On the State Complexity of Complementing Unambiguous Finite Automata

Stefan Kiefer

University of Oxford, UK

IRIF Verification Seminar
23 January 2023

State Complexity

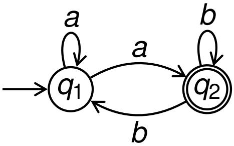
Given two finite automata $\mathcal{A}_1, \mathcal{A}_2$, recognizing L_1, L_2 respectively, how many states are needed (in terms of the number of states in $\mathcal{A}_1, \mathcal{A}_2$) in the worst case for an automaton that recognizes $L_1 \cup L_2$ (or $L_1 \cap L_2$, or $\Sigma^* \setminus L_1$, etc.)?

The state complexity is well understood for many automaton models and many language operations.

For example, complementing an NFA with n states may require 2^n states [Birget'93], even for automata with binary alphabet [Jirásková'05].

Unambiguous Finite Automata

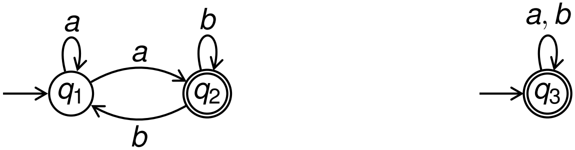
An unambiguous finite automaton (UFA) is an NFA $(Q, \Sigma, \delta, I, F)$ in which every word has at most one accepting run.



For general NFAs, inclusion, equivalence and universality are PSPACE-complete.

For UFAs these operations are in P (even in NC).

Equivalence: via Linear Algebra



$$M(a) = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad M(b) = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Extend M to words: $M(a_1 \cdots a_k) := M(a_1) \cdot \ldots \cdot M(a_k)$.
The two automata are equivalent if and only if

$$\begin{pmatrix} 1 & 0 & -1 \end{pmatrix} \cdot M(w) \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = 0 \qquad \text{for all } w \in \Sigma^* .$$

Compute a basis of the vector space spanned by

$$\left\{ \begin{pmatrix} 1 & 0 & -1 \end{pmatrix} \cdot M(w) \;\middle|\; w \in \Sigma^* \right\} .$$

Can be done in time $O(|\Sigma| n^3)$, inductively over $|w|$.

Mathematical Appeal of UFA: Combinatorics

If a UFA is diamond-free (which one can assume), then $M(w)$ is a 0-1 matrix for all $w \in \Sigma^*$.

So $M(a), M(b)$ generate a finite monoid of matrices (over nonnegative integers). Leads to combinatorics / theory of codes.

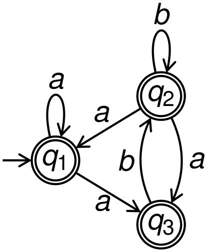
Theorem (K., Mascle, SIAM J. Discret. Math. 2021)

Let $\mathcal{M}$ be a set of $n \times n$-matrices over the nonnegative integers such that the joint spectral radius of $\mathcal{M}$ is at most one. If the zero matrix 0 is a product of matrices in $\mathcal{M}$, then there are $M_1, \ldots, M_{n^5} \in \mathcal{M}$ with $M_1 \cdots M_{n^5} = 0$.

Unambiguousness
$\Longrightarrow$ finiteness of the monoid
$\Longrightarrow$ joint spectral radius at most one

Efficient Probabilistic Model Checking via Unambiguous Büchi Automata

Given an unambiguous Büchi automaton, what is the probability that a random infinite word over $\{a, b\}$ is accepted?



$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} := \begin{pmatrix} \Pr(q_1 \text{ accepts a random word}) \\ \Pr(q_2 \text{ accepts a random word}) \\ \Pr(q_3 \text{ accepts a random word}) \end{pmatrix}$$

$x_1 = \frac{1}{2} \cdot (x_1 + x_3) + \frac{1}{2} \cdot 0$ etc. Overall, $\vec{x} = \left(\frac{1}{2} M(a) + \frac{1}{2} M(b)\right) \vec{x}$.
To determine $\vec{x}$ uniquely, another equation is needed.

