prove the proposition. Let us mention that Part (2) of the proposition is inspired by the proof of [4, Theorem 4.7], which relied partly on the fact that return sets of the form $\mathcal{R}_{u,\varepsilon}$ are prefix codes. However, this property no longer holds for general return sets and we had to find a way to avoid it. This is essentially what is accomplished by the very last paragraph of the proof.

Proof of Proposition 6.2: **(1)** Let $k = |u|$, $m = |u| + |v|$, and fix $r \in \mathcal{R}_{u,v}$. Then $urv \in L$ is the label of a positive path p in $G_{m,k}$ by Part (1) of Lemma 6.3. Consider the factorization $p = q_1 p' q_2$, where $\lambda(q_1) = u$, $\lambda(p') = r$ and $\lambda(q_2) = v$. Write $\alpha(p') = x_1$ and $\omega(p') = x_2$. Consider the factorization $x_1 = u_1 v_1$, where $|u_1| = k$. Since $\omega(q_1) = x_1$, it follows from Part (3) of Lemma 6.3 that u_1 is suffix-comparable with $\lambda(q_1) = u$. As $|u| = k = |u_1|$, we conclude that $u_1 = u$. Similarly, $\alpha(p'q_2) = x_1$, so Part (4) implies that v_1 is prefix-comparable with $\lambda(p'q_2) = rv$. Since $r \in \mathcal{R}_{u,v}$, the word rv starts with v, and since $|v| = m - k = |v_1|$, we conclude that $v_1 = v$. Thus, $x_1 = uv$. A similar argument yields $x_2 = uv$, so p' is a loop over uv, and $\mathcal{R}_{u,v} \subseteq H_{u,v}$. This proves (1).

(2) Let $m' = |u| + |s| + |v|$, and consider a positive path in $G_{m',k}$ of the form $p\colon uvx \to uvy$. We start by proving the following claim: $w = \lambda(p)$ is a concatenation of elements of $\mathcal{R}_{u,v}$.

To prove this claim, let us first consider two positive paths $q_1\colon x_1 \xrightarrow{u} uvx$ and $q_2\colon uvy \xrightarrow{v} x_2$, whose existence is a consequence of Part (3) and (4) of

Lemma 6.3. Since $\alpha(pq_2) = uvx$, Part (4) of Lemma 6.3 implies that $\lambda(pq_2)$ is prefix-comparable with vx; thus, it starts with v. Similarly, since $\omega(q_1p) = uvy$, Part(3) of Lemma 6.3 implies that $\lambda(q_1p)$ is suffix-comparable with u; thus, it ends with u. In particular, this implies

$$uwv = \lambda(q_1)\lambda(pq_2) = \lambda(q_1p)\lambda(q_2) \in uvA^* \cap A^*uv.$$

We now prove the claim by induction on $|w| = |p|$. If $|w| \leq |s|$, then uwv is the label of the positive path q_1pq_2 in $G_{m',k}$, which has length at most m'. Hence, Part (2) of Lemma 6.3 implies that $uwv \in L$. This, taken together with the fact that uwv belongs to $uvA^* \cap A^*uv$, implies that w is a concatenation of elements of $\mathcal{R}_{u,v}$. This establishes the basis of the induction.

For the inductive step, let us suppose that $|w| > |s|$. Let p' be the prefix of p of length m' of q_1pq_2, and let $z = \lambda(p')$. By Part (2) of Lemma 6.3, $z \in L$. Moreover, $z \in uvA^*$, so it is prefix-comparable with some element of $u\mathcal{R}_{u,v}v$. But by assumption, $|z|$ is the maximal length of an element of $u\mathcal{R}_{u,v}v$. Therefore, it follows that z has at least two occurrences of uv. Since z is a proper prefix of uwv, we deduce that uwv has an occurrence of uv at position $0 < j < |s|$. Consider the factorization $p = p_1p_2$ where $|p_1| = j$, and let $x' = \omega(p_1) = \alpha(p_2)$. Since j is an occurrence of uv in $uwv = \lambda(q_1pq_2)$ and $|q_1p_1| = |u| + j$, it follows that $\lambda(q_1p_1)$ ends with u. Consider the factorization $x' = u'x''$, where $|u'| = |u|$. By Part (3) of Lemma 6.3, u' is suffix-comparable with $\lambda(q_1p_1)$, and since $|u'| = |u|$, it follows that $u' = u$. Similarly, the fact that j is an occurrence of uv in uwv, with $uwv = \lambda(q_1p_1p_2q_2)$ and $|q_1p_1| = |u| + j$, implies that $\lambda(p_2q_2)$ starts with v. By Part (4) of Lemma 6.3, it follows that x'' is prefix-comparable with $\lambda(p_2q_2)$, and hence with v. However, recall that $x' \in L \cap A^{m'}$ where $m' = |u| + |s| + |v|$:

