# The branching point approach to Conway's Problem ⋆

Juhani Karhumäki and Ion Petre

Department of Mathematics, University of Turku and
Turku Centre for Computer Science (TUCS)
Turku 20014, Finland
Email: {karhumak,ipetre}@cs.utu.fi

**Abstract.** A word $u$ is a branching point for a set of words $X$ if there are two different letters $a$ and $b$ such that both $ua$ and $ub$ can be extended to words in $X^+$. A branching point $u$ is critical for $X$ if $u \notin X^+$. Using these notions, we give an elementary solution for Conway's Problem in the case of finite biprefixes. We also discuss a possible extension of this approach towards a complete solution for Conway's Problem.

## 1 Introduction

The *centralizer* of a set of words $X$ is the maximal set with respect to union, commuting with $X$. As it can be readily seen, the notion of centralizer is well defined for all sets of words; we denote in this paper the centralizer of $X$ by $\mathcal{C}(X)$. For any $X$, $\mathcal{C}(X)$ is a monoid or more exactly, it is the union of all sets commuting with $L$. As a matter of fact, one can also define a notion of semigroup centralizer and it is an open problem whether or not the two types of centralizer always coincide modulo a finite (or at least rational) set of words. We refer to [8] for more details. In this paper, we always consider semigroup centralizers.

The best known problem connected to the notion of centralizer is the one proposed by Conway ([6], 1971), asking whether or not the centralizer of a rational language is always rational. Surprisingly enough, very little is known on the answer to Conway's Problem. E.g., it is not even known whether the centralizer of any finite set is recursively enumerable. We know however, that Conway's Problem has an affirmative answer for periodic, binary, and ternary sets, as well as for rational $\omega$-codes, see [4], [7], [9], and [13], as well as [10] for a recent survey. We recall that a set of words $X$ is called *periodic* if there is a word $u$ such that $X \subseteq u^*$. A set of words $X$ is called *binary* (*ternary*, resp.) if $X$ consists of two (three, resp.) words. These results have been generally obtained as consequences of characterizing some special cases of the commutation of two sets of words. Interestingly, the above results were obtained using very different approaches: combinatorial properties of finite and infinite words, equations on languages, and algebraic results on the commutation of formal power series. Still another one

⋆ Supported by the Academy of Finland under grant 44087

- the so called fixed point approach - is presented in [8]. We interpret this as an evidence of the challenging nature of the problem.

Following ideas of [8], we propose here still another approach to Conway's Problem. We define the notions of *branching* and *critical* points and prove that Conway's problem can be reduced to those sets of words having at least two words starting with different letters. In turn, for these sets of words, one only has to establish the rationality of the set of critical points to obtain a solution to Conway's Problem. As an illustration of the approach, we give a simple, elementary solution to this problem for binary sets and for finite biprefixes. For binary sets, this gives a simpler solution than that presented in [4]. The result for finite biprefixes is obtained also in [15], through some involved combinatorial arguments, as well as in [7], using some algebraic result on the commutation of two formal power series.

We conclude the paper with a discussion on a possible extension of this approach towards a complete solution for Conway's Problem.

## 2 Branching points

For elementary definitions and results on Combinatorics on Words, we refer to [3], [11], and [12]. For definitions and results on Theory of Codes, we refer to [2].

We say that a word $u$ is a prefix of a word $v$ if $v = ut$, for some word $t$, and we denote $u \leq v$. For a set of words $L$ we denote by $\mathrm{Pref}(L)$ the set of all prefixes of words from $L$: $\mathrm{Pref}(L) = \{x \mid \exists u \in L \text{ such that } x \leq u\}$. We say that $u$ is a suffix of $v$ if $v = tu$, for some word $t$, and we denote $u \leq_s v$. For a set of words $L$ we denote by $\mathrm{Suf}(L)$ the set of all suffixes of words from $L$.

Let $X$ be a set of words and $u$ a word. We say that $u$ is a *branching point* of $X$ if there are two distinct letters $a$ and $b$ such that both $ua$ and $ub$ are prefixes of some words in $X^+$. In other words, $u$ can be extended in two different ways to words of $X^+$. We denote by $\mathcal{B}(X)$ the set of branching points of $X$. A branching point $u$ of $X$ is called *critical* if $u \notin X^+$.

