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

Abstract. 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 closure is either Kleene closure or positive closure. We classify languages according to the structure of the algebra 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.

1 Introduction

In 1922, 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 [11, 7]. 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 [8]. For the analogous result on relations, see [9].

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 [10]. Our point of view most closely matches that of Peleg [13], who briefly observed that Kleene and positive closure are closure operators, and hence Kuratowski's theorem holds for them.

Positive and Kleene closures are discussed in Section 3. In Section 4 we reconsider Kuratowski's theorem in the context of formal languages, where closure is replaced by Kleene closure 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 5 and 7).

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 14) 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.

Because of space limitations, some proofs are omitted or only sketched. For more complete versions, see [2, 3].

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 $\Box$ operating on a set $X \subseteq S$ will be 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:

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

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

Note the duality between $\Box$ and $\circ$: $X^{\circ} = X^{-\Box-}$ and $X^{\Box} = X^{-\circ-}$. This duality also applies to (1), since we have

$$\begin{array}{rl} X \supseteq X^{\circ} & (\circ \text{ is intensive}); \\ X \subseteq Y \text{ implies } X^{\circ} \subseteq Y^{\circ} & (\circ \text{ is isotone}); \\ X^{\circ\circ} = X^{\circ} & (\circ \text{ is idempotent}). \end{array} \tag{2}$$

Moreover, it is equivalent to define $(S,\Box)$ via an *interior operator* satisfying (2).

We now list some fundamental properties of closure systems.

Proposition 1. *The intersection of an arbitrary family of closed sets is closed.*

Proposition 2. *For $X \subseteq S$, the following are identical: (a) $X^{\Box}$; (b) $\bigcap\{Y \subseteq S : Y \supseteq X$ and Y is closed$\}$; (c) $\{a \in S :$ for all open $Y \subseteq S, a \in Y$ implies $Y \cap X \neq \emptyset\}$; (d) $X^{-\circ-}$.*

Proposition 3. *Let $X, Y \subseteq S$. Then the following hold:*

(a) $X^{\Box}$ is closed.
(b) $(X \cup Y)^{\Box} = (X^{\Box} \cup Y^{\Box})^{\Box}$.
(c) $(X \cap Y)^{\Box} \subseteq X^{\Box} \cap Y^{\Box}$.

Duals of Propositions 1–3 also hold [2]. For example, the union of an arbitrary family of open sets is open.

We now state two versions of Kuratowski's theorem. The first [11] is equivalent to Kuratowski's original result generalized to an arbitrary closure system, not necessarily topological:

Theorem 1. *Let $(S, \Box)$ 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*

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

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

In 1983, Peleg [13] defined a closure operator to be *compact* if it satisfies Eq. (4) below. He showed that at most 10 different sets are generated if $\Box$ is compact, and proved that $\Box$ preserves openness if and only if it is compact. The following theorem is a modified version of Peleg's result:

Theorem 2. *Let $(S, \Box)$ be a closure system such that $\Box$ 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*

$$X^{\Box-\Box-\Box} = X^{\Box-\Box-}. \tag{4}$$

3 Positive and Kleene closures of languages

We deal now with closures in the setting of formal languages. Our universal set is Σ^*, the set of all finite words over a finite non-empty alphabet Σ. We consider two closure operators: positive closure and Kleene closure. For $L \subseteq \Sigma^*$, we define $L^- = \Sigma^* \setminus L$, $L^+ = \bigcup_{i \geq 1} L^i$, and $L^* = \bigcup_{i \geq 0} L^i$.

Proposition 4. *Positive closure and Kleene closure are both closure operators.*

We note, importantly, that the positive and Kleene closures are *not* topological (the union of two closed languages is not necessarily closed). As a counterexample, observe that $(aa)^+ \cup (aaa)^+ \subsetneq (aa \cup aaa)^+$, as a^5 belongs to the right-hand side but not the left. Consequently, 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 similarly. The *positive interior* of a language L is $L^{\oplus} = L^{-+-}$; the *Kleene interior* is $L^{\circledast} = L^{-*-}$.

