Orbits of Language Operations: Finiteness and Upper Bounds

Émilie Charlier\textsuperscript{1} Mike Domaratzki\textsuperscript{2} Tero Harju\textsuperscript{3} Jeffrey Shallit\textsuperscript{1}

\textsuperscript{1}University of Waterloo \hspace{0.5cm} \textsuperscript{2}University of Manitoba \hspace{0.5cm} \textsuperscript{3}University of Turku

Theory Seminar
Toronto, August 12, 2011
Closure operations

Let \( x : 2^\Sigma^* \to 2^\Sigma^* \) be an operation on languages. Suppose \( x \) satisfies the following three properties:

1. \( L \subseteq x(L) \) (expanding);
2. If \( L \subseteq M \) then \( x(L) \subseteq x(M) \) (inclusion-preserving);
3. \( x(x(L)) = x(L) \) (idempotent).

Then we say that \( x \) is a closure operation.

Example
Kleene closure, positive closure, prefix, suffix, factor, subword.
Some notation and a first result

If $x(L) = y(L)$ for all languages $L$, then we write $x \equiv y$.

We write $\epsilon(L) = L$ and $xy = x \circ y$, that is, $xy(L) = x(y(L))$.

Define $c$ to be the complementation: $c(L) = \Sigma^* - L$. In particular, we have $cc \equiv \epsilon$.

**Theorem**

*Let $x, y$ be closure operations. Then $xcycxcy \equiv xcy$.*
Proof of the previous result

∀L, xcyxcy(L) ⊆ xcy(L):

We have: ∀L, L ⊆ y(L).
Then: ∀L, cy(L) ⊆ c(L).
Then: ∀L, xcy(L) ⊆ xc(L).
Then: ∀L, xcy(cxcy(L)) ⊆ xc(cxcy(L)) = xcy(L).

∀L, xcy(L) ⊆ xycxcycy(L):

We have: ∀L, L ⊆ x(L).
Then: ∀L, cy(L) ⊆ x(cy(L)).
Then: ∀L, cxcy(L) ⊆ ccy(L) = y(L).
Then: ∀L, ycxcy(L) ⊆ yy(L) = y(L).
Then: ∀L, cy(L) ⊆ cycxcy(L).
Finally: ∀L, xcy(L) ⊆ xycxcycy(L).
Corollary (Peleg 1984, Brzozowski-Grant-Shallit 2009)

Let $x$ be any closure operation and $L$ be any language. If $S = \{x, c\}$, then the orbit $O_S(L) = \{y(L) : y \in S^*\}$ contains at most 14 languages, which are given by the images of $L$ under the 14 operations

$$
\epsilon, \ x, \ c, \ xc, \ cx, \ xcx, \ cxc, \ xcxc, \ cxcx, \\
xcccx, \ cxcxc, \ xccxcx, \ ccxcxx, \ ccxxcc.
$$

NB: This result is the analogous for languages of Kuratowski-14 sets-theorem for topological spaces.
Orbits of languages

Given a set $S$ of operations, we consider the orbit of languages
$O_S(L) = \{x(L) : x \in S^*\}$ under the monoid generated by $S$.

So compositions of operations in $S$ are considered as “words over the alphabet $S$”.

We are interested in the following questions:

- When is the monoid $S^*/\equiv$ finite?
- Is the cardinality of $O_S(L)$ bounded, independently of $L$?
Operations with infinite orbit

The orbit of \( L \) under an arbitrary operation need not be finite.

Example
Consider the operation \( q \) defined by

\[
q(L) = \{ x \in \Sigma^* : x \text{ is a proper prefix of some } y \in L \}.
\]

Let \( M = \{ a^n b^n : n \geq 1 \} \). Then the orbit

\[
O_{\{q\}}(M) = \{ M, q(M), q^2(M), q^3(M), \ldots \}
\]

is infinite, since we have

\[
a^{i+1}b \in q^i(M) \text{ and } a^{i+1}b \notin q^j(M) \text{ for } j > i.
\]
The situation is somewhat different if $L$ is regular:

**Theorem**

Let $L$ be a regular language accepted by a DFA of $n$ states. Then $|\mathcal{O}_{\{q\}}(L)| \leq n$, and this bound is tight.

