

Generation of DFA Minimization Problems

Gregor Hans Christian Sönnichsen

February 16, 2020

Abstract

The theory of deterministic finite automata (DFAs) is a classical topic of computer science-related courses. A typical task for students is to minimize a DFA by deleting irrelevant elements. However generation of those DFAs that shall be minimized is often done manually by the exercise instructor. This work presents approaches to automatize the generation of DFA minimization tasks.

Generierung von DFA Minimierungsproblemen

-

Zusammenfassung

Zusammenfassung...

Table of Contents

1	Introduction	2
1.1	Preliminaries	2
1.1.1	Deterministic Finite Automatons	2
1.1.2	Minimal DFAs	3
1.1.3	Practical Isomorphy of DFAs	4
1.1.4	Equivalent and distinguishable state pairs	4
1.1.5	The minimization algorithm	5
1.1.6	m -COMEQUIVPAIRS.	6
1.1.7	Another equivalence automaton	6
1.2	Requirements analysis	7
1.2.1	Example of a DFA minimization task for students	7
1.2.2	Difficulty adjustment possibilities	7
1.2.3	Summary of found criteria	9
1.3	Approach and general algorithm	10
2	Generating minimal DFAs	11
2.1	Using trial and error	11
2.1.1	Ensuring A_{test} is minimal and $D(A_{test})$ is correct	12
2.1.2	Ensuring A_{test} is planar	12
2.1.3	Ensuring A_{test} is new	13
2.1.4	Option 1: Generating A_{test} via Randomness	13
2.1.5	Option 2: Generating A_{test} via Enumeration	14
2.2	Alternative approach: Building $m(i)$ bottom up	16
2.3	Related research on DFA generation	17
2.4	Empirical and combinatorial results	18
3	Extending minimal DFAs	20
3.1	Creating equivalent state pairs	20
3.1.1	Adding outgoing transitions	21
3.1.2	Adding ingoing transitions	21
3.1.3	The algorithm	21
3.1.4	Creating equivalent state pairs does not change D	22
3.2	Adding unreachable states	25
4	Conclusion	26
A	Bibliography	27
B	Erklärung	29

Chapter 1

Introduction

Automata theory is recommended as part of a standard computer science curriculum [12, pp. 5-6]. As other such theories it provides the chance to gain a precise cognitive model yielding new perspectives on problems and givens. This may thus lead to increased problem solving skills and more accurate thinking.

A typical task in automata theory is the minimization of a given deterministic finite automaton (DFA). The classic textbook “Introduction to automata theory, languages, and computation” by Hopcroft et. al. [15] presents a practicable minimization algorithm. In this work we will confine ourselves to look at DFA minimizations using that algorithm.

In an introduction course to theoretical computer science minimization tasks are thus likely to occur in supplementary exercises or exams. As of the creation of such tasks, one may assume, that it is done mostly manually. Automation would yield here the following advantages:

- freeing time for other things, e.g. research, helping students face-to-face, designing the whole exercise sheet
- generation of tasks which lie in a well-defined range
- increased predictability and consistency of the generated task properties, which can be adjusted accurately through various parameters
- saves human operators from the generating task which involves monotonous work

Gregor: [Delete or find extern from wikipedia](#) Engagement on this topic promises moreover increased clarification which kind of minimization tasks can be generated, and where difficulties of those tasks lie.

This work aims to provide theoretical foundations for a DFA minimization task generator. What requirements a user has towards such a program will be discussed in a short requirements analysis. Based on this work a DFA minimization generator has been implemented. It can be found at <https://github.com/bt701607/Generation-of-DFA-Minimization-Problems>.

1.1 Preliminaries

We start with defining preliminary theoretical foundations.

1.1.1 Deterministic Finite Automatons

A 5-tuple $A = (Q, \Sigma, \delta, s, F)$ with Q being a finite set of *states*, Σ a finite set of *alphabet symbols*, $\delta: Q \times \Sigma \rightarrow Q$ a *transition function*, $s \in Q$ a *start state* and $F \subseteq Q$ *final states*

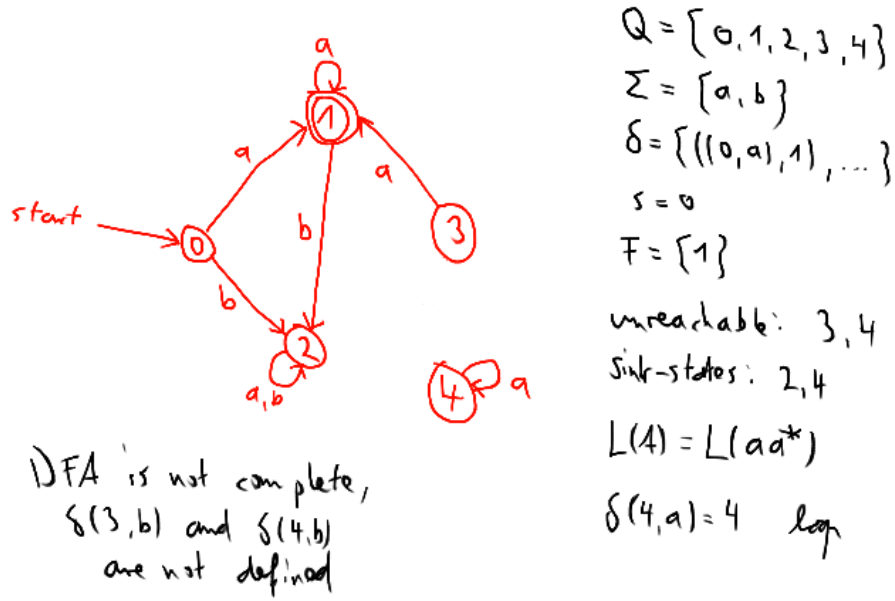


Figure 1.1: An example DFA and its properties.

is called *deterministic finite automaton* (DFA) [15, p. 46]. From now on \mathfrak{A} shall denote the set of all DFAs.

We say $\delta(q, \sigma) = p$ is a transition from q to p using symbol σ . We define the *extended transition function* $\delta^* : Q \times \Sigma^* \rightarrow Q$ of a DFA $A = (Q, \Sigma, \delta, s, F)$ as:

- $\delta^*(q, \varepsilon) = q$
- $\delta^*(q, w\sigma) = \delta(\delta^*(q, w), \sigma)$ for all $q \in Q, w \in \Sigma^*, \sigma \in \Sigma$

Then, the *language* of that DFA is defined as $L(A) = \{ w \mid \delta^*(s, w) \in F \}$ [15, pp. 49-50. 52].

Given a state $q \in Q$. With $d^-(q)$ we denote the set of all *ingoing* transitions $\delta(q', \sigma) = q$ of q . With $d^+(q)$ we denote the set of all *outgoing* transitions $\delta(q, \sigma) = q'$ of q [9, pp. 2-3]. If a transition is of the form $\delta(q, \sigma) = q$, then we say that q has a *loop*.

Definition 1. We say a state q is (*un*-)reachable in a DFA A , iff there is (no) a word $w \in \Sigma^*$ such that $\delta^*(s, w) = q$.

If all states of a DFA A are reachable, then we say A is *accessible* [9, p. 2].

A DFA is called *complete* iff for all states, every symbol of the alphabet is used on an outgoing transition: $\forall q \in Q: \forall \sigma \in \Sigma: \exists p \in Q: \delta(q, \sigma) = p$. Note, that every incomplete DFA can be converted to a complete one by adding a so called *dead state* [15, p. 67]. The resulting automaton has the same language.

1.1.2 Minimal DFAs

This section closely follows [24, pp. 42-45]. We call a DFA A *minimal*, if there exists no other automaton with the same language using less states. With \mathfrak{A}_{min} we shall denote the set of all minimal DFAs.

The *Nerode-relation* $\equiv_L \subseteq \Sigma^* \times \Sigma^*$ of a language L with alphabet Σ is defined as follows:

$$x \equiv_L y \Leftrightarrow_{def} \forall z \in \Sigma^*: (xz \in L \Leftrightarrow yz \in L)$$

The Nerode-relation of a DFA A is the the Nerode-relation of its language: $\equiv_{L(A)}$. If the context makes it clear, than we will shorten the notation of a equivalence class $[x]_{\equiv_L}$ with $[x]$.