Proposition 5. *Let $L \subseteq \Sigma^*$. The following are equivalent:*

(a) L is positive-closed.
(b) $L \cup \{\epsilon\}$ is Kleene-closed.
(c) $L = L^+$.
(d) $L = M^+$ for some $M \subseteq \Sigma^*$.
(e) For all $u, v \in L$, we have $uv \in L$.

The dual of Proposition 5 also holds [2]. For example, (a) L is positive-open is equivalent to (e) For all $u, v \in \Sigma^*$ such that $uv \in L$, we have $u \in L$ or $v \in L$.

In algebraic terms, $L \subseteq \Sigma^*$ is a *semigroup* if $uv \in L$ for all $u, v \in L$. Proposition 5 states that a language is positive-closed if and only if it is a semigroup. Also, $L \subseteq \Sigma^*$ is Kleene-closed if and only if it is a monoid. We verify that if L is positive-closed, then so are $L \setminus \{\epsilon\}$ and $L \cup \{\epsilon\}$. So 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 closure and Kleene closure are so similar, we restrict our attention to positive closure from this point on. This allows us to state our theorems more elegantly, as we need not worry 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.

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

To prove this, let $u, v \in L$. Then $|u|_1 < k|u|_2$ and $|v|_1 < k|v|_2$. But $|uv|_1 = |u|_1 + |v|_1 < k|u|_2 + k|v|_2 = k|uv|_2$, so $uv \in L$, and thus L is closed. By a similar argument, we can prove that $L^- = \{w \in \Sigma^* : |w|_1 \geq k|w|_2\}$ is closed. Thus L is clopen. ■

Example 2. 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.

For prefix-closed languages, we show that L satisfies the dual of Proposition 5 (e), which states that $uv \in L$ implies $u \in L$ or $v \in L$. Let $w \in L$ and suppose $w = uv$. Then $u \in L$ if L is prefix-closed, so our characterization holds and L is open. The proof is similar if L is suffix-closed. Since factor- and subword-closed languages are also prefix-closed, the claim holds. ■

Example 3. Closed languages: Left ideals (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 positive closed. ■

In the 1970's, D. Forkes proved Eq. (3) with the Kleene closure as $\square$, and the first author then proved that Eq. (4) holds when $\square$ is positive closure. (They

were both unaware of [11].) Peleg [13] proved this over a wider class of operators. Here, we state an equivalent fact: positive closure preserves openness.

Theorem 3. *Let $L \subseteq \Sigma^*$ be open. Then L^+ is open.*

Proof. This follows from the facts that Eq. (4) holds for positive closure, and that Eq. (4) is equivalent to compactness. ■

By arguments similar to those in the proof of Theorem 2, we may conclude:

Corollary 1. *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 results are false; for example, there exist languages such as $\{a, aaaa\}$ which are not open, but have clopen closures. We discuss such possibilities extensively in the next section. For now, we give a characterization of the languages with clopen closures and clopen interiors.

Theorem 4. *Let $L \subseteq \Sigma^*$.*

(a) L^+ is clopen iff there exists an open language M with $L \subseteq M \subseteq L^+$.
(b) $L^\oplus$ is clopen iff there exists a closed language M with $L \supseteq M \supseteq L^\oplus$.

Proof. We prove only (a); (b) can be proved using a similar argument. The forward direction of (a) is trivial since we can take $M = L^+$. For the converse, we note that $L \subseteq M$ implies $L^+ \subseteq M^+$ by isotonicity, and $M \subseteq L^+$ implies $M^+ \subseteq L^{++} = L^+$ by isotonicity and idempotency. Thus $M^+ = L^+$, and since M^+ is the closure of an open language, it is clopen and the result follows. ■

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 2 implies that $A(L)$ contains at most 10 languages. As we shall see, this upper bound is tight. Moreover, there are precisely 9 distinct finite algebras $(A(L),^+,^-)$. Since the languages in $A(L)$ must occur in complementary pairs, there can only exist algebras containing 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 1, which implies that $D(L)$ contains at most 14 languages, and we will show that this bound is also tight. There are precisely 12 distinct finite algebras $(D(L),^*,^-)$. We shall describe these algebras by relating them to those in the positive case.