$$|x''| = |x'| - |u| = |s| + |v| \geq |v|.$$

Therefore, v is a prefix of x'', and $x'' = vt$ for some word t. Hence, we conclude that p_1, p_2 satisfy:

$$p_1 \colon uvx \to uvt, \quad p_2 \colon uvt \to uvy.$$

Since $0 < j < |s| < |w|$, we have $|p_1| < |p|$ and $|p_2| < |p|$. Thus, by the induction hypothesis, both $\lambda(p_1)$ and $\lambda(p_2)$ are product of words in $\mathcal{R}_{u,v}$. And, therefore, so is w. This finishes the proof of the claim.

We finish the proof of Part (2) of the proposition by showing that it follows from that claim. First, recall that $G_{m',k}$ is *strongly connected*, in the sense that any two vertices can be joined by a positive path. Moreover, the groups of a strongly connected digraph are generated by the labels of positive loops [12, Corollary 4.5]. Since the claim above shows in particular that the labels of positive loops over usv in $G_{m',k}$ lie in $K_{u,v}$, the result follows. ∎

7. Proof of the main result

Let us first recall the statement of our main result, Theorem 1.1: if L is a suffix-connected uniformly recurrent language on A, then all the return groups of L lie in the same conjugacy class and their rank is $n-c+1$, where $n = \mathrm{Card}(A)$ and c is the number of connected components of $\mathbf{E}(\varepsilon)$.

The proof is split in two lemmas. In the first one, we apply the results obtained in the previous sections to show that (under the assumptions of Theorem 1.1) all the return groups of L belong to the same conjugacy class. The second lemma finishes the proof by showing that the groups in this conjugacy class have rank $n - c + 1$.

Lemma 7.1. *Let L be a uniformly recurrent suffix-connected language. Then, the return groups of L lie in the conjugacy class of subgroups of $F(A)$ generated by the groups of the Rauzy graph $G_{1,1}$.*

Proof: Consider a pair (u, v) such that $uv \in L$ and $uv \neq \varepsilon$. By Corollary 5.5, $H_{uv,\varepsilon} = H_{b,\varepsilon}$, where b is the last letter of uv. Using the conjugacy relation between Rauzy groups, we then have:

$$H_{u,v} = vH_{uv,\varepsilon}v^{-1} = vH_{b,\varepsilon}v^{-1},$$

and this equality holds for any such pair (u, v).

Since L is uniformly recurrent, we may choose an element $s \in \mathcal{R}_{u,v}$ of maximum length and by Proposition 6.2:

$$H_{u,sv} \leq K_{u,v} \leq H_{u,v}.$$

Applying the conclusion of the previous paragraph to the pair u, sv while noting that $s \in K_{u,v} \leq H_{u,v}$, we get:

$$H_{u,sv} = svH_{b,\varepsilon}v^{-1}s^{-1} = sH_{u,v}s^{-1} = H_{u,v}.$$

Hence, $K_{u,v} = H_{u,v} = vH_{b,\varepsilon}v^{-1}$. ∎

The next lemma concludes the proof of Theorem 1.1. It also gives an effective way of computing the Stallings equivalence of $G_{1,1}$ and, in turn, the Stallings equivalence can be used to find a basis for any of the groups defined by $G_{1,1}$.