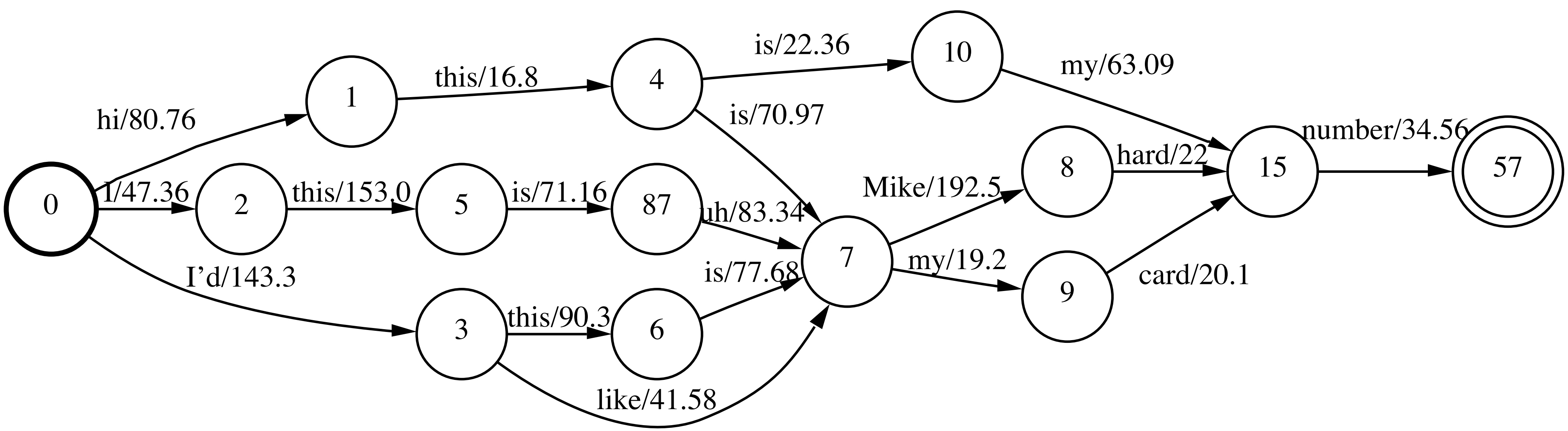
Weighted Transducers Theory and Algorithms

Mehryar Mohri
Courant Institute of Mathematical Sciences
Google Research
mohri@cims.nyu.edu

Tutorial joint work with Corinna Cortes (Google Research).

Speech Recognition

- Problem: assign a category (e.g., *referral*, *pre-certification*) to each speech utterance.
- Example:
  - Spoken utterance: *"Hi this is my number"*
  - Speech recognizer's output ('word lattice'):



Computational Biology: Similar Situation

- Problem: decide which class, e.g., *protein families*, *CpG islands*, a biosequence, or a group of biosequences, belongs to.
- Objects to classify:
  - Single protein sequence.
  - Protein clusters: represented or modeled by weighted automata.

General Problem

- Spoken-dialog classification
- Computational biology
- Information extraction
- Text mining
- Document classification
- Database queries

Motivation

- The objects to analyze in many modern applications are:
  - variable-length sequences.
  - distributions of sequences, typically weighted automata.
- How do we generalize learning algorithms originally designed for fixed-size vectors?
  - weighted automata and transducers.
  - sequence kernels, weighted automata kernels.

This Tutorial

- Weighted transducers theory and algorithms
- Kernels for computational biology and text and speech processing

Software Libraries

- FSM Library: Finite-State Machine Library. General software utilities for building, combining, optimizing, and searching weighted automata and transducers (MM, Pereira, and Riley, 2000).

  http://www.research.att.com/projects/mohri/fsm

- OpenFst Library: Open-source Finite-State Transducer Library. Jointly designed by Courant and Google (Allauzen, Riley, Schalkwyk, Skut, and MM, 2007).

  http://www.openfst.org

This Talk

- Definitions
- Composition
- Shortest-distance algorithms
- Epsilon-removal
- Determinization
- Pushing
- Minimization

Weight Sets: Semirings

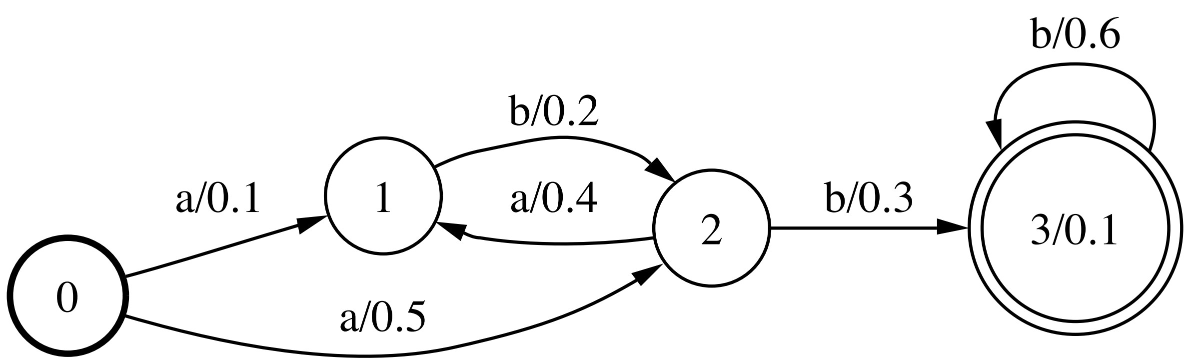
- A semiring $(\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1})$ is a ring that may lack negation.
  - sum: to compute the weight of a sequence (sum of the weights of the paths labeled with that sequence).
  - product: to compute the weight of a path (product of the weights of constituent transitions).

Semirings - Examples

Semiring	Set	$\oplus$	$\otimes$	$\overline{0}$	$\overline{1}$
Boolean	$\{0, 1\}$	$\vee$	$\wedge$	0	1
Probability	$\mathbb{R}_+$	$+$	$\times$	0	1
Log	$\mathbb{R} \cup \{-\infty, +\infty\}$	$\oplus_{\log}$	$+$	$+\infty$	0
Tropical	$\mathbb{R} \cup \{-\infty, +\infty\}$	$\min$	$+$	$+\infty$	0

with $\oplus_{\log}$ defined by: $x \oplus_{\log} y = -\log(e^{-x} + e^{-y})$.

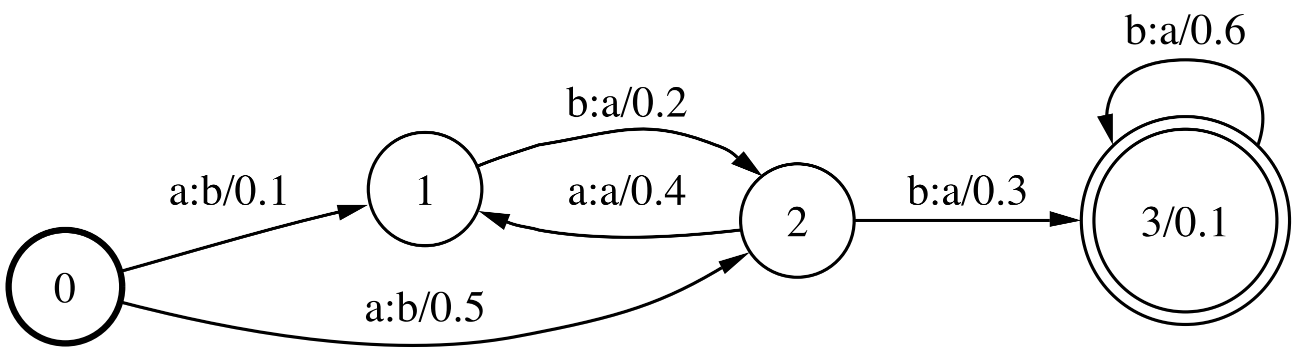
Weighted Automata



$[[A]](x) =$ Sum of the weights of all successful paths labeled with *x*

$$[[A]](abb) = .1 \times .2 \times .3 \times .1 + .5 \times .3 \times .6 \times .1$$

Weighted Transducers



$[[T]](x,y) =$ Sum of the weights of all successful paths with input *x* and output *y*.

$$[[T]](abb, baa) = .1 \times .2 \times .3 \times .1 + .5 \times .3 \times .6 \times .1$$

Rational Operations

- Sum

$$[\![T_1 \oplus T_2]\!](x,y) = [\![T_1]\!](x,y) \oplus [\![T_2]\!](x,y)$$

- Product

$$[\![T_1 \otimes T_2]\!](x,y) = \bigoplus_{\substack{x=x_1x_2\\ y=y_1y_2}} [\![T_1]\!](x_1,y_1) \otimes [\![T_2]\!](x_2,y_2).$$

- Closure

$$[\![T^*]\!](x,y) = \bigoplus_{n=0}^{\infty} [\![T]\!]^n(x,y)$$

This Talk

- Definitions
- Composition
- Shortest-distance algorithms
- Epsilon-removal
- Determinization
- Pushing
- Minimization

Composition

- **Definition**: given two weighted transducer T_1 and T_2 over a commutative semiring, the composed transducer $T = T_1 \circ T_2$ is defined by

$$(T_1 \circ T_2)(x, y) = \bigoplus_z T_1(x, z) \otimes T_2(z, y).$$

- **Algorithm**:
  - Epsilon-free case: matching transitions.
  - General case: ϵ-filter transducer.
  - Complexity: quadratic, $O(|T_1||T_2|)$.
  - On-demand construction.

