A Proof of the Factorization Forest Theorem

Manfred Kufleitner[1,2]

[1]LaBRI, Université de Bordeaux and CNRS
351, Cours de la Libération
F-33405 Talence cedex, France

[2]FMI, Universität Stuttgart
Universitätsstr. 38
D-70569 Stuttgart, Germany

Abstract

We show that for every homomorphism $\Gamma^+ \to S$ where S is a finite semigroup there exists a factorization forest of height $\leq 3\,|S|$. The proof is based on Green's relations.

1 Introduction

Factorization forests where introduced by Simon [5, 7]. An important property of finite semigroups is that they admit factorization forests of finite height. This fact is called the Factorization Forest Theorem. It can be considered as an Ramsey-type property of finite semigroups. There exists different proofs of this fact of different difficulty and with different bounds on the height. The first proof of the Factorization Forest Theorem is due to Simon [7]. He showed that for every finite semigroup S there exists a factorization forest of height $\leq 9\,|S|$. The proof relies on several different techniques. It uses graph colorings, Green's relations, and a decomposition technique inspired by the Rees-Suschkewitsch Theorem on completely 0-simple semigroups. In [8] Simon gave a simplified proof relying on the Krohn-Rhodes decomposition. The bound shown is $2^{|S|+1} - 2$. A concise proof has been given by Chalopin and Leung [1]. The proof relies on Green's relations and yields the bound $7\,|S|$ on the height. Independently of this work, Colcombet has also shown a bound of $3\,|S|$ for the height of factorization forests [2]. He uses a generalization of the Factorization Forest Theorem in terms of *Ramseyan splits*. The proof also relies on Green's relations. A variant of our proof for the special case of aperiodic monoids has been shown in [3] with a bound of $3\,|S|$. The main benefit of that proof is that is uses very little machinery. The proof in this paper can be seen as an extension of that proof. The main tool are again Green's relations. We only require basic results from the theory of finite semigroups which can be found in standard textbooks such as [4].

A lower bound of $|S|$ was shown for rectangular bands in [6] and also in [1]. The same bound has also been shown for groups [1]. Therefore, the upper bound of $3\,|S|$ reduces the gap between the lower and the upper bound.

2 The Factorization Forest Theorem

Let S be a finite semigroup. A *factorization forest* of a homomorphism $\varphi : \Gamma^+ \to S$ is a function d which maps every word w with length $|w| \geq 2$ to a factorization $d(w) = (w_1, \ldots, w_n)$ of $w = w_1 \cdots w_n$ with $n \geq 2$ and $w_i \in \Gamma^+$ and such that $n \geq 3$ implies $\varphi(w_1) = \cdots = \varphi(w_n)$ is idempotent in S. The *height* h of a word w is defined as

$$h(w) = \begin{cases} 0 & \text{if } |w| \leq 1 \\ 1 + \max\{h(w_1), \ldots, h(w_n)\} & \text{if } d(w) = (w_1, \ldots, w_n) \end{cases}$$

We call the tree defined by the "branching" d for the word w the *factorization tree* of w. The height $h(w)$ is the height of this tree. The height of a factorization forest is the supremum over the heights of all words.

Factoriztion Forest Theorem (Simon [7]). *Let S be a finite monoid. Every homomorphism $\varphi : \Gamma^+ \to S$ has a factorization forest of height $\leq 3\,|S|$.*

Proof: Let $[w] = \varphi(w)$. We show that for every $w \in \Gamma^+$ there exists a factorization tree of height $h(w) \leq 3\,|\{x \in S \mid [w] \leq_{\mathcal{J}} x\}|$. First, we perform an induction on the cardinality of the set $\{x \in S \mid [w] \leq_{\mathcal{J}} x\}$; then within one $\mathcal{J}$-class we refine this parameter. Let $w \in \Gamma^+$ with $|w| \geq 2$. Then w has a unique factorization

$$w = w_0 a_1 w_1 \cdots a_m w_m$$

with $a_i \in \Gamma$ and $w_i \in \Gamma^*$ satisfying the following two conditions:

$$\forall\, 1 \leq i \leq m\colon [a_i w_i] \;\mathcal{J}\; [w] \quad \text{and} \quad \forall\, 0 \leq i \leq m\colon w_i = \varepsilon \;\vee\; [w] <_{\mathcal{J}} [w_i]$$

