International Journal of Foundations of Computer Science
Vol. 22, No. 2 (2011) 301–321

DOI: 10.1142/S0129054111008052

World Scientific
www.worldscientific.com

CLOSURES IN FORMAL LANGUAGES AND KURATOWSKI'S THEOREM

JANUSZ BRZOZOWSKI, ELYOT GRANT and JEFFREY SHALLIT
David R. Cheriton School of Computer Science, University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
{*brzozo,egrant,shallit*}*@cs.uwaterloo.ca*

Received 1 June 2009
Accepted 14 July 2010
Communicated by Volker Diekert and Dirk Nowotka

A famous theorem of Kuratowski states that, in a topological space, at most 14 distinct sets can be produced by repeatedly applying the operations of closure and complement to a given set. We re-examine this theorem in the setting of formal languages, where by "closure" we mean either Kleene closure or positive closure. We classify languages according to the structure of the algebras they generate under iterations of complement and closure. There are precisely 9 such algebras in the case of positive closure, and 12 in the case of Kleene closure. We study how the properties of being open and closed are preserved under concatenation. We investigate analogues, in formal languages, of the separation axioms in topological spaces; one of our main results is that there is a clopen partition separating two words if and only if the words do not commute. We can decide in quadratic time if the language specified by a DFA is closed, but if the language is specified by an NFA, the problem is PSPACE-complete.

Keywords: Closure system; complement; complexity; concatenation; decision problem; formal language; Kleene closure; Kuratowski's theorem; positive closure; separating two words.

2010 Mathematics Subject Classification: 68Q45, 54A05

1. Introduction

In a famous 1922 paper, Kuratowski proved that, if S is any set in a topological space, then at most 14 distinct sets can be produced by repeatedly applying the operations of topological closure and complement to S [10, 5]. Furthermore, there exist sets achieving this bound of 14 in many common topological spaces. There is a large and scattered literature on Kuratowski's theorem, most of which focuses on topological spaces; an admirable survey is the paper of Gardner and Jackson [6]. For analogous result on relations, see Graham, Knuth, and Motzkin [7].

The basic properties of closure systems and a version of Kuratowski's theorem in a general setting are presented in Section 2; this version can be found in Hammer [8]. Our point of view most closely matches that of Peleg [12], who briefly observed that

Kleene and positive closure are closure operators, and hence Kuratowski's theorem holds for them.

We discuss positive and Kleene closures in Section 3. In Section 4, we reconsider Kuratowski's theorem in the context of formal languages, where closure is replaced by Kleene or positive closure. We describe all possible algebras of languages generated by a language under the operations of complement and closure. We classify languages according to the structure of the algebras they generate, and give a language of each type (Theorems 18 and 23).

In Section 5 we study how the properties of being open and closed are preserved under concatenation. In Section 6 we investigate analogues, in formal languages, of the separation axioms in topological spaces; one of our main results (Theorem 32) is that there is a clopen partition separating two words if and only if the words do not commute. In Section 7 we show that we can decide in quadratic time if the language specified by a DFA is closed, but if the language is specified by an NFA, the problem is PSPACE-complete.

An earlier version of this work was presented at the 2009 DLT conference [2].

2. Closure Systems and Kuratowski's Theorem

We recall the definitions and properties of closures in general. Let S be a set, which we call the *universal set*. An operator $\Box$ applied to $X \subseteq S$ is denoted by $X^{\Box}$. Then a mapping $\Box : 2^S \to 2^S$ is a *closure operator* if and only if it satisfies the following, for all subsets X and Y of S:

$$\begin{array}{rl} X \subseteq X^{\Box} & (\Box \text{ is } \textit{extensive}); \\ X \subseteq Y \text{ implies } X^{\Box} \subseteq Y^{\Box} & (\Box \text{ is } \textit{isotone}); \\ X^{\Box\Box} = X^{\Box} & (\Box \text{ is } \textit{idempotent}). \end{array} \tag{1}$$

A pair $(S,^{\Box})$ satisfying (1) is a *closure system*. The set $X^{\Box}$ is the *closure* of X. We say X is *closed* if $X = X^{\Box}$. The *complement* $S \setminus X$ of a set $X \subseteq S$ is denoted X^{-}. The set X is *open* if its complement X^{-} is closed, and X is *clopen* if it is both open and closed. The *interior* of X, denoted X°, is defined to be $X^{-\Box-}$.

We now list some fundamental properties of closure systems; proofs are routine. Note the duality between closure and interior.

Proposition 1.

(a) The intersection of an arbitrary family of closed sets is closed.
(b) The union of an arbitrary family of open sets is open.

Proposition 2. *For $X \subseteq S$, the following sets are identical:*

(a) $X^{\Box} = X^{-\circ-}$;
(b) $\bigcap\{Y \subseteq S : Y \supseteq X \text{ and } Y \text{ is closed}\}$;
(c) $\{a \in S : \text{for all open } Y \subseteq S, a \in Y \text{ implies } Y \cap X \neq \emptyset\}$.

Proposition 3. *For $X \subseteq S$, the following sets are identical:*

(a) $X^\circ = X^{-\square-}$;
(b) $\bigcup\{Y \subseteq S : Y \subseteq X \text{ and } Y \text{ is open}\}$;
(c) $\{a \in S : \text{there exists an open } Y \subseteq X, \text{with } a \in Y\}$.

Proposition 4. *Let* $X, Y \subseteq S$. *Then the following hold:*

(a) $X^\square$ *is closed and* X° *is open.*
(b) $(X \cup Y)^\square = (X^\square \cup Y^\square)^\square$ *and* $(X \cap Y)^\circ = (X^\circ \cap Y^\circ)^\circ$.
(c) $(X \cap Y)^\square \subseteq X^\square \cap Y^\square$ *and* $(X \cup Y)^\circ \supseteq X^\circ \cup Y^\circ$.

A closure operator is *topological* if the union of a finite family of closed sets is always closed. We shall not assume that this property holds in general.

We now state two versions of Kuratowski's theorem. The first is equivalent to Kuratowski's original result [10]. With no additional work, this result can be generalized [8] to an arbitrary closure system that is not necessarily topological:

Theorem 5. *Let* $(S, \square)$ *be a closure system, and let* $X \subseteq S$. *Starting with* X, *apply the operations of closure and complement in any order, any number of times. Then at most* 14 *distinct sets are generated. Also, any* $X \subseteq S$ *satisfies*

$$X^{\square-\square-\square-\square} = X^{\square-\square}. \tag{2}$$

A closure operator $\square$ *preserves openness* if $X^\square$ is open for all open sets X, or equivalently, if Y° is closed for all closed sets Y. Hence, if $\square$ preserves openness, then $X^{\square\circ}$ and $X^{\circ\square}$ are clopen for all sets X. We will see later that the positive closure of languages preserves openness.

In 1983, Peleg [12] defined a closure operator to be *compact* if it satisfies Eq. (3) below. He showed that at most 10 different sets are generated if $\square$ is compact, and proved that $\square$ preserves openness if and only if it is compact. The following theorem is a modified version of Peleg's result:

Theorem 6. *Let* $(S, \square)$ *be a closure system such that* $\square$ *preserves openness, and let* $X \subseteq S$. *Starting with* X, *apply the operations of closure and complement in any order, any number of times. Then at most* 10 *distinct sets are generated. Also, any* $X \subseteq S$ *satisfies*

$$X^{\square-\square-\square} = X^{\square-\square-}. \tag{3}$$

3. Positive and Kleene Closures of Languages

We deal now with closures in the setting of formal languages. Our universal set is Σ^* for a finite non-empty alphabet Σ. For $L \subseteq \Sigma^*$, we define the complement $L^- = \Sigma^* \setminus L$, the positive closure $L^+ = \bigcup_{i \geq 1} L^i$, and the Kleene closure $L^* = \bigcup_{i \geq 0} L^i$. One may easily verify that positive and Kleene closure are both closure operators.

We emphasize that the positive and Kleene closures are *not* topological, as the union of two closed languages is not necessarily closed. For example, observe that

$(aa)^+ \cup (aaa)^+ \subsetneq (aa \cup aaa)^+$, as a^5 belongs to the right-hand side but not the left. Hence languages *do not* form a topology under positive or Kleene closure.

