Asymptotic Approximation in Formal Languages

17th Workshop Computational Logic and Applications
@the Jagiellonian University in Kraków

Akita University

Ryoma Sin'ya (Akita University), 2023 Dec 14.

CIAA 2024 in Akita (Septermber 3 to 6)

28th International Conference on Implementation and Application of Automata

- Topics: algorithms on automata, automata and logic, bioinformatics, complexity of automata operations, compilers, computer-aided verification, concurrency, data structure design for automata, data and image compression, design and architecture of automata software, digital libraries, DNA/molecular/membrane computing, document engineering, editors, environments, experimental studies and practical experience, implementation of verification methods and model checking, industrial applications, natural language and speech processing, networking, new algorithms for manipulating automata, object-oriented modeling, pattern-matching, pushdown automata and context-free grammars, quantum computing, structured and semi-structured documents, symbolic manipulation environments for automata, transducers and multi-tape automata, techniques for graphical display of automata, VLSI, viruses and related phenomena, and the world-wide web.

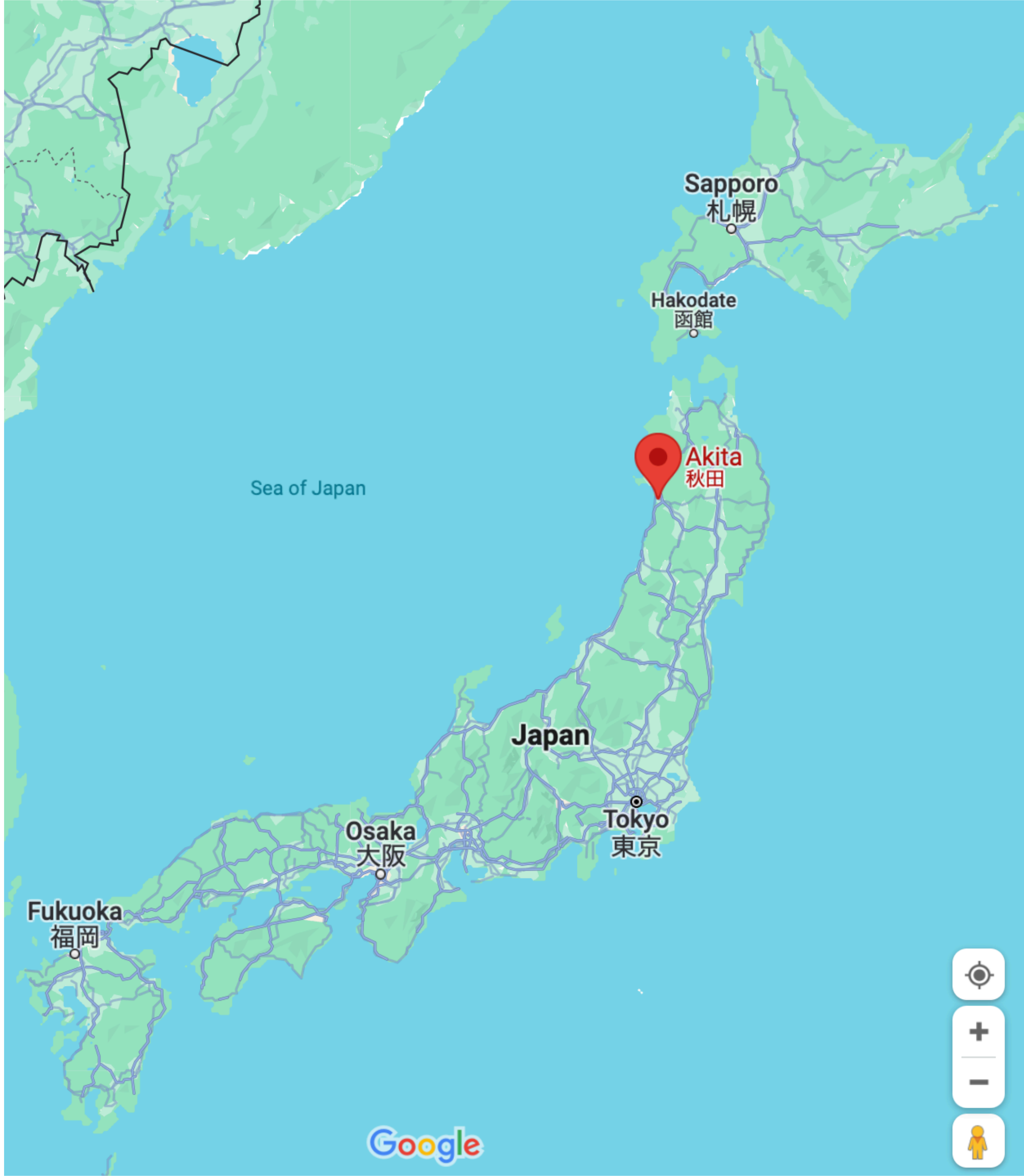
- The proceedings will be published by in the series of LNCS
  - Tentative submission deadline: April 14, 2024