The idea is that we successively choose $a_i w_i \in \Gamma^+$ from right to left to be the shortest non-empty word such that $[a_i w_i] \;\mathcal{J}\; [w]$. Let $w_i' = a_i w_i$ for $1 \leq i \leq m$. For each $1 \leq i < m$ define a pair (L_i, R_i) where L_i is the $\mathcal{L}$-class of $[w_i']$ and R_i is the $\mathcal{R}$-class of $[w_{i+1}']$. Every such pair represents an $\mathcal{H}$-class within the $\mathcal{J}$-class of $[w]$. All $\mathcal{H}$-classes within this $\mathcal{J}$-class contain the same number n of elements. Let

$$h'(w) \;=\; h(w) - 3 \cdot |\{x \in S \mid [w] <_{\mathcal{J}} x\}|$$

We can think of h' as the height of a tree where we additionally allow words $v \in \Gamma^+$ with $[w] <_{\mathcal{J}} [v]$ as leafs. Within the $\mathcal{J}$-class of w we perform an induction on the cardinality of the set $\{(L_i, R_i) \mid 1 \leq i < m\}$ in order to show

$$h'(w) \;\leq\; 3n \cdot |\{(L_i, R_i) \mid 1 \leq i < m\}|$$

Since $n \cdot |\{(L_i, R_i) \mid 1 \leq i < m\}| \;\leq\; |\{x \in S \mid [w] \;\mathcal{J}\; x\}|$ this yields the desired bound for the height $h(w)$. If every pair (L, R) occurs at most twice then we have $m - 1 \leq 2 \cdot |\{(L_i, R_i) \mid 1 \leq i < m\}|$. We define a factorization tree for w by

$$\begin{aligned} d(w) &= (w_0 w_1',\; w_2' \cdots w_m') \\ d(w_0 w_1') &= (w_0, w_1') \\ d(w_i' \cdots w_m') &= (w_i', w_{i+1}' \cdots w_m') && \text{for } 2 \leq i < m \\ d(w_i') &= (a_i, w_i) && \text{for } 1 \leq i \leq m \end{aligned}$$

Since $[w] <_{\mathcal{J}} [w_i]$, by induction every w_i has a factorization tree of height $h(w_i) \;<\; 3\,|\{x \mid [w_i] \leq_{\mathcal{J}} x\}| \;\leq\; 3\,|\{x \mid [w] <_{\mathcal{J}} x\}|$. This yields:

$$h'(w) \;\leq\; m \;\leq\; 3n \cdot |\{(L_i, R_i) \mid 1 \leq i < m\}|$$

Note that the height does not increase if some of the w_i are empty. Now suppose there exists a pair $(L, R) \in \{(L_i, R_i) \mid 1 \leq i < m\}$ occurring (at least) three times. Let $i_0 < \cdots < i_k$ be the sequence of all positions with $(L, R) = (L_{i_j}, R_{i_j})$. By construction we have $k \geq 2$. Let $\widehat{w_j} = w_{i_{j-1}+1}' \cdots w_{i_j}'$ for $1 \leq j \leq k$. For all $1 \leq j \leq \ell \leq k$ we have

- $[\widehat{w_j} \cdots \widehat{w_\ell}] \leq_{\mathcal{L}} [w'_{i_\ell}] \;\mathcal{L}\; [w'_{i_0}]$.
- $[\widehat{w_j} \cdots \widehat{w_\ell}] \leq_{\mathcal{R}} [w'_{i_{j-1}+1}] \;\mathcal{R}\; [w'_{i_0+1}]$.
- $[w'_{i_\ell}] \leq_{\mathcal{J}} [\widehat{w_j} \cdots \widehat{w_\ell}] \leq_{\mathcal{J}} [w] \;\mathcal{J}\; [w'_{i_\ell}] \;\mathcal{J}\; [w'_{i_0}] \;\mathcal{J}\; [w'_{i_0+1}]$ by assumption on the factorization.