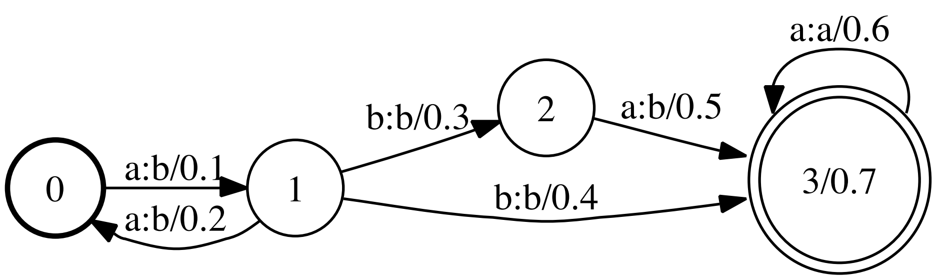
Epsilon-Free Composition

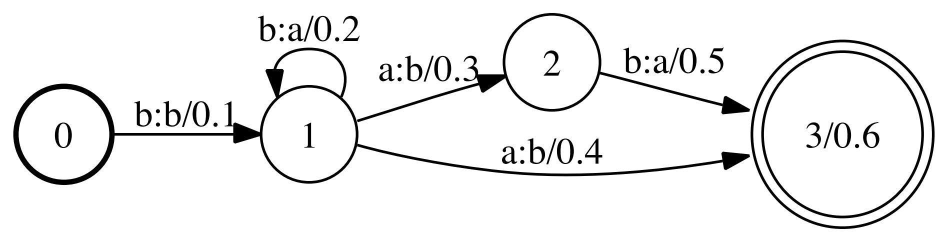
- States $Q \subseteq Q_1 \times Q_2$.
- Initial states $I = I_1 \times I_2$.
- Final states $F = Q \cap F_1 \times F_2$.
- Transitions

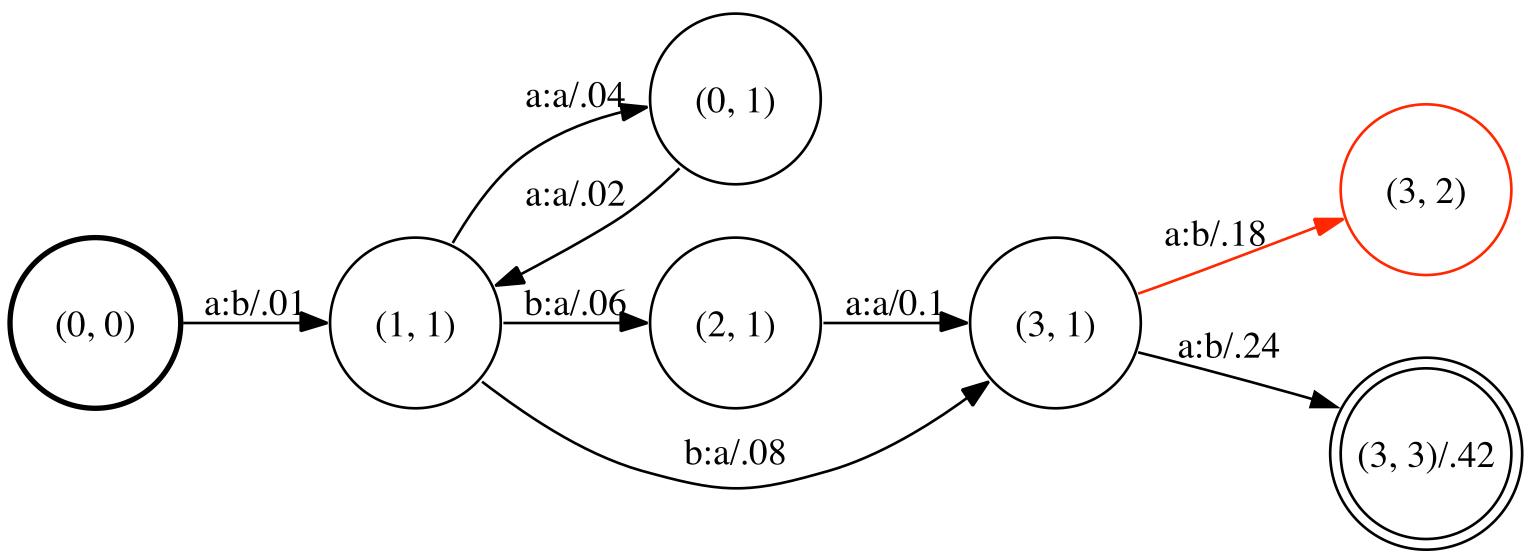
$$
\begin{aligned}
E = \{((q_1, q_1'), a, c, w_1 \otimes w_2, (q_2, q_2')) :& \\
(q_1, a, b, w_1, q_2), (q_1', b, c, w_2, q_2') \in Q\}&.
\end{aligned}
$$

Illustration

- **Program:** fsmcompose A.fsm B.fsm >C.fsm
  fstcompose A.fsm B.fsm >C.fsm

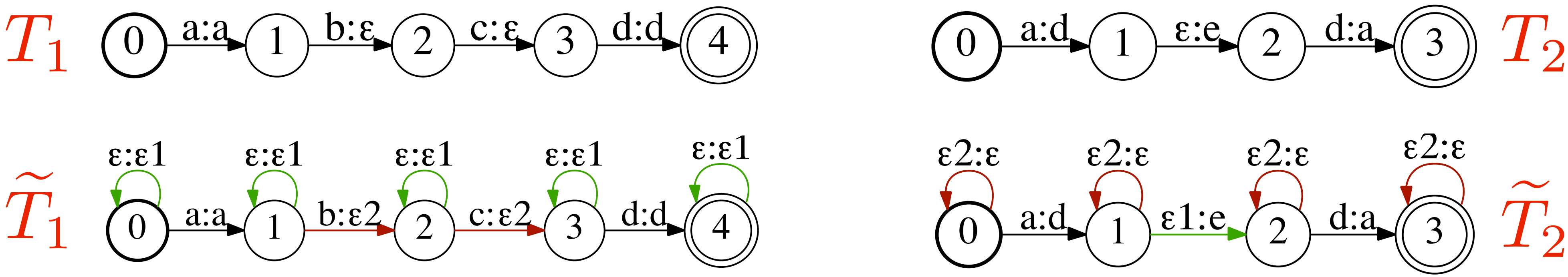


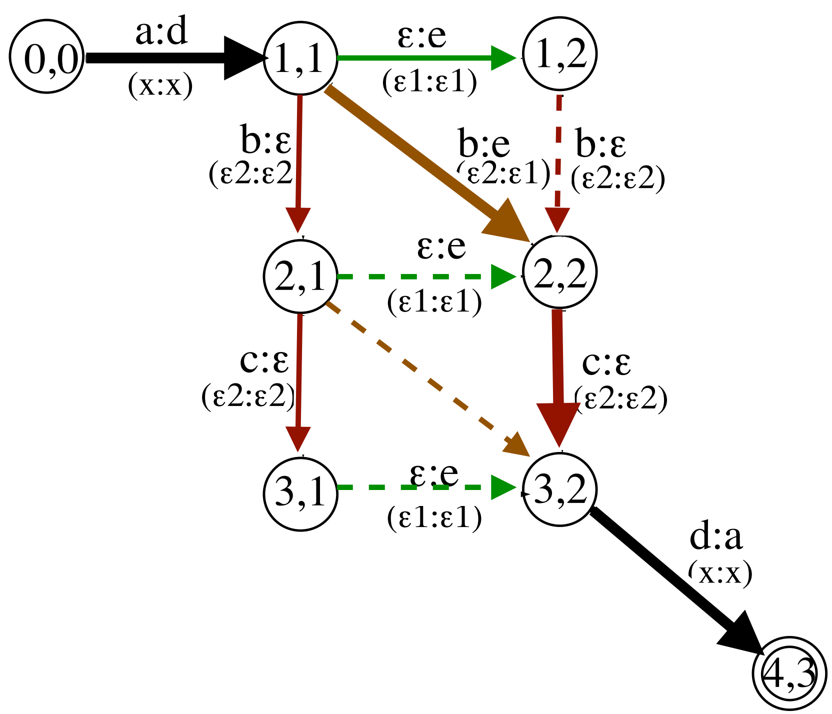


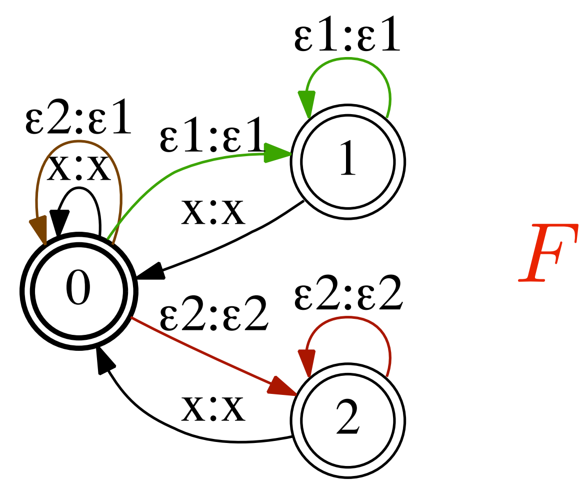


Redundant ϵ-Paths Problem

(MM, Pereira, Riley 1996)