Leads to efficient probabilistic model checking with PRISM [Baier, K., Klein, Klüppelholz, Müller, Worrell, CAV 2016].

UFAs have appealing connections.

- linear algebra
- combinatorics
- probabilistic model checking
- weighted automata
- communication complexity

Back to state complexity

Thomas Colcombet in a DCFS'15 invited paper:

Theorem 3. **([40,41]).** *The problems of universality and equivalence of unambiguous automata as well as containment of a non-deterministic automaton in an unambiguous automaton are solvable in polynomial time.*

We shall see in this section a complete proof of this result, which is a good excuse for introducing several important techniques.

Of course, knowing this complexity result, and since universality amounts to checking the emptiness of the complement, one might think that another proof of this result could be as follows: complement the unambiguous automaton with a polynomial blowup of states, and then test for emptiness in polynomial time. However, the question of whether unambiguous automata can be complemented with a polynomial blowup in the number of states is an open problem.

Conjecture 1. It is possible to complement unambiguous automata of size n into unambiguous automata of size polynomial in n.

In fact, even whether we can complement an unambiguous automaton into a non-deterministic automaton of polynomial size is open. We lack techniques for addressing this question. In particular, how can we prove a lower bound on the size of an unambiguous automaton for a given language?

Back to state complexity

Mikhail Raskin refuted Colcombet's conjecture in 2018:

Theorem (Mikhail Raskin, ICALP 2018)

For any $n \in \mathbb{N}$ there exists a unary (i.e., $|\Sigma| = 1$) UFA $\mathcal{A}_n$ with n states such that any NFA that recognizes $\Sigma^* \setminus L(\mathcal{A}_n)$ has at least $n^{(\log\log\log n)^{\Theta(1)}}$ states.

Upper Bounds

[Jirásek, Jirásková, Šebej, 2018] proposed to take the smaller of two UFAs for the complement:

1. standard subset construction for determinization; then swap accepting and non-accepting states
2. subset construction backwards, starting from accepting states; then swap initial and non-initial states

Both are UFAs for the complement. This leads to:

Theorem (Jirásek, Jirásková, Šebej, International Journal of Foundations of Computer Science 2018)

Let $\mathcal{A}$ be a UFA with $n \geq 7$ states that recognizes a language $L \subseteq \Sigma^*$. Then there exists a UFA with at most $n \cdot 2^{0.786n}$ states that recognizes the language $\Sigma^* \setminus L$.

Upper Bound

Emil Indzhev (former Oxford undergrad) and I looked again at Jirásek et al.'s construction.

Lemma (Indzhev, K., IPL 2022)

Let $\mathcal{A} = (Q, \Sigma, \delta, I, F)$ be a UFA. Suppose that its forward determinization has k states, and its backward determinization has ℓ states. Then there exists an undirected graph with $|Q|$ vertices that has at least k cliques and at least ℓ independent sets.

Proof sketch.

Construct the graph with Q as vertex set, and an edge between q and q' if they are reachable in $\mathcal{A}$ from initial states via the same word. Then every state in the forward determinization is a clique. By unambiguousness, every state in the backward determinization is an independent set. □

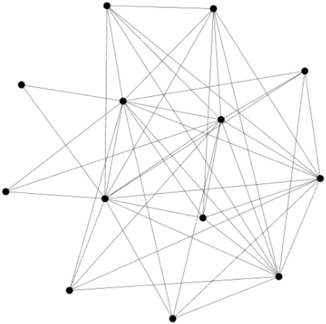
Upper Bound

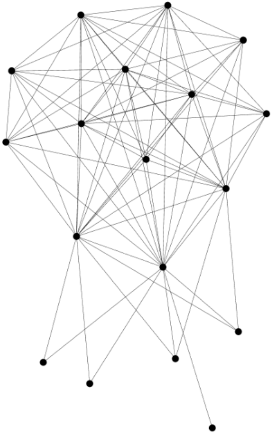
Perhaps there is no graph that has “many” cliques and independent sets at the same time? Another undergrad ran experiments to heuristically search (via simulated annealing) for graphs with both many cliques and many isets.