To see that the bound is tight, consider the language $L_n = \{\epsilon, a, a^2, \ldots, a^{n-2}\}$, which is accepted by a $n$ state DFA. Then $q(L_n) = L_{n-1}$, so this shows $|\mathcal{O}_{\{q\}}(L_n)| = n$. 
It is possible for the orbit under a single operation to be infinite even if the operation is expanding and inclusion-preserving.

Example
Consider the operation of fractional exponentiation, defined by

\[ n(L) = \{ x^\alpha : x \in L \text{ and } \alpha \geq 1 \text{ rational} \} = \bigcup_{x \in L} x^+ p(\{x\}). \]

Let \( M = \{ab\} \). Then the orbit

\[ O_{\{n\}}(M) = \{M, n(M), n^2(M), n^3(M), \ldots\} \]

is infinite, since we have

\[ aba^i \in n^i(M) \quad \text{and} \quad aba^j \not\in n^j(M) \quad \text{for } j < i. \]
Some notation and definitions

If $t, x, y, z$ are words with $t = xyz$, we say

- $x$ is a prefix of $t$;
- $z$ is a suffix of $t$; and
- $y$ is a factor of $t$.

If $t = x_1y_1x_2y_2 \cdots x_ny_nx_{n+1}$ for some words $x_i$ and $y_j$, we say

- $y_1 \cdots y_n$ is a subword of $t$.

Thus a factor is a contiguous block, while a subword can be “scattered”.

Further, $x^R$ denotes the reverse of the word $x$. 
8 natural operations on languages

\[ k: L \mapsto L^* \]
\[ e: L \mapsto L^+ \]
\[ c: L \mapsto \overline{L} = \Sigma^* - L \]
\[ p: L \mapsto \text{pref}(L) \]

\[ s: L \mapsto \text{suff}(L) \]
\[ f: L \mapsto \text{fact}(L) \]
\[ w: L \mapsto \text{subw}(L) \]
\[ r: L \mapsto L^R \]

where

\[ L^* = \bigcup_{n \geq 0} L^n \quad \text{and} \quad L^+ = \bigcup_{n \geq 1} L^n \]
\[ \text{pref}(L) = \{ x \in \Sigma^* : x \text{ is a prefix of some } y \in L \} \]
\[ \text{suff}(L) = \{ x \in \Sigma^* : x \text{ is a suffix of some } y \in L \} \]
\[ \text{fact}(L) = \{ x \in \Sigma^* : x \text{ is a factor of some } y \in L \} \]
\[ \text{subw}(L) = \{ x \in \Sigma^* : x \text{ is a subword of some } y \in L \} \]
\[ L^R = \{ x \in \Sigma^* : x^R \in L \} \]
Kuratowski identities

We now consider the set $S = \{k, e, c, p, s, f, w, r\}$.

**Lemma**

The 14 operations $k, e, p, s, f, w, kp, ks, kf, kw, ep, es, ef, and ew$ are closure operations.

**Theorem (mentioned above)**

Let $x, y$ be closure operations. Then $xcycxcy \equiv xcy$.

Together, these two results thus give $196 = 14^2$ separate identities.
Further identities

**Lemma**

Let \( a \in \{ k, e \} \) and \( b \in \{ p, s, f, w \} \). Then \( aba \equiv bab \equiv ab \).

In a similar fashion, we obtain many kinds of Kuratowski-style identities involving the operations \( k, e, c, p, s, f, w, \) and \( r \).

