Tackling the Generalized Star-Height Problem

Tom Bourne

School of Mathematics and Statistics
University of St Andrews

NBSAN, St Andrews
22nd April 2015

Regular Expressions

Given a finite alphabet A, we define $\emptyset$, ε (the empty word), and a in A to be basic regular expressions.

If E and F are regular expressions then we recursively define new regular expressions by using the following operations:

- EF (concatenation)
- $E \cup F$ (set union)
- E^* (Kleene star)

We use regular expressions to represent regular languages, where a language is any subset of the free monoid generated by A.

For example, if $A = \{a, b\}$ then $A^*a = (a \cup b)^*a$ represents the language in which all words end with the letter a.

Star-Height

The star-height $h(E)$ of a regular expression E is defined recursively as follows:

- $h(\emptyset) = h(\varepsilon) = h(a) = 0$, where $a \in A$;
- $h(EF) = h(E \cup F) = \max\{h(E), h(F)\}$;
- $h(E^*) = h(E) + 1$.

Then, for a language $L \subseteq A^*$, we define the star-height of L by

$$h(L) = \min\{h(E) \mid E \text{ is a regular expression for } L\}.$$

It is best to think of the star-height of L as the nesting depth of Kleene stars in the regular expression representing L that features the fewest Kleene stars.

Generalized Star-Height

Now suppose that in addition to the aforementioned operations for defining regular expressions, we also allow complementation; that is, if E is a regular expression then so is E^c.

Including the complement operation leads us to refer to E as a generalized regular expression.

We then define $h(E^c) = h(E)$, and define the generalized star-height of a language L as in the restricted case.

Note that, by De Morgan's laws, we can now freely use the intersection ($\cap$) and set difference ($\setminus$) operations when dealing with regular expressions. It follows that

$$h(E \cap F) = h(E \setminus F) = \max\{h(E), h(F)\}.$$

The Generalized Star-Height Problem

A language which has (generalized) star-height 0 is said to be star-free. We have the following result:

Theorem (Schützenberger (1965))

A language is star-free if and only if its syntactic monoid is finite and aperiodic.

This theorem gives us an algorithm for deciding whether a language has (generalized) star-height 0.

The Generalized Star-Height Problem

Does there exist a language of generalized star-height greater than 1?

Known Results

Theorem (Eggan (1963))

For every natural number n, there exists a regular language of restricted star-height n.

Theorem (Henneman (1971))

A regular language recognized by a finite commutative group is of generalized star-height at most 1.

Known Results

Theorem (Eggan (1963))

For every natural number n, there exists a regular language of restricted star-height n.

Theorem (Pin, Straubing, Thérien (1989))

A regular language recognized by a finite nilpotent group of class $0, 1$ *or* 2 *is of generalized star-height at most 1.*

Theorem (Pin, Straubing, Thérien (1989))

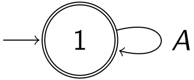
Every regular language recognized by a group of order less than 12 is of generalized star-height at most 1.

Removing Stars I

Lemma

For any finite alphabet A, the language $L = A^*$ is star-free.

The minimal automaton recognizing L is



The syntactic monoid of L is the trivial monoid, which is finite and aperiodic, so L must be star-free.

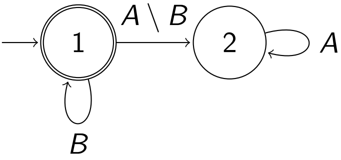
A star-free expression for L is $\emptyset^c$.

Removing Stars II

Lemma

For any finite alphabet A and any subset B of A, we have $h(B^*) = 0$.

The minimal automaton recognizing B^* is



The syntactic monoid of B^* is $M(B^*) = \langle x \mid x^2 = x \rangle$, which is finite and aperiodic, so B^* must be star-free.

A star-free expression for B^* is $(\emptyset^c\,(A \setminus B)\,\emptyset^c)^c$.

Counting Subwords of Length Two: Case I

Let A be a finite alphabet. For every word v in A^* and for any integers k and n such that $0 \leq k < n$ we define

$$L(v, k, n) = \{w \in A^* \mid |w|_v \equiv k \bmod n\}.$$

For $a, b \in A$ with $a \neq b$, define $U \subset A^*$ to be the set of all words that do not feature ab as a subword.

A generalized regular expression for U is $(\emptyset^c ab \emptyset^c)^c$, which implies that U is star-free.

Knowing this, we can obtain an expression for $L(ab, k, n)$ of star-height one:

$$L(ab, k, n) = (Uab)^k \left((Uab)^n\right)^* U.$$

Counting Subwords of Length Two: Case II

Define

$$B = A \setminus \{a\},$$
$$U = A^* \setminus A^* a^2 A^* = \left(\emptyset^c a^2 \emptyset^c\right)^c,$$

both of which are star-free. Let $W = B \cup BUB = B(\varepsilon \cup UB)$.

Let $L'(a^2, k)$ be the set of words that begin and end with a^2 or a higher power of a and contain precisely k occurrences of a^2.

k	$L'(a^2, k)$
1	a^2
2	$a^3 \cup a^2 W a^2$
3	$a^4 \cup a^3 W a^2 \cup a^2 W a^3 \cup a^2 W a^2 W a^2$

Counting Subwords of Length Two: Case II

In general, we have that

$$L'(a^2, k) = \bigcup_{r=1}^{k} \bigcup_{\substack{k_1, k_2, \ldots, k_r \geq 2 \\ k_1 + k_2 + \cdots + k_r = k + r}} a^{k_1} W a^{k_2} W \cdots W a^{k_r}.$$

Note that this expression is star-free.

Now, a star-free expression for all words that have precisely k occurrences of a^2 as a subword, denoted by $L(a^2, k)$, is

$$L(a^2, k) = (\varepsilon \cup UB) \cdot L'(a^2, k) \cdot (BU \cup \varepsilon).$$

Counting Subwords of Length Two: Case II

Let $M(a^2, n)$ denote the set of words such that $a \cdot M(a^2, n)$ contains precisely n occurrences of a^2.

n	$M(a^2, n)$
2	$a^2 \cup aWa^2 \cup W(a^3 \cup a^2Wa^2)$
3	$a^3 \cup a^2Wa^2 \cup aW(a^3 \cup a^2Wa^2)$ $\cup\ W(a^4 \cup a^3Wa^2 \cup a^2Wa^3 \cup a^2Wa^2Wa^2)$

In general, we have that

$$M(a^2, n) = a^n \cup \left(\bigcup_{i=1}^{n} a^{n-i} W \cdot L'(a^2, i) \right).$$

Counting Subwords of Length Two: Case II

For $L(a^2, k, n)$, where $0 < k < n$, we have

$$L(a^2, k, n) = (\varepsilon \cup UB) \cdot L'(a^2, k) \cdot M(a^2, n)^* \cdot (BU \cup \varepsilon)\,.$$

When $k = 0$ we have

$$L(a^2, 0, n) = U \cup (\varepsilon \cup UB) \cdot L'(a^2, n) \cdot M(a^2, n)^* \cdot (BU \cup \varepsilon)\,.$$

Both of these expressions are of star-height one, so the languages that they represent are of star-height at most one.

Current and Future Research

Theorem (proof under construction)

Regular languages recognized by Rees matrix semigroups over cyclic groups are of star-height at most 1.

Three Month Plan

- For any alphabet A, can we describe all $B \subseteq A^n$, where $n \in \mathbb{N}$, that satisfy $h(B^*) = 0$?
- What star-height do languages recognized by Rees matrix semigroups over abelian groups have?
- What about Rees 0-matrix semigroups?