The *equivalence class automaton* $A_L = (Q_L, \Sigma_L, \delta_L, s_L, F_L)$ to a regular language L with alphabet Σ is defined as follows:

- $Q_L = \{ [x] \mid x \in \Sigma^* \}$
- $\Sigma_L = \Sigma$
- $\delta_L([x], \sigma) = [x\sigma], \forall x \in \Sigma^*, \forall \sigma \in \Sigma$
- $s = [\varepsilon]$
- $F = \{ [x] \mid x \in L \}$

Theorem 1. *Given a language L , then the equivalence class automaton A_L is minimal.*

1.1.3 Practical Isomorphism of DFAs

Given two DFAs $A_1 = (Q_1, \Sigma_1, \delta_1, s_1, F_1)$ and $A_2 = (Q_2, \Sigma_2, \delta_2, s_2, F_2)$. We say A_1 and A_2 are *practical isomorph*, iff:

- $|Q_1| = |Q_2|, |\Sigma_1| = |\Sigma_2|$ and
- there exists a bijection $\phi: \Sigma_1 \rightarrow \Sigma_2$ such that:
- there exists a bijection $\pi: Q_1 \rightarrow Q_2$ such that:
 - $\pi(s_1) = s_2$
 - $\forall q \in Q_1: (q \in F_1 \iff \pi(q) \in F_2)$
 - $\forall q \in Q_1: \forall \sigma \in \Sigma_1: \pi(\delta_1(q, \sigma)) = \delta_2(\pi(q), \phi(\sigma))$

Note that practical isomorphism between two DFAs A_1, A_2 does not imply $L(A_1) = L(A_2)$. This would be given, if $\Sigma_1 = \Sigma_2$ were the case (see [24, p. 45]). However the language of such DFAs is equivalent except for an exchange of alphabet symbols:

$$\{ \phi(\sigma_0) \dots \phi(\sigma_n) \mid \sigma_0 \dots \sigma_n \in L(A_1) \} = L(A_2)$$

Practical isomorphism will be sufficient for our case. **Gregor:** why? **Gregor:** Write down isomorphism test. Maybe discuss faster methods here? Look for faster methods in general?

1.1.4 Equivalent and distinguishable state pairs

Definition 2 (Equivalent and Distinguishable State Pairs). [15, p. 154] A state pair $q_1, q_2 \in Q$ of a DFA $A = (Q, \Sigma, \delta, s, F)$ is called *equivalent*, iff $\sim_A(q_1, q_2)$ is true, whereas

$$q_1 \sim_A q_2 \Leftrightarrow_{def} \forall z \in \Sigma^*: (\delta^*(q_1, z) \in F \Leftrightarrow \delta^*(q_2, z) \in F)$$

If $q_0 \not\sim_A q_1$, then q_0 and q_1 are called a *distinguishable* state pair. The relation \sim_A is an equivalence relation

1.1.5 The minimization algorithm

This minimization algorithm MINIMIZEDFA works in four major steps, removing essentially states in such a way, that no unreachable states and no equivalent state pairs are left.

1. Compute all unreachable states via breadth-first search.

```

1: function COMUNREACHABLES( $A$ )
2:    $U \leftarrow Q \setminus \{s\}$  ▷ undiscovered states
3:    $O \leftarrow \{s\}$  ▷ observed states
4:    $D \leftarrow \{\}$  ▷ discovered states
5:   while  $|O| > 0$  do
6:      $N \leftarrow \{ p \mid \exists q \in O \ \sigma \in \Sigma: \delta(q, \sigma) = p \ \wedge \ p \notin O \cup D \}$ 
7:      $U \leftarrow U \cup N$ 
8:      $D \leftarrow D \cup O$ 
9:      $O \leftarrow N$ 
10:  return  $U$ 

```

2. Remove all unreachable states and their transitions.

```

1: function REMUNREACHABLES( $A, U$ )
2:    $\delta' \leftarrow \delta \setminus \{ ((q, \sigma), p) \mid q \in U \vee p \in U \}$ 
3:   return  $(Q \setminus U, \Sigma, \delta', s, F \setminus U)$ 

```

3. Compute all equivalent state pairs (\sim_A). Inspired by Schönig [24, p. 46] **Gregor:** And TI-VL.

```

1: function COMEQUIVPAIRS( $A$ )
2:    $M \leftarrow \{(p, q), (q, p) \mid p \in F, q \notin F\}$ 
3:   do
4:      $M' \leftarrow \{(p, q) \mid (p, q) \notin M \wedge \exists \sigma \in \Sigma: (\delta(p, \sigma), \delta(q, \sigma)) \in M\}$ 
5:      $M \leftarrow M \cup M'$ 
6:   while  $M' \neq \emptyset$ 
7:   return  $Q^2 \setminus M$ 

```

Note that COMEQUIVPAIRS requires its input automaton to be complete. **Gregor:** Why?

4. Merge all equivalent state pairs, which are exactly those in \sim_A . Inspired by Högborg [17, p. 10].

```

1: function REMEQUIVPAIRS( $A, \sim_A$ ) ▷  $[\cdot]_{\sim_A}$  shall be abbreviated  $[\cdot]$ 
2:    $Q_E \leftarrow \emptyset$ 
3:    $\delta_E \leftarrow \emptyset$ 
4:    $F_E \leftarrow \emptyset$ 
5:   for  $q$  in  $Q$  do
6:     Add  $[q]$  to  $Q_E$ 
7:     for  $\sigma$  in  $\Sigma$  do
8:        $\delta_E([q], \sigma) = [\delta(q, \sigma)]$ 
9:     if  $q \in F$  then
10:      Add  $[q]$  to  $F_E$ 
11:  return  $(Q_E, \Sigma, \delta_E, [s], F_E)$ 

```

Note that REMEQUIVPAIRS creates complete automata.


```

1: function MINIMIZEDFA( $A_{task}$ )
2:    $A_{re} \leftarrow \text{REUNREACHABLES}(A_{task}, \text{COMUNREACHABLES}(A_{task}))$ 
3:    $A_{sol} \leftarrow \text{REMEQUIVPAIRS}(A_{re}, \text{COMEQUIVPAIRS}(A_{re}))$ 
4:   return  $A_{sol}$ 

```

This DFA minimization algorithm has been found by Hopcroft [14] and was established in teaching by Hopcroft et al. [15, pp. 154-164].

Theorem 2. [15, pp. 162-164] MINIMIZEDFA *computes a minimal DFA to its input DFA.*

1.1.6 m -ComEquivPairs.

When looking at COMEQUIVPAIRS, one notes, that it computes distinct subsets of $Q \times Q$ on the way. Indeed, one could write the algorithm in such a way, that these subsets are explicitly computed in form of a function $m: \mathbb{N} \rightarrow \mathcal{P}(Q \times Q)$:

```

1: function  $m$ -COMEQUIVPAIRS( $A$ )
2:    $i \leftarrow 0$ 
3:    $m(0) \leftarrow \{(p, q), (q, p) \mid p \in F, q \notin F\}$ 
4:   do
5:      $i \leftarrow i + 1$ 
6:      $m(i) \leftarrow \{(p, q), (q, p) \mid (p, q) \notin \bigcup m(\cdot) \wedge \exists \sigma \in \Sigma: (\delta(p, \sigma), \delta(q, \sigma)) \in m(i-1)\}$ 
7:   while  $m(i) \neq \emptyset$ 
8:   return  $\bigcup m(\cdot)$ 

```

Using this redefinition, we can easier refer to the state pairs marked in a certain iteration. We will use both variants in exchange.

We will denote the number of iterations done by COMEQUIVPAIRS on an DFA A as $D(A)$. Note that $D(A) = \max n \in \mathbb{N} \mid m(n) \neq \emptyset$. **Gregor:** Does that note maybe fit very well to the proof of lemma 2?

1.1.7 Another equivalence automaton

Consider step 3 and 4 of MINIMIZEDFA. A view on these algorithms reveal, that they are essentially collapsing each \sim -equivalence class of a DFA A into one state.

Definition 3 (\sim_A -equivalence automaton). We will call the automaton $E_A = (Q_E, \Sigma_E, \delta_E, s_E, F_E)$ created by COMEQUIVPAIRS and REMEQUIVPAIRS(A) the \sim_A -equivalence automaton of A . It has the following properties:

- $Q_E = \{ [q] \mid q \in Q \}$
- $\Sigma_E = \Sigma$
- $\delta_E([q], \sigma) = [\delta(q, \sigma)], \forall q \in Q, \sigma \in \Sigma$
- $s_E = [s]_{\sim}$
- $F_E = \{ [q] \mid q \in Q \}$

Note, that we know by theorem 2 that E_A is guaranteed to be a minimal automaton to A , if A has no unreachable states.

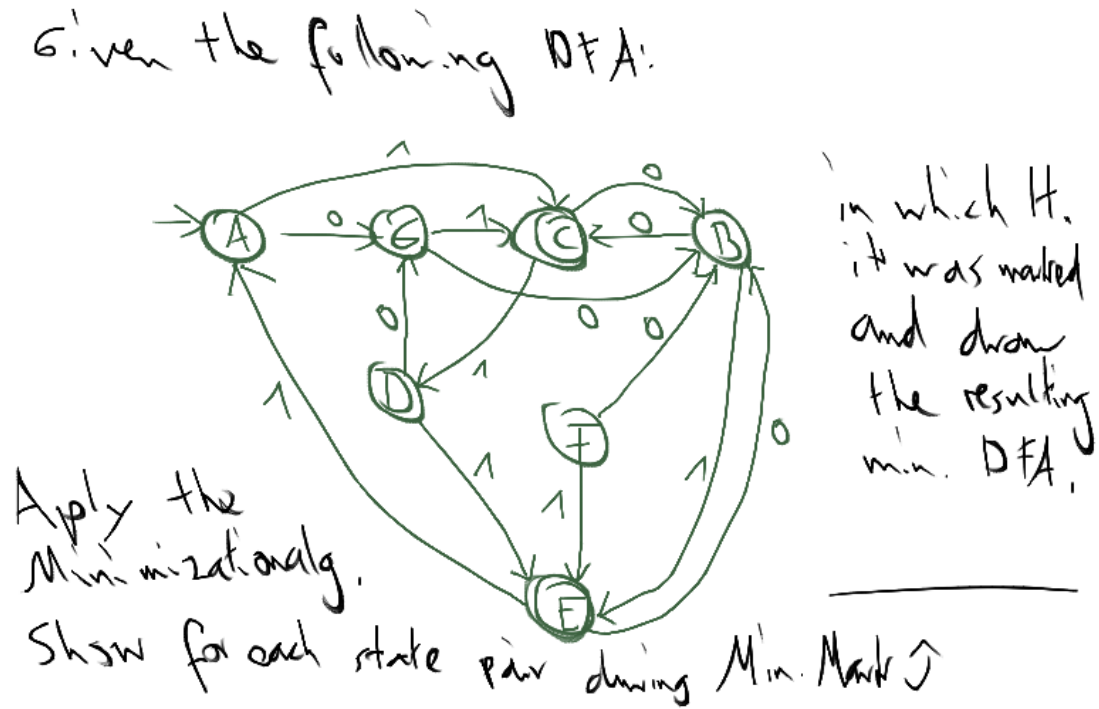


Figure 1.2: An example DFA minimization task.