A language is *positive-closed* if it is a closed set under positive closure. It is *positive-open* if its complement is positive-closed. The terms *Kleene-closed*, and *Kleene-open* are defined analogously.

Proposition 7. *Let $L \subseteq \Sigma^*$. The following are equivalent:*

(a) L is positive-closed.
(b) $L \cup \{\epsilon\}$ is Kleene-closed.
(c) For all $u, v \in L$, we have $uv \in L$.

If L is positive-closed, then so are $L \setminus \{\epsilon\}$ and $L \cup \{\epsilon\}$. Consequently, there is an obvious 2-to-1 mapping between positive-closed and Kleene-closed languages—positive-closed languages may or may not contain ϵ, and Kleene-closed languages must. Since positive and Kleene closure are similar, hereafter we restrict our attention to positive closure since our theorems can then be stated without worrying about ϵ. For the remainder of this article, a language is *closed* if it is positive-closed, *open* if it is positive-open, and *clopen* if it is both positive-closed and positive-open.

We define idempotent interior operators as well. The *positive interior* of a language L is $L^{\oplus} = L^{-+-}$; the *Kleene interior* is $L^{\circledast} = L^{-*-}$. We note the following:

Proposition 8. *Let $L \subseteq \Sigma^*$. The following are equivalent:*

(a) L is positive-open (in other words, $L = L^{\oplus}$).
(b) $L \setminus \{\epsilon\}$ is Kleene-open.
(c) For all $u, v \in \Sigma^*$ such that $uv \in L$, we have $u \in L$ or $v \in L$.

In the 1970's, D. Forkes proved Eq. (2) with the Kleene closure as $\Box$, and the first author then proved that Eq. (3) holds when $\Box$ is positive closure. (They were both unaware of [10].) Peleg [12] proved this over a wider class of operators. Since Eq. (3) holds if and only if a closure operator preserves openness, we have:

Theorem 9. *Let $L \subseteq \Sigma^*$ be open. Then L^+ is open.*

Corollary 10. *Let $L \subseteq \Sigma^*$. Then $L^{+\oplus}$ and $L^{\oplus +}$ are clopen. Moreover, if L is open, then L^+ is clopen, and if L is closed, then $L^{\oplus}$ is clopen.*

The converses of the above are false. For example, the language $\{a, aaaa\}$ is not open, but its closure is clopen. We discuss such possibilities in the next section.

We now present several examples of clopen, open and closed languages.

Example 11. Clopen languages: *Let Σ be an alphabet and let $\Sigma_1, \Sigma_2 \subseteq \Sigma$ be subalphabets. For $w \in \Sigma^*$, let $w[i]$ denote the i'th letter of w, and for $a \in \{1, 2\}$, let $|w|_a$ denote the number of distinct values of i for which $w[i] \in \Sigma_a$. Suppose $k \geq 0$. Then $L = \{w \in \Sigma^* : |w|_1 < k|w|_2\}$ is clopen.*

To see this, observe that if $|u|_1 < k|u|_2$ *and* $|v|_1 < k|v|_2$*, then* $|uv|_1 <= k|uv|_2$*, and thus* L *is closed. By a similar argument,* L^- *is closed, so* L *is clopen.*

Example 12. Open languages: *A language* L *is* prefix-closed *if and only if for every* $w \in L$*, each prefix of* w *is in* L*. We analogously define* suffix-closed, subword-closed, *and* factor-closed *languages. Here by subword, we mean an arbitrary subsequence, and by factor, we mean a contiguous subsequence. For any* $L \subseteq \Sigma^*$*, if* L *is prefix-, suffix-, factor-, or subword-closed, then* L *is open by Proposition 8.*

Example 13. Closed languages: *Left ideals (those languages* L *satisfying* $L = \Sigma^* L$*), right ideals (*$L = L\Sigma^*$*), two-sided ideals (*$L = \Sigma^* L \Sigma^*$*), or languages of the form* $L = \bigcup_{a_1 \cdots a_n \in L} \Sigma^* a_1 \Sigma^* \cdots \Sigma^* a_n \Sigma^*$*, all satisfy* $L = L^+$*, and so are all positive-closed.*

Example 14. Closures of open languages: *By Corollary 10, the closure of any open language is clopen. Consequently, we may obtain clopen languages by applying positive closure to any prefix-, suffix-, factor-, or subword-closed language. For example, if* w *is any (possibly infinite) word, then the set of all words that can be factored into prefixes of* w *is a clopen language.*

Our next example requires some explanation. Suppose we are given a word w, and we wish to determine *all* the open languages containing w that are minimal with respect to set inclusion. For simplicity, we only analyze the case in which $w = a_1 a_2 \cdots a_n$ where each $a_i \in \Sigma$ and $a_i \neq a_j$ for all $i \neq j$.

By part (c) of Proposition 8, if L is an open language containing w, then for all $1 \leq i < n$, either $a_1 \cdots a_i \in L$ or $a_{i+1} \cdots a_n \in L$. Additionally, for all $1 \leq i < j \leq n$, if $a_1 \cdots a_i \notin L$ and $a_j \cdots a_n \notin L$, then $a_{i+1} \cdots a_{j-1} \in L$ (we obtain this result by applying part (c) of Proposition 8 to $a_{i+1} \cdots a_n$). Motivated by these observations, we define a language $L \subseteq \Sigma^+$ to be a *w-core* if all of the following hold:

(a) $w \in L$ and every word in L is a factor of w.
(b) For all $1 \leq i < n$, exactly one of $a_1 \cdots a_i$ or $a_{i+1} \cdots a_n$ is in L.
(c) For all $1 \leq i < j \leq n$, L contains the word $a_{i+1} \cdots a_{j-1}$ if and only if $a_1 \cdots a_i \notin L$ and $a_j \cdots a_n \notin L$.

By our argument above, all open languages containing w must contain some w-core as a sublanguage. Moreover, there are exactly 2^{n-1} unique w-cores, determined exclusively by the $n-1$ choices made in item (b) of our definition above. One possible w-core is the set of all prefixes of w—this occurs precisely when we always include $a_1 \cdots a_i$, but never $a_{i+1} \cdots a_n$. Analogously, the set of all suffixes of w is a w-core.

Moreover, we can show that every w-core is, in fact, open. Again, we employ part (b) of Proposition 8. Let L be a w-core and let $x \in L$. We write $x = uv$ for words u and v, and write $u = a_i \cdots a_k$ and $v = a_{k+1} \cdots a_j$ for some $1 \leq i \leq k < j \leq n$, since uv must be a factor of w. If $u \notin L$, then by items (b) and (c) of our definition of a w-core, we must have either $i > 1$ and $a_1 \cdots a_{i-1} \in L$, or $a_{k+1} \cdots a_n \in L$.

However, if $i > 1$ and $a_1 \cdots a_{i-1} \in L$, then we cannot have $uv \in L$ by items (b) and (c), so we must instead have $a_{k+1} \cdots a_n \in L$. Similarly, if $v \notin L$, then we must have $a_1 \cdots a_k \in L$. Item (b) of our definition implies that we cannot have both $a_1 \cdots a_k \in L$ and $a_{k+1} \cdots a_n \in L$, so we must have one of $u \in L$ or $v \in L$. Thus L is open. We may therefore conclude the following:

Example 15. Minimal open languages containing a word: *Let $w = a_1 a_2 \cdots a_n$ where $a_i \in \Sigma$ for all i and $a_i \neq a_j$ for all $i \neq j$. Then the open languages containing w that are minimal with respect to set inclusion are the 2^{n-1} w-cores.*

4. Kuratowski's Theorem for Languages

For any language L, let $A(L)$ be the family of all languages generated from L by complementation and positive closure. Since positive closure preserves openness, Theorem 6 implies that $A(L)$ contains at most 10 languages. As we will see, this upper bound is tight. Moreover, we will show that there are precisely 9 distinct finite algebras $(A(L),^+,^-)$. Since the languages in $A(L)$ occur in complementary pairs, $A(L)$ must contain 2, 4, 6, 8, or 10 distinct languages. We will provide a list of conditions that classify languages according to the structure of $(A(L),^+,^-)$, and thus completely describe the circumstances under which $|A(L)|$ is equal to 2, 4, 6, 8, or 10.