In a sense, our results are the formal language analogue of topological results obtained by Chagrov [5] and discussed in [8]. Peleg [13] 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

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. Recall that the closure of an open language is clopen and the interior of a closed language is clopen by Corollary 1. Since the closure and interior operators are idempotent on the clopen languages $L^{+\oplus}$ and $L^{\oplus+}$, it follows that $B(L) = \{L, L^+, L^{+\oplus}, L^\oplus, L^{\oplus+}\}$. Of course, these five languages may not all be distinct; we will address this later. At the moment, we provide the following proposition, which demonstrates that it suffices to analyze the structure of $B(L)$ to determine the structure of $A(L)$.

Proposition 6. *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, by 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 constuct 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 $M \in C(L)$, and thus $A(L) = B(L) \cup C(L)$.

We now prove that $B(L) \cap C(L) = \emptyset$. We assume otherwise to obtain a contradiction; $B(L)$ must then contain some complementary pair of languages M and M^-. We note that $L^\oplus \subseteq L^{\oplus+}$ by extensivity, $L^\oplus \subseteq L \subseteq L^+$ by intensivity and extensivity, and $L^\oplus \subseteq L^{+\oplus}$ by isotonicity, and 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 hence L^+ and $L^{+\oplus}$ do not either. But then no language in $B(L)$ contains a string of length 1, and thus no pair of languages in $B(L)$ are complements, and we have our contradiction. ■

Proposition 6 implies that $|A(L)| = 2|B(L)|$, and moreover that 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 with 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 1. *Suppose $L \subseteq \Sigma^*$. If L^+ and $L^\oplus$ are both clopen, then L must be open or closed.*

Proof. Seeking a contradiction, we assume that both L^+ and $L^\oplus$ are clopen but L is neither open nor closed. If L is not open, then $L \setminus L^\oplus$ is non-empty.

Let w be the shortest word in $L \setminus L^\oplus$. Consider $M = L^\oplus \cup \{w\}$. It must not be open, because if it were, we would have $M \subseteq L^\oplus$ by the dual to Proposition 2. Then the dual of Proposition 5 (e) 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 structures 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 1 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 shall 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

Theorem 5. *Start with any language L, and apply the operators of positive closure and complement in any order, any number of times. Then at most* 10 *distinct languages are generated, and this bound is optimal. Furthermore, Table 1 classifies languages according to the algebra they generate and gives a language generating each algebra.*