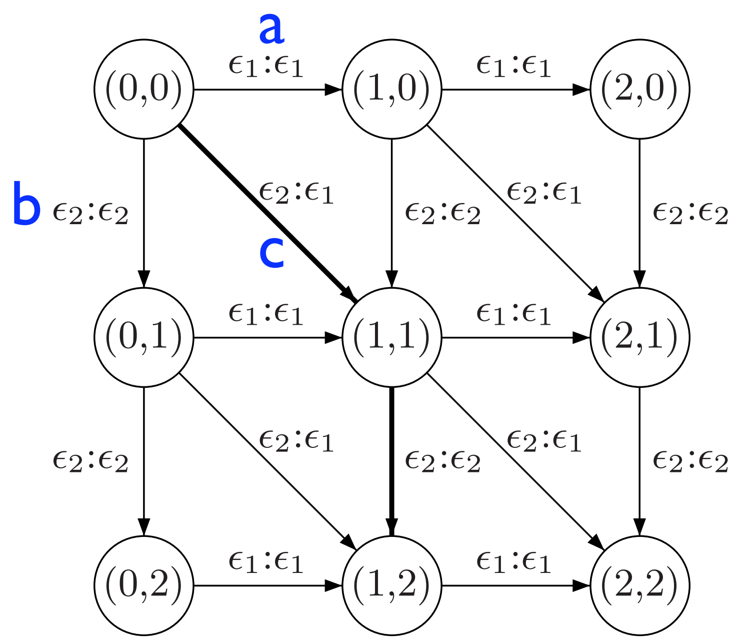




$$T = \widetilde{T}_1 \circ F \circ \widetilde{T}_2.$$

Correctness of Filter

- **Proposition**: filter F allows a unique path between two states of the following grid.



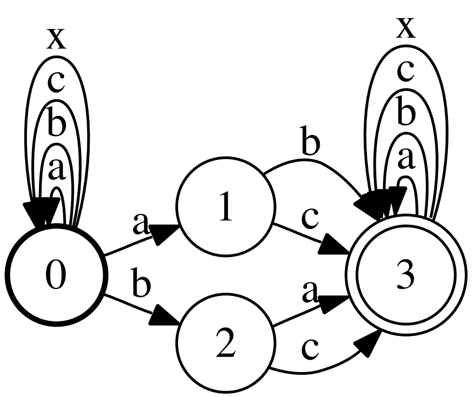
- **Proof**: Observe that a necessary and sufficient condition is that the following sequences be forbidden: ab, ba, ac, and bc.

Correctness of Filter

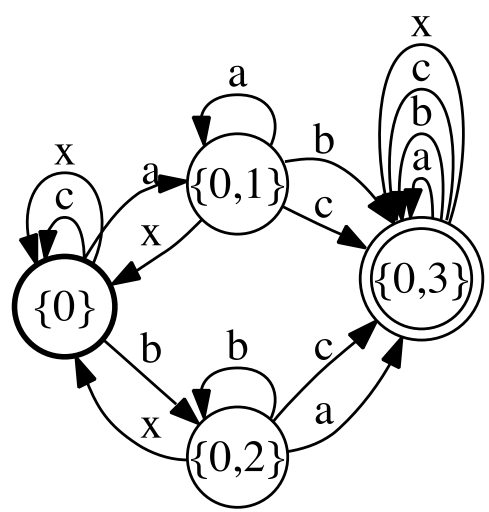
- Proof (cont.): Let $\sigma = \{a, b, c, x\}$, then set of sequences forbidden is exactly

$$L = \sigma^*(ab + ba + ac + bc)\sigma^*.$$

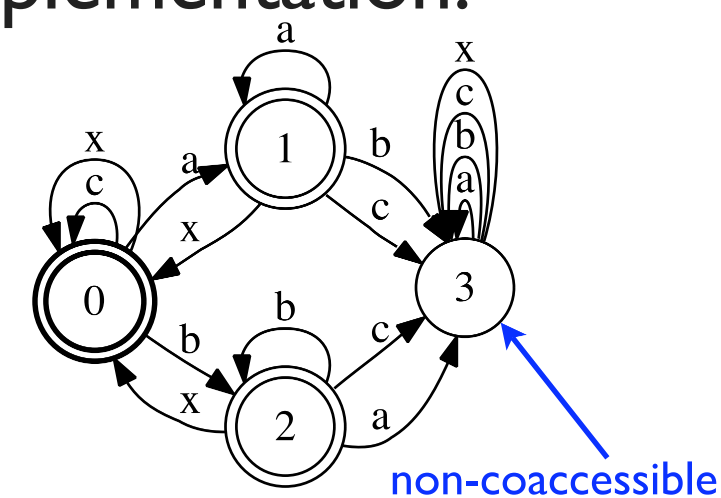
- An automaton representing the complement can be constructed by determ. and complementation.



A



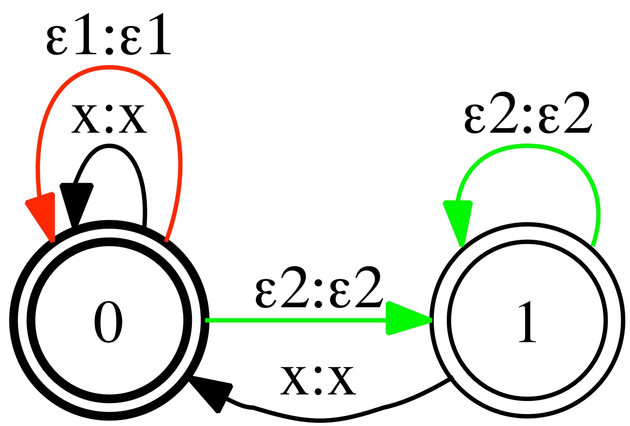
$det(A)$



$\overline{det(A)}$

Other Filters

(Pereira and Riley, 1997)



Sequential Filter.

This Talk

- Definitions
- Composition
- Shortest-distance algorithms
- Epsilon-removal
- Determinization
- Pushing
- Minimization

Shortest-Distance Problem

- Definition: for any regulated weighted transducer T, define the shortest distance from state q to F as

$$d(q, F) = \bigoplus_{\pi \in P(q,F)} w[\pi].$$

- Problem: compute $d(q, F)$ for all states $q \in Q$.
- Algorithms:
  - Generalization of Floyd-Warshall.
  - Single-source shortest-distance algorithm.

All-Pairs Shortest-Distance Algorithm

(MM, 2002)

- Assumption: closed semiring (not necessarily idempotent).
- Idea: generalization of Floyd-Warshall algorithm.
- Properties:
  - Time complexity: $\Omega(|Q|^3(T_\oplus + T_\otimes + T_\star))$.
  - Space complexity: $\Omega(|Q|^2)$ with an in-place implementation.

Closed Semirings

(Lehmann, 1977)

- Definition: a semiring is closed if the closure is well defined for all elements and if associativity, commutativity, and distributivity apply to countable sums.
- Examples:
  - Tropical semiring.
  - Probability semiring when including infinity or when restricted to well-defined closures.

Pseudocode

GENERIC-ALL-PAIRS-SHORTEST-DISTANCE (G)

for $i \leftarrow 1$ **to** $|Q|$
    do for $j \leftarrow 1$ **to** $|Q|$
        do $d[i, j] \leftarrow \bigoplus_{e \in P(i,j)} w[e]$
for $k \leftarrow 1$ **to** $|Q|$
    do for $i \leftarrow 1$ **to** $|Q|$
        do for $j \leftarrow 1$ **to** $|Q|$
            do $d[i, j] \leftarrow d[i, j] \oplus (d[i, k] \otimes d[k, k]^* \otimes d[k, j])$
for $k \leftarrow 1$ **to** $|Q|$
    do $d[k, k] \leftarrow \overline{1}$
return d

Single-Source Shortest-Distance Algorithm

(MM, 2002)

- Assumption: k-closed semiring.

$$\forall x \in \mathbb{K}, \quad \bigoplus_{i=0}^{k+1} x^i = \bigoplus_{i=0}^{k} x^i.$$

- Idea: generalization of relaxation, but must keep track of weight added to $d[q]$ since the last time q was enqueued.

- Properties:
  - works with any queue discipline and any k-closed semiring.
  - Classical algorithms are special instances.

Pseudocode

Generic-Single-Source-Shortest-Distance (G, s)

for $i \leftarrow 1$ **to** $|Q|$
    do $d[i] \leftarrow r[i] \leftarrow \overline{0}$
$d[s] \leftarrow r[s] \leftarrow \overline{1}$
$S \leftarrow \{s\}$
while $S \neq \emptyset$
    do $q \leftarrow head(S)$
        Dequeue(S)
        $r' \leftarrow r[q]$
        $r[q] \leftarrow \overline{0}$
        for each $e \in E[q]$
            do if $d[n[e]] \neq d[n[e]] \oplus (r' \otimes w[e])$
                then $d[n[e]] \leftarrow d[n[e]] \oplus (r' \otimes w[e])$
                    $r[n[e]] \leftarrow r[n[e]] \oplus (r' \otimes w[e])$
                    if $n[e] \notin S$
                        then Enqueue$(S, n[e])$
$d[s] \leftarrow \overline{1}$

Notes

- Complexity:
  - depends on queue discipline used.

$$O(|Q| + (T_\oplus + T_\otimes + C(A))|E| \max_{q \in Q} N(q) + (C(I) + C(E)) \sum_{q \in Q} N(q))$$

  - coincides with that of Dijkstra and Bellman-Ford for shortest-first and FIFO orders.
  - linear for acyclic graphs using topological order.

$$O(|Q| + (T_\oplus + T_\otimes)|E|)$$

- Approximation: ϵ-k-closed semiring, e.g., for graphs in probability semiring.