We will also explore Kleene closure, where there are subtle differences. Let $D(L)$ be the family of all languages generated from L by complementation and Kleene closure. Kleene closure does not preserve openness, since Kleene-closed languages contain ϵ and Kleene-open languages do not. Therefore we must fall back to Theorem 5, which implies that $D(L)$ contains at most 14 languages, and we will show that this bound is also tight. We will show that there are precisely 12 distinct finite algebras $(D(L),^*,^-)$. We describe these algebras by relating them to those in the positive case.

In a sense, our results are the formal language analogues of topological results obtained by Chagrov [3] and discussed in [6]. Peleg [12] noted the tightness of the bounds of 10 and 14 in the positive and Kleene cases, but went no further.

4.1. *Structures of the algebras with positive closure and complement*

We may better understand the structure of $A(L)$ by first analyzing a related algebra of languages. Let $B(L)$ be the family of all languages generated from L by positive closure and positive interior, and let $C(L) = \{M : M^- \in B(L)\}$ be their complements. Since the closure and interior operators fix the languages $L^{+\oplus}$ and $L^{\oplus+}$ (which are clopen by Corollary 10), it follows that $B(L) = \{L, L^+, L^{+\oplus}, L^{\oplus}, L^{\oplus+}\}$. Of course, these five languages may not all be distinct. However, we can show that it suffices to analyze the structure of $B(L)$ to determine the structure of $A(L)$.

Lemma 16. *Let $L \subseteq \Sigma^*$. Then $A(L) = B(L) \cup C(L)$, and the union is disjoint.*

Proof. Clearly $A(L) \supseteq B(L) \cup C(L)$, since any language generated from L by closure, interior, and complement can be generated using only closure and complement via the identity $L^{\oplus} = L^{-+-}$. To prove the reverse inclusion, we let $M \in A(L)$. Then there is some string of symbols $z \in \{+,-\}^*$ such that $M = L^z$. We construct a string $z' \in \{+,-,\oplus\}^*$ by starting with z and repeatedly replacing all instances of $-+$ by $\oplus-$ and all instances of $-\oplus$ by $+-$, until no such replacements are possible. Since $L^{-+} = L^{\oplus-}$ and $L^{-\oplus} = L^{+-}$, we have $M = L^{z'}$. However, in producing z', we effectively shuffle all complements to the right. Consequently, the operation performed by z' is a series of positive closures and interiors followed by an even or odd number of complements. Hence either $M \in B(L)$ or $M \in C(L)$.

We now prove that $B(L) \cap C(L) = \emptyset$. We note that $L^{\oplus} \subseteq L^{\oplus+}$, $L^{\oplus} \subseteq L \subseteq L^{+}$, and $L^{\oplus} \subseteq L^{+\oplus}$ by isotonicity. Hence $L^{\oplus} \subseteq M$ for all $M \in B(L)$. Thus for two languages in $B(L)$ to be complements, $L^{\oplus}$ must be empty. Then L contains no strings of length 1, and thus neither do L^{+} and $L^{+\oplus}$. Then no language in $B(L)$ contains a string of length 1, so no two languages in $B(L)$ are complements. ■

The fact that $B(L) \cap C(L) = \emptyset$ can be proven in a more straightforward manner, but the supplied proof generalizes to Kleene closure, which we require later.

The disjointness of $B(L)$ and $C(L)$ is a property of formal languages that is crucial to our analysis. In general closure systems, the intersection of $B(L)$ and $C(L)$ may be non-empty. For example, consider the real numbers under the usual topology. The rational numbers are then a set whose interior is the complement of its closure.

Lemma 16 implies that $|A(L)| = 2|B(L)|$. Moreover, there is an exact 1-to-2 correspondence between the languages in $B(L)$ and $A(L)$: each language in $B(L)$ can be associated to itself and its complement. Hence the algebra $(A(L),^{+},^{-})$ can be constructed by simply merging the two algebras $(B(L),^{+},^{\oplus})$ and $(C(L),^{+},^{\oplus})$ and adding the complement operator. Thus we have reduced the problem of describing all algebras $(A(L),^{+},^{-})$ to the simpler task of describing the algebras $(B(L),^{+},^{\oplus})$. Before we proceed, we need to exclude a possible case via the following:

Lemma 17. *Suppose $L \subseteq \Sigma^*$. If L^{+} and $L^{\oplus}$ are both clopen, then L must be clopen.*

Proof. Seeking a contradiction, we assume that both L^{+} and $L^{\oplus}$ are clopen but L is not clopen. Then L is neither open nor closed (otherwise L is L^{+} or $L^{\oplus}$, both of which are clopen.) If L is not open, then $L \setminus L^{\oplus}$ is non-empty.

Let w be a shortest word in $L \setminus L^{\oplus}$. Consider $M = L^{\oplus} \cup \{w\}$. M is clearly a subset of L since $w \in L$, and $L^{\oplus}$ contains all open subsets of L by Proposition 3. Since $w \in M$ but $w \notin L^{\oplus}$, it follows that M is not open. Then Proposition 8 (c) must fail to hold for some word in M. But it holds for all words in $L^{\oplus}$ and thus must fail for w. Then there exist non-empty words x and y with $xy = w$, but $x \notin M$ and $y \notin M$. Then neither x nor y is in $L^{\oplus}$.

By our assumption that L^{+} is open, the fact that $w \in L^{+}$ implies that either $x \in L^{+}$ or $y \in L^{+}$. Without loss of generality, suppose that $x \in L^{+}$. Then x is the

concatenation of a list of words from L; we write $x = u_1 u_2 \cdots u_n$ with $u_i \in L$ for all $1 \leq i \leq n$. Then $|u_i| \leq |x| < |w|$ for all i, and thus $u_i \in L^{\oplus}$ for all i by our definition of w as the shortest word in $L \setminus L^{\oplus}$. However, x is then the concatenation of a list of words from $L^{\oplus}$ and is thus an element of $L^{\oplus +}$, which is $L^{\oplus}$ since we assumed $L^{\oplus}$ was closed. This is a contradiction since $x \notin L^{\oplus}$. ∎

Finally, we characterize the 9 possible algebras $(B(L),^{+},^{\oplus})$. Table 1 classifies all languages according to the structure of the algebras they generate and gives an example of each type. Here, we briefly explain our analysis. Clearly $B(L) = \{L\}$ if and only if L is clopen, giving Case (1). If L is open but not closed, then $B(L) = \{L, L^{+}\}$ since L^{+} must then be clopen. Similarly, if L is closed but not open, then $B(L) = \{L, L^{\oplus}\}$. These situations yield Cases (2) and (3). We henceforth assume that L is neither open nor closed, and thus L, $L^{\oplus}$, and L^{+} are all different. The remaining cases depend on the values of $L^{\oplus +}$ and $L^{+\oplus}$. Both must be clopen, so neither can equal L. Lemma 17 proves that $L^{\oplus}$ and L^{+} cannot both be clopen. If neither $L^{\oplus}$ nor L^{+} are clopen, then we have Case (8) if $L^{\oplus +}$ and $L^{+\oplus}$ are equal, and Case (9) if they are not. The remaining cases occur when one of L^{+} and $L^{\oplus}$ is clopen and the other is not. If L^{+} is clopen and $L^{\oplus}$ is not, then we get Case (4) if $L^{\oplus +} = L^{+}$ and Case (6) otherwise. Analogously, if $L^{\oplus}$ is clopen and L^{+} is not, then we get Case (5) if $L^{+\oplus} = L^{\oplus}$ and Case (7) otherwise.

We see that if $(B(L),^{+},^{\oplus})$ has algebraic structure (2), then $(C(L),^{+},^{\oplus})$ has structure (3); thus we say that Case (3) is the *dual* of Case (2). By examining the conditions under which each case holds, we can easily see that Cases (4) and (5) are also duals, as are Cases (6) and (7). Cases (1), (8), and (9) are self-dual. This notion is useful in constructing the algebra $(A(L),^{+},^{-})$: we connect an instance of $(B(L),^{+},^{\oplus})$ to its dual structure in the obvious way via the complement operator. Figure 1 gives an example of this for Case (6).