**Proposition**

Let \( a \in \{ k, e \} \) and \( b \in \{ p, s, f, w \} \). Then we have the following identities:

- \( abcacaca \equiv abca \)
- \( bcbcbcab \equiv bcab \)
- \( abcbcabcab \equiv abcab \)
Additional identities (I)

We obtain many additional identities connecting the operations $k, e, c, p, s, f, w,$ and $r$.

**Proposition**

We have the following identities:

- $rp ≡ sr; \ rs ≡ pr$
- $rf ≡ fr; \ rw ≡ wr; \ rc ≡ cr; \ rk ≡ kr$
- $ps ≡ sp ≡ pf ≡ fp ≡ sf ≡ fs ≡ f$
- $pw ≡ wp ≡ sw ≡ ws ≡ fw ≡ wf ≡ w$
- $rkw ≡ kw ≡ wk$
- $ek ≡ ke ≡ k$
- $fks ≡ pks; \ fkp ≡ skp$
- $rkf ≡ skf ≡ pkf ≡ fkk ≡ kf$
Additional identities (II)

Proposition

For all languages $L$, we have

- $pcs(L) = \Sigma^*$ or $\emptyset$.
- The same result holds for $pcf$, $fcs$, $fcf$, $scp$, $scf$, $fcp$, $wcp$, $wcs$, $wcf$, $pcw$, $scw$, $fcw$, and $wcw$.

Let's prove this for $pcs$:

If $s(L) = \Sigma^*$, then $cs(L) = \emptyset$ and $pcs(L) = \emptyset$.

Otherwise, $s(L)$ omits some word $w$.
Then $s(L) \cap \Sigma^* w = \emptyset$.
Then $\Sigma^* w \subseteq cs(L)$.
Then $\Sigma^* = p(\Sigma^* w) \subseteq pcs(L)$, hence $pcs(L) = \Sigma^*$. 
Proposition

For all languages $L$, we have

★ $sckp(L) = \Sigma^* \text{ or } \emptyset$.

★ The same result holds for $fckp$, $pcks$, $fcks$, $pckf$, $sckf$, $fckf$, $wckp$, $wcks$, $wckf$, $wckw$, $pckw$, $sckw$, $fckw$.

Proposition

For all languages $L$, we have

★ $scskp(L) = \Sigma^* \text{ or } \emptyset$.

★ The same result holds for $pcpks$. 
Additional identities (IV)

**Proposition**

*For all languages* \( L \) *and for all* \( b \in \{ p, s, f, w \} \), *we have*

\[
\begin{align*}
\text{kcb}(L) &= \text{cb}(L) \cup \{\epsilon\} \\
\text{kckb}(L) &= \text{ckb}(L) \cup \{\epsilon\} \\
\text{kckck}(L) &= \text{ckck}(L) \cup \{\epsilon\} \\
\text{kbcbckb}(L) &= \text{bcbckb}(L) \cup \{\epsilon\}.
\end{align*}
\]

Let’s prove \( \text{kcp}(L) \subseteq \text{cp}(L) \cup \{\epsilon\} \):

Assume \( x \in \text{kcp}(L) \) and \( x \neq \epsilon \).

We have \( x = x_1x_2 \cdots x_n \) for some \( n \geq 1 \), where each \( x_i \in \text{cp}(L) \).

Then \( x_1x_2 \cdots x_n \notin \text{p}(L) \), because if it were, then \( x_1 \in \text{p}(L) \).

Hence \( x \in \text{cp}(L) \).
Main Result

Theorem (C-Domaratzki-Harju-Shallit 2011)

Let $S = \{k, e, c, p, f, s, w, r\}$. Then for every language $L$, the orbit $\mathcal{O}_S(L)$ contains at most 5676 distinct languages.
Sketch of the proof

We used breadth-first search to examine the set $S^* = \{k, e, c, p, f, s, w, r\}^*$ w.r.t. the radix order with $k < e < c < p < f < s < w < r$.

As each new word $x$ is examined, we test it to see if any factor is of the form given by “certain identities”.