This Talk

- Definitions
- Composition
- Shortest-distance algorithms
- Epsilon-removal
- Determinization
- Pushing
- Minimization

Epsilon-Removal

(MM, 2002)

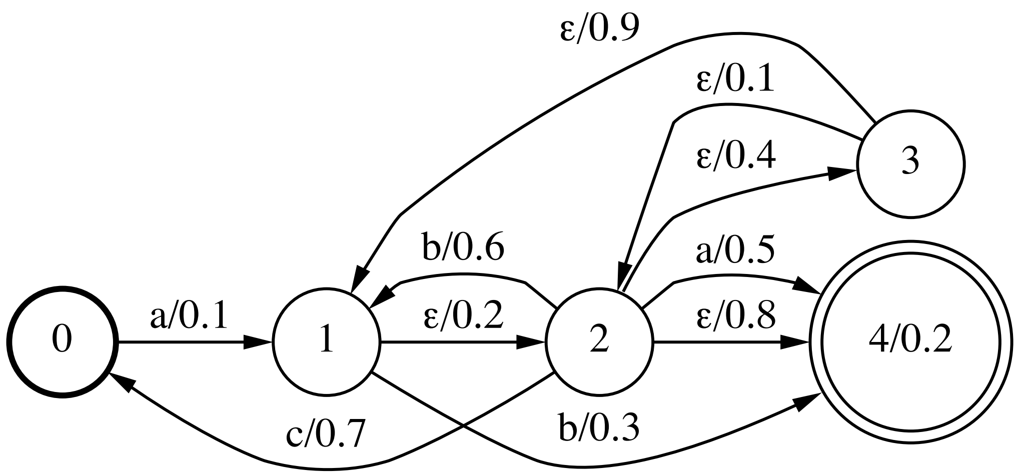
- **Definition**: given weighted transducer T, create equivalent weighted transducer with no epsilon-transition.

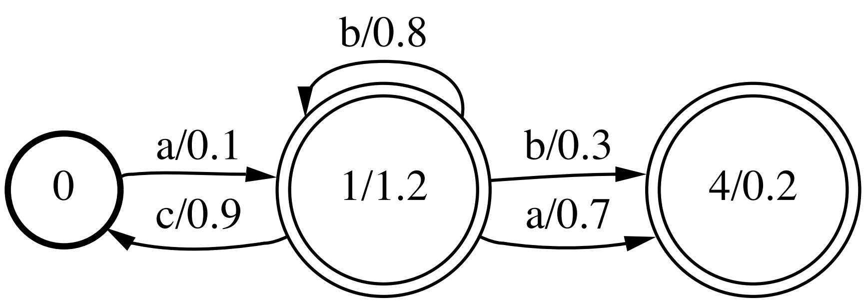
- **Algorithm components**:
  - Computation of the ϵ-closure at each state:

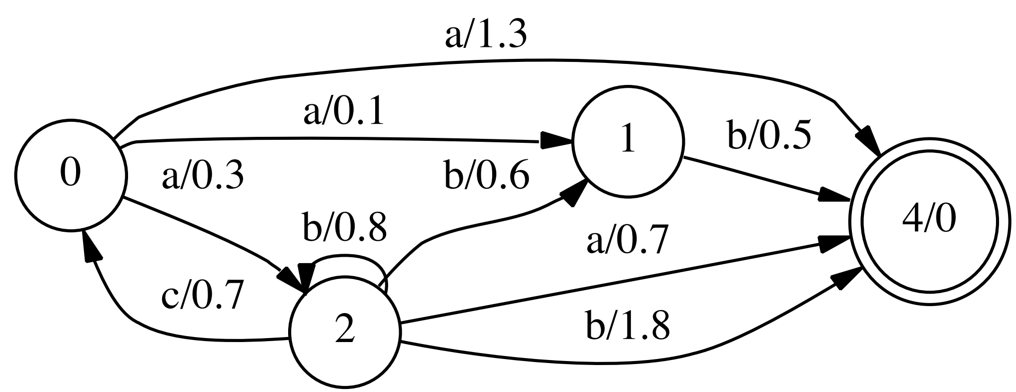
$$C[p] = \left\{(q, d_\epsilon[p,q]) : d_\epsilon[p,q] \neq \overline{0})\right\} \text{ with } d_\epsilon[p,q] = \bigoplus_{\pi \in P(p,\epsilon,q)} w[\pi].$$

  - Removal of ϵs.
  - On-demand construction.

Illustration







Main Algorithm

- Shortest-distance algorithms:
  - closed semirings: generalization of Floyd-Warshall algorithm.
  - *k*-closed semirings: single-source shortest-distance algorithm.
- Complexity: shortest-distance and removal.
  - Acyclic T_ϵ: $O(|Q|^2 + |Q||E|(T_\oplus + T_\otimes))$.
  - General case, tropical semiring:

$$O(|Q||E| + |Q|^2 \log |Q|).$$

This Talk

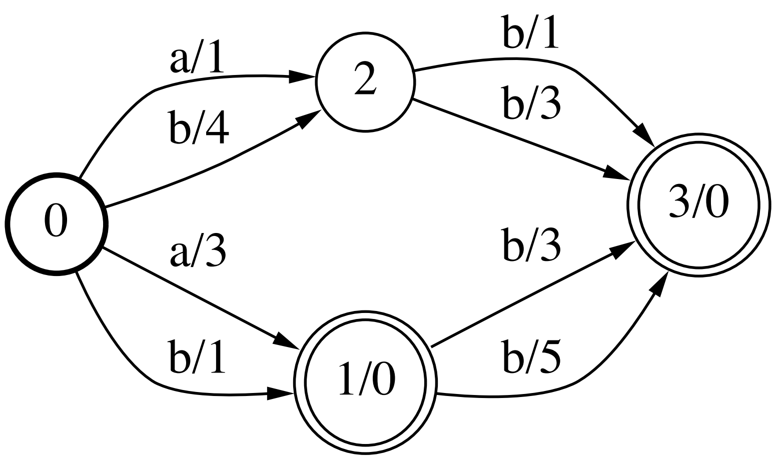
- Definitions
- Composition
- Shortest-distance algorithms
- Epsilon-removal
- Determinization
- Pushing
- Minimization

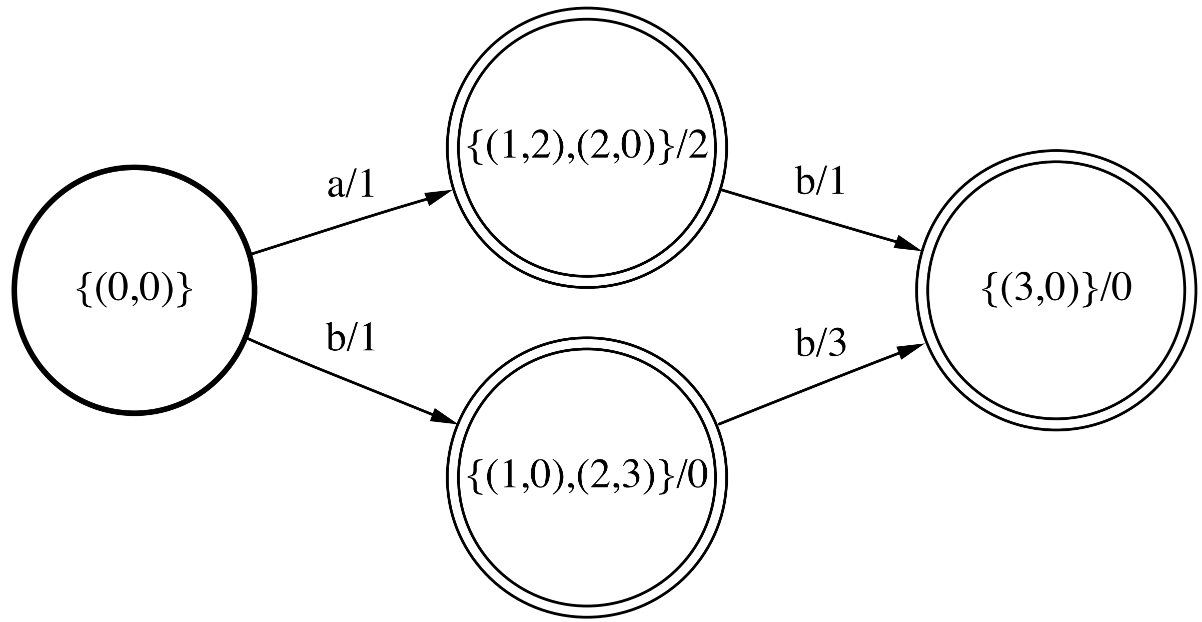
Determinization

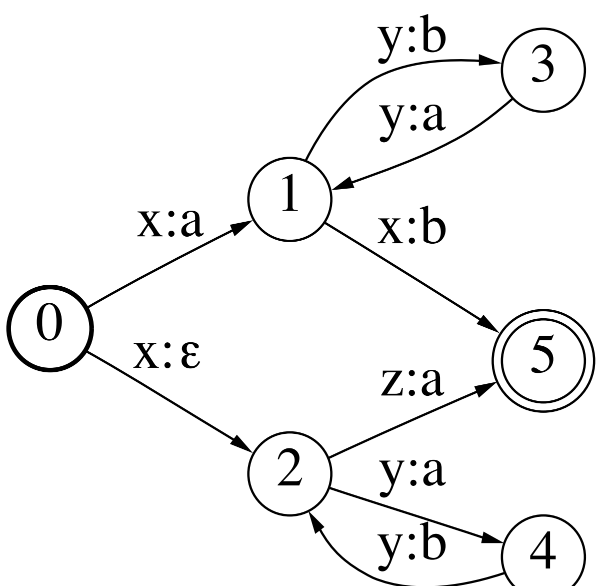
(MM, 1997)