1.2 Requirements analysis

Now that we have introduced all necessary basic definitions, we shall do a short analysis of an example DFA minimization task and its sample solution, as it could have been given to students in an introductory course to automata theory.

1.2.1 Example of a DFA minimization task for students

Gregor: search for typical task in standard text books

Figures 1.2 and 1.3 show such a task and solution. The students are confronted with a *task DFA* A_{task} . Firstly, unreachable states have to be eliminated, we then gain the *reachable DFA* A_{re} . Secondly equivalent state pairs of A_{re} are merged such that the minimal *solution DFA* A_{sol} is found. The table T displayed in figure 1.3 is nothing else but a visualization of the function m , whereas $T(q_0, q_1) = i \Leftrightarrow (q_0, q_1) \in m(i)$.

We do some rather formal statements and requirements. Firstly, we can state that

- $A_{re} = \text{REMUNREACHABLES}(A_{task}, \text{COMUNREACHABLES}(A_{task}))$ and
- $A_{sol} = \text{REMEQUIVPAIRS}(A_{re}, \text{COMEQUIVPAIRS}(A_{re}))$

Therefore A_{sol} is minimal regarding A_{re} and A_{task} . Secondly the languages of A_{task} , A_{re} and A_{sol} are be equal. We know that COMEQUIVPAIRS requires A_{re} to be complete and that REMEQUIVPAIRS creates complete DFAs, so A_{sol} is complete too. Furthermore we know that every state of A_{re} is reachable since it is the output of REMUNREACHABLES.

Gregor: How to define 'already found DFA sol' as requirement. Final states argument missing.

1.2.2 Difficulty adjustment possibilities

Concerning the execution of MINIMIZEDFA we find that its difficulty can be classified through various classification numbers.

Step 1: Delete unreachable states.
F is unreachable.

Step 2: Min Mark & merge dupl. states

	A	B	C	D	E	F
A	1	1	0		0	2
B	1	1	0	1	0	1
C	1	1	1	0		0
D	1	1	1	1	0	2
E	1	1	1	1	1	0
F	1	1	1	1	1	1

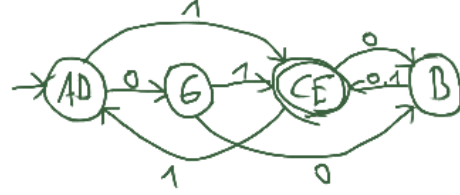


Figure 1.3: Solution to the DFA minimization task in fig. 1.2.

ComEquivPairs-depth ($D(A_{task})$). Consider the computation of the sets $m(i)$ in COMEQUIVPAIRS. Determining $m(0)$ is quite straightforward, because it consists simply of tests whether two states are in $F \times Q \setminus F$ (see 0, line 3). Determining $m(1)$ is less easy: The rule for determining all $m(i), i > 0$ is different to that for $m(0)$ and more complicated (see 0, line 6). Determining $m(2)$ requires the same rule. It shows nonetheless a students understanding of the terminating behavior of COMEQUIVPAIRS: It does not stop after computing $m(1)$, but only when no more distinguishable state pairs were found. Concerning the sets $m(i), i > 2$ however no additional understanding can be shown.

It would therefore be sensible if $D(A_{task})$ could be adjusted for example by parameters m_{min}, m_{max} which give lower and upper bounds for that value.

Number of states (Q_{sol}, Q_{eq}, Q_{unr}). To control the number of states in A_{task}, A_{re} and A_{sol} , we will introduce three parameters: $Q_{sol}, Q_{eq}, Q_{unr} \in \mathbb{N}$. These parameters get their meaning by the following equations:

$$\begin{aligned} |Q_{sol}| &= Q_{sol} \\ |Q_{re}| &= Q_{sol} + Q_{eq} \\ |Q_{task}| &= Q_{sol} + Q_{eq} + Q_{unr} \end{aligned}$$

It is sensible to have $Q_{unr} > 1, Q_{eq} > 1$, such that REMUNREACHABLES and REMEQUIVPAIRS will not be skipped. To not make the task trivial, $Q_{sol} > 2$ is sensible. An exercise instructor will find it useful, to control exactly how big Q_{unr}, Q_{eq} and Q_{sol} are: The higher Q_{unr}, Q_{eq} , the more states have to be eliminated and merged. The higher $Q_{sol} + Q_{eq}$, the more state pairs have to be checked during COMEQUIVPAIRS.

Gregor: Concept for min max.

Alphabet size. The more symbols the alphabet of A_{task}, A_{re} and A_{sol} has (note how MINIMIZEDFA does not change the alphabet), the more transitions have to be followed when checking whether $(\delta(q, \sigma), \delta(p, \sigma)) \in m(i - 1)$ is true for each state pair p, q .

Completeness of A_{task} . Even though COMUNREACHABLES and REMUNREACHABLES do not require their input DFA A_{task} to be complete, it is sensible to build it that way. The implications of the completeness-property are - in comparison to the other concepts involved here - rather subtle. This is especially due to its purely representational nature, a DFA has the same language and D -value, whether it is represented in its complete form or not. Nonetheless we shall introduce a parameter c , that determines if there exist unreachable states, that make A_{task} incomplete. Thus an exercise lecturer could showcase this matter on a DFA and generate according exercises.

Planar drawing of A_{task} . A graph G is *planar* if it can be represented by a drawing in the plane such that its edges do not cross. Such a drawing is then called *planar drawing* of G . A visual aid for students would be given, if the task DFA were planar and presented as a planar drawing. In this work libraries and parameters $p_1, p_2 \in \{0, 1\}$ (toggling planarity of A_{sol}, A_{task}) will be used to allow the option of planarity, but neither ensuring planarity nor planar drawing will be investigated further theoretically.

Maximum degree of any state in A_{task} . The *degree* $deg(q)$ of a state $q \in Q$ in a DFA A is defined as $deg(q) = |d^-(q)| + |d^+(q)|$, so the total number of transitions in which q participates. By capping the maximum degree for all states, the graphical representation of the DFA would be more clear. In this work the inclusion of a maximum degree parameter is omitted.

1.2.3 Summary of found criteria

Gregor: TODO

Accepted general criteria:

- > $L(A_{sol}) = L(A_{re}) = L(A_{task})$
- > $D(A_{sol}) = D(A_{re}) = D(A_{task})$

Accepted solution DFA criteria:

- > has to be minimal, complete
- > number of essential states
- > number of COMEQUIVPAIRS iterations ($D(A_{sol})$)
- > alphabet size
- > number of accepting states
- > planarity
- > A_{sol} is new

Definition 4 (New DFAs). A DFA A_{sol} is *new* if it is not practically isomorph to any previously generated solution DFA.

Accepted reachable DFA criteria:

- > has to be complete
- > number of unreachable states
- > planarity (can be checked in $O(|Q_{task}|)$)

Accepted task DFA criteria:

- > number of states that are added to create equivalent state pairs
- > planarity
- > completeness

1.3 Approach and general algorithm

In this work we will first build the solution DFA (step 1), and - based on that - the task DFA by creating equivalent states and adding unreachable states (step 2). Both steps will fulfill all criteria chosen above and are covered in depth in chapter 2 respectively chapter 3.

It follows that D and L of both DFAs will be set when building A_{sol} . We know that creating equivalent states and adding unreachable does not change $L(A_{task})$ in comparison to A_{sol} , else MINIMIZEDFA would not work (a minimal DFA has in particular the same language as the original DFA). However we must ensure, that adding those states does not change D . Since unreachable states are eliminated before COMEQUIVPAIRS is applied, we need only to ensure, that creating equivalent states does not change the D -value. We will do this during the discussion of step 2, more specifically in section 3.1.4.