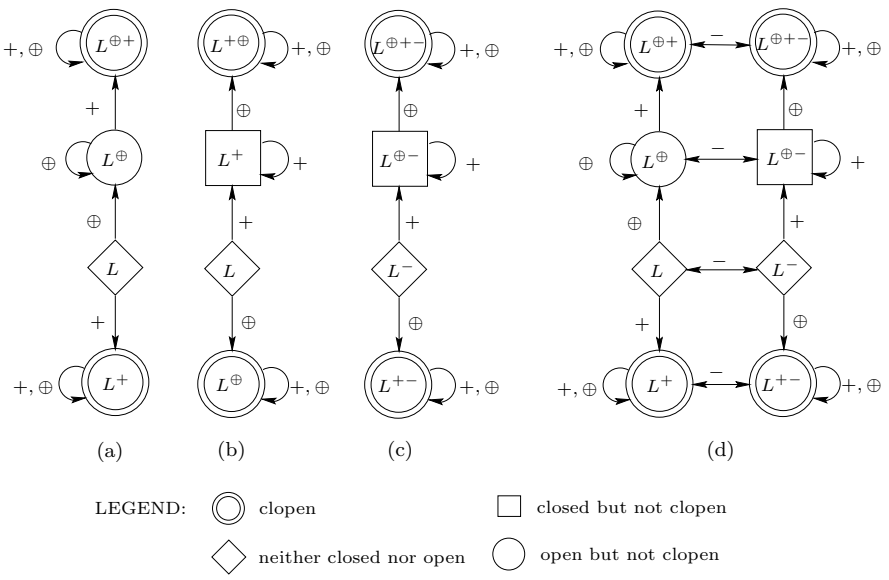

Fig. 1. Construction of $A(L)$, Case (6): (a) $(B(L),^{+},^{\oplus})$, Case (6); (b) $(B(L),^{+},^{\oplus})$, Case (7), the dual of Case (6) obtained by interchanging $+$ with $\oplus$, and "open" with "closed"; (c) $(C(L),^{+},^{\oplus})$, that is, $(B(L),^{+},^{\oplus})$, Case (7), with elements renamed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

In the unary case, we obtain the following:

Theorem 6. *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* 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.*

Note that all the example languages are regular. Hence Theorems 5 and 6 also hold for any regular language and any regular unary language, respectively.

4.2 Structures of the algebras with Kleene closure

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 *clopen* will still mean positive-clopen). We first claim the following, which can be proven in the same manner as Proposition 6:

Proposition 7. *Let $L \subseteq \Sigma^{*}$. Then $D(L) = E(L) \cup F(L)$, and the union is disjoint.*

Case	Necessary and Sufficient Conditions	$\|B(L)\|$	$\|A(L)\|$	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	$a \cup abaa$	(7)
(7)	L is neither open nor closed; $L^\oplus$ is closed but L^+ is not open; $L^{+\oplus} \neq L^\oplus$.	4	8	$(a \cup b)^* \setminus (a \cup abaa)$	(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)

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

Next, we give a way of relating $E(L)$ to $B(L)$. We recall that $L^* = L^+ \cup \{\epsilon\}$ and $L^\circledast = L^\oplus \setminus \{\epsilon\}$. 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:

Proposition 8. *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 shall 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 thus $E(L) = \{L, L \cup \{\epsilon\}, L^{\circledast}, L^{\circledast} \cup \{\epsilon\}\}$. 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 2. *Let $L \subseteq \Sigma^*$ be neither open nor closed. Then*

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

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 2 allows us to describe the structure of the algebra $(E(L), ^*, ^{\circledast})$ in Cases (4) through (9). 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.

We now summarize our results for the Kleene case:

Theorem 7. *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.*

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	$a \cup abaa$	(7)
(7)	L is neither open nor closed; $L^\oplus$ is closed but L^+ is not open; $L^{+\oplus} \neq L^\oplus$.	6	12	$(a \cup b)^* \setminus (a \cup abaa)$	(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)

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

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

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.

Here, we deal mainly with positive closure, but most of our theorems have obvious analogues for the Kleene closure. However, the presence or absence of ϵ can be crucial when dealing with concatenation of languages, so we mention a few exceptional cases where the choice of positive or Kleene closure is important.

Theorem 9. *Let $L, M \subseteq \Sigma^*$.*

(a) Suppose L is positive-closed, and let k be a positive integer. Then $L^k \subseteq L$ and L^k is positive-closed.
(b) Suppose L is Kleene-closed, and let k be a positive integer. Then $L^k = L$.
(c) Suppose L and M are positive-closed (respectively, Kleene-closed) and satisfy $LM = ML$. Then LM is positive-closed (respectively, Kleene-closed).
(d) Suppose L and M are positive-closed (respectively, Kleene-closed) unary languages. Then LM is positive-closed (respectively, Kleene-closed).

Proof. (a) If L is positive-closed then $L = L^+$, and $L^k \subseteq L^+ = L$. Also, for $k > 1$, $L^k L^k = L^{k-1} L^{k+1} \subseteq L^{k-1} L = L^k$ and L^k is positive-closed.
(b) If L is Kleene-closed, then $L^k = (L^*)^k \subseteq (L^*)^* = L^* = L$, and $L \subseteq L^* = (L^*)^k$.
(c) For positive closure, $LMLM = LLMM \subseteq LM$; hence LM is positive-closed.
(d) This is a special case of part (c), since unary languages commute.