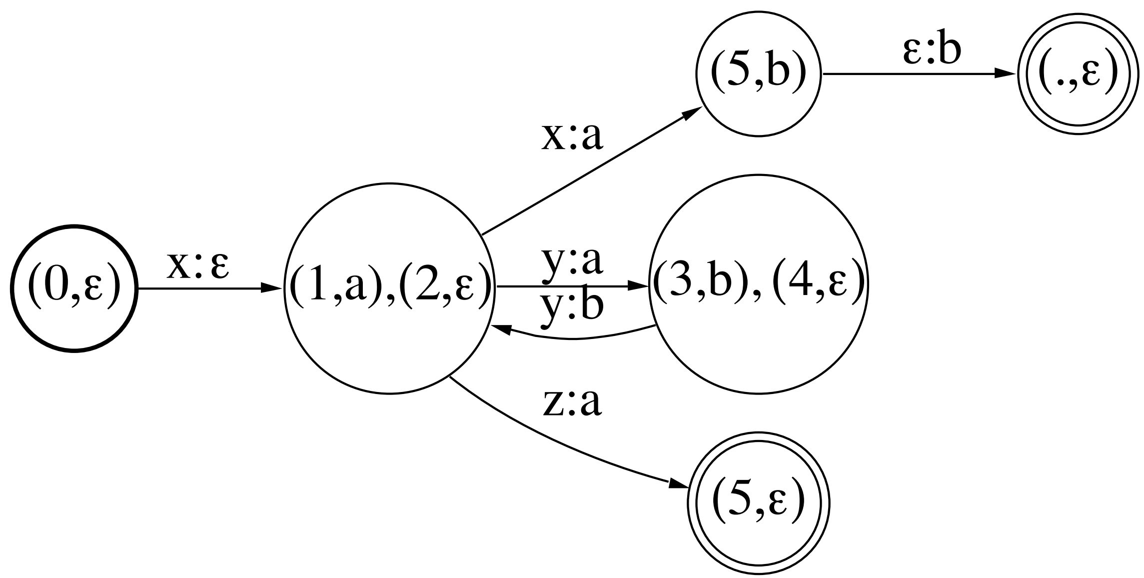
- Definition: given weighted transducer T, create equivalent non-deterministic weighted transducer.
- Algorithm (weakly left divisible semirings):
  - generalization of subset constructions to weighted labeled subsets
    $$\{(q_1, x_1, w_1), \ldots, (q_m, x_m, w_m)\}\,.$$
  - complexity: exponential, but lazy implementation.
  - not all weighted transducers are determinizable but all acyclic weighted transducers are. Test? For some cases, using the twins property.