n	value = max {min (# cliques, # isets)}	log2(value) / n
5	11	0.691886
6	17	0.681244
7	25	0.663408
8	37	0.651182
9	55	0.642373
10	79	0.630378
11	127	0.635335
12	164	0.613129
13	262	0.617956
14	331	0.597906
15	523	0.602044
16	667	0.586346

Results of simulation

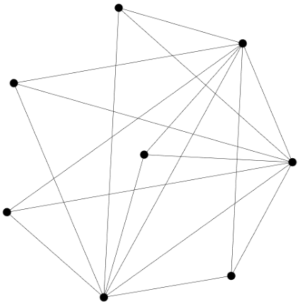
graphs from his report:

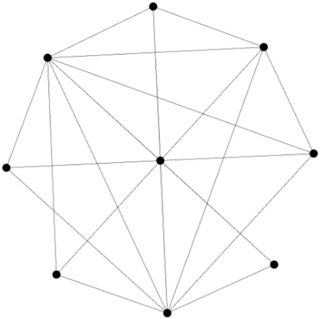




Results of simulation

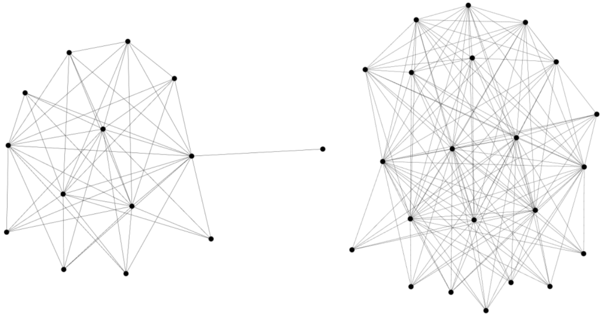
n = 8, n = 9:





Results of simulation

n = 16, n = 26:



They are currently not very helpful... **I think I need some help on how to detect patterns from these graphical results.**

Upper Bound: Extremal Graph Theory

This leads to extremal graph theory, a branch of combinatorics.

Lemma (Indzhev, K., IPL 2022)

Let (V, E) be a graph with $|V| = n$. Then

$$|\{X \subseteq V \mid X \text{ is a clique}\}| \cdot |\{Y \subseteq V \mid Y \text{ is an iset}\}| \;\leq\; (n+1)2^n\,.$$

This implies the following upper bound:

Theorem (Indzhev, K., IPL 2022)

Any graph with n vertices has at most $\sqrt{n+1} \cdot 2^{n/2}$ cliques or at most $\sqrt{n+1} \cdot 2^{n/2}$ isets. Moreover, for any $n \geq 0$ there is a graph with n vertices that has at least $\frac{1}{2}\sqrt{n+1} \cdot 2^{n/2}$ cliques and at least $\frac{1}{2}\sqrt{n+1} \cdot 2^{n/2}$ isets.

Upper Bound

Together with the previous lemma, we get:

Theorem (Indzhev, K., IPL 2022)

Let $\mathcal{A}$ be a UFA with $n \geq 0$ states that recognizes a language $L \subseteq \Sigma^*$. Then there exists a UFA with at most $\sqrt{n+1} \cdot 2^{n/2}$ states that recognizes the language $\Sigma^* \setminus L$.

This analysis of this particular complementation procedure is tight up to a factor 2:

Proposition (Indzhev, K., IPL 2022)

For every $n \geq 0$ there is a UFA with n states such that both its forward and its backward determinization have at least $\frac{1}{2}\sqrt{n+1} \cdot 2^{n/2}$ states.

Other complementation procedures might be better.