*Example 1.* (i) Let $F = \{a, aba, bb\}$. Then $a$ is a branching point of $F$: $aa, aba \in F^+$, while $b$ is not a branching point. Indeed, $ba$ is not a prefix of any word in $F^+$. Also, $ab$ is a critical point: $aba, abb \in F^+$ and $ab \notin F^+$. Note that $ab$ is not in $\mathcal{C}(F)$ since, as it is easy to see, $\mathcal{C}(F) = F^+$.

(ii) Let $F = \{aa, ab, ba, bb\}$. Then both $a$ and $b$ are critical points of $F$. Moreover, $a$ and $b$ are both in $\mathcal{C}(F)$. Indeed, $\mathcal{C}(F) = \{a, b\}^+$.

We say that a set of words $L$ is *branching* if $L$ has words starting with different letters. We say that $L$ is *marked* if no two words of $L$ start with the same letter.

For a branching set $L$ the critical points are the only potential nontrivial elements of the centralizer $\mathcal{C}(L)$, e.g., elements outside $L^+$. This follows from the following simple lemma, cf. [8].

**Lemma 1.** *For any language $L$, $1 \notin L$, we have the following:*

*(i)* $\mathcal{C}(L)$ *is a semigroup.*
*(ii)* $L^+ \subseteq \mathcal{C}(L) \subseteq \mathrm{Pref}(L^+) \cap \mathrm{Suf}(L^+)$.
*(iii) If* $L$ *is branching, then* $\mathcal{C}(L) \subseteq \mathcal{B}(L)$.

By Lemma 1(iii), if $L$ is a branching set of words, then all words in $\mathcal{C}(L)$ are branching points of $L$. We prove in the next result that for any rational language $L$, $\mathcal{B}(L)$ is rational, thus supporting a possible affirmative answer to Conway's Problem.

**Theorem 1.** *For any rational language* $R$*, the set of its branching points* $\mathcal{B}(R)$ *is rational.*

*Proof.* If $R$ is rational, then $R^*$ is rational and so is $\mathrm{Pref}(R^*)$. Let $\mathcal{A}$ be a complete deterministic finite automaton accepting $\mathrm{Pref}(R^*)$, with $\delta$ its transition mapping, $Q$ the set of states, $F$ the set of final states, and $q_0$ its initial state. For each $q \in Q$ and each letter $a$, let $q_a = \delta(q, a)$.

We construct an automaton $\mathcal{A}'$ accepting $\mathcal{B}(A)$. Intuitively, to accept a word $u$, we walk in $\mathcal{A}$ with $u$ and then we check whether or not both letters $a$ and $b$ lead to final states. Formally, let $Q'$ be a set isomorphic to $Q$: $Q' = \{q' \mid q \in Q\}$ and let $r, s$ be two new states, $r, s \notin Q \cup Q'$. The set of states of $\mathcal{A}'$ is $Q \times (Q' \cup \{r, s\}) \times (Q' \times \{r, s\})$, the initial state is $(q_0, r, r)$ and the transition mapping is defined as follows:

(i) $\delta'((q, r, r), x) = (\delta(q, x), r, r)$, for all letters $x$;
(ii) $\delta'((q, r, r), 1) = (q, s, s)$;
(iii) $\delta'((q, s, s), 1) = (q, q'_a, q'_b)$, for all letters $a \neq b$.

The set of final states of $\mathcal{A}'$ is $Q \times F' \times F'$, where $F = \{q' \mid q \in F\}$. Consequently, $\mathcal{A}'$ is a so called generalized finite automaton and hence, is accepting a rational language. As it is easy to see, the language accepted by $\mathcal{A}'$ is $\mathcal{B}(R)$. Indeed, a word $u \in \mathcal{B}(R)$ if and only if $ua, ub \in \mathrm{Pref}(R^+)$.

We prove in the next result that Conway's Problem can be reduced to the special case of branching sets.

**Theorem 2.** *For any non-periodic set of words* $L$*,* $1 \notin L$*, there is a branching set of words* $L'$ *such that* $\mathcal{C}(L)$ *is rational if and only if* $\mathcal{C}(L')$ *is rational. Moreover,* $\mathcal{C}(L) = L^+$ *if and only if* $\mathcal{C}(L') = L'^+$.