Illustration

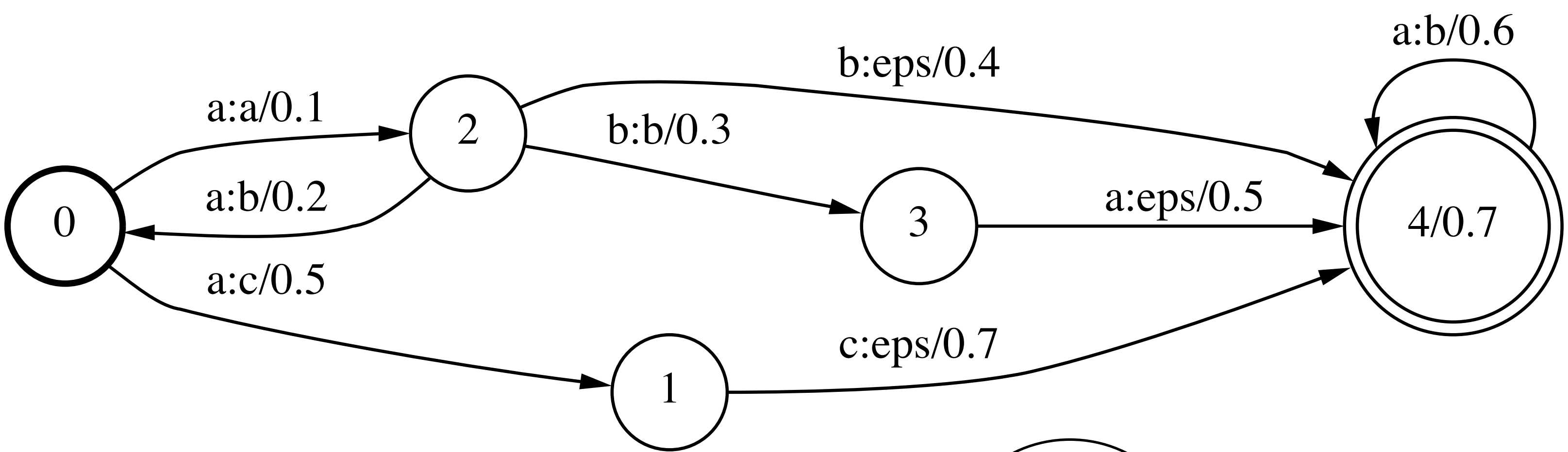


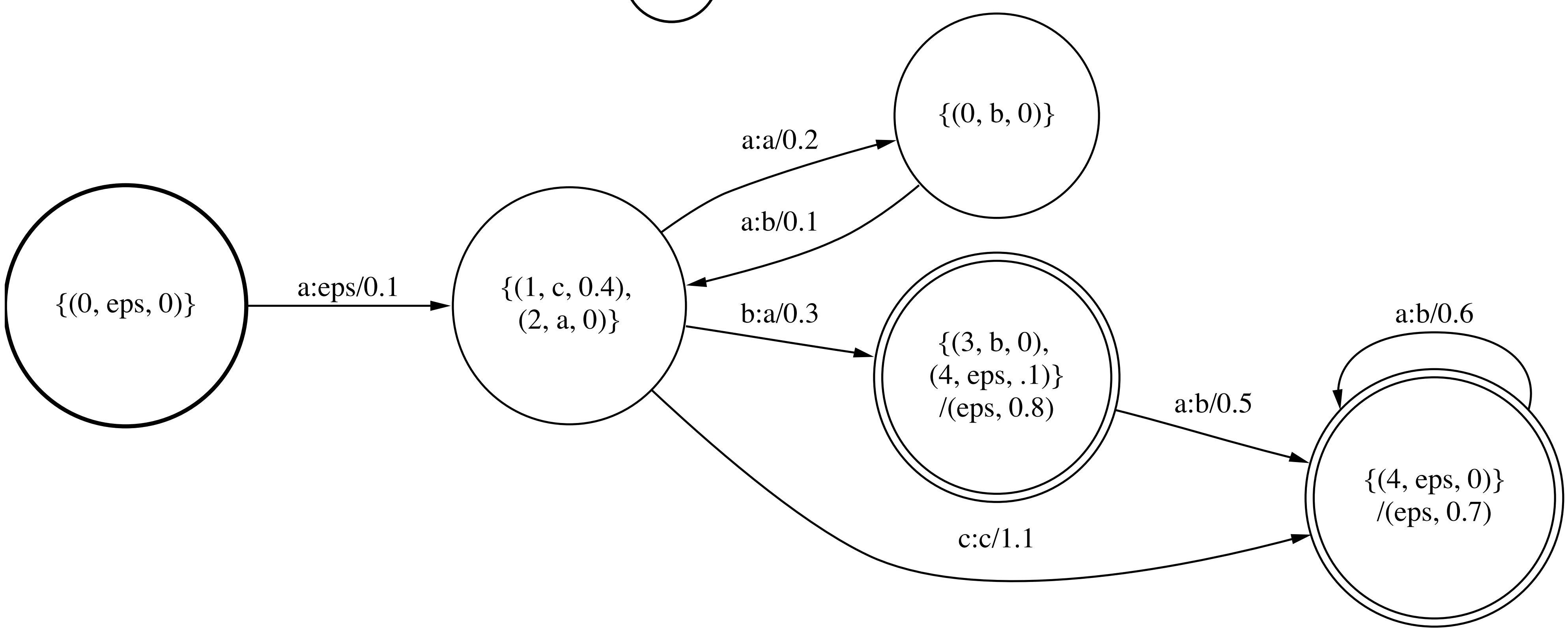




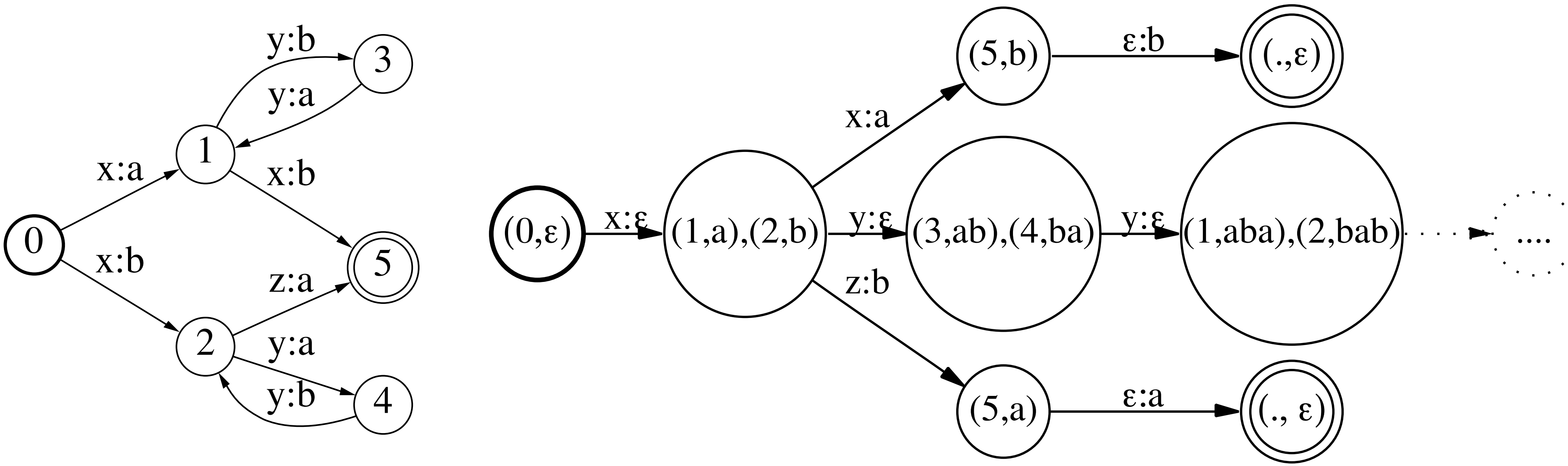


Illustration





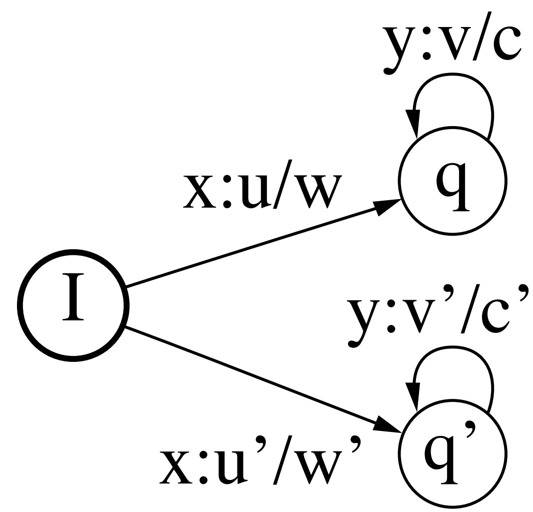
Non-Determinizable Transducer

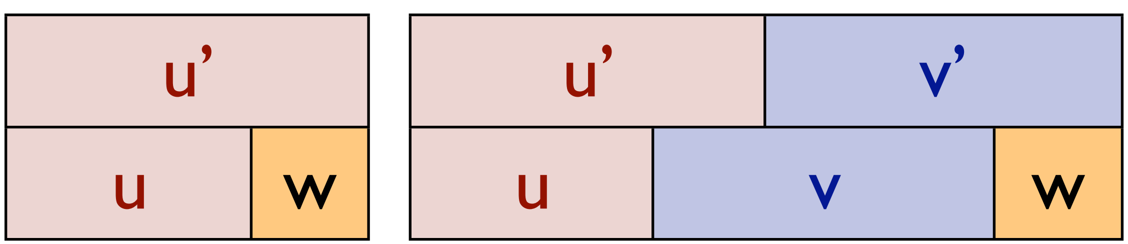


Twins Property

(Choffrut, 1978; MM 1997)

- Definition: a weighted transducer T over the tropical semiring has the twins property if for any two states q and q' as in the figure, the following holds:
  - $c = c'$;
  - $u^{-1}u' = (uv)^{-1}u'v'$.





Determinizability

(Choffrut, 1978; MM 1997; Allauzen and MM, 2002)

- Theorem: a trim unambiguous weighted automaton over the tropical semiring is determinizable iff it has the twins property.
- Theorem: let T be a weighted transducer over the tropical semiring. Then, if T has the twins property, then it is determinizable.
- Algorithm for testing the twins property:
  - unambiguous automata: $O(|Q|^2 + |E|^2)$.
  - unweighted transducers: $O(|Q|^2(|Q|^2 + |E|^2))$.

This Talk

- Definitions
- Composition
- Shortest-distance algorithms
- Epsilon-removal
- Determinization
- Pushing
- Minimization

Pushing

(MM, 1997; 2004)

- **Definition**: given weighted transducer T, create equivalent weighted transducer such the sum (longest common prefix) of the weights (output strings) of all outgoing paths be $\overline{1}$ (ϵ) at all states, modulo initial states.

- **Algorithm components**:
  - Single-source shortest-distance computation

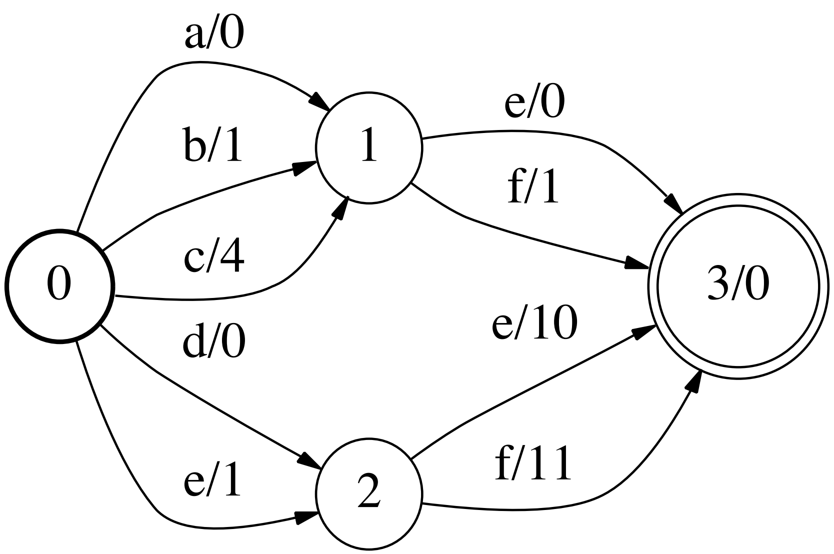
$$d[q] = \bigoplus_{\pi \in P(q,F)} w[\pi].$$

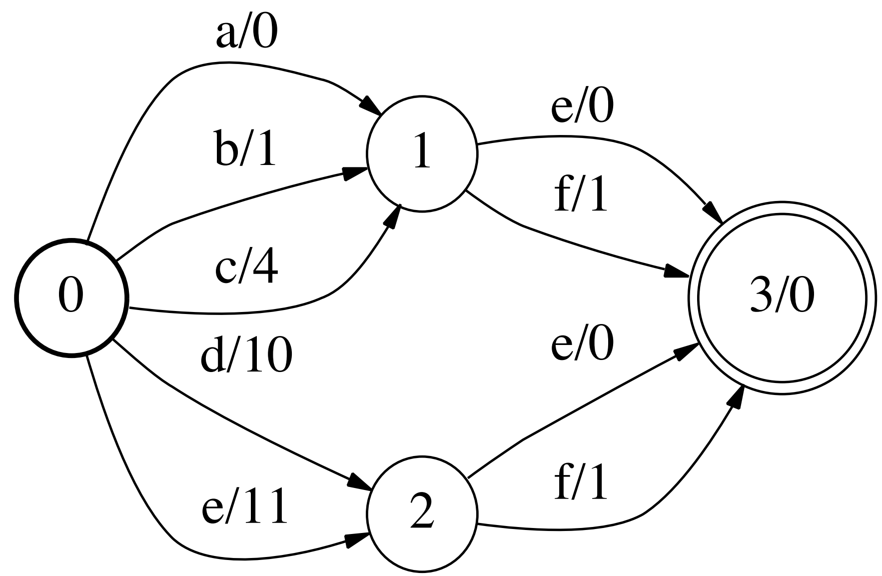
  - Reweighting: $w[e] \leftarrow (d[p[e]])^{-1}(w[e] \otimes d[n[e]])$ for each transition e.