At the beginning of chapter 2 and 3, we will provide formal problem definitions for both steps, that specify precisely all requirements. Here we shall content ourselves with the definition of the main algorithm:

- 1: **function** GENERATEDFAMINIMIZATIONPROBLEM($Q_{sol}, a, f, m_{min}, m_{max}, p_1, p_2, Q_{eq}, Q_{unr}, c$)
- 2: $A_{sol} \leftarrow \text{GENERATENEWMINIMALDFA}(Q_{sol}, a, f, m_{min}, m_{max}, p_1)$
- 3: $A_{task} \leftarrow \text{EXTENDMINIMALDFA}(A_{sol}, p_2, Q_{eq}, Q_{unr}, c)$
- 4: **return** A_{sol}, A_{task}

Chapter 2

Generating minimal DFAs

We seek algorithms for generation of minimal DFAs that fulfilling the conditions defined in the requirements analysis section 1.2.3. We formally subsume these conditions via the GenerateNewMinimalDFA-problem:

Definition 5 (GenerateNewMinimalDFA).

Given:

$$\begin{aligned} Q_{sol} &\in \mathbb{N} && \text{number of states} \\ a &\in \mathbb{N} && \text{alphabet size} \\ f &\in \mathbb{N} && \text{number of final states} \\ m_{min}, m_{max} &\in \mathbb{N} && \text{lower and upper bound for } D\text{-value} \\ p &\in \{0, 1\} && \text{planarity-bit} \end{aligned}$$

Task: Compute, if it exists, a solution DFA A_{sol} with

- $|Q_{sol}| = Q_{sol}$, $|\Sigma_{sol}| = a$, $|F_{sol}| = f$
- $m_{min} \leq D(A_{sol}) \leq m_{max}$
- A_{sol} being planar iff $p = 1$
- A_{sol} being new

We consider different approaches to solve this problem, of which those using trial-and-error will be discussed most broadly.

Note that the presented algorithms will not be able to compute all of \mathfrak{A}_{min} since we are going to exclude minimal DFAs that are practical isomorph to already found ones.

2.1 Using trial and error

We will develop an algorithm that makes partly use of the trial-and-error paradigm to find matching DFAs. The approach here is as follows:

Firstly a *test* DFA A_{test} is generated by use of either randomness or enumeration. Alphabet size and number of (final) states will already be correct. On this DFA then tests will be executed, to check if it is minimal, planar (if wished) and new. If this is the case, A_{test} will be returned, if not, new test DFAs are generated until all tests pass.

By constructing test DFAs with already correct alphabet size and number of (final) states we are able to subdivide the search space of DFAs in advance into much smaller pieces which are in particular finite.

Gregor: How much smaller? Why now finite?

```

1: function BUILDNEWMINIMALDFA-1 ( $\mathcal{Q}_{sol}, a, f, m_{min}, m_{max} \in \mathbb{N}, p \in \{0, 1\}$ )
2:   while True do
3:     generate DFA  $A_{test}$  with  $|Q|, |\Sigma|, |F|$  matching  $\mathcal{Q}_{sol}, a, f$ 
4:     if  $A_{test}$  not minimal or not  $m_{min} \leq D(A_{test}) \leq m_{max}$  then
5:       continue
6:     if  $p = 1$  and  $A_{test}$  is not planar then
7:       continue
8:     if  $A_{test}$  is not new then
9:       continue
10:    return  $A_{test}$ 

```

We will complete this algorithm by resolving how the tests in lines 4, 6 and 8 work and by showing two methods for generation of automata with given restrictions of $|Q|$, $|\Sigma|$ and $|F|$.

2.1.1 Ensuring A_{test} is minimal and $D(A_{test})$ is correct

In order to test, whether A_{test} is minimal, we could simply use the minimization algorithm and compare the resulting DFA and A_{test} using an isomorphism test. However it is sufficient to ensure, that no equivalent or unreachable states exist. **Gregor: minimality planarity complete under isomorphism**

To get $D(A_{test})$, we have to run COMEQUIVPAIRS entirely anyway. Hence we can combine the test for equivalent states with computing the DFA's D -value:

```

1: function HASEQUIVALENTSTATES( $A$ )
2:    $depth \leftarrow 0$ 
3:    $M \leftarrow \{(p, q), (q, p) \mid p \in F, q \notin F\}$ 
4:   do
5:      $depth \leftarrow depth + 1$ 
6:      $M' \leftarrow \{(p, q) \mid (p, q) \notin M \wedge \exists \sigma \in \Sigma: (\delta(p, \sigma), \delta(q, \sigma)) \in M\}$ 
7:      $M \leftarrow M \cup M'$ 
8:   while  $M' \neq \emptyset$ 
9:    $hasDupl \leftarrow |\{(p, q) \mid p \neq q \wedge (p, q) \notin M\}| > 0$ 
10:  return  $hasDupl, depth$ 

```

Since COMEQUIVPAIRS basically computes all distinguishable state pairs $\not\sim_A$, we test in line 9, whether there is a pair of distinguishable states not in $\not\sim_A$.

Regarding the unreachable states, we can just use COMUNREACHABLES and test whether the computed set is empty:

```

1: function HASUNREACHABLESTATES( $A$ )
2:  return  $|\text{COMUNREACHABLES}(A)| > 0$ 

```

2.1.2 Ensuring A_{test} is planar

There exist several algorithms for planarity testing of graphs. In this work, the library *pygraph*¹ has been used, which implements the Hopcroft-Tarjan planarity algorithm. More information on this can be found for example in this [18] introduction from William Kocay. The original paper describing the algorithm is [13].

¹<https://github.com/jciskey/pygraph>

2.1.3 Ensuring A_{test} is new

In our requirements we stated, that we wanted the generated solution DFA to be new, meaning not practically isomorph to any previously generated solution DFA. This implies the need of a database, that allows saving and loading DFAs. We name this database *DB1*. Assuming the database is relational, the following scheme is proposed:

$$|Q_A| \quad |\Sigma_A| \quad |F_A| \quad D(A) \quad isPlanar(A) \quad encode(A)$$

With this scheme we can fetch once all DFAs matching the search parameters. Thus we need not fetch all previously found DFAs every time, but only those that are relevant. Afterwards we must only check whether any practical isomorphy test on the current test DFA and one of the fetched DFAs is positive. If any test DFA passes all tests and is going to be returned, then we have to save that DFA in the database.

A more concrete specification of this proceeding is shown below, embedded in the main algorithm:

```

1: function BUILDNEWMINIMALDFA-2 ( $\mathcal{Q}_{sol}, a, f, m_{min}, m_{max}, p$ )
2:    $l \leftarrow$  all DFAs in DB1 matching  $\mathcal{Q}_{sol}, a, f, m_{min}, m_{max}, p$ 
3:   while True do
4:     ...
5:     if  $A_{test}$  is practical isomorph to any DFA in  $l$  then
6:       continue
7:     save  $A_{test}$  and its respective properties in DB1
8:     return  $A_{test}$ 

```

2.1.4 Option 1: Generating A_{test} via Randomness

We now approach the task of generating a random DFA whereas alphabet and number of (final) states are set.

Corollary ?? tells us, that the states names are irrelevant for the minimality of a DFA, therefore we will give our generated DFAs simply the states $e_1, \dots, e_{\mathcal{Q}_{sol}}$. For alphabet symbols this is not given. But since we [Gregor: TODO minimality and planarity complete under isomorphy](#)

We can state, that our start state is $e_1 \in Q$, since we can apply an isomorphism to every DFA, such that its start state is renamed to e_1 .

The remaining elements that need to be defined are δ and F . The set of final states is supposed to have a size of f and be a subset of Q . Therefore we can simply choose randomly f distinct states from Q .

The transition function has to make the DFA complete, so we have to choose an “end” state for every combination in $Q \times \Sigma$. There is no restriction as to what this end state shall be, so given $e \in Q$ and $\sigma \in \Sigma$ we can randomly choose an end state from Q .

With defining how to compute δ we have covered all elements of a DFA.

```

1: function BUILDNEWMINIMALDFA-3A ( $\mathcal{Q}_{sol}, a, f, m_{min}, m_{max}, p$ )
2:    $l \leftarrow$  all DFAs in DB1 matching  $\mathcal{Q}_{sol}, a, f, m_{min}, m_{max}, p$ 
3:    $Q \leftarrow \{e_1, \dots, e_{\mathcal{Q}_{sol}}\}$ 
4:    $\Sigma \leftarrow \{\sigma_1, \dots, \sigma_a\}$ 
5:   while True do
6:      $\delta \leftarrow \emptyset$ 

```



```

7:      for  $e$  in  $Q$  do
8:          for  $\sigma$  in  $\Sigma$  do
9:               $e' \leftarrow$  random chosen state from  $Q$ 
10:              $\delta \leftarrow \delta \cup \{(e, \sigma), e'\}$ 
11:          $s \leftarrow e_1$ 
12:          $F \leftarrow$  random sample of  $f$  states from  $Q$ 
13:          $A_{test} \leftarrow (Q, \Sigma, \delta, s, F)$ 
14:         if  $A_{test}$  not minimal or not  $m_{min} \leq D(A_{test}) \leq m_{max}$  then
15:             continue
16:         if  $p = 1$  and  $A_{test}$  is not planar then
17:             continue
18:         if  $A_{test}$  is isomorph to any DFA in  $l$  then
19:             continue
20:         save  $A_{test}$  and its respective properties in DB1
21:     return  $A_{test}$ 

```

2.1.5 Option 2: Generating A_{test} via Enumeration

The second method of test DFA generation is based on the idea, that instead of randomly generating F and δ , we could just enumerate through all possible final state sets and transition functions.