In summary, we have proven the following result:

Theorem 18. *Start with any language L, and apply the operators of positive closure and complement in Σ^* in any order, any number of times. Then at most* 10 *distinct languages are generated, and this bound is optimal. Furthermore, Table 1 describes the 9 algebras generated by this process, classifies languages according to the algebra they generate, and gives a language generating each algebra.*

In the unary case, we obtain the following:

Theorem 19. *Start with any unary language L, and apply the operators of positive closure and complement in Σ^* in any order, any number of times. Then at most* 6 *distinct languages are generated, and this bound is optimal. Furthermore, precisely Cases (1) through (5) in Table 1 are possible for a unary language.*

Proof. We assume that $L \subseteq a^*$ and consider the following two possibilities:

Case (i): $a \in L$. Then $L^{+} = aa^*$ or $L^{+} = a^*$, both of which are clopen. Furthermore, $a \in L^{\oplus}$, so $L^{\oplus +} = L^{+}$. Hence one of cases (1), (2), or (4) must hold.

Table 1. Classification of languages by the structure of $(B(L), ^+, ^\oplus)$.

Case	Necessary and Sufficient Conditions	$\lvert B(L)\rvert$	$\lvert A(L)\rvert$	Example	Dual
(1)	L is clopen.	1	2	a^+	(1)
(2)	L is open but not closed.	2	4	a	(3)
(3)	L is closed but not open.	2	4	aaa^*	(2)
(4)	L is neither open nor closed; L^+ is clopen and $L^{\oplus +} = L^+$.	3	6	$a \cup aaaa$	(5)
(5)	L is neither open nor closed; $L^\oplus$ is clopen and $L^{+\oplus} = L^\oplus$.	3	6	aa	(4)
(6)	L is neither open nor closed; L^+ is open but $L^\oplus$ is not closed; $L^{\oplus +} \neq L^+$.	4	8	$G := a \cup abaa$	(7)
(7)	L is neither open nor closed; $L^\oplus$ is closed but L^+ is not open; $L^{+\oplus} \neq L^\oplus$.	4	8	$(a \cup b)^+ \setminus G$	(6)
(8)	L is neither open nor closed; $L^\oplus$ is not closed and L^+ is not open; $L^{+\oplus} = L^{\oplus +}$.	4	8	$a \cup bb$	(8)
(9)	L is neither open nor closed; $L^\oplus$ is not closed and L^+ is not open; $L^{+\oplus} \neq L^{\oplus +}$.	5	10	$a \cup ab \cup bb$	(9)

Case (*ii*): $a \notin L$. Then $L^\oplus = \emptyset$ or $L^\oplus = \{\epsilon\}$, both of which are clopen. Furthermore, $a \notin L^+$, so $L^{+\oplus} = L^+$. Hence one of cases (1), (3), or (5) must hold.

Unary examples for cases (1) through (5) can be found in Table 1. ■

Note that all the example languages are either finite or cofinite and are thus regular. Consequently, Theorems 18 and 19 also hold for any regular language and any regular unary language, respectively.

4.2. *Structures of the algebras with Kleene closure and complement*

Since Kleene closure requires the existence of the identity element ϵ, we consider only the case where complementation is with respect to Σ^*. As we did in the positive case, first we restrict ourselves to closure and interior. Let $E(L)$ be the family of all languages generated from L by Kleene closure and Kleene interior, and let $F(L) = \{M : M^- \in E(L)\}$ be their complements. Our next results relate $D(L)$ and $E(L)$ to $A(L)$ and $B(L)$. Our discussion involves both closure operators, so we will be explicit about which closure properties we are invoking (although the word

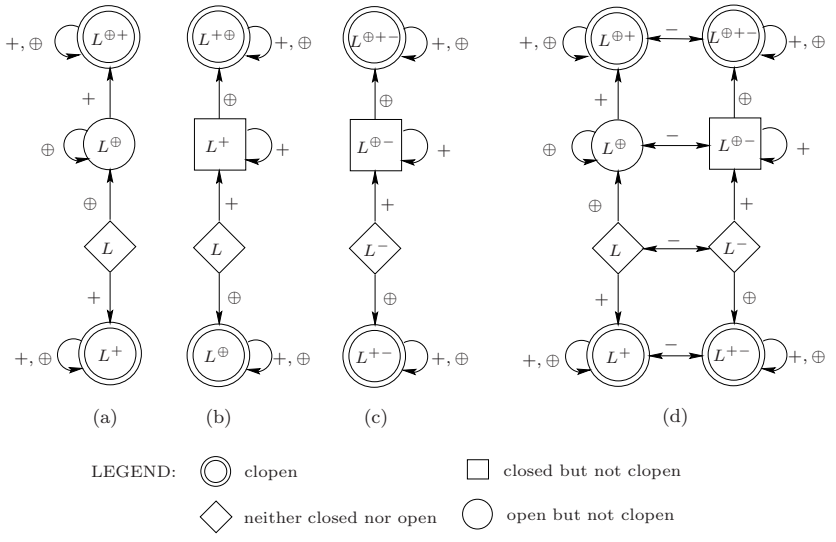

Fig. 1. Construction of $A(L)$, Case (6): (a) $(B(L),^+,^\oplus)$, Case (6); (b) $(B(L),^+,^\oplus)$, Case (7), the dual of Case (6) obtained by interchanging + with $\oplus$, and "open" with "closed"; (c) $(C(L),^+,^\oplus)$, that is, $(B(L),^+,^\oplus)$, Case (7), with elements renamed as complements of those of Case (6); (d) $A(L)$ constructed from $B(L)$ and $C(L)$.

clopen will still mean positive-clopen). We first claim the following, which can again be proven in the same manner as Lemma 16:

Lemma 20. *Let $L \subseteq \Sigma^*$. Then $D(L) = E(L) \cup F(L)$, and the union is disjoint.*

Next, we give a way of relating $E(L)$ to $B(L)$. We recall that $L^* = L^+ \cup \{\epsilon\}$ and $L^\oplus = L^\oplus \setminus \{\epsilon\}$. Consequently, $E(L) \subseteq \bigcup_{M \in B(L)} \{M \cup \{\epsilon\}, M \setminus \{\epsilon\}\}$. We now know enough to explicitly determine $D(L)$ in the following case:

Lemma 21. *Let $L \subseteq \Sigma^*$ be clopen. Then $D(L) = \{L \cup \{\epsilon\}, L \setminus \{\epsilon\}, L^- \cup \{\epsilon\}, L^- \setminus \{\epsilon\}\}$.*

Since the operations of positive closure and positive interior preserve the presence or absence of ϵ in a language, we may also note that if $\epsilon \in L$, then all languages in $B(L)$ contain ϵ, and conversely if $\epsilon \notin L$, then no language in $B(L)$ contains ϵ. For $M \in E(L)$, we write $\phi(M)$ to denote either $M \cup \{\epsilon\}$ or $M \setminus \{\epsilon\}$, whichever lies in $B(L)$. We note that $\phi(M)$, $\phi(M \cup \{\epsilon\})$, and $\phi(M \setminus \{\epsilon\})$ are equal. Moreover, we note that $\phi(M^*) = \phi(M)^+$ and $\phi(M^\circledast) = \phi(M)^\oplus$; ϕ can therefore be thought of as a homomorphism from $E(L)$ to $B(L)$. Consequently, $E(L) \subseteq \{M : \phi(M) \in B(L)\}$. We use this idea and the classifications of Table 1 to determine all possible algebras $(E(L),^*,^\circledast)$. As we will see, there are precisely 12 distinct algebras, each containing at most 14 elements.

