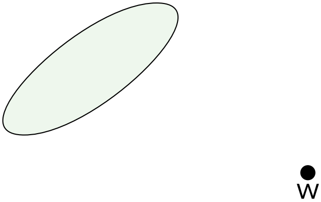
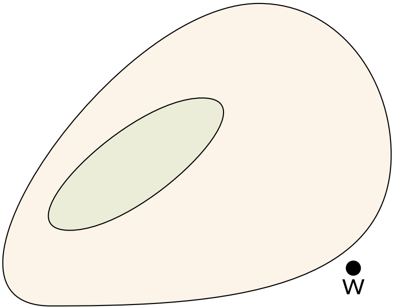
# State Complexity of Suffix Distance

Timothy Ng David Rappaport Kai Salomaa

School of Computing, Queen's University, Kingston, Canada

DCFS 2017, Milan, Italy


w


w

We are interested in the state complexity of neighbourhoods of regular language classes with respect to the suffix distance.

We are interested in the state complexity of neighbourhoods of regular language classes with respect to the suffix distance. We show state complexity bounds for the following classes:

- regular languages
- finite languages
- suffix-closed regular languages

A distance is a function $d : \Sigma^* \times \Sigma^* \to [0, \infty)$ such that

1. $d(x, y) = 0$ if and only if $x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, y) \leq d(x, w) + d(w, y)$

The neighbourhood of a language $L \subseteq \Sigma^*$ of radius $k \geq 0$ with respect to a distance measure $d$ is the set of all words $u$ with $d(w, u) \leq k$ for some $w \in L$,

$$E(L, d, k) = \{u \in \Sigma^* : (\exists w \in L)\, d(w, u) \leq k\}.$$

c oncatenate

co ncatenate

con catenate

⋮

concatenat e

- Additive distances are regularity preserving (Calude, Salomaa, Yu 2002)

- Additive distances are regularity preserving (Calude, Salomaa, Yu 2002)
- The state complexity of these neighbourhoods is $(k+2)^n$
  - Upper bound (Salomaa, Schofield 2007)
  - Lower bound (Ng, Rappaport, Salomaa 2015)

The prefix distance of $x$ and $y$ counts the number of symbols which do not belong to the longest common prefix of $x$ and $y$.

The prefix distance of $x$ and $y$ counts the number of symbols which do not belong to the longest common prefix of $x$ and $y$.

concatenate
concavity

Prefix distance is not additive, but neighbourhoods are still regular.

Prefix distance is not additive, but neighbourhoods are still regular. The state complexity of these neighbourhoods is $O(n^2)$ (Ng, Rappaport, Salomaa 2015).

The suffix distance of $x$ and $y$ counts the number of symbols which do not belong to the longest common suffix of $x$ and $y$.

The suffix distance of $x$ and $y$ counts the number of symbols which do not belong to the longest common suffix of $x$ and $y$.

prefix
suffix

The suffix distance of $x$ and $y$ counts the number of symbols which do not belong to the longest common suffix of $x$ and $y$.

prefix
suffix

$$d_s(x, y) = d_p(x^R, y^R)$$

### Theorem (NRS2015)

*Let $k \geq 0$ and $L$ be a regular language recognized by an NFA with $n$ states. Then there exists an NFA recognizing $E(L, d_s, k)$ with at most $n + k$ states.*

concatenate

concatenate

xzxxconcatenate

concatenate

xzxxconcatenate

atenate

Given a DFA $A = (Q, \Sigma, \delta, q_0, F)$, we want to construct a DFA that recognizes $E(L, d_s, k)$.

Given a DFA $A = (Q, \Sigma, \delta, q_0, F)$, we want to construct a DFA that recognizes $E(L, d_s, k)$.
For each state $q \in Q$, we define the function

$$\psi_A(q) = \min_{w \in \Sigma^*} \{|w| \mid \delta(q_0, w) = q\}.$$

States are of the form

$$(i, P)$$

where $i = 0, \ldots, k$ and $P \subseteq Q$.