■

Theorem 10. *Let $L, M \subseteq \Sigma^*$. Suppose L and M are positive-closed (respectively, Kleene-closed) and such that $L \cup M$ is positive-closed. Then*

(a) LM is positive-closed (respectively, Kleene-closed).
(b) More generally, consider the semigroup of languages $\{L, M\}^+$ generated by L and M. Let $W \in \{L, M\}^+$. Then W is positive-closed (respectively, Kleene-closed) when considered as a language over Σ.

Proof. (a) It suffices to show that $(LM)^k \subseteq LM$; we do this by induction on k. For $k > 1$, $(LM)^k \subseteq L(L \cup M)(L \cup M)M(LM)^{k-2} \subseteq L(L \cup M)M(LM)^{k-2} = (LLM \cup LMM)(LM)^{k-2} \subseteq LM(LM)^{k-2} = (LM)^{k-1} \subseteq LM$.
(b) The cases where $W = L^k$ or $W = M^k$ are proven by Theorem 9, so we may assume that W contains at least one L and one M (when considered as a word in $\{L, M\}^+$.) This implies that either LM or ML is a factor of W. Without loss of generality suppose that LM is a factor of W. Let $W = W_1 W_2 \cdots W_k W_{k+1} \cdots W_n$ where $W_i \in \{L, M\}$ for all i, and specifically $W_k = L$ and $W_{k+1} = M$. Now, to prove that W is closed, we let $u, v \in W$ be words in Σ^*. We show that $uv \in W$. Let $u = u_1 \cdots u_n$ and $v = v_1 \cdots v_n$ where $u_i, v_i \in W_i$ for all i. Consider $x = u_{k+1} \cdots u_n v_1 \cdots v_k$, a factor of uv. We see that $x \in (L \cup M)^+$. But since $L \cup M$ is positive-closed, $(L \cup M)^+ = L \cup M$, and hence either $x \in L$ or $x \in M$. If $x \in L$, then $u_k x \in L = W_k$ by closure of L and thus $uv = u_1 \cdots u_{k-1}(u_k x) v_{k+1} \cdots v_n \in W_1 \cdots W_n = W$. If $x \in M$, then $x v_{k+1} \in M = W_{k+1}$ by closure of M and thus $uv = u_1 \cdots u_k (x v_{k+1}) v_{k+2} \cdots v_n \in W_1 \cdots W_n = W$. So we must have $uv \in W$ in either case, and thus W is closed. For the Kleene-closed case, we again simply note that if $\epsilon \in L$ and $\epsilon \in M$, then $\epsilon \in W$. ■

Theorem 11. *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 $L = \emptyset$.
(d) 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.

Proof. (a) Let $ab \in LM$ where $a \in L$ and $b \in M$. Let $ab = uv$ for some words u and v. To prove that LM is open, we must show that either $u \in LM$ or $v \in LM$. We have two cases: either u is a prefix of a, or v is a suffix of b.

If u is a prefix of a, let $a = ux$, so $ab = uxb$ and hence $v = xb$. Since L is open, applying Proposition 5 (b) to $a \in L$ implies that either $u \in L$ or $x \in L$. If $u \in L$, then, since $\epsilon \in M$, we have $u = u\epsilon \in LM$ and we are done. If $x \in L$, then $v = xb \in LM$ and we are also done.

The case where v is a prefix of b is similar and relies on the fact that $\epsilon \in L$.

(b) If $L = \emptyset$ or $M = \emptyset$, then $LM = \emptyset$, which is open. Conversely, if $\epsilon \notin L$, $\epsilon \notin M$, and neither $L = \emptyset$ nor $M = \emptyset$, then LM is non-empty but contains no words of length 0 or 1 and is thus not open.

(c) This follows immediately from parts (a) and (b).