Thus for all $1 \leq j \leq \ell \leq k$ and $1 \leq j' \leq \ell' \leq k$ we get

- $[\widehat{w_j} \cdots \widehat{w_\ell}] \;\mathcal{L}\; [w'_{i_1}] \;\mathcal{L}\; [\widehat{w_{j'}} \cdots \widehat{w_{\ell'}}]$ and
- $[\widehat{w_j} \cdots \widehat{w_\ell}] \;\mathcal{R}\; [w'_{i_1+1}] \;\mathcal{R}\; [\widehat{w_{j'}} \cdots \widehat{w_{\ell'}}]$ and therefore
- $[\widehat{w_j} \cdots \widehat{w_\ell}] \;\mathcal{H}\; [\widehat{w_{j'}} \cdots \widehat{w_{\ell'}}]$

Therefore, all $[\widehat{w_j}]$ denote elements in the same $\mathcal{H}$-class H and since $k \geq 2$ the class H is a group. We consider the following set of elements in H induced by proper prefixes

$$P(\widehat{w_1} \cdots \widehat{w_k}) = \{[\widehat{w_1} \cdots \widehat{w_j}] \mid 1 \leq j < k\}$$

For the pair (L, R) we show by induction on $|P(\widehat{w_1} \cdots \widehat{w_k})|$ that

$$h'(w) \;\leq\; 3\,|P(\widehat{w_1} \cdots \widehat{w_k})| \;+\; 3n\,|\{(L_i, R_i) \mid 1 \leq i < m\}|$$

Suppose every element $x \in P(\widehat{w_1} \cdots \widehat{w_k}) \subseteq H$ occurs at most twice. Then $k - 1 \leq 2\,|P(\widehat{w_1} \cdots \widehat{w_k})|$. We construct the following factorization tree for w:

$$\begin{aligned} d(w) &= (w_0 w'_1 \cdots w'_{i_1},\; w'_{i_1+1} \cdots w'_m) \\ d(w_0 w'_1 \cdots w'_{i_1}) &= (w_0 w'_1 \cdots w'_{i_0},\; \widehat{w_1}) \\ d(w'_{i_0+1} \cdots w'_m) &= (\widehat{w_2} \cdots \widehat{w_k},\; w'_{i_k+1} \cdots w'_m) \\ d(\widehat{w_i} \cdots \widehat{w_k}) &= (\widehat{w_i},\; \widehat{w_{i+1}} \cdots \widehat{w_k}) \qquad \text{for } 2 \leq i < k \end{aligned}$$

By induction on the number of pairs (L_i, R_i) there exist factorization trees for the words $w_0 w'_1 \cdots w'_{i_0}$, $w'_{i_k+1} \cdots w'_m$, and all $\widehat{w_i}$ of height

$$\leq\; 3n\,|\{(L_i, R_i) \mid 1 \leq i < m\} \setminus \{(L, R)\}| \;+\; 3\,|\{x \mid [w] <_{\mathcal{J}} x\}|$$

This yields

$$h'(w) - 3n\,|\{(L_i, R_i) \mid 1 \leq i < m\}| \;\leq\; k \;\leq\; 3\,|P(\widehat{w_1} \cdots \widehat{w_k})|$$