*Proof.* If $L$ is branching, then the claim is trivially true, with $L' = L$. Thus, let us assume that $L = aL_1$, for some letter $a$ and some set of words $L_1$. Then, as $\mathcal{C}(L)L = L\mathcal{C}(L)$, it follows that $\mathcal{C}(L) = 1 + aX$, for some set of words $X$ and so, $L_1 + XaL_1 = L_1 + L_1aX$. Thus, $(1 + Xa)L_1a = L_1a(1 + Xa)$ i.e., $Xa \subseteq \mathcal{C}(L_1a)$. The other inclusion can be proved similarly and so, $\mathcal{C}(aL_1) = 1 + a(\mathcal{C}(L_1a)a^{-1})$. Note that $\mathcal{C}(aL_1)$ is rational if and only if $\mathcal{C}(L_1a)$ is rational and moreover, $\mathcal{C}(aL_1) = (aL_1)^+$ if and only if $\mathcal{C}(L_1a) = (L_1a)^+$.

If $L_1a$ is not branching, then we repeat the reasoning with $L_1a$ instead of $L$. Since $L$ is not periodic, we find in a finite number of steps a branching set $L'$ such that $\mathcal{C}(L)$ is rational if and only if $\mathcal{C}(L')$ is rational. Moreover, $\mathcal{C}(L) = L^+$ if and only if $\mathcal{C}(L') = L'^+$.

Consequently, Conway's Problem can be reduced to two types of sets: periodic sets and branching sets of words. The case of periodic sets is, however, easy to settle.

**Theorem 3 ([13]).** *Let $u$ be a primitive word and $L \subseteq u^+$. Then $\mathcal{C}(L) = u^+$.*

*Proof.* Sine $L\mathcal{C}(L) = \mathcal{C}(L)L$, for any word $x \in \mathcal{C}(L)$ and any $\alpha \in L$, $x\alpha^\omega \in L^\omega$. Thus, $xu^\omega = u^\omega$ and so, $x = u^n$, for some $n \geq 0$. Due to the maximality of the centralizer, it follows that $\mathcal{C}(L) = u^+$.

From now on, we consider branching sets of words only. Based on this reduction, we give a simple proof for Conway's Problem in the case of binary sets. This result has been originally proved in [4] using somewhat more involved combinatorial arguments.

**Theorem 4.** *The centralizer of any binary set $F$ over the alphabet $\Sigma$ is rational. Moreover,*

*(i) If $1 \in F$, then $\mathcal{C}(F) = \Sigma^+$.*
*(ii) If $F$ is periodic, $F \subseteq u^+$, for some primitive word $u$, then $\mathcal{C}(F) = u^+$.*
*(iii) If $F$ is not periodic and $1 \notin F$, then $\mathcal{C}(F) = F^+$.*

*Proof.* The first case is trivial and it holds more generally for any set of words $F$. Also, Case (ii) is concluded in Theorem 3. For Case (iii), note that by Theorem 2, we can assume without loss of generality that $F$ is a branching set of words. Let $F = \{au, bv\}$, where $a$ and $b$ are distinct letters and $u, v$ some words.

Assume that $F^+ \neq \mathcal{C}(F)$ and let $x \in \mathcal{C}(F) \setminus F^+$. Since $F$ is a prefix and $x \in \mathrm{Pref}(F^+)$, it follows that there are unique words $u_1, \ldots, u_m \in F$, $m \geq 0$ such that $x = u_1 \ldots u_m t$, for a word $t \in \mathrm{Pref}(F)$.

Observe now that $x$ is a branching point. Indeed, since $F$ is a branching set of words, all words of $\mathcal{C}(F)$ are branching points of $F$. Thus, $t$ is also a branching point of $F$ and so, as $t \in \mathrm{Pref}(F)$, there are $\alpha, \beta \in \mathrm{Pref}(F)$ such that $ta \leq \alpha$ and $tb \leq \beta$. If $t \neq 1$ then, since $F$ is marked, it follows that $\alpha = \beta$ and $a = b$, a contradiction. Thus, $t = 1$ and $x \in F^+$, again impossible. Consequently, $\mathcal{C}(F) = F^+$.