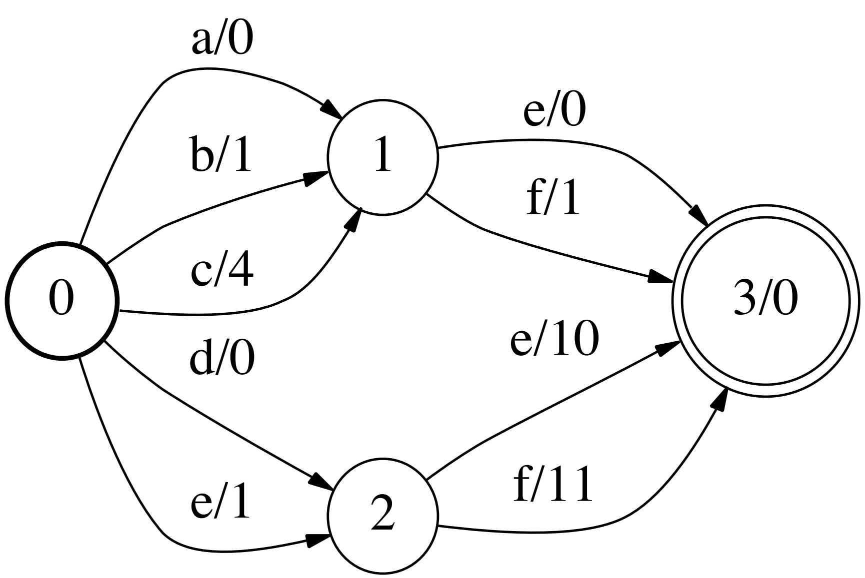
Main Algorithm

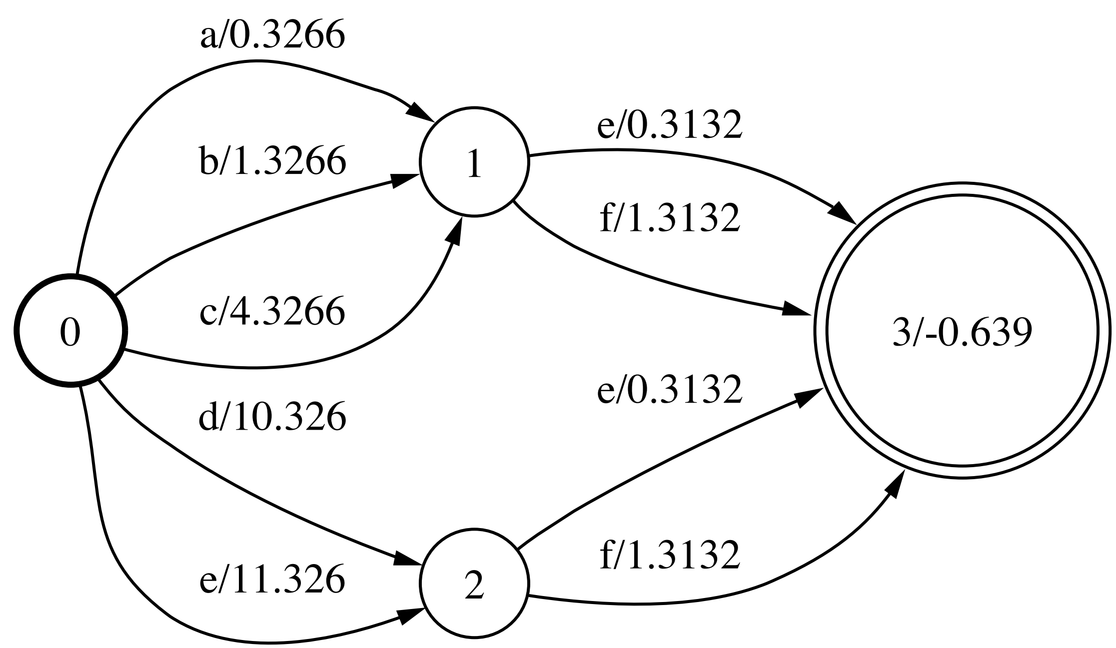
- Automata: single-source shortest-distance.
  - acyclic case: $O(|Q| + |E|(T_\oplus + T_\otimes))$.
  - general case tropical semiring: $O(|Q| \log |Q| + |E|)$.
  - general case *k*-closed semirings
    $$O(|Q| + (T_\oplus + T_\otimes + C(A))|E| \max_{q \in Q} N(q) + (C(I) + C(E)) \sum_{q \in Q} N(q))$$
  - general case closed semirings $\Omega(|Q|^3(T_\oplus + T_\otimes + T_\star))$.
- Transducers: $O((|P_{max}| + 1)\,|E|)$.

Illustration

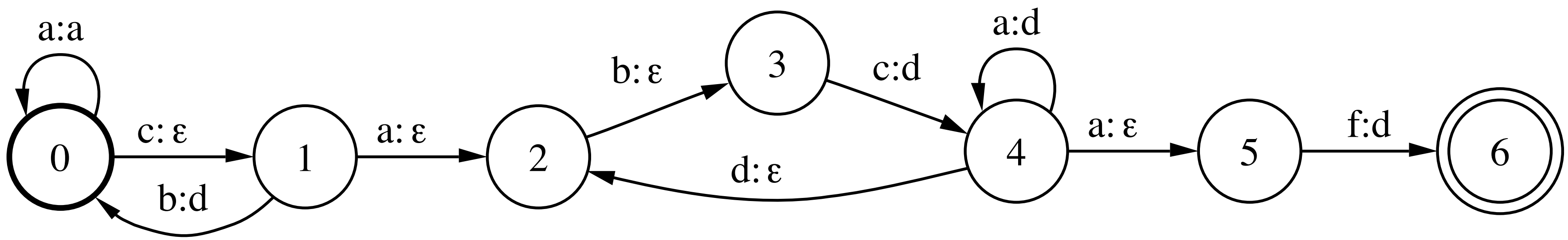


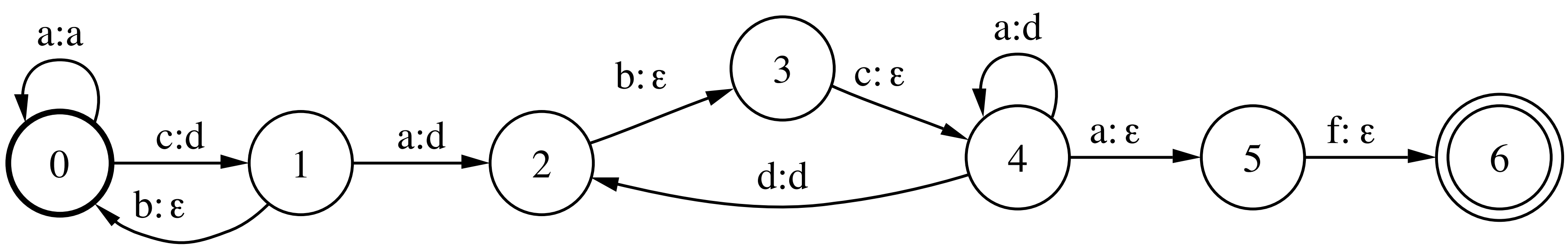






Illustration





This Talk

- Definitions
- Composition
- Shortest-distance algorithms
- Epsilon-removal
- Determinization
- Pushing
- Minimization

Algorithm

(MM, 1997, 2000, 2005)

- Automata: pushing and automata minimization, general (Hopcroft, 1971) and acyclic case (Revuz 1992).
  - acyclic case: $O(|Q| + |E|(T_\oplus + T_\otimes))$.
  - general case tropical semiring: $O(|E| \log |Q|)$.
- Transducers:
  - acyclic case: $O(S + |Q| + |E| \left(|P_{max}| + 1\right))$.
  - general case tropical semiring:

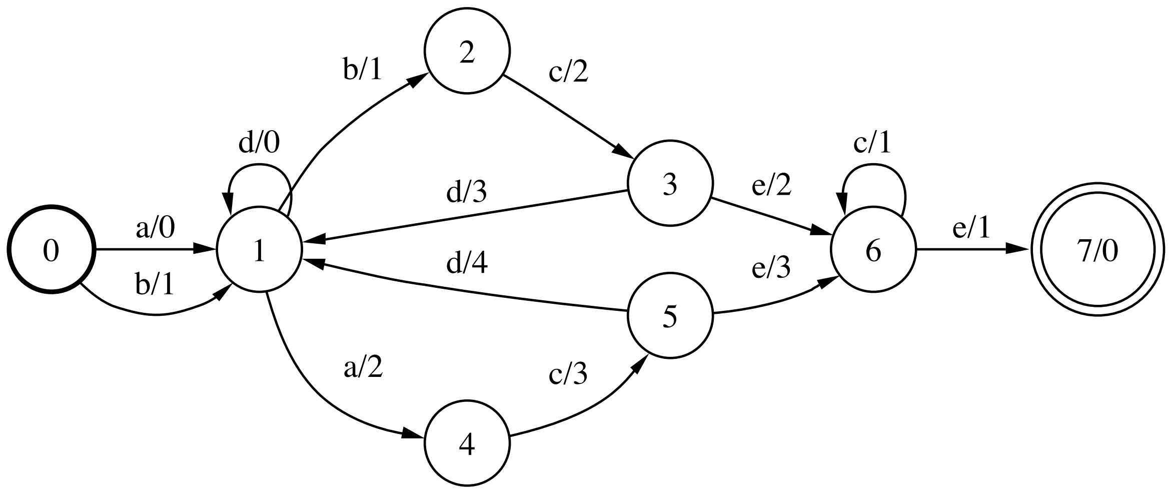
$$O(S + |Q| + |E| \left(\log |Q| + |P_{max}|\right)).$$