Improve the Lower Bound via Communication Complexity

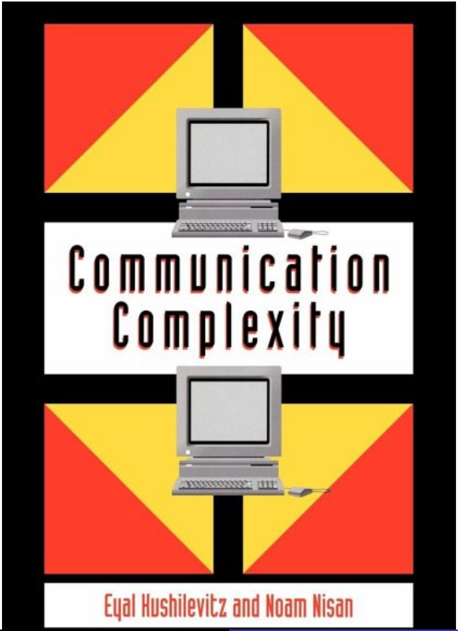
Plan for the rest of the talk: improve Raskin's lower bound.

Theorem (Mikhail Raskin, ICALP 2018)

For any $n \in \mathbb{N}$ there exists a unary (i.e., $|\Sigma| = 1$) UFA $\mathcal{A}_n$ with n states such that any NFA that recognizes $\Sigma^* \setminus L(\mathcal{A}_n)$ has at least $n^{(\log \log \log n)^{\Theta(1)}}$ states.

Communication complexity is the key.

Standard Text on Communication Complexity



Clique vs Independent Set Problem

On page 6 of this book, and in [Yannakakis, STOC'88] the following problem appears:

Clique vs Independent Set (CIS) Problem

Alice and Bob both know a fixed undirected graph (V, E) with $|V| = n$.
Alice holds a clique $x \subseteq V$. Bob holds an iset $y \subseteq V$.

They want to communicate (but as little as possible) to find out whether $x \cap y \neq \emptyset$.

nondeterministic communication complexity:

$$\mathrm{NP}^{\mathrm{cc}}(\mathrm{CIS}) = \log n \quad (\text{guess } x \cap y)$$

unambiguous communication complexity:

$$\mathrm{UP}^{\mathrm{cc}}(\mathrm{CIS}) = \log n \quad (\text{guess } x \cap y)$$

Matrix of the Problem

0	1	1	1	0	1	1	1
0	1	1	1	0	1	1	1
0	1	1	1	0	0	0	0
0	1	1	1	1	1	0	1
0	0	0	1	1	1	0	1
1	1	1	1	1	1	1	1
1	1	1	1	0	1	1	1
0	0	0	0	0	1	1	1

$$F : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$$

Alice holds a row, an element of $\mathcal{X}$.
Bob holds a column, an element of $\mathcal{Y}$.

0	1	1	1	0	1	1	1
0	1	1	1	0	1	1	1
0	1	1	1	0	0	0	0
0	1	1	1	1	1	0	1
0	0	0	1	1	1	0	1
1	1	1	1	1	1	1	1
1	1	1	1	0	1	1	1
0	0	0	0	0	1	1	1

NP^cc

Alice holds a row, an element of $\mathcal{X}$.
Bob holds a column, an element of $\mathcal{Y}$.

$\mathrm{NP}^{\mathrm{cc}}(F)$ is defined as the $\log$ of the least number of rectangles that cover the 1s.