The initial state is

$$(0, \{q \in Q \mid \psi_A(q) \leq k - 1\}).$$

For $0 \leq i < k$ and $a \in \Sigma$,

$$\delta((i, P), a) = (i + 1, X),$$

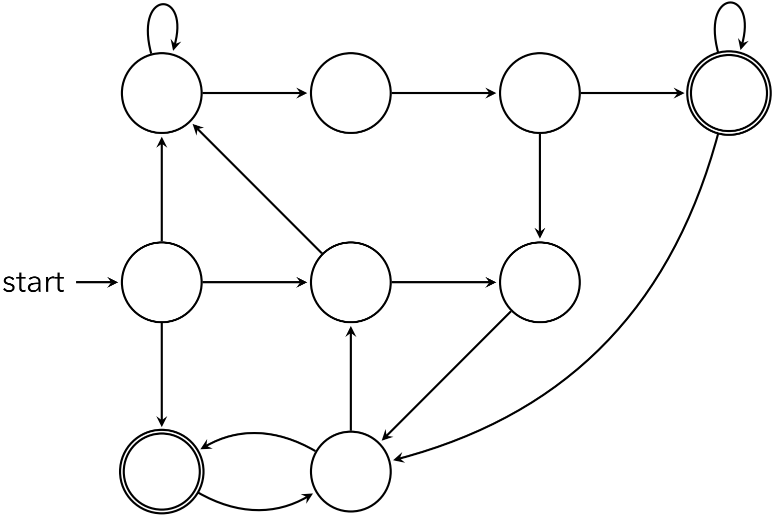
where $X = \{\delta(p, a) \mid p \in P\} \cup \{q \in Q \mid \psi_A(q) \leq k - (i + 1)\}$.

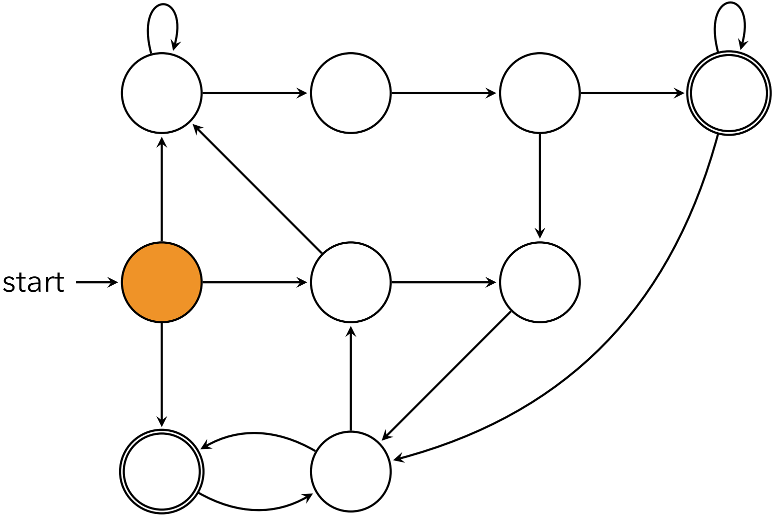
For $i = k$ and $a \in \Sigma$,

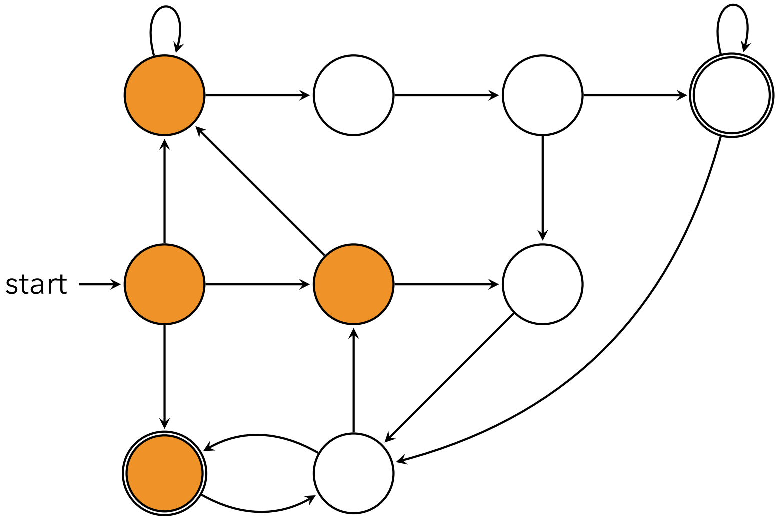
$$\delta((k, P), a) = (k, \{\delta(p, a) \mid p \in P\}).$$