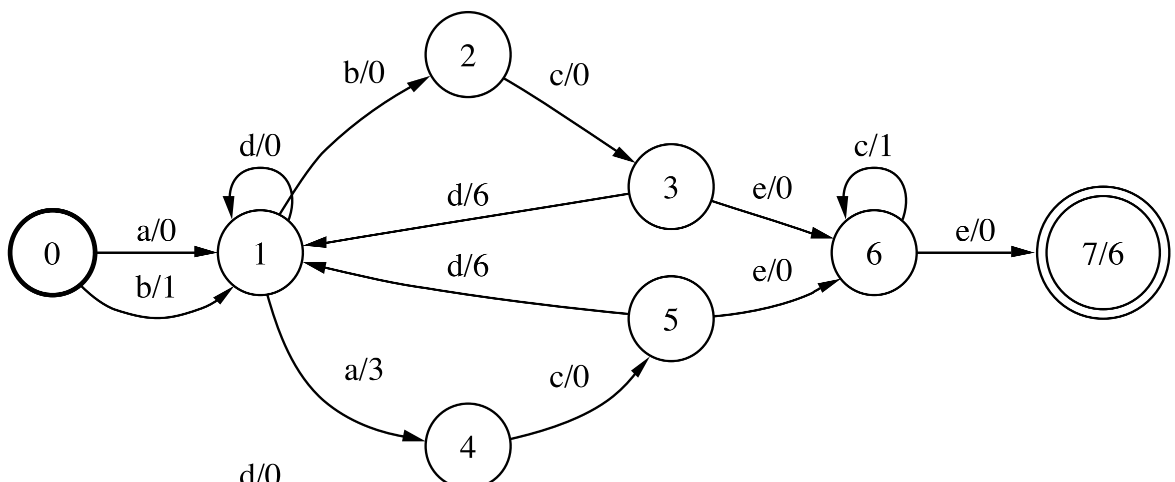
Minimization

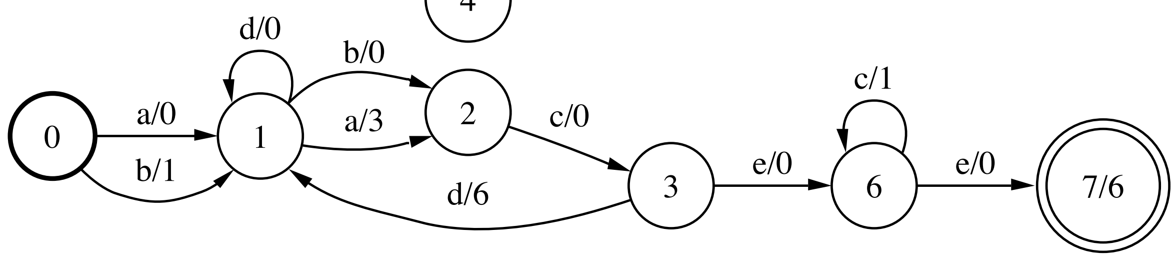
(MM, 1997, 2000, 2005)

- Definition: given deterministic weighted transducer T, create equivalent deterministic weighted transducer with the minimal number of states (and transitions).
- Algorithm components:
  - apply pushing to create canonical representation.
  - apply unweighted automata minimization after encoding (input labels, output label, weight) as a single label.

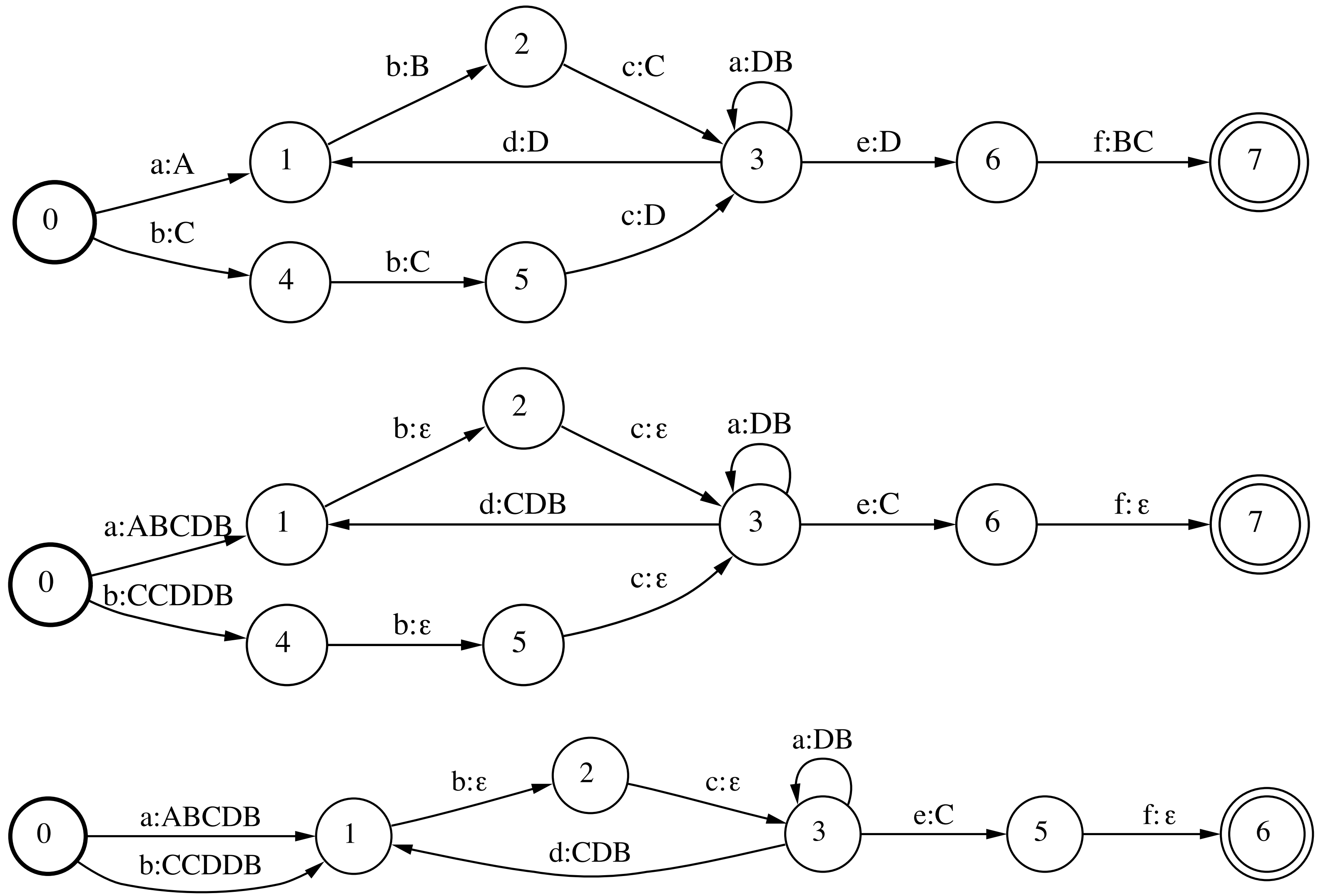
Illustration







Illustration



References

- Cyril Allauzen and Mehryar Mohri. Efficient Algorithms for Testing the Twins Property. *Journal of Automata, Languages and Combinatorics*, 8(2):117-144, 2003.

- John E. Hopcroft. An n log n algorithm for minimizing the states in a finite automaton. In The Theory of Machines and Computations, pages 189-196. Academic Press, 1971.

- Mehryar Mohri. Finite-State Transducers in Language and Speech Processing. *Computational Linguistics*, 23:2, 1997.

- Mehryar Mohri. Minimization Algorithms for Sequential Transducers. Theoretical Computer Science, 234:177-201, March 2000.

- Mehryar Mohri. Semiring Frameworks and Algorithms for Shortest-Distance Problems. Journal of Automata, Languages and Combinatorics, 7(3):321-350, 2002.

- Mehryar Mohri. Generic Epsilon-Removal and Input Epsilon-Normalization Algorithms for Weighted Transducers. *International Journal of Foundations of Computer Science*, 13(1): 129-143, 2002.

- Mehryar Mohri. Statistical Natural Language Processing. In M. Lothaire, editor, *Applied Combinatorics on Words*. Cambridge University Press, 2005.

References

- Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. Weighted Automata in Text and Speech Processing. In *Proceedings of the 12th biennial European Conference on Artificial Intelligence (ECAI-96), Workshop on Extended finite state models of language*. Budapest, Hungary, 1996. John Wiley and Sons, Chichester.

- Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. The Design Principles of a Weighted Finite-State Transducer Library. *Theoretical Computer Science*, 231:17-32, January 2000.

- Mehryar Mohri and Michael Riley. A Weight Pushing Algorithm for Large Vocabulary Speech Recognition. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech'01)*. Aalborg, Denmark, September 2001.

- Fernando Pereira and Michael Riley. *Finite State Language Processing*, chapter Speech Recognition by Composition of Weighted Finite Automata. The MIT Press, 1997.

- Dominique Revuz. Minimisation of Acyclic Deterministic Automata in Linear Time. Theoretical Computer Science 92(1): 181-189, 1992.