Unambiguous Communication Complexity

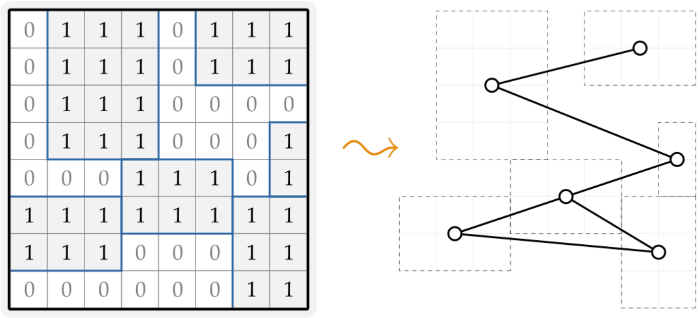
0	1	1	1	0	1	1	1
0	1	1	1	0	1	1	1
0	1	1	1	0	0	0	0
0	1	1	1	0	0	0	1
0	0	0	1	1	1	0	1
1	1	1	1	1	1	1	1
1	1	1	0	0	0	1	1
0	0	0	0	0	0	1	1

UP$^{\text{cc}}$

Alice holds a row, an element of $\mathcal{X}$.
Bob holds a column, an element of $\mathcal{Y}$.

UP$^{\text{cc}}(F)$ is defined as the $\log$ of the least number of rectangles that partition the 1s.

The CIS Problem is Complete for UP^{cc}



There is an edge between any two rectangles that share a row.
Alice maps her input row to the rectangles on that row, a clique.
Bob maps his input col to the rectangles on that col, an iset.
F maps their inputs to 1 $\iff$ they share a rectangle.

Yannakakis's Upper Bound on P^{cc}

0	1	1	1	0	1	1	1
0	1	1	1	0	1	1	1
0	1	1	1	0	0	0	0
0	1	1	1	0	0	0	1
0	0	0	1	1	1	0	1
1	1	1	1	1	1	1	1
1	1	1	0	0	0	1	1
0	0	0	0	0	0	1	1

Suppose $UP^{cc}(F) = \log n$, so there are n rectangles in partition.
For any 2 rectangles, Alice or Bob witnesses their disjointness.

$\Longrightarrow$ For every rectangle, Alice or Bob witnesses its disjointness with at least $\frac{1}{2}$ of the rectangles.

$\Longrightarrow$ deterministic protocol with $\log n$ rounds, with $\log n$ bits of communication per round

$\Longrightarrow$ $P^{cc}(F) \in O(\log^2 n)$

Co-Nondeterministic Communication Complexity

It follows that $\mathrm{P}^{cc}(F) \in O(\mathrm{UP}^{cc}(F)^2)$.
Equivalently, $\mathrm{P}^{cc}(\mathrm{CIS}) \in O(\log^2 n)$.

Since $\mathrm{coNP}^{cc}(F) \leq \mathrm{P}^{cc}(F)$, also $\mathrm{coNP}^{cc}(F) \in O(\mathrm{UP}^{cc}(F)^2)$.
Equivalently, $\mathrm{coNP}^{cc}(\mathrm{CIS}) \in O(\log^2 n)$.

Yannakakis's Question

Is $\mathrm{coNP}^{cc}(\mathrm{CIS}) \in O(\log n)$?

Yannakakis proved a connection to whether TSPs can be expressed with small LPs.

The question was later shown to be equivalent to the polynomial version of the Alon-Saks-Seymour conjecture ($\chi(G) \leq bp(G) + 1$) from graph theory.

Co-Nondeterministic Communication Complexity

Yannakakis's Question

Is $\mathrm{coNP}^{\mathrm{cc}}(\mathrm{CIS}) \in O(\log n)$?

$O(\log^2 n)$	[Yannakakis, 1991]
$\geq 6/5 \cdot \log n$	[Huang and Sudakov, 2010]
$\geq 3/2 \cdot \log n$	[Amano, 2014]
$\geq 2 \cdot \log n$	[Shigeta and Amano, 2014]
$\Omega(\log^{1.12} n)$	[Göös, 2015], so the answer is no
$\Omega(\log^{1.22} n)$	[Ben-David, Hatami, Tal, 2015]
$\tilde{\Omega}(\log^2 n)$	[Balodis, Ben-David, Göös, Jain, Kothari, 2021]

So there is F such that $\mathrm{coNP}^{\mathrm{cc}}(F) \in \tilde{\Theta}(\mathrm{UP}^{\mathrm{cc}}(F)^2)$,
i.e., Yannakakis's upper bound is tight up to logarithmic factors.