## 3 A simple solution to Conway's Problem for finite biprefixes

It is well known, see [14], that the set of prefixes is a free monoid. In particular, this implies that any prefix has a unique primitive root, similarly as words have. It is a consequence of a result of [15] characterizing the commutation with a prefix code that, for any prefix code $X$, we have $\mathcal{C}(X) = \rho(X)^+$, where $\rho(X)$ denotes the primitive root of $X$. This result was extended in [7] to $\omega$-codes: any $\omega$-code $X$ has a unique primitive root $\rho(X)$ and $\mathcal{C}(X) = \rho(X)^+$. However, both these results of [7] and [15] rely on some complex arguments. For prefix codes ([15]), one uses some involved combinatorial arguments, while for $\omega$-codes,

[7], one relies on some results of Bergman and Cohn, [1], [5], characterizing the commutation of two polynomials and of two formal power series, respectively. Using the notion of branching point, we give a simple, elementary solution for Conway's Problem in the case of finite biprefixes. We begin by proving that the centralizer of any biprefix set of words is necessarily of a very special form.

**Theorem 5.** *For any biprefix $L$, there is a set $T$ of nonempty branching points of $\rho(L)$ such that $\mathcal{C}(L) = \rho(L)^+(1 + \sum_{t\in T} \rho(L)^{k_t}t)$, where $\rho(L)$ denotes the primitive root of $L$ and $k_t \geq 0$, for all $t \in T$.*

*Proof.* As in Theorem 2, we can assume without loss of generality that $L$ is a branching set of words.

Let $x \in \mathcal{C}(L) \setminus \rho(L)^+$. By Lemma 1, $x \in \mathrm{Pref}(L^+)$ and thus, since $L$ is a prefix, there are unique words $u_1, \ldots, u_k, t$, such that $x = u_1 \ldots u_k t$, with $u_i \in L$ and $t \in \mathrm{Pref}(L) \setminus \rho(L)^*$. Moreover, since $L$ is branching, $t$ is a branching point of $L$ i.e., $ta, tb \in \mathrm{Pref}(L)$, for distinct letters $a$ and $b$. Because $L$ is a prefix code and $\mathcal{C}(L)L^k = L^k\mathcal{C}(L)$, it follows that $xL^k \subseteq L^k\mathcal{C}(L)$ and so, $tL^k \subseteq \mathcal{C}(L)$. Similarly, since $L$ is also a suffix, it follows that $L^k t \subseteq \mathcal{C}(L)$. Thus, there is $k' \geq 1$ and $t' \in \mathrm{Pref}(\rho(L)) \setminus \rho(L)^*$ such that $\rho(L)^{k'}t' \subseteq \mathcal{C}(L)$.

Let $T$ be the set of all words $t'$ defined above, or more formally

$$T = \{t \in (\rho(L)^+)^{-1}\mathcal{C}(L) \mid u \not\leq t,\ \forall u \in \rho(L)\}.$$

For any $t \in T$, let $k_t = \min\{k \geq 0 \mid \rho(L)^k t \subseteq \mathcal{C}(L)\}$. We claim that

$$\mathcal{C}(L) = \rho(L)^+ \left(1 + \sum_{t\in T} \rho(L)^{k_t} t\right).$$

Clearly, by construction, $\rho(L)^{k_t}t \subseteq \mathcal{C}(L)$ and so, $\rho(L)^+ \left(1 + \sum_{t\in T} \rho(L)^{k_t}t\right) \subseteq \mathcal{C}(L)$, since $\mathcal{C}(L)$ is closed under union and under multiplication by $\rho(L)$.

For the reverse inclusion, let $x \in \mathcal{C}(L) \setminus \rho(L)^+$. Then, as shown above, there is $l \geq 1$ such that $x \in \rho(L)^l t$, with $t \in T$. Since $F$ is a biprefix, it follows as above that $\rho(L)^l t \subseteq \mathcal{C}(L)$. Consequently, $l \geq k_t$ and the claim follows.

As a matter of fact, based on some involved considerations of [15], we know that $T = \emptyset$. Unfortunately, we do not have a simple argument for this.

**Corollary 1.** *The centralizer of any finite biprefix is rational.*

*Proof.* Let $L$ be a finite biprefix. Since $T$ is a set of branching points, $T \subseteq \mathrm{Pref}(L)$. Thus, $T$ is finite and $\mathcal{C}(L)$ is rational.