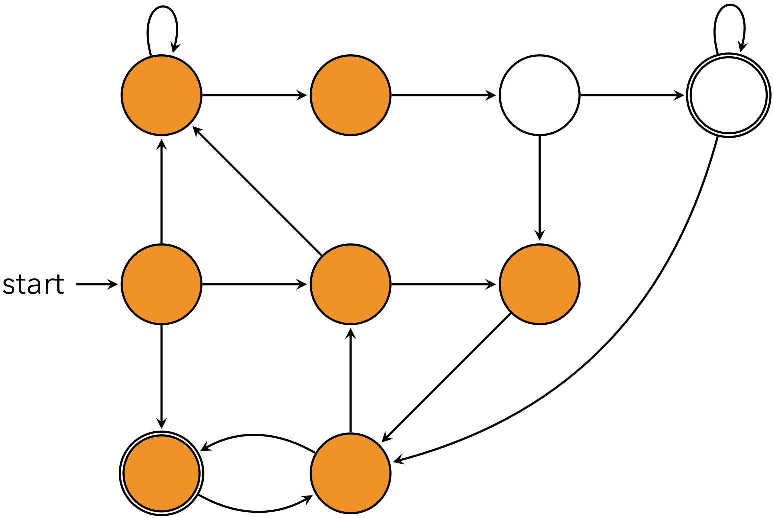
Then a word is accepted when it reaches a final state

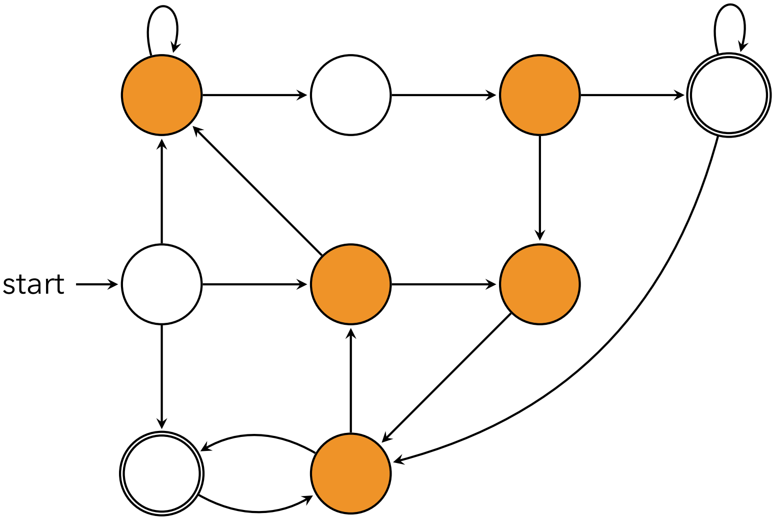
$$(i, \{P \subseteq Q \mid P \cap F \neq \emptyset\}).$$