We have seen what happens in Case (1) when L is clopen; two algebras are possible depending on whether $\epsilon \in L$ or not, and we refer to these as Cases (1a) and (1b) respectively. We next examine Cases (2) and (3), in which L is not clopen but is open or closed. Suppose L is open but not clopen, and hence $B(L) = \{L, L^+\}$. Then L^* is clopen and thus $E(L^*) = \{L^*, L^* \setminus \{\epsilon\}\}$. Since $E(L^*) \subseteq E(L)$ we thus have $\{L, L^*, L^* \setminus \{\epsilon\}\} \subseteq E(L) \subseteq \{M : \phi(M) \in \{L, L^+\}\}$. Therefore, we have two cases; either one or both of $L \setminus \{\epsilon\}$ and $L \cup \{\epsilon\}$ may be in $E(L)$, depending on whether or not $L^{\circledast} = L$. If $\epsilon \notin L$, then $L^{\circledast} = L$ and thus $E(L) = \{L, L^*, L^* \setminus \{\epsilon\}\}$. If $\epsilon \in L$, then $L^{\circledast} = L \setminus \{\epsilon\}$ and thus $E(L) = \{L, L \setminus \{\epsilon\}, L^*, L^* \setminus \{\epsilon\}\}$. We refer to these situations as Cases (2a) and (2b) respectively.

Similar possibilities occur when L is closed but not clopen. If $\epsilon \in L$ then $E(L) = \{L, L^{\circledast}, L^{\circledast} \cup \{\epsilon\}\}$. If $\epsilon \notin L$ then $L^* = L \cup \{\epsilon\}$ and $E(L) = \{L, L \cup \{\epsilon\}, L^{\circledast}, L^{\circledast} \cup \{\epsilon\}\}$. We refer to these situations as Cases (3a) and (3b) respectively.

We now turn to Cases (4)–(9), when L is neither closed nor open.

Lemma 22. *Let $L \subseteq \Sigma^*$ be neither open nor closed. Then*

$$E(L) = \{L\} \cup \{M \cup \{\epsilon\} : M \in B(L) \text{ and } M \text{ is closed}\} \cup \{M \setminus \{\epsilon\} : M \in B(L) \text{ and } M \text{ is open}\}.$$

Proof. Clearly $L \in E(L)$. We claim that no other language M with $\phi(M) = L$ can be in $E(L)$. If we suppose otherwise, then such an M must be generated by taking the Kleene closure or interior of some other language in $E(L)$. This would imply that M is open or closed, which is impossible since $\phi(M) = L$ and L is neither open nor closed.

For each remaining $M \in B(L) \setminus \{L\}$, we wish to show that $M \cup \{\epsilon\} \in E(L)$ if and only if M is closed, and $M \setminus \{\epsilon\} \in E(L)$ if and only if M is open. Let $M \in B(L) \setminus \{L\}$ be generated by some non-empty sequence S of positive closures and positive interiors. If we replace each positive closure by a Kleene closure and each positive interior by a Kleene interior, then we obtain a sequence S' that generates some $M' \in E(L)$ with $\phi(M') = M$. Now M' contains ϵ if and only if the last operation in S' was a Kleene closure. If M is closed, we may append a final positive closure to any such S to obtain one in which the last operation is a closure. Conversely, if there exists an S whose last operation is a closure, then M must be closed. Thus there exists an $M' \in E(L)$ containing ϵ with $\phi(M') = M$ if and only if M is closed. By a similar argument, there exists an $M' \in E(L)$ not containing ϵ with $\phi(M') = M$ if and only if M is open. The result follows. ■

Lemma 22 allows us to describe the structure of the algebra $(E(L), ^*, ^{\circledast})$ in Cases (4) through (9). The algebra $E(L)$ contains $M \cup \{\epsilon\}$ for all closed M in $B(L)$, $M \setminus \{\epsilon\}$ for all open M in $B(L)$, and both for all clopen M in $B(L)$.

We classify the 12 distinct algebras in Table 2. The conditions are identical to those found in Table 1; the only differences lie in Cases (1), (2), and (3), where the initial presence or absence of ϵ can affect the structure of the algebra.

Table 2. Classification of languages by the structure of $(E(L),^*,^{\circledast})$.

Case	Necessary and Sufficient Conditions	$\lvert E(L)\rvert$	$\lvert D(L)\rvert$	Example	Dual
(1a)	L is clopen; $\epsilon \in L$.	2	4	a^*	(1b)
(1b)	L is clopen; $\epsilon \notin L$.	2	4	a^+	(1a)
(2a)	L is open but not clopen; $\epsilon \in L$.	3	6	$a \cup \epsilon$	(3a)
(2b)	L is open but not clopen; $\epsilon \notin L$.	4	8	a	(3b)
(3a)	L is closed but not clopen; $\epsilon \notin L$.	3	6	aaa^*	(2a)
(3b)	L is closed but not clopen; $\epsilon \in L$.	4	8	$aaa^* \cup \epsilon$	(2b)
(4)	L is neither open nor closed; L^+ is clopen and $L^{\oplus +} = L^+$.	4	8	$a \cup aaa$	(5)
(5)	L is neither open nor closed; $L^{\oplus}$ is clopen and $L^{+\oplus} = L^{\oplus}$.	4	8	aa	(4)
(6)	L is neither open nor closed; L^+ is open but $L^{\oplus}$ is not closed; $L^{\oplus +} \neq L^+$.	6	12	$G := a \cup abaa$	(7)
(7)	L is neither open nor closed; $L^{\oplus}$ is closed but L^+ is not open; $L^{+\oplus} \neq L^{\oplus}$.	6	12	$(a \cup b)^+ \setminus G$	(6)
(8)	L is neither open nor closed; $L^{\oplus}$ is not closed and L^+ is not open; $L^{+\oplus} = L^{\oplus +}$.	5	10	$a \cup bb$	(8)
(9)	L is neither open nor closed; $L^{\oplus}$ is not closed and L^+ is not open; $L^{+\oplus} \neq L^{\oplus +}$.	7	14	$a \cup ab \cup bb$	(9)

We now summarize our results for the Kleene case:

Theorem 23. *Start with any language L, and apply the operators of Kleene closure and complement in any order, any number of times. Then at most* 14 *distinct languages are generated, and this bound is optimal. Furthermore, Table 2 describes the 12 algebras generated by this process, classifies languages according to the algebra they generate, and gives a language generating each algebra.*

Theorem 24. *Start with any unary language L, and apply the operators of positive closure and complement in any order, any number of times. Then at most* 8 *distinct languages are generated, and this bound is optimal. Furthermore, pre-*

cisely Cases (1a) through (5) in Table 2 describe the 8 *possible algebras that can be generated from a unary language by this process.*

The proof of Theorem 24 is analogous to the proof of Theorem 19, and again, all of our example languages are regular, so Theorems 23 and 24 also hold for any regular language and any regular unary language, respectively.

5. Closure Operators and Concatenation

We note that the concatenation of two closed languages need not be closed, and that the concatenation of two open languages need not be open. For example, consider the languages $L = \{a\}^+$ and $M = \{b\}^+$ for $a, b \in \Sigma$, which are both clopen (under positive closure). Then $ab \in LM$ but $abab \notin LM$, so LM is not closed. Additionally, $ab \in LM$, but neither a nor b is in LM, so LM is not open. However, we do have several results regarding cases when the concatenation of closed or open languages must be closed or open. All proofs are elementary, and analogous results hold for Kleene closure.

Theorem 25. *If L and M are closed and $LM = ML$, then LM is closed. In particular, the concatenation of any two closed unary languages is closed.*

Proof. Immediate from the fact that $LMLM = LLMM$ whenever $LM = ML$. ■

Theorem 26. *Let $L, M \subseteq \Sigma^*$. Suppose L and M are closed and $L \cup M$ is closed. Then LM is closed. More generally, if $W \in \{L, M\}^+$ is any sequence of concatenations of L and M, then W is closed. In particular, L^k is closed for all $k \geq 1$.*

Proof. To show that LM is closed, it suffices to show that $(LM)^k \subseteq LM$; we do this by induction on k. For $k > 1$, $(LM)^k \subseteq L(L \cup M)(L \cup M)M(LM)^{k-2} \subseteq L(L \cup M)M(LM)^{k-2} = (LLM \cup LMM)(LM)^{k-2} \subseteq LM(LM)^{k-2} = (LM)^{k-1} \subseteq LM$. The generalization of this proof to arbitrary W is straightforward (one may consider $W = L^k$ and $W = M^k$ as special cases and note that in all other cases, the sequence W contains LM or ML as a subsequence.) ■

The next results follow directly from Proposition 8 (c).