## 4 Biprefixes with at most one critical point

We consider in the following some simple cases of finite biprefixes to further illustrate the branching point approach. Namely, we consider the case of finite biprefixes with at most one critical point.

**Theorem 6.** *Let $L$ be a biprefix code. Then $L$ has no critical points if and only if $L$ is marked. Moreover, in this case, $\mathcal{C}(L) = L^+$.*

*Proof.* Let $L$ be a biprefix code and assume that $L$ has no critical points. If $L$ is not marked, then there are two words $u$ and $v$ starting with the same letter. Thus, as $L$ is a prefix, the longest common prefix of $u$ and $v$ is a critical point of $L$, a contradiction.

If $L$ is marked, let us assume that $L$ has a critical point $x$. Thus, $xa, xb \in \mathrm{Pref}(L^+)$, for distinct letters $a$ and $b$, and $x \notin L^+$. Since $L$ is a biprefix code, it follows that there are unique words $u_1, \ldots, u_m \in L$ and $t \in \mathrm{Pref}(L) \setminus L^*$ such that $x = u_1 \ldots u_m t$. Since $ta, tb \in \mathrm{Pref}(L^+)$, there are $\alpha, \beta \in L$ such that $ta \leq \alpha$ and $tb \leq \beta$. However, $t \neq 1$ and so, as $L$ is marked, it follows that $\alpha = \beta$ and $a = b$, a contradiction.

If $L$ is marked, then all points of $\mathcal{C}(L)$ are branching points. Since $L$ has no critical points, it follows that $\mathcal{C}(L) = L^+$.

Observe that for any critical point $u$ of a prefix code $L$, all words in $\rho(L)^*u$ are also critical points of $L$. We say that $v$ is a *minimal* critical point of a code $L$ if there is no critical point $u$ of $L$ such that $v \in \rho(L)^*u$.

*Example 2.* Let $F = \{aa, ab\}$. Then the set of critical points of $F$ is $F^*a$. However, the only minimal critical point of $F$ is $a$.

**Theorem 7.** *Let $L$ be a biprefix with at most one minimal critical point. Then $\mathcal{C}(L) = \rho(L)^+$.*

*Proof.* By Theorem 2, we can assume without loss of generality that $L$ is a branching set of words. If $L$ has no critical point, then the claim follows by Theorem 6.

Assume that $L$ has one minimal critical point. By Theorem 5, $\mathcal{C}(L) = \rho(L)^+(1 + \sum_{t\in T} \rho(L)^{k_t} t)$, for a set $T$ of critical points of $\rho(L)$. Since $L$ has only one minimal critical point, $T \subseteq \{t\}$, with $t \in \mathrm{Pref}(\rho(L)) \setminus \{1\}$. If $T = \emptyset$, then the claim follows. Assuming that $T = \{t\}$, we obtain $\mathcal{C}(L) = \rho(L)^+(1 + \rho(L)^k t)$ and so,

$$\rho(L)^+(1 + \rho(L)^k t)L = L\rho(L)^+(1 + \rho(L)^k t). \tag{1}$$

We prove that $\rho(L)^+ t$ commutes with $L$.

If $\rho(L)^+\rho(L)^k tL \cap L\rho(L)^+ \neq \emptyset$ then, as $L$ and $\rho(L)$ are prefixes, it follows that $tL \cap \rho(L)^+ \neq \emptyset$. Since $L$ and $\rho(L)$ are also suffixes, it follows that $t \in \rho(L)^+$, a contradiction.

If $\rho(L)^+ L \cap L\rho(L)^+\rho(L)^k t \neq \emptyset$ then, as $L$ and $\rho(L)$ are prefixes, it follows that $t \in \rho(L)^+$, again a contradiction.

Consequently, it follows from (1) that $\rho(L)^+\rho(L)^k tL = L\rho(L)^+\rho(L)^k t$. Moreover, since $\rho(L)$ is a prefix, it follows that

$$\rho(L)^+ tL = L\rho(L)^+ t. \tag{2}$$

Since $L$ has only one minimal critical point, $\rho(L)$ is marked, with just one exception: there are $u, v \in \rho(L)$ such that the common prefix of $u$ and $v$ is $t$. Let $L = \rho(L)^n$. Clearly, since $L$ is branching, there is a word $w \in \rho(L) \setminus \{u, v\}$.