Both enumerations are finite, given Q_{sol} and a . Having a requirement of f final states, then Q_{sol} choose f is the number of possible F -configurations. On the other hand there are $Q_{sol}^{Q_{sol} * a}$ possible δ -configurations: We have to choose one of Q_{sol} possible end states for every combination in $Q \times \Sigma$ - so $Q_{sol} * a$ times.

Again we will call our states and symbols w.l.o.g. $e_1, \dots, e_{Q_{sol}}$ resp. $\sigma_1, \dots, \sigma_a$. We will represent the state of an enumeration with two fields F_F and F_δ . The first field shall have Q_{sol} Bits, whereas Bit $F_F[i] \in [0, 1]$ represents the information, whether e_i is a final state or not. The second field shall have $Q_{sol} * a$ entries containing state indices, such that entry $F_\delta[i * a + j] = k, k \in [1, Q_{sol}]$ says, that $\delta(e_i, \sigma_j) = e_k$. These semantics are illustrated in figure 2.1.

Given an enumeration state b_f, b_t and Q_{sol}, a, f we will then compute the next DFA based on this state as follows. We will treat both fields as numbers, F_f as binary and F_δ as Q_{sol} -ary. To get to the next DFA, we will first increment F_δ by 1. If $F_\delta = 1 \dots 1$, then we increment F_F until it contains f ones (again) and set F_δ to $0 \dots 0$. This behavior is summarized in the following algorithm: **Gregor: Clarify what happens at 11111...**

```

1: function INCREMENTENUMPROGRESS ( $F_F, F_\delta, Q_{sol}, a, f$ )
2:     add 1 to  $(F_\delta)_{Q_{sol}}$ 
3:     if  $F_\delta = 0 \dots 0$  then
4:         while  $\#_1(F_F) \neq f$  do                                 $\triangleright$  if the number of 1s in  $F_F$  is not  $f$ 
5:             add 1 to  $(F_F)_2$ 
6:             if  $F_F = 0 \dots 0$  then
7:                 return  $\perp$ 
8:              $F_\delta = 0 \dots 0$ 
9:     return  $F_F, F_\delta$ 

```

The example in figure 2.2 illustrates such increments.

Based on the incremented bit-fields the new DFA can be build according to the semantics defined above:

Given: $q=4, a=2, f=3$

example b_f :

1101

so

$$b_f[0] = 1$$

$$b_f[1] = 1$$

$$b_f[2] = 0$$

$$b_f[3] = 1$$

$$F = \{q_0, q_2, q_3\}$$

example b_+ :

q_0	q_1	q_2	q_3
$\sigma_0 \sigma_1$	$\sigma_0 \sigma_1$	$\sigma_0 \sigma_1$	$\sigma_0 \sigma_1$
$q_1 q_2$	$q_0 q_1$	$q_3 q_2$	$q_0 q_1$
01	10	00	01 11 10 00 01



so e.g.

$$\delta(q_0, \sigma_0) = q_1$$

$$\delta(q_2, \sigma_1) = q_3$$

...

Figure 2.1: Example for two possible configurations of the fields F_F and F_δ given Q_{sol}, a and f . Below the corresponding DFA is drawn.

Given: $q=4, a=2, f=3$

example b_+ :

example b_f :

1101

q_0	q_1	q_2	q_3
$\sigma_0 \sigma_1$	$\sigma_0 \sigma_1$	$\sigma_0 \sigma_1$	$\sigma_0 \sigma_1$
$q_1 q_2$	$q_0 q_1$	$q_3 q_2$	$q_0 q_1$
01	10	00	01 11 10 00 01
			1
			= 01 10 00 01 11 10 00 10
q_1	q_2	q_0	$q_1 q_3 q_2 q_0 q_2$

Assuming $b_+ = 11 11 11 11 11 11 11 11$ before mar.
then b_f is incr. until it has f 1's again

$$1101 \rightarrow 1110$$

and b_+ is set to 00 00 00 00 00 00 00 00

Figure 2.2: The upper half shows how a F_δ -increment results in a change in the resulting DFAs transition function: $\delta(e_3, \sigma_1) = e_1$ becomes $\delta(e_3, \sigma_1) = e_2$. The lower half shows what happens, if F_δ has reached its end.

```

1: function DFAFROMENUMPROGRESS ( $F_F, F_\delta, \mathcal{Q}_{sol}, a, f$ )
2:    $Q \leftarrow \{e_1, \dots, e_{\mathcal{Q}_{sol}}\}$ 
3:    $\Sigma \leftarrow \{\sigma_1, \dots, \sigma_a\}$ 
4:    $\delta \leftarrow \emptyset$ 
5:   for  $i$  in  $[1, \dots, \mathcal{Q}_{sol}]$  do
6:     for  $j$  in  $[1, \dots, a]$  do
7:        $\delta(e_i, \sigma_j) = e_{F_\delta[i*a+j]}$ 
8:    $s \leftarrow e_1$ 
9:    $F \leftarrow \{ e_i \mid i \in [1, \dots, \mathcal{Q}_{sol}] \wedge F_F[i] = 1 \}$ 
10:  return ( $Q, \Sigma, \delta, s, F$ )

```

The initial field values are each time $0 \dots 0$. Note how construction and use of these fields results in DFAs with correct alphabet size and number of (final) states. An enumeration can finish either because a matching DFA has been found or all DFAs have been enumerated.

Once the enumeration within a call of BUILDNEWMINIMALDFA has been finished, it is reasonable to *save* the enumeration progress (meaning the current content of F_F, F_δ), such that during the next call enumeration can be resumed from that point on. The alternative would mean, that the enumeration is run in its entirety until that point again, whereas all so far found DFAs would be found to be not new. Thus we introduce a second database *DB2* with the following table:

$$|Q_A| \quad |\Sigma_A| \quad F_F \quad F_\delta$$

We reduce the enumeration room for each calculation.

```

1: function BUILDNEWMINIMALDFA-3B ( $\mathcal{Q}_{sol}, a, f, m_{min}, m_{max}, p$ )
2:    $l \leftarrow$  all DFAs in DB1 matching  $\mathcal{Q}_{sol}, a, f, m_{min}, m_{max}, p$ 
3:    $F_F, F_\delta \leftarrow$  load enumeration progress for  $\mathcal{Q}_{sol}, a, f, p$  from DB2
4:   while True do
5:     if  $F_F, F_\delta$  is finished then
6:       save  $F_F, F_\delta$ 
7:       return  $\perp$ 
8:      $A_{test} \leftarrow$  next DFA based on  $F_F, F_\delta$ 
9:     if  $A_{test}$  not minimal or not  $m_{min} \leq D(A_{test}) \leq m_{max}$  then
10:      continue
11:     if  $p = 1$  and  $A_{test}$  is not planar then
12:       continue
13:     if  $A_{test}$  is isomorph to any DFA in  $l$  then
14:       continue
15:     save  $F_F, F_\delta$  in DB2
16:     save  $A_{test}$  and its respective properties in DB1
17:     return  $A_{test}$ 

```

2.2 Alternative approach: Building $m(i)$ bottom up

Build m from m -COMEQUIVPAIRS iteratively. (Why would this basically result in running COMEQUIVPAIRS all the time?)

2.3 Related research on DFA generation

Nicaud provides an overview of results on random generation and combinatorial properties of DFAs in [21]. We will outline relevant related research.

Nicaud’s summary indicates, that research has focused on randomized generation of accessible, but not minimal DFAs so far. In the following we will sketch some approaches that have come up.

Using the recursive method. Champarnaud and Paranthoën [9] continue ideas started by Nicaud in his thesis [20]. Let $\mathfrak{F}_{n,m}$ be the set of extended m -ary trees of order n . These trees are characterized by a partitioning $V = N \uplus L$ with $|N| = n$ and $v \in N \Rightarrow d^+(v) = m$ and $v \in L \Rightarrow d^+(v) = 0$. We define the following set of tuples using $s = n(m - 1)$:

$$\mathfrak{R}_{m,n} = \{ (k_1, \dots, k_s) \in \mathbb{N}^s \mid \forall i \in [2, s]: k_i \geq \left\lceil \frac{i}{m-1} \right\rceil \text{ and } k_i \geq k_{i-1} \}$$

In [9, p. 6] it is shown that there exists a bijection φ between $\mathfrak{F}_{n,m}$ and $\mathfrak{R}_{m,n}$ which maps to k_i , $i \in [1, s]$ of a tuple the number of leaves visited before the i th leaf in a tree. The connection to accessible DFAs is established by proving that “transition structures²” with $|Q| = n$, $|\Sigma| = m$ reduced to the set of the smallest paths from the s to each other state are in bijection with extended m -ary trees of order n (see [9, p. 8]).

As a consequence they are able to construct a random generation of accessible complete DFAs using the “recursive method” from [22] which generates n -tuples [9, p. 10]. Nicaud states in his survey that the algorithm’s runtime is $\mathcal{O}(n^2)$ but notes, that generation of DFAs with more than “a few thousand states” is practically hard to do [21, pp. 10-11].

Almeida et. al. [1, 2, 23] present and implement methods using a string-encoding of DFAs for exact enumeration and random generation of DFAs. Nicaud [21, p. 11] states in a remark, that this approach uses the same recursive method and differs only in the DFA encoding.