(d) If $L = \{\epsilon, a, aaa, aaaaa\}$ and $M = \{a\}$ (which are both easily verified to be open), then we have $aaaaaa \in LM$, but $aaa \notin LM$, and thus LM is not open. On the other hand, if $L = \{\epsilon, a, aaa\}$ and $M = \{a\}$, then $LM = \{a, aa, aaaa\}$, which is clearly open. ■

Theorem 12. *Let $L, M \subseteq \Sigma^*$ both be clopen.*

(a) If $L \cup M = \Sigma^*$, then LM is clopen.
(b) Suppose that $L \cup M = \Sigma^*$ and consider the semigroup of languages $\{L, M\}^+$ generated by L and M. Let $W \in \{L, M\}^+$. Then W is clopen if and only if $W = \emptyset$ or W contains at most one occurrence of a language which does not contain ϵ.
(c) The converses of the above statements are false; indeed, it is possible that LM is clopen, but $L \cup M$ is not even positive-closed.

Proof. (a) From Theorem 10 (a) we have that LM is closed, since Σ^* is closed. To show that LM is open, let $ab \in LM$ where $a \in L$ and $b \in M$. Let $ab = uv$ for some words u and v. To prove that LM is open, we must show that either $u \in LM$ or $v \in LM$. There are two cases: either u is a prefix of a, or v is a suffix of b.

Without loss of generality, we assume that u is a prefix of a and let $a = ux$, so $ab = uxb$ and hence $v = xb$. Since L is open, applying Proposition 5 (b) to $a \in L$ implies that 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 $L \cup M = \Sigma^*$. In this case, we note that $xb \in M$ since $x \in M$, $b \in M$, and M is closed. Then $\epsilon xb = v \in LM$. So in all cases, we have either $u \in LM$ or $v \in LM$. Thus LM is open and hence is clopen.

(b) Let $W = W_1W_2 \cdots W_n$ where $W_i \in \{L, M\}$ for all i. By Theorem 10 (b), W is closed. If each W_i contains ϵ, then W is open by repeated applications of Theorem 11 (a) and is thus clopen.

If there exist i and j with $i \neq j$, $\epsilon \notin W_i$, and $\epsilon \notin W_j$, then W contains no words of length 1, so either $W = \emptyset$ or W is not open (and thus not clopen).

Finally, we deal with the case where there exists a unique i such that $\epsilon \notin W_i$. Suppose, without loss of generality, that $W_i = M$. Then $W = L^{i-1}ML^{n-i}$. Since $L \cup M$ is Kleene-closed, it must contain ϵ, so $\epsilon \in L$. Thus $L^k = L$ for all positive k by Theorem 9 (b), so we must have $W = M$, $W = LM$, $W = ML$, or $W = LML$. In the first case, $W = M$ is known to be clopen, and in the second and third cases, W is clopen by part (a). Thus we must only consider the case where $W = LML$. We know that LM is clopen by part (a). Furthermore, $M \subseteq LM$ since $\epsilon \in L$, so $LM \cup L \supseteq M \cup L = \Sigma^*$ and thus $LM \cup L = \Sigma^*$. Thus we can apply part (a) on LM and L, proving that LML is clopen.

(c) 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. As we proved in Example 1, L and M are both clopen. Furthermore, L and M both contain ϵ, so LM is open by Theorem 11.

Next, we show that LM is closed. Let $u, v \in LM$, then let $u = u_1u_2$ and $v = v_1v_2$, where $u_1, v_1 \in L$ and $u_2, v_2 \in M$. We observe that $|u_1|_a < |u_1|_b$ and $|v_2|_a > |v_2|_b$. We examine the factor u_2v_1 and consider two cases. If $|u_2v_1|_a \geq |u_2v_1|_b$, then $|u_2v_1v_2|_a > |u_2v_1v_2|_b$ and thus $u_2v_1v_2 \in M$. Since $u_1 \in L$, we must then have $uv = u_1u_2v_1v_2 \in LM$. Similarly, if $|u_2v_1|_a \leq |u_2v_1|_b$, then $|u_1u_2v_1|_a < |u_1u_2v_1|_b$ and thus $u_1u_2v_1 \in L$. Since $v_2 \in M$, we must then have $uv = u_1u_2v_1v_2 \in LM$. So in all cases, $uv \in LM$, and LM is closed. Hence LM is clopen.