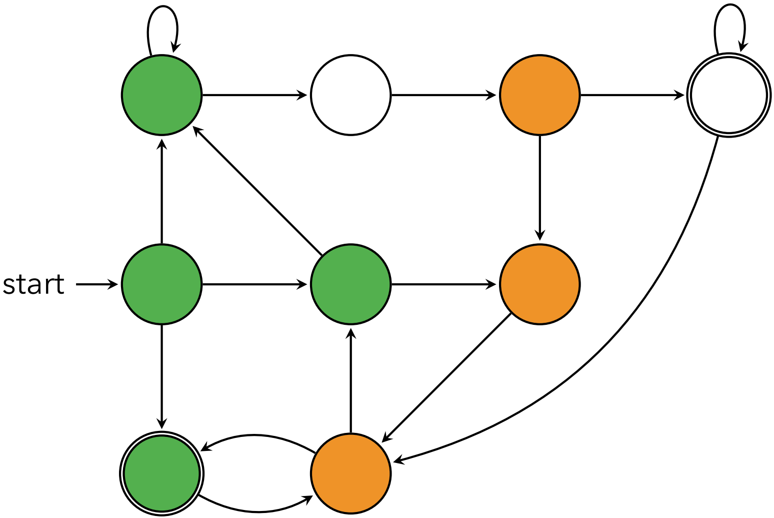

start

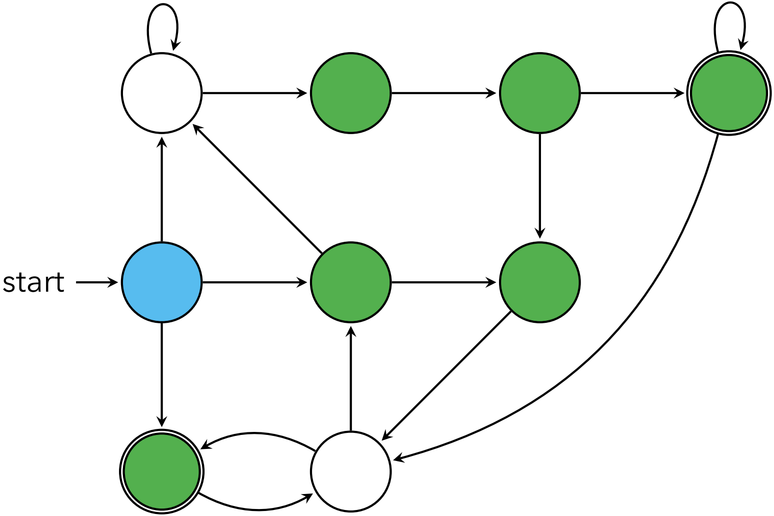

start


start


start


start


start


start

We have defined roughly $(k+1) \cdot 2^n - 1$ states. However, not all of these states are reachable.

We have defined roughly $(k+1) \cdot 2^n - 1$ states. However, not all of these states are reachable.
In total, there are at most

$$\frac{|\Sigma|^k - 1}{|\Sigma| - 1} + 2^n - 1$$

reachable states.

The state complexity of suffix distance neighbourhoods of radius $k$ given a DFA with $n$ states is

| | $n > k$ | $k \geq n$ |
|---|---|---|
| Regular | $\frac{\lvert\Sigma\rvert^k - 1}{\lvert\Sigma\rvert - 1} + 2^n - 1$ | $(k-n) + 2^{n+1} - 2$ |
| Finite | $2^k + k \cdot 2^{\lfloor \frac{n}{2} \rfloor} - 1$ | $(k-n) + 2^{n+1} - 2$ |
| Suffix-closed | $n + k + 1$ | |

The state complexity of suffix distance neighbourhoods of radius $k$ given a DFA with $n$ states is

| | $n > k$ | $k \geq n$ |
|---|---|---|
| Regular | $\frac{\lvert\Sigma\rvert^k - 1}{\lvert\Sigma\rvert - 1} + 2^n - 1$ | $(k-n) + 2^{n+1} - 2$ |
| Finite | $2^k + k \cdot 2^{\lfloor \frac{n}{2} \rfloor} - 1$ | $(k-n) + 2^{n+1} - 2$ |
| Suffix-closed | $n + k + 1$ | |

Given a DFA with $n$ states, there exists a DFA that recognizes subword distance neighbourhoods of radius $k$ with at most $\frac{\lvert\Sigma\rvert^k - 1}{\lvert\Sigma\rvert - 1} + (k+2) \cdot 2^{n \cdot (k+1)}$ states.

Future work

- Lower bounds for suffix distance neighbourhoods that don't depend on the alphabet or radius
- Lower bounds for subword distance neighbourhoods
- State complexity of suffix distance neighbourhoods for suffix-closed, -free, -convex languages