Using Boltzmann sampler. Bassino, David and Nicaud present and implement a more efficient random generator of accessible complete DFAs in [4, 6]. Their idea is based on so called Boltzmann samplers. This framework of samplers is characterized in particular by the fact that the size of its generated objects are not fixed but in an interval around a given input size - this stands in opposition to most random generators in literature [11, p. 2].

In [6] the authors use a Boltzmann sampler to generate set partitions that are shown to be in bijection with so called box diagrams [6, p. 8] which are in turn in bijection to accessible complete DFAs [6, p. 4]. They thus acquire an average runtime complexity of $\mathcal{O}(n^{3/2})$ for a single random generation.

Using a rejection algorithm. Carayol and Nicaud [8] give a simple algorithm with the same runtime complexity. They use a result stating that the size of accessible DFAs is concentrated around some computable value. In the end random possibly inaccessible DFAs of a specific size are generated, of which afterwards all unreachable states are deleted. This is thus essentially a rejection algorithm with clever generation of test DFAs. They furthermore show that allowing approximate sampling with the number of states being in $[n - \varepsilon\sqrt{n}, n + \varepsilon\sqrt{n}]$ results in linear expected runtime.

²Those are essentially DFAs without final state sets.

Others and comparison to algorithm presented in this work. In his survey Nicaud mentions a paper by Bassino and Sportiello [3] that yields random generation of accessible DFAs in expected linear time. This work will not be discussed further here.

In this work we use a rejection algorithm that generates test DFAs either by randomization or by enumeration. Both methods implement a naive approach. The generated test DFAs are not necessary minimal and in particular not necessary accessible as in [8]. The enumeration method uses encodings of DFAs similar to those used by Almeida et. al. [23].

2.4 Empirical and combinatorial results

Concerning combinatorial properties of DFAs, several authors (e.g. [6, 10, 16]) consider a work from Vyssotsky [25] in the Bell laboratories to be the first on this subject. A contribution by Korshunov [19] is often cited in this regard, for he firstly “determines an asymptotic estimate of the number of accessible complete and deterministic n -state automata over a finite alphabet” [5].

Implementations (e.g. [1, 4]) of various random and enumeration generation methods have given rise to several empirical observations concerning the number of minimal DFAs, their fraction among all DFAs and so forth.

Domaratzki, Kisman, and Shallit [10] give some asymptotic estimates and explicit computations for the number of each several types of languages and automata that are distinct. The here relevant results have been subsumed and extended in [2, p. 8] by means of exact enumeration and are confirmed in [4].

$ \Sigma (k)$	$ Q (n)$	$ \mathcal{A}_{min,n,k} $	$ \mathcal{A}_{n,k} $	Minimal %
$k = 2$	2	24	64	0.38
	3	1028	5832	0.18
	4	56014	1048576	0.05
	5	3705306	312500000	0.01
	6	286717796	139314069504	0.0
	7	25493886852	86812553324672	0.0
$k = 3$	2	112	256	0.44
	3	41928	157464	0.27
	4	26617614	268435456	0.1
	5	25184560134	976562500000	0.03
$k = 4$	2	480	1024	0.47
	3	1352732	4251528	0.32
	4	7756763336	68719476736	0.11
$k = 5$	2	1984	4096	0.48
	3	36818904	114791256	0.32

Figure 2.3: Table depicting the amount of minimal complete DFAs among all complete DFAs for various sizes of Q, Σ . The numbers of minimal DFAs (bold numbers) are taken from [2, p. 8].

In table 2.3 we use these results to determine the ratios of minimal complete DFAs among all complete DFAs for given $|Q|$ and $|\Sigma|$. The number of all DFAs is computed as follows:

$$|\mathcal{A}_{n,k}| = \underbrace{n^{n*k}}_{\text{\#possible } \delta\text{'s}} * \underbrace{2^n}_{\text{\#possible sets } F}$$

Thus we gain an insight into how probable the generation of a distinct minimal test DFA is without applying further constraints. For our proposed default parameters $n \in [4 - 5]$

and $k \in [2 - 3]$ the probabilities of successful generation range from 1% to 5%. Practical tests have shown that this leads to sufficient short run times for our implementation.

Further interesting results in this area include the determination of the fraction of minimal automata among accessible complete DFAs [5] and asymptotic estimates for the number of states that a random minimized DFA has [7].

Chapter 3

Extending minimal DFAs

We firstly define a formal problem for extending a minimal DFA A_{sol} to a task DFA A_{task} based on our requirements analysis (see 1.2.3):

Definition 6 (ExtendMinimalDFA).

Given:

$$\begin{aligned}
 A_{sol} = (Q, \Sigma, \delta, s, F) &\in \mathfrak{A}_{min} && \text{solution DFA} \\
 Q_{eq} &\in \mathbb{N} && \text{number of states creating equivalent state pairs} \\
 Q_{unr} &\in \mathbb{N} && \text{number of unreachable states} \\
 p &\in \{0, 1\} && \text{planarity-bit} \\
 c &\in \{0, 1\} && \text{completeness-bit}
 \end{aligned}$$

Task: Compute, if it exists, a task DFA A_{task} with

- $Q_{task} = Q_{sol} \cup \{r_1, \dots, r_{Q_{eq}}, u_1, \dots, u_{Q_{unr}}\}$
- $r_1, \dots, r_{Q_{eq}}$ each creating an equivalent state pair
- $u_1, \dots, u_{Q_{unr}}$ unreachable
- $\Sigma_{task} = \Sigma_{sol}, s_{task} = s_{sol}, F_{task} \subseteq F_{sol}$
- A_{task} being planar iff $p = 1$
- A_{task} being complete iff $c = 1$
- A_{sol} being isomorph to $\text{MINIMIZEDFA}(A_{task})$

In order to fulfill these requirements we will deduce for both kinds of states how they may be added by examining their desired properties. We will show for the action of adding equivalent states, that this does not change a DFA's D -value.

3.1 Creating equivalent state pairs

Step 3 and 4 of the minimization algorithm are concerned with detection and elimination of equivalent state pairs. We now want to add states $r_1, \dots, r_{Q_{eq}}$ to a DFA A_{sol} , gaining A_{re} with $Q_{re} = Q_{sol} \cup \{r_1, \dots, r_{Q_{eq}}\}$, such that each of these states is equivalent to a state in A_{re} . Note that, for reasons of clarity, we are going to abbreviate from now on $A_{re} = A$, $Q_{re} = Q$, $\sim_{A_{re}} = \sim_A$ etc.

Consider the properties $r_1, \dots, r_{Q_{eq}}$ must have. They are equivalent to states $o_1, \dots, o_{Q_{eq}}$ of A .

$$\exists r_1, \dots, r_{Q_{eq}} \in Q: \exists o_1, \dots, o_{Q_{eq}} \in Q: \forall i \in [1, Q_{eq}]: r_i \sim_A o_i$$

But we know also and in particular, that each of them is equivalent a state e of A_{sol} .

$$\exists r_1, \dots, r_{Q_{eq}} \in Q: \forall i \in [1, Q_{eq}]: \exists e \in Q_{sol}: r_i \sim_A e$$

In our algorithm, we will choose the state e for each state we add.

3.1.1 Adding outgoing transitions

Regarding the outgoing transitions of any r_i equivalent to a state e , we are directly restricted by the relationship $\forall \sigma \in \Sigma: [\delta(r_i, \sigma)]_{\sim_A} = [\delta(e, \sigma)]_{\sim_A}$. Thus, when adding some r_i , we have to choose for each symbol $\sigma \in \Sigma$ at exactly one transition (completeness requirement for A) from the following set:

$$O_{e,\sigma} = \{ ((r_i, \sigma), q) \mid q \in [\delta(e, \sigma)]_{\sim_A} \}$$

Since the solution DFA is complete and since every here added state gets a transition for every alphabet symbol, we know that every $O_{e,\sigma} \neq \emptyset$.

Gregor: Why does this not affect the eq. class of any other state?

3.1.2 Adding ingoing transitions

First of all, we know, that r_i is reachable, since every state of A must be reachable, so we need to give r_i at least one ingoing transition. Doing this, we have to ensure, that any state q , that gets such an outgoing transition to r_i remains in its \sim -equivalence class.

Thus a fitting state q has to have a transition to some state in $[r_i]_{\sim_A} = [e]_{\sim_A}$ already. So, given a state q with $\delta(q, \sigma) = p$ and $p \in [e]_{\sim_A}$, we can set $\delta(q, \sigma) = r_i$ and thus “steal” q its ingoing transition.

We see here, that q must have at least 2 ingoing transitions, else it would become unreachable. Thus we summarize:

$$I_e = \{ ((q, \sigma), p) \mid \delta(q, \sigma) = p \wedge p \in [e] \wedge d^-(p) \geq 2 \}$$

Choose at least one $((q, \sigma), p) \in I_e$, remove $((q, \sigma), p)$ from δ and add $((q, \sigma), r_i)$.

These finding lead us to a general requirement regarding the choice of a state e for an r_i : The equivalence class of any e has to contain at least one state with at least 2 ingoing transitions (see fig. 3.1). We establish the following notion to pin down this restriction:

$$duplicatable(q) \Leftrightarrow_{def} (\exists p \in [q]_{\sim_A}: |d^-(p)| \geq 2)$$

The number of duplicatable states in any accessible DFA A is 0 for $|\Sigma| \leq 1$ (due to the restriction $|d^-(p)| \geq 2$) and greater than 0 for $|\Sigma| > 1$ due to the pigeonhole principle: An accessible complete DFA has $|Q||\Sigma|$ transitions which have to be spread across $|Q|$ states.