About Akita

- In Tohoku (northern Japan) region
  - The latitude of Akita City is 39.72°N

- Reachable from Tokyo, Nagoya, Osaka:
  - By plane (1~1.5 hours)
  - By bullet train (3.5~6 hours)

- Lot of snow in winter:
  There is a ski resort **15 minutes by car** from Akita university



Snowboarding in Akita (Febrary 16, 2023)

Outline

1. Background I: density and measurability (7 min.)
2. Background II: known properties (5 min.)
3. Complexity results (8 min.)
4. Conclusion (3 min.)

Outline

1. Background I: density and measurability (7 min.)
2. Background II: known properties (5 min.)
3. Complexity results (8 min.)
4. Conclusion (3 min.)

Density of formal languages

The density of a language L over A is defined as

$$\delta_A(L) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \frac{\#(L \cap A^i)}{\#(A^i)}.$$

Here $\#(X)$ denotes the cardinality of X.

$\delta_A(L)$ can be regarded as the (average) **probability** that a randomly chosen word is in L.

Density of formal languages

The density of a language L over A is defined as

$$\delta_A(L) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \frac{\#(L \cap A^i)}{\#(A^i)}.$$

Example 1: $\delta_A((AA)^*) = \dfrac{1}{2}$.

Example 2: $\delta_A(A^*wA^*) = 1$ for any w.

Example 3: $L_\perp = \{w \in A^* \mid 3^n \leq |w| < 3^{n+1}$ for some even $n\}$ does **not** have a density.

Theorem [Berstel 1973]:
Every regular language *do* have a *rational density*, and It is computable.
Every unambiguous context-free language do have an *algebraic density*.

Theorem [Kemp 1980]:
There is a context-free language *whose density is transcendental*.

Density of formal languages

The density of a language L over A is defined as

$$\delta_A(L) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \frac{\#(L \cap A^i)}{\#(A^i)}.$$

Example 1: $\delta_A((AA)^*) = \dfrac{1}{2}$.

Example 2: $\delta_A(A^*wA^*) = 1$ for any w.

Example 3: $L_\perp = \{w \in A^* \mid 3^n \leq |w| < 3^{n+1}$ for some even $n\}$
does **not** have a density.

Problem:

Does every context-free language L have a density? ($\delta_A(L)$ converges?)

Theorem [Nakamura 2019]:

it is undecidable whether a given context-free language L satisfies $\delta_A(L) = 1$.