However, $L \cup M$ 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 3. *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 $M = L^- \cap \{x \in \Sigma^+ : |x| \leq |w|\}$. This is clearly finite, and is open by the dual to part (e) of Proposition 5. ■

Theorem 13. *Let $u, v \in \Sigma^+$.*

(a) 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 $u \neq v^k$.
(b) If $u \neq v$, then either there exists an open language L with $u \in L$ and $v \notin L$, or there exists an open language L with $u \notin L$ and $v \in L$. (In other words, all words are distinguishable by open languages.)

Proof. (a) 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 5 (b). For the reverse direction, we apply Lemma 3 to u and $\{v\}^+$, which is closed.
(b) Without loss of generality, let $|u| \leq |v|$. This implies that, for all k, $u \neq v^k$, and hence the claim follows from (a). ■

We now recall a basic result from combinatorics on words (see, e.g., [12]). Recall that a word w is *primitive* if it cannot be expressed in the form x^k for a word x and an integer $k \geq 2$.

Lemma 4. *Let $u, v \in \Sigma^+$. The following are equivalent:*

(1) $uv = vu$, that is, u and v commute.
(2) There exists a word x and integers $p \geq 1$ and $q \geq 1$ such that $u = x^p$ and $v = x^q$.
(3) There exists a word y and integers $p \geq 1$ and $q \geq 1$ such that $y = u^p$ and $y = v^q$.
(4) u and v are each a power of the same primitive word.

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*.

Theorem 14. *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. We handle the forward direction first. Suppose a clopen language L exists with $u \in L$ and $v \notin L$. If u and v commute, then there exists a word x and integers p and q such that $u = x^p$ and $v = x^q$. In particular, this implies that any open set containing u will also contain x, and any open set containing v will also contain x. Then we must have both $x \in L$ (since L is open and contains u) and $x \in L^-$, since L^- is open and contains v. Thus we have a contradiction, and u and v must not commute.

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 1) 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 $uv = \phi(p)\phi(q) = \phi(pq) = \phi(qp) = \phi(q)\phi(p) = vu$, since ϕ is a morphism. This is impossible since $uv \neq vu$, so p and q do not commute. 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 will 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. ■

Corollary 2. *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. As in the proof of Theorem 14, we note that if u and v commute, then there is some x such that $u = x^p$ and $v = x^q$, implying that every open language containing u or v must contain x, and thus there is no open language containing u but not v. If u and v do not commute, then by our theorem, 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 our Proposition 5 (b) since K and K^- are both open. ■

We can also use Theorem 14 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 created by any closure operator; it need not be topological). Since Theorem 14 implies that $u \sim v$ if and only if $u = x^p$ and $v = x^q$ for some integers p and q, it follows that each connected component of Σ^+ consists of a primitive word and all of its powers. Connected components of other languages will simply consist of collections of words sharing a common primitive root.

Note that connected components must be closed, but they need not be clopen. In fact, the only clopen components of Σ^+ are the languages $\{a\}^+$ for each $a \in \Sigma$.

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

Theorem 15. *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 it follows that L^+ and M^+ are disjoint. ■

Corollary 3. *Let $L, M \subseteq \Sigma^*$ 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 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 16. *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 $q_0 \in F$. For the open case, we start with a DFA for $\overline{L}$.

It is easy to verify that $L(M)L(M)$ can be accepted by an NFA with $2n$ states, and therefore the language $(L(M))^2 \setminus L(M)$ can be accepted by an NFA with $O(n^2)$ states. For details of the construction, see [3]. ■

From Proposition 5 (a), we know that L is not closed if and only if there exists a word $uv \notin L$ such that $u, v \in L$. We call such a word a *counterexample*.