3.1.3 The algorithm

- 1: **function** CREATEEQUIVALENTSTATEPAIRS(A, Q_{eq})
- 2: $Q \leftarrow Q_{sol}$
- 3: $\delta \leftarrow \delta_{sol}$
- 4: $F \leftarrow F_{sol}$
- 5: $K \leftarrow \{ \{q\} \mid q \in Q \}$ ▷ tracks the equivalence classes of A

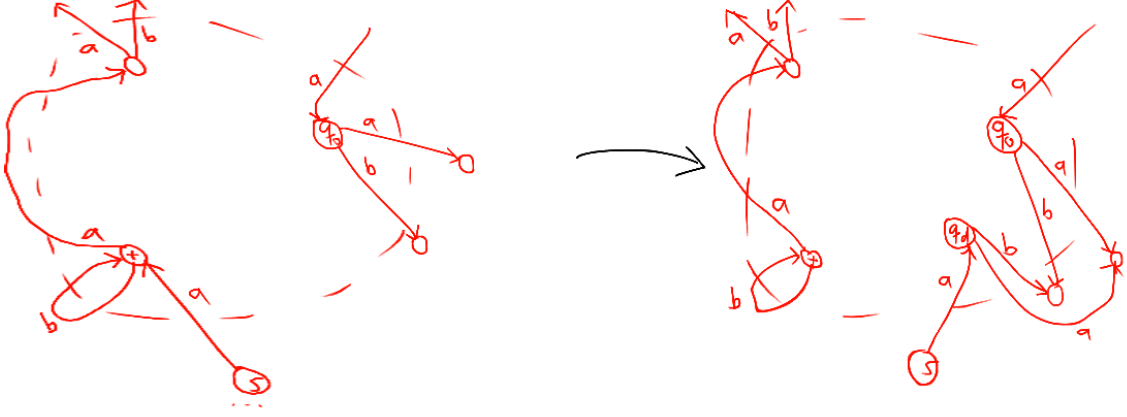


Figure 3.1: If an equivalence class (here denoted by the states in the dashed area) contains a state with 2 or more ingoing transitions (in this case p), then a state equivalent to any of the classes states may be added. Here r is equivalent to o and is “stealing” the ingoing transition $\delta(q, a)$ from p .

```

6:    $k(q) = C$  such that  $q \in C$  and  $C \in K$            ▷ returns the equivalence class to  $q$ 
7:    $in(q) = |d^-(q)|$  for all  $q \in Q$                    ▷ tracks the number of ingoing  $t$ .
8:   for  $i$  in  $[1, Q_{eq}]$  do
9:       for  $q$  in  $Q$  do                               ▷ find a duplicatable state  $e$ 
10:          if  $in(q) \geq 2$  then
11:               $e \leftarrow$  random chosen state from  $k(q)$ 
12:              break
13:          Add  $r_i$  to  $Q$                                 ▷ create to  $e$  equivalent state  $r_i$ 
14:          Add  $r_i$  to  $k(e)$ 
15:          for  $\sigma$  in  $\Sigma$  do                           ▷ add  $d^+(r_i)$ 
16:               $\delta(r_i, \sigma) =$  random chosen state from  $k(\delta(e, \sigma))$ 
17:           $P \leftarrow \{ ((s, \sigma), t) \in \delta \mid t \in k(e), in(t) \geq 2 \}$    ▷ add  $d^-(r_i)$ 
18:           $C \leftarrow$  random nonempty subset of  $P$ 
19:          for  $((s, \sigma), t)$  in  $C$  do
20:               $in(t) \leftarrow in(t) - 1$ 
21:               $in(r_i) \leftarrow 1$ 
22:               $\delta(s, \sigma) = r_i$ 
23:   return  $(Q, \Sigma_{sol}, \delta, s_{sol}, F)$ 

```

3.1.4 Creating equivalent state pairs does not change D

To prove this statement, we will prove two minor propositions first. In this context we will call a word w *distinguishing word* of p, q , iff *distinguishes*(w, p, q) whereas

$$\begin{aligned}
 \text{distinguishes}_A(w, p, q) &\Leftrightarrow \\
 &(\delta^*(p, w) \in F \Leftrightarrow \delta^*(q, w) \notin F) \vee \\
 &(\delta^*(q, w) \in F \Leftrightarrow \delta^*(p, w) \notin F) \\
 &\Leftrightarrow p \not\sim_A q
 \end{aligned}$$

The following lemma and its proof are based on [26, ch. 4 p. 18].

Lemma 1. *In the context of COMEQUIVPAIRS the following is true:*

$$(p, q) \in m(n) \iff \exists w \in \Sigma^*: (|w| = n \wedge \text{distinguishes}_A(w, p, q)) \\ \wedge \nexists v \in \Sigma^*: (|v| < |w| \wedge \text{distinguishes}_A(v, p, q))$$

Proof. Per induction on n .

$n = 0$, “ \Leftarrow ”.

$$(p, q) \in m(0) = \{(p, q), (q, p) \mid p \in F, q \notin F\} \text{ (see 0, line 2)} \\ \iff \text{one of } p, q \text{ in } F, \text{ one not} \\ \iff \text{one of } \delta^*(p, \varepsilon), \delta^*(q, \varepsilon) \text{ in } F, \text{ one not} \\ \iff \exists w \in \Sigma^*: (|w| = 0 \wedge \text{one of } \delta^*(p, w), \delta^*(q, w) \text{ in } F, \text{ one not}) \\ \iff \exists w \in \Sigma^*: (|w| = 0 \wedge \text{distinguishes}_A(w, p, q)) \\ \text{and there is no word shorter than 0 } \checkmark$$

$n - 1 \rightarrow n$, “ \Rightarrow ”. Then the following holds for some states p, q (see 0, line 5):

$$(p, q) \in m(n) = \{(p, q), (q, p) \mid (p, q) \notin \bigcup m(\cdot) \wedge \exists \sigma \in \Sigma: (\delta(p, \sigma), \delta(q, \sigma)) \in m(n - 1)\}$$

So in particular there exists a symbol σ such that $(\delta(p, \sigma), \delta(q, \sigma)) \in m(n - 1)$. Per induction there exist states $(p', q') = (\delta(p, \sigma), \delta(q, \sigma))$ in $m(n - 1)$ and a word w' , $|w'| = n - 1$ such that one of $\delta^*(p', w')$, $\delta^*(q', w')$ is not in F and there is no shorter word. Thus one of $\delta^*(p, \sigma w')$, $\delta^*(q, \sigma w')$ is not in F , which makes $\sigma w'$ a distinguishing word of length n for p, q .

$n - 1 \rightarrow n$, “ \Leftarrow ”. Then the following holds for some states p, q (see 0, line 5):

$$\exists w \in \Sigma^*: (|w| = n \wedge \text{distinguishes}_A(w, p, q))$$

So there exist p, q such that one of $\delta^*(p, w)$, $\delta^*(q, w)$ is not in F .

So in particular there exists a symbol σ such that $(\delta(p, \sigma), \delta(q, \sigma)) \in m(n - 1)$. Per induction there exist states $(p', q') = (\delta(p, \sigma), \delta(q, \sigma))$ in $m(n - 1)$ and a word w' , $|w'| = n - 1$ such that one of $\delta^*(p', w')$, $\delta^*(q', w')$ is not in F . Thus one of $\delta^*(p, \sigma w')$, $\delta^*(q, \sigma w')$ is not in F , which makes $\sigma w'$ a distinguishing word of length n for p, q . \square

Lemma 2. *If COMEQUIVPAIRS has done $D(A)$ iterations and terminated, then the longest distinguishing word w of a A has length $D(A) - 1$.*

Proof. Via direct proof. Assume $m\text{-COMEQUIVPAIRS}(A)$ has done n iterations (so $D(A) = n$). We then know, that

1. $\forall i \in [0, n - 1]: m(i) \neq \emptyset$
2. $\forall i > n - 1: m(i) = \emptyset$ **Gregor: fix this**

$m\text{-COMEQUIVPAIRS}(A)$ terminates iff $m(i) = \emptyset$. If the first point would not hold, then the algorithm would have stopped before.

Since the algorithm did n iterations, the internal variable i must be n at the end of the last iteration. The terminating condition is $m(i) \neq \emptyset$; thus follows the second point.

We prove that then there exists a distinguishing word of length $n - 1$, but the existence of a longer word leads to a contradiction. Recall lemma 1:

$$(p, q) \in m(n) \iff \exists w \in \Sigma^*: (|w| = n \wedge \text{distinguishes}_A(w, p, q))$$

Following this lemma and point 1, we can deduce that there exists at least one distinguishing word w with $|w| = n - 1 = D(A) - 1$ for some $p, q \in Q$.