From (2), we obtain that $Lt \subseteq \rho(L)^+tL$. In particular, $w^nt \in \rho(L)^+tL$. Since $t \not\leq w$ and $w \not\leq t$, we obtain that $t \in \rho(L)^+tL$, a contradiction.

## 5 Conclusions

Based on the simple notions of branching and critical points, we have proposed a new approach - the branching point approach - to attack Conway's Problem. We have demonstrated its usefulness by giving very simple solutions of the problem in the case of binary sets and finite biprefix sets. As a matter of fact, our result for biprefix sets can be easily extended to codes with bounded decoding delay in both directions, proving that also in this case, the centralizer has a simple, rational form.

We believe that the branching point approach can be used to derive also other results on Conway's problem, maybe even an affirmative answer for the case of all finite sets. To support this idea, let us denote by $T_{\mathcal{B}(L)}$ the tree of all words in $\mathcal{B}(L)$, for a rational language $L$. By Theorem 1, $\mathcal{B}(L)$ is rational and so, this tree - let us call it the *branching tree* of $L$ - is of a regular type: it contains only a finite number of (maximal) subtrees. For branching sets of words $L$, to which Conway's Problem can be reduced, all words in $\mathcal{C}(L)$ are branching and thus, nodes in $T_{\mathcal{B}(L)}$. Let $Z = \mathcal{C}(L) \setminus L^+$. As it is easy to see, $L^+Z \cup Z^+ \cup ZL^+ \subseteq \mathcal{C}(L)$. Consequently, a single node from $Z$ determines many other nodes of $T_{\mathcal{B}(L)}$ to be in $\mathcal{C}(L)$. Thus, for a solution to Conway's Problem, an important question is: can one "saturate" $T_{\mathcal{B}(L)}$ with nodes from $\mathcal{C}(L)$ in a finite number of steps of this type, at least for some types of (finite) sets $X$ ? Intuitively, the regularity of $T_{\mathcal{B}(L)}$ supports this view.

## References

1. G.Bergman, Centralizers in free associative algebras, *Transactions of the American Mathematical Society* 137: 327–344, 1969.
2. J.Berstel, D.Perrin, *Theory of codes*, Academic Press, New York, 1985.
3. C.Choffrut, J.Karhumäki, Combinatorics on Words. In G.Rozenberg, A.Salomaa (eds.), *Handbook of Formal Languages*, vol. 1: 329-438, Springer-Verlag, 1997.
4. C.Choffrut, J.Karhumäki, N.Ollinger, The commutation of finite sets: a challenging problem, *Theoret. Comput. Sci.*, to appear.
5. P.M.Cohn, Centralisateurs dans les corps libres, in J.Berstel (Ed.), *Séries formelles*: 45–54, Paris, 1978.
6. J.H.Conway, *Regular Algebra and Finite Machines*, Chapman Hall, 1971.
7. T.Harju, I.Petre, On commutation and primitive roots of codes, submitted.
8. J.Karhumäki, Challenges of commutation: an advertisement, in *Proc. of FCT 2001*, LNCS 2138, 15–23, Springer, 2001.
9. J.Karhumäki, I.Petre, Conway's Problem for three word sets, *Theoret. Comput. Sci.*, to appear; preliminary version in *Proc. ICALP 2000*, LNCS 1853 536–546, Springer, 2000.
10. J.Karhumäki, I.Petre, Conway's problem and the commutation of languages, *Bulletin of EATCS* 74, 171–177, 2001.

11. M.Lothaire, *Combinatorics on Words*, Addison-Wesley, Reading, MA., 1983.
12. M.Lothaire, *Algebraic Combinatorics on Words*, Cambridge University Press, to appear.
13. A.Mateescu, A.Salomaa, S.Yu, On the decomposition of finite languages, TUCS Technical Report 222, http://www.tucs.fi/, 1998.
14. D.Perrin, Codes conjugués, Information and Control 20: 222–231, 1972.
15. B.Ratoandromanana, Codes et motifs, *RAIRO Inform. Theor.*, 23(4): 425-444, 1989.