If it is, then the corresponding language must be either $\Sigma^*$, $\emptyset$, $\{\epsilon\}$, or $\Sigma^+$; furthermore, each descendant language will be of this form. In this case the word $x$ is discarded.

Otherwise, we use the remaining identities to try to reduce $x$ to an equivalent word that we have previously encountered. If we succeed, then $x$ is discarded.

Otherwise we append all the words in $Sx$ to the end of the queue.
If the process terminates, then $\mathcal{O}_S(L)$ is of finite cardinality.

For $S = \{k, e, c, p, f, s, w, r\}$, the process terminated with 5672 nodes that could not be simplified using our identities. We did not count $\emptyset, \{\epsilon\}, \Sigma^+, \text{ and } \Sigma^*$. The total is thus 5676.

(The longest word examined was $ckpcscpckpckpcpckckcr$, of length 25, and the same word with $p$ replaced by $s$.)
If we use two arbitrary closure operations $a$ and $b$ with no relation between them, then the monoid generated by $\{a, b\}$ is infinite, since any two finite prefixes of $ababab \cdots$ are distinct.

Example
Define the exponentiation operation

$$t(L) = \{x^i : x \in L \text{ and } i \text{ is an integer } \geq 1\}.$$  

Then $t$ is a closure operation.

Hence the orbits $O_{\{p\}}(L)$ and $O_{\{t\}}(L)$ are finite, for all $L$.