There cannot be any distinguishing word w' with $|w'| = k > n - 1$ for any two states $p', q' \in Q$ fulfilling this property. Following the lemma again, $m(k), k > n - 1$ would be non-empty, which is contradictory to point 2. \square

Lemma 3.

$$\text{distinguishes}_A(w, p, q) \wedge q \sim_A q' \Rightarrow \text{distinguishes}_A(w, p, q')$$

Proof. Via direct proof.

$$\begin{aligned} & \text{distinguishes}_A(w, p, q) \wedge q \sim_A q' \\ \Rightarrow & (\delta^*(p, w) \in F \Leftrightarrow \delta^*(q, w) \notin F) \wedge (\forall z \in \Sigma^*: (\delta^*(q, z) \in F \Leftrightarrow \delta^*(q', z) \in F)) \\ \Rightarrow & (\delta^*(p, w) \in F \Leftrightarrow \delta^*(q, w) \notin F) \wedge (\delta^*(q, w) \in F \Leftrightarrow \delta^*(q', w) \in F) \\ \Rightarrow & (\delta^*(p, w) \in F \Leftrightarrow \delta^*(q', w) \notin F) \\ \Rightarrow & \text{distinguishes}_A(w, p, q') \end{aligned}$$

\square

Theorem 3. *Creating equivalent state pairs in a minimal DFA A does not increase the number of iterations when the COMEQUIVPAIRS-algorithm is applied on it.*

Proof. Per contradiction. Let us assume there were n states r_0, \dots, r_n added to a given minimal DFA $A = (Q, \Sigma, \delta, s, F)$ resulting in a DFA $A' = (Q', \Sigma, \delta', s, F')$ such that:

- $Q' = Q \cup \{r_1, \dots, r_n\}$
- $\forall i \in [1, n]: \exists q \in Q: r_i \sim_{A'} q$
- $D(A) < D(A')$

According to lemma 2 the longest distinguishing word w of A has length $|w| = D(A) - 1$, while its counterpart w' in A' has length $D(A') - 1$. Consequently $|w| < |w'|$.

There exist states $p', q' \in Q'$ such that w' distinguishes p', q' . We differentiate between three cases regarding belonging of p', q' .

Both p', q' in Q : Since the longest distinguishing word any state pair in Q can have is guaranteed shorter than w' , we may conclude that $p', q' \in Q$ and p', q' have w' as distinguishing word is a contradiction. \nmid

One of p', q' in Q , one in $Q' \setminus Q$: W.l.o.g. $p' \in Q, q' \in Q' \setminus Q$. We then know, that $q' \in \{r_1, \dots, r_n\}$. This implies that $q' \sim_{A'} q$ for an $q \in Q$.

By lemma 3 w' is a distinguishing word for p', q . But this is contradictory to $p', q \in Q$. \nmid

None of p', q' in Q : Since p', q' both have to be in $Q' \setminus Q = \{r_1, \dots, r_n\}$, we can find states $p, q \in Q$ equivalent to p', q' . These states would be distinguished by w' , which is contradictory again. \nmid \square

3.2 Adding unreachable states

From step 1 of the minimization algorithm we can deduce how to add unreachable states. These can easily be added to a DFA by adding non-start states with no ingoing transitions (see def. 1). Number and nature of outgoing transitions may be arbitrary.

```
1: function ADDUNREACHABLESTATES ( $A, \mathcal{Q}_{unr}, c$ )
2:   for  $\mathcal{Q}_{unr}$  times do
3:      $q \leftarrow \max Q + 1$ 
4:      $Q \leftarrow Q \cup \{q\}$ 
5:      $outSymbols \leftarrow c = 1 ? \Sigma : \text{random subset of } \Sigma$ 
6:      $R \leftarrow \text{random chosen sample of } |outSymbols| \text{ states from } Q \setminus \{q\}$ 
7:     for  $\sigma$  in  $outSymbols$  do
8:        $q' \in R$ 
9:        $R \leftarrow R \setminus \{q'\}$ 
10:       $\delta \leftarrow \delta \cup \{(q, \sigma), q'\}$ 
11:   return  $A$ 
```

Chapter 4

Conclusion

What happens, if we change start and accepting states?

What happens, if we add transitions only?

dfa specific planarity test?

use planarity test information for better drawing?

Appendix A

Bibliography

- [1] André Almeida, Marco Almeida, José Alves, Nelma Moreira, and Rogério Reis. Fado and guitar. volume 5642, pages 65–74, 07 2009. doi:10.1007/978-3-642-02979-0_10.
- [2] Marco Almeida, Nelma Moreira, and Rogério Reis. Aspects of enumeration and generation with a string automata representation. *Theoretical Computer Science*, 387:93–102, 06 2009. doi:10.1016/j.tcs.2007.07.029.
- [3] Frederique Bassino and Andrea Sportiello. Linear-time generation of specifiable combinatorial structures: general theory and first examples, 2013. arXiv:1307.1728.
- [4] Frédérique Bassino, Julien David, and Cyril Nicaud. Regal: A library to randomly and exhaustively generate automata. 07 2007. doi:10.1007/978-3-540-76336-9_28.
- [5] Frédérique Bassino, Julien David, and Andrea Sportiello. Asymptotic enumeration of minimal automata. *Leibniz International Proceedings in Informatics, LIPIcs*, 14, 09 2011. doi:10.4230/LIPIcs.STACS.2012.88.
- [6] Frédérique Bassino and Cyril Nicaud. Enumeration and random generation of accessible automata. *Theoretical Computer Science*, 381:86–104, 08 2007. doi:10.1016/j.tcs.2007.04.001.
- [7] Daniel Berend and Aryeh Kontorovich. The state complexity of random dfas. *Theoretical Computer Science*, 652, 07 2013. doi:10.1016/j.tcs.2016.09.012.
- [8] Arnaud Carayol and Cyril Nicaud. Distribution of the number of accessible states in a random deterministic automaton. volume 14, pages 194–205, 02 2012. doi:10.4230/LIPIcs.STACS.2012.194.
- [9] Jean-Marc Champarnaud and Thomas Paranthoën. Random generation of dfas. *Theoretical Computer Science*, 330:221–235, 02 2005. doi:10.1016/j.tcs.2004.03.072.
- [10] Michael Domaratzki, Derek Kisman, and Jeffrey Shallit. On the number of distinct languages accepted by finite automata with n states. *J. Autom. Lang. Comb.*, 7(4):469–486, September 2002.
- [11] Philippe Duchon, Philippe Flajolet, Guy Louchard, and Gilles Schaeffer. Boltzmann samplers for the random generation of combinatorial structures. *Comb. Probab. Comput.*, 13(4–5):577–625, July 2004. URL: <https://doi.org/10.1017/S0963548304006315>, doi:10.1017/S0963548304006315.

- [12] Gesellschaft für Informatik. Empfehlungen für bachelor- und masterprogramme im studienfach informatik an hochschulen (juli 2016), 2016. Accessed: 2020-02-12. URL: https://dl.gi.de/bitstream/handle/20.500.12116/2351/58-GI-Empfehlungen_Bachelor-Master-Informatik2016.pdf.
- [13] John Hopcroft and Robert Tarjan. Efficient planarity testing. *J. ACM*, 21(4):549–568, October 1974.
- [14] John E. Hopcroft. An $n \log n$ algorithm for minimizing states in a finite automaton. Technical report, Stanford, CA, USA, 1971.
- [15] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to automata theory, languages, and computation - (2. ed.)*. Addison-Wesley series in computer science. Addison-Wesley-Longman, 2001.
- [16] Pierre-Cyrille Héam and Jean-Luc Joly. On the uniform random generation of deterministic partially ordered automata using monte carlo techniques. 12 2014.
- [17] Johanna Högberg and Lars Larsson. Dfa minimisation using the myhill-nerode theorem. Accessed: 2020-02-09. URL: <http://www8.cs.umu.se/kurser/TDBC92/VT06/final/1.pdf>.
- [18] William Kocay. The hopcroft-tarjan planarity algorithm. October 1993.
- [19] A. Korshunov. Enumeration of finite automata. *Problemy Kibernetiki*, page 34:5–82, 1959. In russian.
- [20] C. Nicaud. *Étude du comportement en moyenne des automates finis et des langages rationnels*. PhD thesis, Université Paris 7, 2000.
- [21] Cyril Nicaud. Random deterministic automata. pages 5–23, 08 2014. doi:10.1007/978-3-662-44522-8_2.
- [22] Albert Nijenhuis and Herbert S. Wilf. *Combinatorial Algorithms*. Academic Press, 1978.
- [23] Rogério Reis, Nelma Moreira, and Marco Almeida. On the representation of finite automata. *7th International Workshop on Descriptive Complexity of Formal Systems, DCFS 2005 - Proceedings*, 06 2005.
- [24] Uwe Schöning. *Theoretische Informatik - kurzgefasst*. Spektrum, Akad. Verl, Heidelberg Berlin, 2001.
- [25] V. Vyssotsky. A counting problem for finite automata. Technical report, 1959. Bell Telephon Laboratories.
- [26] Thomas Schwentick Wim Martens. Theoretische informatik i ss18. Lecture notes, 2018.

Appendix B

Erklärung

Hiermit versichere ich, Gregor Hans Christian Sönnichsen, dass ich die vorliegende Arbeit selbständig verfasst habe, keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt habe und die Arbeit nicht bereits zur Erlangung eines akademischen Grades eingereicht habe.

Bayreuth, den 8. Februar 2020.

Gregor Hans Christian Sönnichsen