Theorem 27. *Let L and M be open.*

(a) Suppose $\epsilon \in L$ and $\epsilon \in M$. Then LM is open.
(b) Suppose $\epsilon \notin L$ and $\epsilon \notin M$. Then LM is open if and only if $L = \emptyset$ or $M = \emptyset$.
(c) LL is open if and only if $\epsilon \in L$ or $L = \emptyset$.

Additionally, if neither L nor M is empty and $\epsilon \in L \cup M$ but $\epsilon \notin L \cap M$, then we may or may not have LM open, even in the unary case. As examples, consider $L = \{\epsilon, a, aaa, aaaaa\}$, $M = \{a\}$ and $L = \{\epsilon, a, aaa\}$, $M = \{a\}$.

Theorem 28. *Let $L, M \subseteq \Sigma^*$ both be clopen with $L \cup M = \Sigma^*$. Then LM is clopen. More generally, if $W \in \{L, M\}^+$ is any sequence of concatenations of L and M,*

then W is clopen if and only if $W = \emptyset$ or W contains at most one occurrence of a language that does not contain ϵ.

Proof. Theorem 26 immediately implies that LM is closed. To show that LM is open, let $ab \in LM$ where $a \in L$ and $b \in M$ and let $ab = uv$ for arbitrary words u and v. We must show that either $u \in LM$ or $v \in LM$. Without loss of generality, we assume that u is a prefix of a and let $a = ux$, so $ab = uxb$ and hence $v = xb$. Then either $u \in L$ or $x \in L$. If $x \in L$, then $v = xb \in LM$ and we are done. Otherwise, we have $x \notin L$, implying $u \in L$ and $x \in M$ since $L \cup M = \Sigma^*$. If $\epsilon \in M$, $u = u\epsilon \in LM$ and we are done. Otherwise, we have $\epsilon \notin M$, and thus $\epsilon \in L$ since $L \cup M = \Sigma^*$. Then $xb \in M$ since $x \in M$, $b \in M$, and M is closed. Then $\epsilon xb = v \in LM$ and thus LM is open and hence is clopen.

The generalization to arbitrary W is straightforward by repeated applications of the above. If W contains multiple occurrences of a language not containing ϵ, then W contains no words of length 1, so either $W = \emptyset$ or W is not open. ■

Note that the converse of the above theorem is false; indeed, it is possible that LM is clopen, but $L \cup M$ is not even positive-closed. As a counterexample, we let $L = \{\epsilon\} \cup \{w \in \{a, b\}^* : |w|_a < |w|_b\}$ and let $M = \{\epsilon\} \cup \{w \in \{a, b\}^* : |w|_a > |w|_b\}$, where by $|w|_c$ for a letter c, we mean the number of occurrences of c in w. It is not hard to see that LM is clopen, but $L \cup M$ is not closed since we have $b \in L \subseteq L \cup M$ and $a \in M \subseteq L \cup M$, but $ba \notin L \cup M$.

6. Separation of Words and Languages

Next, we discuss analogies of the separation axioms of topology in the realm of languages. Although languages do not form a topology under Kleene or positive closure, there are many interesting results describing when there exist open, closed, and clopen languages that separate given words or languages. In most of these theorems, we only consider words in Σ^+, as ϵ is always a trivial case.

Lemma 29. *Let $w \in \Sigma^+$, and let $L \subseteq \Sigma^*$ be closed with $w \notin L$. Then there exists a finite open language M such that $w \in M$ but $M \cap L = \emptyset$.*

Proof. We simply take $M = L^- \cap \{x \in \Sigma^+ : |x| \leq |w|\}$. ■

Theorem 30. *Let $u, v \in \Sigma^+$. There exists an open language L with $u \in L$ and $v \notin L$ if and only if for all natural numbers k, we have $u \neq v^k$. Consequently, if $u \neq v$, then there exists an open language L with either $u \in L$ and $v \notin L$, or $u \notin L$ and $v \in L$. In other words, all words are distinguishable by open languages.*

Proof. For the forward direction, we note that if $u = v^k$ for some positive k, then any open language containing u must contain v by Proposition 8 (c). For the reverse direction, we apply Lemma 29 to u and $\{v\}^+$, which is closed. ■

We now recall a basic result from combinatorics on words (see, e.g., [11]).

Lemma 31. *Let $u, v \in \Sigma^+$. The following are equivalent:*

(1) $uv = vu$.
(2) There exists a word x and integers $p \geq 1$ and $q \geq 1$ such that $u = x^p$ and $v = x^q$.

If any of the above hold, then we say that u and v *commute*.

Let $u, v \in \Sigma^+$. Suppose there exists a clopen language $L \subseteq \Sigma^*$ with $u \in L$ and $v \notin L$. We note that L^- is also clopen whenever L is, and we call the pair (L, L^-) a *clopen partition separating u and v*. Motivated by the desire to extend the topological notion of connected components to formal languages, we have the following result, which characterizes precisely when clopen partitions exist:

Theorem 32. *Let $u, v \in \Sigma^+$. There exists a clopen partition separating u and v if and only if u and v do not commute.*

Proof. The forward direction is quite straightforward. If u and v commute, then there exists a word x and integers p and q such that $u = x^p$ and $v = x^q$. Then any open language containing u will also contain x, and any open language containing v will also contain x. It follows that no clopen partition can separate u and v.

For the reverse direction, we proceed by induction on $|u| + |v|$. We will apply the induction hypothesis on words in various alphabets, so we make no assumption that $|\Sigma|$ is constant.

For our base case, suppose $|u| + |v| = 2$. If u and v do not commute, then they must be distinct words of length 1, and thus the language $\{u\}^+$ is a clopen language separating u from v.

Suppose, as a hypothesis, that for some $k \geq 2$, the result holds for all finite alphabets Σ and for all $u, v \in \Sigma^+$ such that $2 \leq |u| + |v| \leq k$. Now, given any Σ, let $u, v \in \Sigma^+$ be such that u and v do not commute and $|u| + |v| = k + 1$. Let Σ_u and Σ_v, respectively, be the symbols that occur one or more times in u and v. If $\Sigma_u \cap \Sigma_v = \emptyset$, then Σ_u^+ is a clopen language containing u but not v, and our result holds. If not, suppose $a \in \Sigma_u \cap \Sigma_v$. Let $\lambda_u = \frac{|u|_a}{|u|}$ and $\lambda_v = \frac{|v|_a}{|v|}$ be the respective relative frequencies of a in u and v. If $\lambda_u > \lambda_v$, then $\{w \in \Sigma^* : |w|_a \geq \lambda_u |w|\}$ is clopen (by Example 11) and contains u but not v, and we are done. Similarly, if $\lambda_u < \lambda_v$, then $\{w \in \Sigma^* : |w|_a \leq \lambda_u |w|\}$ is a clopen language containing u but not v. Thus it remains to show that the result holds when $\lambda_u = \lambda_v$.

Assume $\lambda_u = \lambda_v = \lambda$. If $\lambda = 1$, then $u = a^i$ and $v = a^j$ for some positive integers i and j, and thus u and v commute, contradicting our original assumption. Hence we must have $0 < \lambda < 1$. Let $n = \frac{|u|}{\gcd(|u|_a, |u|)} = \frac{|v|}{\gcd(|v|_a, |v|)}$ be the denominator of λ when it is expressed in lowest terms. We must have $n > 1$ since λ is not an integer.

Next, we consider a new alphabet Δ with $|\Sigma|^n$ symbols, each corresponding to a word of length n in Σ^*. We consider the bijective morphism ϕ mapping words in Δ^* to words in $(\Sigma^n)^*$ by replacing each symbol in Δ with its corresponding word in

Σ^n. Since n divides both $|u|$ and $|v|$, there must then exist unique words $p, q \in \Delta^*$ such that $\phi(p) = u$ and $\phi(q) = v$.

Our plan is now to inductively create a clopen language L over Δ which contains p but not q, and then use this language to construct our clopen partition over Σ separating u and v. We must check that p and q do not commute. If $pq = qp$ then we would have $uv = \phi(p)\phi(q) = \phi(pq) = \phi(qp) = \phi(q)\phi(p) = vu$, which is impossible since $uv \neq vu$. We also have $n|p| + n|q| = |u| + |v|$. Since $n > 1$ implies $|p| + |q| < |u| + |v| = k + 1$, the induction hypothesis can be applied to p and q. Thus there exists a clopen language $L \subseteq \Delta^*$ with $p \in L$ and $q \notin L$.