However, if $M = \{ab\}$, then the orbit $O_{\{p,t\}}(M)$ is infinite, as

$$aba^i \in (pt)^i(M) \text{ and } aba^i \not\in (pt)^j(M) \text{ for } j < i.$$  

In this case at most 14 distinct languages can be generated. The bound of 14 can be achieved, e.g., by the regular language over $\Sigma = \{a, b, c, d\}$ accepted by the following DFA:

```
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td>a</td>
<td>a</td>
<td>a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
The following table gives the appropriate set of final states under the operations.

<table>
<thead>
<tr>
<th>language</th>
<th>final states</th>
<th>language</th>
<th>final states</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>3,7,8</td>
<td>$pcpc(L)$</td>
<td>1,5,6,7</td>
</tr>
<tr>
<td>$c(L)$</td>
<td>1,2,4,5,6</td>
<td>$cpcp(L)$</td>
<td>2,3,6,7</td>
</tr>
<tr>
<td>$p(L)$</td>
<td>1,2,3,5,6,7,8</td>
<td>$cpcpc(L)$</td>
<td>2,3,4,8</td>
</tr>
<tr>
<td>$pc(L)$</td>
<td>1,2,3,4,5,6,8</td>
<td>$pcpcp(L)$</td>
<td>1,2,3,5,6,7</td>
</tr>
<tr>
<td>$cp(L)$</td>
<td>4</td>
<td>$pcpcpc(L)$</td>
<td>1,2,3,4,5,8</td>
</tr>
<tr>
<td>$cpc(L)$</td>
<td>7</td>
<td>$cpcpcp(L)$</td>
<td>4, 8</td>
</tr>
<tr>
<td>$pcp(L)$</td>
<td>1,4,5,8</td>
<td>$cpcpcpc(L)$</td>
<td>6, 7</td>
</tr>
</tbody>
</table>
Here there are at most 13 distinct languages, given by the action of
\[ \{ \epsilon, k, p, s, f, kp, ks, kf, pk, sk, fk, pks, skp \} . \]
The bound of 13 is achieved, for example, by \( L = \{ abc \} . \)
Here breadth-first search gives 78 languages, so our bound is $78 + 4 = 82$. We can improve this bound by considering new kinds of arguments.

**Lemma**

*For all languages $L$, we have either $f(L) = \Sigma^*$ or $fc(L) = \Sigma^*$.***

**Theorem (C-Domaratzki-Harju-Shallit 2011)**

*Let $L$ be an arbitrary language. Then 50 is a tight upper bound for the size of $O\{k,c,f\}(L)$.***
Sketch of the proof

The languages in $O_{\{k,c,f\}}(L)$ that may differ from $\Sigma^*, \emptyset, \Sigma^+$, and $\{\epsilon\}$ are among the images of $L$ and $c(L)$ under the 16 operations

$$f, kf, kckf, kcf, fk, kcfk, fck, kfck, kckfck, kcfck, \quad (1)$$

$$fkck, kcfkck, fckck, kfckck, kckfckck, kcfckck,$$

the complements of these images, together with the 14 languages in $O_{\{k,c\}}(L)$.

We show that there are at most 32 distinct languages among the $64 = 16 \cdot 4$ languages given by the images of $L$ and $c(L)$ under the 16 operations (1) and their complements.

Adding the 14 languages in $O_{\{k,c\}}(L)$, and $\Sigma^*, \emptyset, \Sigma^+$, and $\{\epsilon\}$, we obtain that $50 = 32 + 14 + 4$ is an upper bound for the size of the orbit of $\{k, c, f\}$. 
This DFA accepts a language $L$ with orbit size 50 under $\{k, c, f\}^*$. 
### Summary of results

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>r</strong></td>
<td>2</td>
<td><strong>w</strong></td>
<td>2</td>
<td><strong>f</strong></td>
<td>2</td>
</tr>
<tr>
<td><strong>s</strong></td>
<td>2</td>
<td><strong>p</strong></td>
<td>2</td>
<td><strong>c</strong></td>
<td>2</td>
</tr>
<tr>
<td><strong>k</strong></td>
<td>2</td>
<td><strong>w, r</strong></td>
<td>4</td>
<td><strong>f, r</strong></td>
<td>4</td>
</tr>
<tr>
<td><strong>f, w</strong></td>
<td>3</td>
<td><strong>s, w</strong></td>
<td>3</td>
<td><strong>s, f</strong></td>
<td>3</td>
</tr>
<tr>
<td><strong>p, w</strong></td>
<td>3</td>
<td><strong>p, f</strong></td>
<td>3</td>
<td><strong>c, r</strong></td>
<td>4</td>
</tr>
<tr>
<td><strong>c, w</strong></td>
<td>6*</td>
<td><strong>c, f</strong></td>
<td>6*</td>
<td><strong>c, s</strong></td>
<td>14</td>
</tr>
<tr>
<td><strong>c, p</strong></td>
<td>14</td>
<td><strong>k, r</strong></td>
<td>4</td>
<td><strong>k, w</strong></td>
<td>4</td>
</tr>
<tr>
<td><strong>k, f</strong></td>