Density of formal languages

The density of a language L over A is defined as

$$\delta_A(L) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \frac{\#(L \cap A^i)}{\#(A^i)}.$$

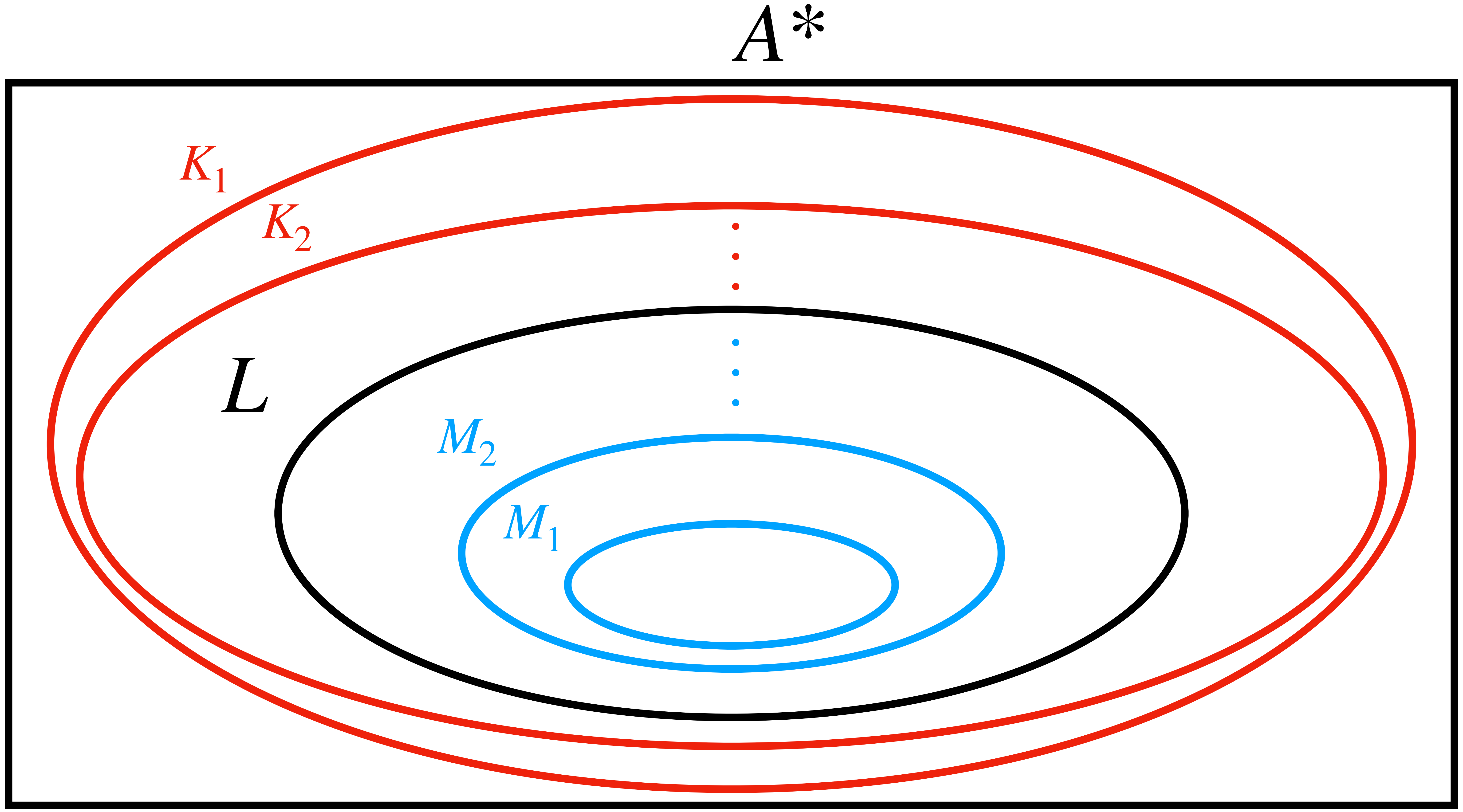
Example 1: $\delta_A((AA)^*) = \dfrac{1}{2}$.

Example 2: $\delta_A(A^*wA^*) = 1$ for any w.

Example 3: $L_\perp = \{w \in A^* \mid 3^n \leq |w| < 3^{n+1}$ for some even $n\}$
does **not** have a density.