Construction of the UFA

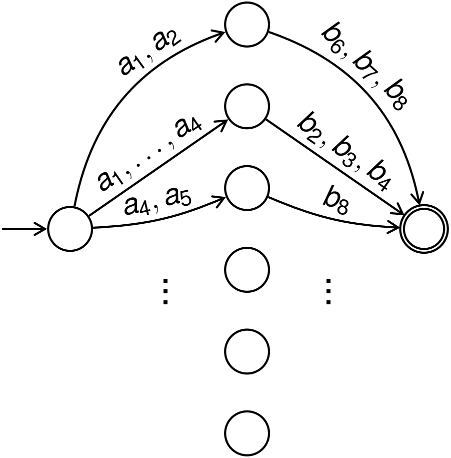
$\mathrm{coNP}^{cc}(F) \in \tilde{\Omega}(\mathrm{UP}^{cc}(F)^2)$ means “many” rectangles are needed to cover the 0s.

0	1	1	1	0	1	1	1
0	1	1	1	0	1	1	1
0	1	1	1	0	0	0	0
0	1	1	1	0	0	0	1
0	0	0	1	1	1	0	1
1	1	1	1	1	1	1	1
1	1	1	0	0	0	1	1
0	0	0	0	0	0	1	1

Construction of the UFA

$\mathrm{coNP}^{\mathrm{cc}}(F) \in \tilde{\Omega}(\mathrm{UP}^{\mathrm{cc}}(F)^2)$ means "many" rectangles are needed to cover the 0s.

	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8
a_1	0	1	1	1	0	1	1	1
a_2	0	1	1	1	0	1	1	1
a_3	0	1	1	1	0	0	0	0
a_4	0	1	1	1	0	0	0	1
a_5	0	0	0	1	1	1	0	1
a_6	1	1	1	1	1	1	1	1
a_7	1	1	1	0	0	0	1	1
a_8	0	0	0	0	0	0	1	1



Any NFA for the complement language has $n^{\tilde{\Omega}(\log n)}$ states.

Lower Bounds on Complement, Union, Separation

One can make the alphabet have size 2.

Theorem (Göös, K., Yuan, ICALP'22)

For every $n \in \mathbb{N}$ there is a language $L \subseteq \{0,1\}^*$ recognized by an n-state UFA such that any NFA that recognizes $\Sigma^* \setminus L$ has $n^{\tilde{\Omega}(\log n)}$ states.

Theorem (Göös, K., Yuan, ICALP'22)

For every $n \in \mathbb{N}$ there are languages $L_1, L_2 \subseteq \{0,1\}^*$ recognized by n-state UFAs such that any UFA that recognizes $L_1 \cup L_2$ has $n^{\tilde{\Omega}(\log n)}$ states.

Theorem (Göös, K., Yuan, ICALP'22)

For every $n \in \mathbb{N}$ there is a language $L \subseteq \{0,1\}^*$ such that both L and $\Sigma^* \setminus L$ are recognized by n-state NFAs but any UFA that recognizes L has $n^{\Omega(\log n)}$ states.

How did Balodis et al. find F with $\mathrm{coNP}^{\mathrm{cc}}(F) \in \tilde{\Omega}(\mathrm{UP}^{\mathrm{cc}}(F)^2)$?

Query Complexity: for $f : \{0,1\}^n \to \{0,1\}$ define

- $\mathrm{C}_1(f)$ as the least k such that f can be written as DNF of width k;
- $\mathrm{C}_0(f)$ as the least k such that f can be written as CNF of width k;
- $\mathrm{UC}_1(f)$ as the least k such that f can be written as unambiguous DNF of width k.

1. Find f with $\mathrm{C}_0 \in \tilde{\Omega}(\mathrm{UC}_1(f)^2)$;
2. Use a lifting gadget to transfer high query complexity of f to high communication complexity of $F : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$.