Corollary 4. *If L is a regular language, accepted by a n-state DFA, that is not closed, then the smallest counterexample is of length $\leq n^2 + n - 1$.*

This $O(n^2)$ upper bound on the length of the shortest counterexample is matched by a corresponding $\Omega(n^2)$ lower bound:

Theorem 17. *There exists a class of DFA's M_n with $2n + 5$ states, having the following property: a shortest word $x \notin L(M_n)$ such that there exist $u, v \in L(M_n)$ with $x = uv$ is of length $n^2 + 2n + 2$.*

Proof. It is easier to describe DFA $M'_n = (Q, \Sigma, \delta, q_0, F)$ that accepts the complement of $L(M_n)$. In other words, we will show that a shortest word $x \in L(M'_n)$ such that there exist $u, v \notin L(M_n)$ with $x = uv$ is of length $n^2 + 2n + 2$. 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 the prefix ending at r. Thus either x' is of the form $10^{in+n-1}110^k$

$a\backslash q$	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

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

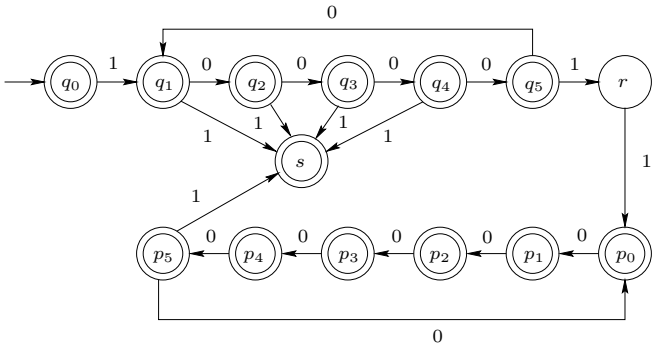

Fig. 2. Example of DFA M_n for $n = 5$. Unspecified transitions go to the dead state d.

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 $j(n+1)+n$ is of the form $in+n-1$. If $in+n-1 = j(n+1)+n$, then 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 easily be accomplished via a trivial modification of the proof given in [1], since the word w encodes a configuration of the Turing machine T.

Theorem 18. *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 4 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 $uv \notin L$. While we guess u and v, 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 [6], 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 18.

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. Closures in formal languages and Kuratowski's theorem. Preprint, `http://arxiv.org/abs/0901.3761`, January 2009.
3. J. Brzozowski, E. Grant, and J. Shallit. Closures in formal languages: concatenation, separation, and algorithms. `http://arxiv.org/abs/0901.3763`, January 2009.
4. S. N. Burris and H. P. Sankappanavar. *A Course in Universal Algebra*, 2nd edition. Available at `http://www.math.uwaterloo.ca/~snburris/htdocs/ualg.html`.
5. A. V. Chagrov. Kuratowski numbers. In *Application of Functional Analysis in Approximation Theory*, Kalinin. Gos. Univ., Kalinin, 1982, pp. 186–190. In Russian.
6. 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.
7. J. H. Fife. The Kuratowski closure-complement problem. *Math. Mag.* **64** (1991), 180—182.
8. B. J. Gardner and M. Jackson. The Kuratowski closure-complement theorem. *New Zealand J. Math.,* to appear. Preprint available at `http://www.latrobe.edu.au/mathstats/department/algebra-research-group/Papers/GJ_Kuratowski.pdf`
9. R. L. Graham, D. E. Knuth, and T. S. Motzkin. Complements and transitive closures. *Discrete Math.* **2** (1972), 17–29.
10. P. C. Hammer. Kuratowski's closure theorem. *Nieuw Archief v. Wiskunde* **7** (1960), 74–80.
11. C. Kuratowski. Sur l'opération $\overline{A}$ de l'analysis situs. *Fund. Math.* **3** (1922), 182–199.
12. 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.
13. D. Peleg. A generalized closure and complement phenomenon. *Discrete Math.* **50** (1984), 285–293.