Now suppose there exists an element $x \in P(\widehat{w_1} \cdots \widehat{w_k}) \subseteq H$ that occurs at least three times. Let $j_0 < \cdots < j_t$ be the sequence of all positions with $x = [\widehat{w_1} \cdots \widehat{w_{j_i}}]$. By construction we have $t \geq 2$. It follows that $[\widehat{w_{j_i+1} \cdots w_{j_{i+1}}}] = e = e^2$ where e is the neutral element of the group H. Let $v_i = \widehat{w_{j_{i-1}+1} \cdots w_{j_i}}$ for $1 \leq i \leq t$. We construct the following factorization tree for w:

$$\begin{aligned} d(w) &= (w_0 \cdots \widehat{w_{i_0}},\; \widehat{w_{i_0+1}} \cdots w'_m) \\ d(\widehat{w_{i_0+1}} \cdots w'_m) &= (v_1 \cdots v_t,\; \widehat{w_{i_t+1}} \cdots w'_m) \\ d(v_1 \cdots v_t) &= (v_1, \ldots, v_t) \end{aligned}$$

We have $x \in P(\widehat{w_1} \cdots \widehat{w_k}) \setminus P(\widehat{w_1} \cdots \widehat{w_{i_0}})$ and $xP(\widehat{w_{j_{i-1}+1} \cdots w_{j_i}}) \subseteq P(\widehat{w_1} \cdots \widehat{w_k})$ but $x \notin xP(\widehat{w_{j_{i-1}+1} \cdots w_{j_i}})$. Hence, by induction on the cardinality of the prefix sets, there exist factorization forests for $w_0 \cdots \widehat{w_{i_0}}$, $\widehat{w_{i_t+1}} \cdots w'_m$ and the v_i of height

$$\begin{aligned} &\leq 3\,|P(\widehat{w_1} \cdots \widehat{w_k})| \;-\; 3 \\ &\quad + 3n\,|\{(L_i, R_i) \mid 1 \leq i < m\} \setminus \{(L, R)\}| \\ &\quad + 3\,|\{x \mid [w] <_{\mathcal{J}} x\}| \end{aligned}$$

This yields a factorization tree for w with the desired height bound. $\square$

Acknowledgement. I would like to thank Volker Diekert for numerous discussions on this topic.

References

[1] Jérémie Chalopin and Hing Leung. On factorization forests of finite height. *Theoretical Computer Science*, 310(1-3):489–499, 2004.

[2] Thomas Colcombet. Factorisation forests for infinite words. In Erzsébet Csuhaj-Varjú and Zoltán Ésik, editors, *Fundamentals of Computation Theory, 16th International Symposium, FCT 2007, Budapest, Hungary, August 27-30, 2007, Proceedings*, volume 4639 of *Lecture Notes in Computer Science*, pages 226–237. Springer-Verlag, 2007.

[3] Volker Diekert and Manfred Kufleitner. On first-order fragments for words and Mazurkiewicz traces: A survey. In Tero Harju, Juhani Karhumäki, and Arto Lepistö, editors, *Developments in Language Theory, 11th International Conference, DLT 2007, Turku, Finland, July 3-6, 2007, Proceedings*, volume 4588 of *Lecture Notes in Computer Science*, pages 1–19. Springer-Verlag, 2007.

[4] Jean-Éric Pin. *Varieties of Formal Languages*. North Oxford Academic, London, 1986.

[5] Imre Simon. Factorization forests of finite height. Technical Report 87-73, Laboratoire d'Informatique Théorique et Programmation, Paris, 1987.

[6] Imre Simon. Properties of factorization forests. In Jean-Éric Pin, editor, *Formal Properties of Finite Automata and Applications: LITP Spring School on Theoretical Computer Science*, volume 386 of *Lecture Notes in Computer Science*, pages 65–72. Springer-Verlag, 1988.

[7] Imre Simon. Factorization forests of finite height. *Theoretical Computer Science*, 72(1):65–94, 1990.

[8] Imre Simon. A short proof of the factorization forest theorem. In Maurice Nivat and Andreas Podelski, editors, *Tree Automata and Languages*, pages 433–438. Elsevier, 1992.