Theorem [Kozik, CLA'05]: it is undecidable whether $\displaystyle\lim_{n\to\infty} \frac{\#(L \cap A^n)}{\#(A^n)}$ converges or not for a given context-free language L.

$\mathscr{C}$-measurability [S., SOFSEM'21] (cf. [Buck, 1946])



L is said to be $\mathscr{C}$**-measurable** if there exists an *infinite sequence of pairs of languages* $(M_n, K_n)_{n \in \mathbb{N}}$ in $\mathscr{C}$ such that $M_n \subseteq L \subseteq K_n$ and $\lim_{n \to \infty} \delta_A(K_n \setminus M_n) = 0$.

Example of a regular measurable language

Theorem:

B $= \{w \in A \mid |w|_a = |w|_b\}$ over $A = \{a, b\}$ is regular measurable.

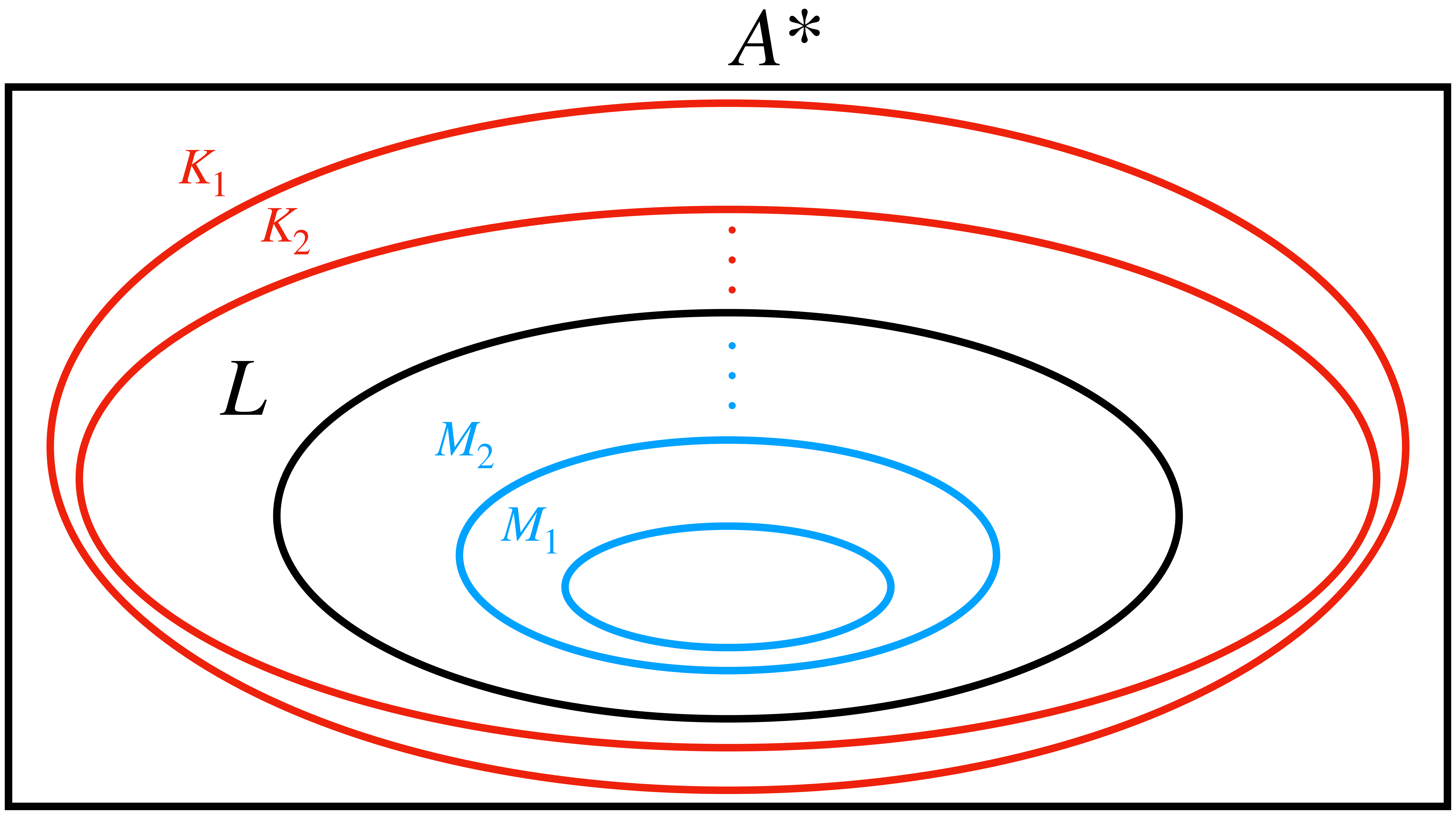
the # of occurrences of a in w

Proof: Let $L_k = \{w \in A^* \mid |w|_a = |w|_b \mod k\}$ for each $k \geq 1$.

Then, for each $k \geq 1$, B $\subseteq L_k$ and $\delta_A(L_k) = \dfrac{1}{k} \to 0$ (if $k \to \infty$) .

Thus the infinite sequence $(\emptyset, L_k)_{k \geq 1}$ converges to B.

$\mathscr{C}$-measurability [S., SOFSEM'21] (cf. [Buck, 1946])



L is said to be $\mathscr{C}$**-measurable** if there exists an *infinite sequence of pairs of languages* $(M_n, K_n)_{n\in\mathbb{N}}$ in $\mathscr{C}$ such that $M_n \subseteq L \subseteq K_n$ and $\lim_{n\to\infty} \delta_A(K_n \setminus M_n) = 0.$

Original motivation of mesurability

- A non-empty word w is said to be ***primitive*** if it can not be represented as a power of shorter words, i.e., $w = u^n \Rightarrow u = w$ (and $n = 1$).
  Q denotes the set of all primitive words over $\{a, b\}$.

Example : $ababa \in \mathrm{Q}$ $\quad ababab = (ab)^3 \notin \mathrm{Q}$

Primitive worsd conjecture [Dömösi-Horvath-Ito 1991]:
Q is *not* context-free.

My naive idea: while every context-free language is
regular measurable, Q is regular **im**measurable.

Summary of [S., SOFSEM'21]

Regular measurable languages

A (simple) deterministic CFL

M

$= \{w \in \{a, b\}^* \mid |w|_a > |w|_b\}$

Q

The set of all primitive words

B $= \{w \in A \mid |w|_a = |w|_b\}$

Many complex context-free languages.

There are **uncountably many** regular measurable languages.

Outline

1. Background I: density and measurability (7 min.)
2. Background II: known properties (5 min.)
3. Complexity results (8 min.)
4. Conclusion (3 min.)

Some properties of $\mathscr{C}$-measurability [S. DLT'21]

Notation: $\mathscr{M}_A(\mathscr{C}) = \{L \subseteq A^* \mid L \text{ is } \mathscr{C}\text{-measurable}\}$

- $\mathscr{M}_A(\mathscr{C})$ can be defined as the *Carathéodory extension* of $\mathscr{C}$, a standard notion from measure theory.

- $\mathscr{M}_A(\mathscr{C})$ is closed under Boolean operations and left-and-right quotients if $\mathscr{C}$ is closed under these operations.

- "Is a given CFG generates a regular measurable languages?" is **undecidable**.

Some properties of $\mathscr{C}$-measurability [S. DLT'21]

Notation: $\mathscr{M}_A(\mathscr{C}) = \{L \subseteq A^* \mid L \text{ is } \mathscr{C}\text{-measurable}\}$

Q: How about the decidability of $\mathscr{C}$-measurability for some **subclass $\mathscr{C}$ of regular languages?**

- "Is a given CFG generates a regular measurable languages?" is **undecidable**.

Decidability

- PT-measurability for DFAs is **decidable in linear time** [SYN 2022], where PT is the class of all *piecewise testable* languages.