We now construct our clopen partition over Σ separating u and v. We introduce some notation to make this easier. As usual, define $\phi(L) = \{w \in \Sigma^* : w = \phi(r) \text{ for some } r \in L\}$. Let $A^< = \{w \in \Sigma^* : |w|_a < \lambda|w|\}$ and let $A^= = \{w \in \Sigma^* : |w|_a = \lambda|w|\}$. Additionally, let $A^\leq = A^< \cup A^=$. It is easy to verify that $A^<$, $A^\leq$, and $A^=$ are all closed, and both $A^<$ and $A^\leq$ are open as well. Finally, we let $M = (\phi(L) \cap A^=) \cup A^<$. Since $p \in L$ and $q \notin L$, we must have $u \in \phi(L)$ and $v \notin \phi(L)$. Then since u and v are both contained in $A^=$ but not $A^<$, we must have $u \in M$ and $v \notin M$.

We now finish the proof by showing that M is clopen. We first show that M is closed. Let $x, y \in M$. We must show that $xy \in M$. There are two cases to consider:

Case (A1): $x, y \in (\phi(L) \cap A^=)$. We see that $\phi(L)\phi(L) = \phi(LL) \subseteq \phi(L)$, so $\phi(L)$ is closed. Then since $A^=$ is closed, $\phi(L) \cap A^=$ is the intersection of two closed languages, and hence closed. Thus $xy \in \phi(L) \cap A^= \subseteq M$.

Case (A2): One or more of x or y is not in $\phi(L) \cap A^=$. Without loss of generality, suppose $x \notin \phi(L) \cap A^=$. Then $x \in A^<$, so $|x|_a < \lambda|x|$. Furthermore, $y \in M \subseteq A^\leq$, so $|y|_a \leq \lambda|y|$. Adding these two inequalities yields $|x|_a + |y|_a < \lambda|x| + \lambda|y|$, so $|xy|_a < \lambda|xy|$ and thus $xy \in A^< \subseteq M$.

Lastly, we show that M is open. Let $z \in M$ and suppose $z = xy$ for some $x, y \in \Sigma^+$. We show that $x \in M$ or $y \in M$. Again, we have two cases to consider:

Case (B1): $z \in A^<$. Since $A^<$ is open, at least one of x or y is in $A^<$. Since $A^< \subseteq M$, we are done.

Case (B2): $z \in \phi(L) \cap A^=$. If either x or y is in $A^<$, then we are done, so assume otherwise. Then $|x|_a \geq \lambda|x|$ and $|y|_a \geq \lambda|y|$. But $|xy|_a = \lambda|xy|$, so we must have $|x|_a = \lambda|x|$ and $|y|_a = \lambda|y|$ and thus $x, y \in A^=$. Then $\lambda|x|$ and $\lambda|y|$ must be integers and hence n divides both $|x|$ and $|y|$. Then there exist $s, t \in \Delta^*$ such that $\phi(s) = x$ and $\phi(t) = y$. But since ϕ is a morphism, we must then have $\phi(st) = \phi(s)\phi(t) = xy = z$. But z in $\phi(L)$, so $st \in L$. Since L is open, we must then have either $s \in L$ or $t \in L$. Thus we must have either $x = \phi(s) \in \phi(L)$ or $y = \phi(t) \in \phi(L)$. Then one of x or y is in $\phi(L) \cap A^= \subseteq M$.

Thus M is both closed and open, and the result follows by induction. ■

Holub and Kortelainen have recently provided an alternate proof of Theorem 32 [9]. They extend our result by showing how to obtain a *regular* clopen partition separating two words if they do not commute.

Corollary 33. *Let $u, v \in \Sigma^+$. There exist non-intersecting finite open languages L and M with $u \in L$ and $v \in M$ if and only if u and v do not commute.*

Proof. The forward direction is identical to that of Theorem 32. For the reverse direction, we suppose u and v do not commute and let K be a clopen language containing u but not v. We then take $L = \{w \in K : |w| \leq |u|\}$ and $M = \{w \in K^- : |w| \leq |v|\}$. These are open by Proposition 8 (c) since K and K^- are both open. ■

We can also use Theorem 32 to extend the topological notion of *connected components* to the setting of formal languages. We say that words $u, v \in \Sigma^+$ are *disconnected* if there exists a clopen partition separating u from v, and *connected* otherwise. We write $u \sim v$ if u and v are connected, and note that $\sim$ is an equivalence relation (indeed, this is the case when we consider the clopen partitions induced by any closure operator; it need not be topological). Recall that w is *primitive* if there is no word x with $w = x^k$ for some $k \geq 2$. By Theorem 32, $u \sim v$ if and only if u and v are both powers of some primitive word x. It follows that each connected component of Σ^+ consists of a primitive word and its powers. Dividing an arbitrary language into connected components simply sorts its words by primitive root.

The following theorem holds for all closure operators that preserve openness.

Theorem 34. *If $L, M \subseteq \Sigma^*$ are disjoint and open, then L^+ and M^+ are disjoint.*

Proof. If $L \cap M = \emptyset$, then $M \subseteq L^-$. Then by isotonicity, $M^+ \subseteq L^{-+} = L^-$ since L^- is closed. But then $L \subseteq M^{+-}$. Applying isotonicity again yields $L^+ \subseteq M^{+-+}$. But M^+ is the closure of an open language and is thus clopen, so M^{+-} is also clopen and thus $M^{+-+} = M^{+-}$. Hence $L^+ \subseteq M^{+-}$ and the result follows. ■

Corollary 35. *Let $L, M \subseteq \Sigma^*$ both be closed and such that $L \cup M = \Sigma^*$. Then $L^\oplus \cup M^\oplus = \Sigma^*$.*

In our setting, it is not true that a single "point" x and a closed set S can be separated by two open sets. As a counterexample, consider $x = ab$ and $y = \{aa, bb\}^*$. Furthermore, it is not true that arbitrary disjoint sets, even ones whose closures are disjoint, can be clopen separated. As an example, consider $\{ab\}^*$ and $\{aa, bb\}^*$.

7. Algorithms

We now consider the computational complexity of determining if a given language L is closed or open. Of course, the answer depends on how L is represented.

Theorem 36. *Given an n-state DFA $M = (Q, \Sigma, \delta, q_0, F)$ accepting the regular language L, we can determine in $O(n^2)$ time if L is closed or open.*

Proof. We prove the result when L is positive-closed. For Kleene-closed, we have the additional check whether q_0 is in F. For the open case, we start with a DFA for $\overline{L}$.

We know from Proposition 7 that L is closed if and only if, for all $u, v \in L$ we have $uv \in L$. Given M, we create an NFA-ϵ M' that accepts all words $x \notin L$ such that there exists a decomposition $x = uv$ with $u, v \in L$. Then $L(M')$ is empty if and only if L is closed.

Here is the construction of M': $M' = (Q', \Sigma, \delta', q_0', F')$, where $Q' = Q \cup Q \times Q$, $q_0' = q_0$, $F' = (Q - F) \times F$, and δ' is defined as follows:

$$\begin{aligned} \delta'(p, a) &= \{\delta(p, a)\} \text{ for } p \in Q, a \in \Sigma; \\ \delta'(p, \epsilon) &= \{[p, q_0]\}, \text{ if } p \in F; \\ \delta'([p, q], a) &= \{[\delta(p, a), \delta(q, a)]\} \text{ for } p, q \in Q, a \in \Sigma. \end{aligned}$$

M' functions as follows: on input u, it simulates the computation of M. If and only if a final state is reached (and so $u \in L$), M' has the option to use its ϵ-transition to enter a state specified by two components, the second of which is q_0. Now M' processes v, determining $\delta(q_0, uv)$ in its first component and $\delta(q_0, v)$ in the second. If $uv \notin L$, but $v \in L$, then M' accepts. Thus M' accepts uv if and only if $u, v \in L$ and $uv \notin L$.