<td>5</td>
<td><strong>k, s</strong></td>
<td>5</td>
<td><strong>k, p</strong></td>
<td>5</td>
</tr>
<tr>
<td><strong>k, c</strong></td>
<td>14</td>
<td><strong>f, w, r</strong></td>
<td>6</td>
<td><strong>s, f, w</strong></td>
<td>4</td>
</tr>
<tr>
<td><strong>p, f, w</strong></td>
<td>4</td>
<td><strong>p, s, f</strong></td>
<td>4</td>
<td><strong>c, w, r</strong></td>
<td>10*</td>
</tr>
<tr>
<td><strong>c, f, r</strong></td>
<td>10*</td>
<td><strong>c, f, w</strong></td>
<td>8*</td>
<td><strong>c, s, w</strong></td>
<td>16*</td>
</tr>
<tr>
<td><strong>c, s, f</strong></td>
<td>16*</td>
<td><strong>c, p, w</strong></td>
<td>16*</td>
<td><strong>c, p, f</strong></td>
<td>16*</td>
</tr>
<tr>
<td><strong>k, w, r</strong></td>
<td>7</td>
<td><strong>k, f, r</strong></td>
<td>9</td>
<td><strong>k, f, w</strong></td>
<td>6</td>
</tr>
<tr>
<td><strong>k, s, w</strong></td>
<td>7</td>
<td><strong>k, s, f</strong></td>
<td>9</td>
<td><strong>k, p, w</strong></td>
<td>7</td>
</tr>
<tr>
<td><strong>k, p, f</strong></td>
<td>9</td>
<td><strong>k, c, r</strong></td>
<td>28</td>
<td><strong>k, c, w</strong></td>
<td>38*</td>
</tr>
<tr>
<td><strong>k, c, f</strong></td>
<td>50*</td>
<td><strong>k, c, s</strong></td>
<td>1070</td>
<td><strong>k, c, p</strong></td>
<td>1070</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>-----</td>
<td>------------------</td>
<td>-----</td>
<td>------------------</td>
<td>-----</td>
</tr>
<tr>
<td>$p, s, f, r$</td>
<td><strong>8</strong></td>
<td>$p, s, f, w$</td>
<td><strong>5</strong></td>
<td>$c, f, w, r$</td>
<td><strong>12</strong></td>
</tr>
<tr>
<td>$c, s, f, w$</td>
<td><strong>16</strong></td>
<td>$c, p, f, w$</td>
<td><strong>16</strong></td>
<td>$c, p, s, f$</td>
<td><strong>16</strong></td>
</tr>
<tr>
<td>$k, f, w, r$</td>
<td><strong>11</strong></td>
<td>$k, s, f, w$</td>
<td><strong>10</strong></td>
<td>$k, p, f, w$</td>
<td><strong>10</strong></td>
</tr>
<tr>
<td>$k, p, s, f$</td>
<td><strong>13</strong></td>
<td>$k, c, w, r$</td>
<td><strong>72</strong></td>
<td>$k, c, f, r$</td>
<td><strong>84</strong></td>
</tr>
<tr>
<td>$k, c, f, w$</td>
<td><strong>66</strong></td>
<td>$k, c, s, w$</td>
<td><strong>1114</strong></td>
<td>$k, c, s, f$</td>
<td><strong>1450</strong></td>
</tr>
<tr>
<td>$k, c, p, w$</td>
<td><strong>1114</strong></td>
<td>$k, c, p, f$</td>
<td><strong>1450</strong></td>
<td>$p, s, f, w, r$</td>
<td><strong>10</strong></td>
</tr>
<tr>
<td>$c, p, s, f, r$</td>
<td><strong>30</strong></td>
<td>$c, p, s, f, w$</td>
<td><strong>16</strong></td>
<td>$k, p, s, f, r$</td>
<td><strong>25</strong></td>
</tr>
<tr>
<td>$k, p, s, f, w$</td>
<td><strong>14</strong></td>
<td>$k, c, f, w, r$</td>
<td><strong>120</strong></td>
<td>$k, c, s, f, w$</td>
<td><strong>1474</strong></td>
</tr>
<tr>
<td>$k, c, p, f, w$</td>
<td><strong>1474</strong></td>
<td>$k, c, p, s, f$</td>
<td><strong>2818</strong></td>
<td>$c, p, s, f, w, r$</td>
<td><strong>30</strong></td>
</tr>
<tr>
<td>$k, p, s, f, w, r$</td>
<td><strong>27</strong></td>
<td>$k, c, p, s, f, r$</td>
<td><strong>5628</strong></td>
<td>$k, c, p, s, f, w$</td>
<td><strong>2842</strong></td>
</tr>
<tr>
<td>$k, c, p, s, f, w, r$</td>
<td><strong>5676</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Further work

We plan to continue to refine our estimates of the previous tables, and pursue the status of other sets of operations.

For example, if $t$ is the exponentiation operation, then, using the identities $kt \equiv tk \equiv k$, and the inclusion $t \subseteq k$, we get the additional Kuratowski-style identities

- $kctckck \equiv kck$,
- $kckctck \equiv kck$,
- $kctctck \equiv kck$,
- $tctctck \equiv tck$,
- $kctctct \equiv kct$.

This allows us to prove that $O_{\{k,c,t\}}(L)$ is finite and of cardinality at most 126.