We now use the usual depth-first search technique to determine if $L(M')$ is empty, which uses time proportional to the number of states and transitions of M'. Since M' has $|Q||\Sigma| + |F| + |Q|^2|\Sigma|$ transitions and $|Q| + |Q|^2$ states, our depth-first search can be done in $O(n^2)$ time. ■

From Proposition 7, we know that L is not closed if and only if there exists a word $uv \notin L$ such that $u, v \in L$. In the following proposition we give an upper bound on the length of such a word.

Corollary 37. *If L is a regular language that is not closed, and accepted by a n-state DFA, then there exist $u, v \in L$ with $uv \notin L$ such that $|uv| \leq n^2 + n - 1$.*

This $O(n^2)$ upper bound is matched by a corresponding $\Omega(n^2)$ lower bound:

Theorem 38. *For each positive integer n there exists a DFA M_n with $2n + 5$ states satisfying the following property: for any $u, v \in L(M_n)$, if $uv \notin L(M_n)$, then $|uv| \geq n^2 + 2n + 2$.*

Proof. It is easier to describe DFA $M_n' = (Q, \Sigma, \delta, q_0, F)$ that accepts the complement of $L(M_n)$. Let $Q = \{q_0, q_1, \ldots, q_n, r, p_0, p_1, \ldots, p_n, s, d\}$, let δ be given by Table 3, and let $F = \{q_0, q_1, \ldots, q_n, p_0, p_1, \ldots, p_n, s\}$. The case $n = 5$ is shown in Fig. 2.

First, we observe that $x = 10^{n-1}110^{n^2+n-1}1$ is accepted by M_n', but neither $u = 10^{n-1}1$ nor $v = 10^{n^2+n-1}1$ is. Next, take any word x' accepted by M_n'. If the acceptance path does not pass through r, then by examining the DFA we see that every prefix of x' is also accepted. Otherwise, the acceptance path passes through r. Again, we see that every prefix of x' is accepted, with the possible exception of

Table 3. Transition function $\delta(q, a)$ of M'_n.

$a\backslash q$	q_0	q_1	q_2	$\ldots$	q_{n-1}	q_n	r	p_0	p_1	$\ldots$	p_{n-1}	p_n	s	d
0	d	q_2	q_3	$\ldots$	q_n	q_1	d	p_1	p_2	$\ldots$	p_n	p_0	d	d
1	q_1	s	s	$\ldots$	s	r	p_0	d	d	$\ldots$	d	s	d	d

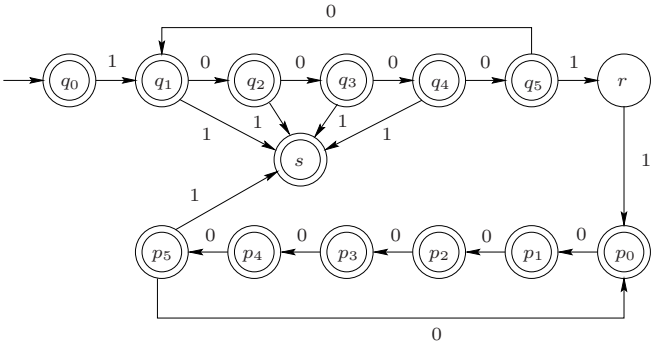

Fig. 2. DFA M_5. Unspecified transitions go to the dead state d (not shown).

the prefix ending at r. Thus either x' is of the form $10^{in+n-1}110^k$ for some $i, k \geq 0$, or x' is of the form $10^{in+n-1}110^{j(n+1)+n}1$ for some $i, j \geq 0$. In both cases the prefix ending at r is $10^{in+n-1}1$, so in the first case, the corresponding suffix is 10^k for some $k \geq 0$, and this suffix is accepted by M'_n. In the latter case, the corresponding suffix is $10^{j(n+1)+n}1$. This is accepted unless $j(n+1)+n = in+n-1$ for some $i \geq 0$. By taking both sides modulo n, we see that $j \equiv -1 \pmod{n}$. Thus $j \geq n-1$. Thus $|x'| \geq 1+n-1+1+1+(n-1)(n+1)+n+1 = n^2+2n+2$. ∎

We now turn to the case where M is represented as an NFA or regular expression. For the following theorem, we actually require the word w exhibited in the theorem above to have length ≥ 2. However, this can be accomplished easily using a trivial modification of the proof given in [1], since the word w encodes a configuration of the Turing machine T.

Theorem 39. *The following problem is PSPACE-complete: given an NFA M, decide if $L(M)$ is closed.*

Proof. First, we observe that the problem is in PSPACE. We give a nondeterministic polynomial-space algorithm to decide if $L(M)$ is not closed, and use Savitch's theorem to conclude the result.

If M has n states, then there is an equivalent DFA M' with $N \leq 2^n$ states. From Corollary 37 we know that if $L = L(M) = L(M')$ is not closed, then there exist

words u, v with $u, v \in L$ but $uv \notin L$, and $|uv| \leq N^2 + N - 1 = 2^{2n} + 2^n - 1$. We now guess u, processing it symbol-by-symbol, arriving in a set of states S of M. Next, we guess v, processing it symbol-by-symbol starting from both q_0 and S, respectively and ending in sets of states T and U. If U contains a state of F and T does not, then we have found $u, v \in L$ such that $uv \notin L$. As we proceed, we count the number of symbols guessed, and reject if that number is greater than $2^{2n} + 2^n - 1$.

To show that the problem is PSPACE-hard, we note that Δ^* is closed, but $\Delta^* \setminus \{w\}$ for w with $|w| \geq 2$ is not. With the aid of Lemma 10.2 of [1] we could use an algorithm solving the problem of whether a language is closed to solve the membership problem for polynomial-space bounded Turing machines. ■

If L is not closed and is accepted by an n-state NFA, then a minimal-length word uv, with $u, v \in L$ but $uv \notin L$, may be exponentially long. Such an example is given in [4], where it is shown that for some constant c, there exist NFA's with n states such that a shortest word not accepted is of length $> 2^{cn}$. We note also that the problem of deciding, for a given NFA M, whether $L(M)$ is open is PSPACE-complete. The proof is similar to that of Theorem 39.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada. We thank the anonymous referees for suggesting ways to shorten some of our proofs.

References

[1] A. Aho, J. Hopcroft, and J. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, 1974.

[2] J. Brzozowski, E. Grant, and J. Shallit. In V. Diekert and D. Nowotka, Eds. *Developments in Language Theory, 13th International Conference, DLT 2009*. Springer, Lecture Notes in Computer Science, Vol. 5583, 2009, pp. 125–144.

[3] A. V. Chagrov. Kuratowski numbers. In *Application of Functional Analysis in Approximation Theory*, Kalinin. Gos. Univ., Kalinin, 1982, pp. 186–190. In Russian.

[4] K. Ellul, B. Krawetz, J. Shallit, and M.-w. Wang. Regular expressions: new results and open problems. *J. Autom. Lang. Combin.* **10** (2005), 407–437.

[5] J. H. Fife. The Kuratowski closure-complement problem. *Math. Mag.* **64** (1991), 180–182.

[6] B. J. Gardner and M. Jackson. The Kuratowski closure-complement theorem. *New Zealand*
J. Math. **38** (2008), 9–44. Preprint available at `http://www.latrobe.edu.au/mathstats/department/algebra-research-group/Papers/GJ_Kuratowski.pdf`

[7] R. L. Graham, D. E. Knuth, and T. S. Motzkin. Complements and transitive closures. *Discrete Math.* **2** (1972), 17–29.

[8] P. C. Hammer. Kuratowski's closure theorem. *Nieuw Archief v. Wiskunde* **7** (1960), 74–80.

[9] S. Holub and J. Kortelainen. On partitions separating two words. In *7th International Conference on Words*, 2009.

[10] C. Kuratowski. Sur l'opération $\overline{A}$ de l'analysis situs. *Fund. Math.* **3** (1922), 182–199.

[11] R. C. Lyndon and M. P. Schützenberger. The equation $a^M = b^N c^P$ in a free group. *Michigan Math. J.* **9** (1962), 289–298.

[12] D. Peleg. A generalized closure and complement phenomenon. *Discrete Math.* **50** (1984), 285–293.