  Definition: L is ***piecewise testable*** [Simon 1972] if it can be represented as a finite Boolean combination of languages of the form $L_w = A^*a_1A^*a_2\ldots A^*a_nA^*$ where $w = a_1a_2\cdots a_n$.

Decidability

- PT-measurability for DFAs is **decidable in linear time** [SYN 2022], where PT is the class of all *piecewise testable* languages.
- AT-measurability for DFAs is **coNP-complete** [SYN 2022], where AT is the class of all *alphabet testable* languages.

Definition: L is ***alphabet testable*** if it can be represented as a finite Boolean combination of languages of the form A^*aA^* (where $a \in A$) .

Decidability

- PT-measurability for DFAs is **decidable in linear time** [SYN 2022], where PT is the class of all *piecewise testable* languages.
- AT-measurability for DFAs is **coNP-complete** [SYN 2022], where AT is the class of all *alphabet testable* languages.
- The decidability of GD-measurability for DFAs is **decidable** [S. CIAA'23], where GD is the class of all *generalised definite* languages.

  Definition: L is ***generalised defintie*** if it can be represented as a finite Boolean combination of languages of the form uA^*, A^*v .

Decidability

Notation: $\mathscr{M}_A(\mathscr{C}) = \{L \subseteq A^* \mid L \text{ is } \mathscr{C}\text{-measurable}\}$

- PT-measurability for DFAs is **decidable in linear time** [SYN 2022], where PT is the class of all *piecewise testable* languages.
- AT-measurability for DFAs is **coNP-complete** [SYN 2022], where AT is the class of all *alphabet testable* languages.
- The decidability of GD-measurability for DFAs is **decidable** [S. CIAA'23], where GD is the class of all *generalised definite* languages.
- Hierarchy is strict [S. DLT'22]: $\mathscr{M}_A(\mathrm{AT}) \subsetneq \mathscr{M}_A(\mathrm{PT}) \subsetneq \mathscr{M}_A(\mathrm{GD})$.

Outline

Characterisation of AT- and PT-meaurability

Theorem [S. DLT'22]:

A language L is AT-measurable if and only if L or $\overline{L}$ contains the language $\bigcap_{a \in A} A^*aA^*$.

A language L is PT-measurable if and only if L or $\overline{L}$ contains the language L_w for some $w \in A^*$, where L_w is the set of all words containing w as a subsequence (scattered subword).

Corollary:

If a given language L is regular, then AT-measurability is decidable (because the inclusion problem is decidable for regular languages).

Characterisation of AT- and PT-meaurability

Theorem [S. DLT'22]:

A language L is AT-measurable if and only if L or $\overline{L}$ contains the language $\bigcap_{a \in A} A^*aA^*$.

A language L is PT-measurable if and only if L or $\overline{L}$ contains the language L_w for some $w \in A^*$, where L_w is the set of all words containing w as a subsequence (scattered subword).

Theorem [SYN 2022]:

The AT-measurability for DFAs is **coNP-complete**, while the PT-measurability for DFAs is decidable in **linear time**.

Definite languages [Brzozowski 1962] [Ginzburg 1966]

Notation: $\mathscr{B}(\mathscr{C})$ denotes the (finite) Boolean closure of $\mathscr{C}$.

Definite:

$$\mathrm{D} = \mathscr{B}\{A^*w \mid w \in A^*\}$$

Reverse definite:

$$\mathrm{RD} = \mathscr{B}\{wA^* \mid w \in A^*\}$$

Generalised definite:

$$\mathrm{GD} = \mathscr{B}\{uA^*v \mid u, v \in A^*\}$$

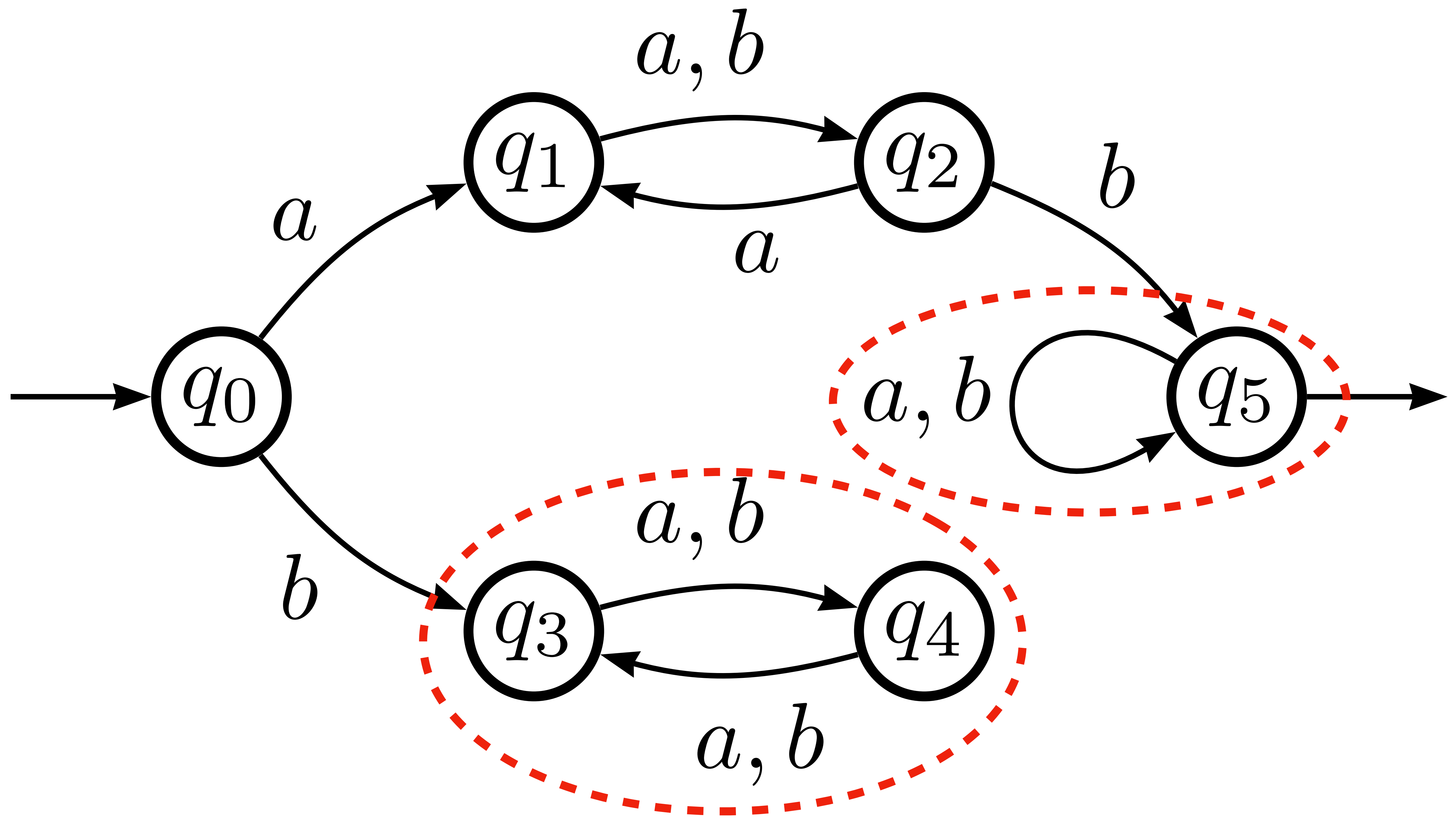
Sink component (a.k.a bottom strongly connected component)

Definition: Let $\mathscr{A} = (Q, \cdot, q_0, F)$ be a deterministic automaton.

A subset $S \subseteq Q$ is called sink if it satisfies following:

(1) S is strongly connected: $\forall p, q \in S \, \exists w \in A^* \, p \cdot w = q$

(2) S has no outgoing transition: $\forall p \in S \, \forall w \in A^* \, p \cdot w \in S$.



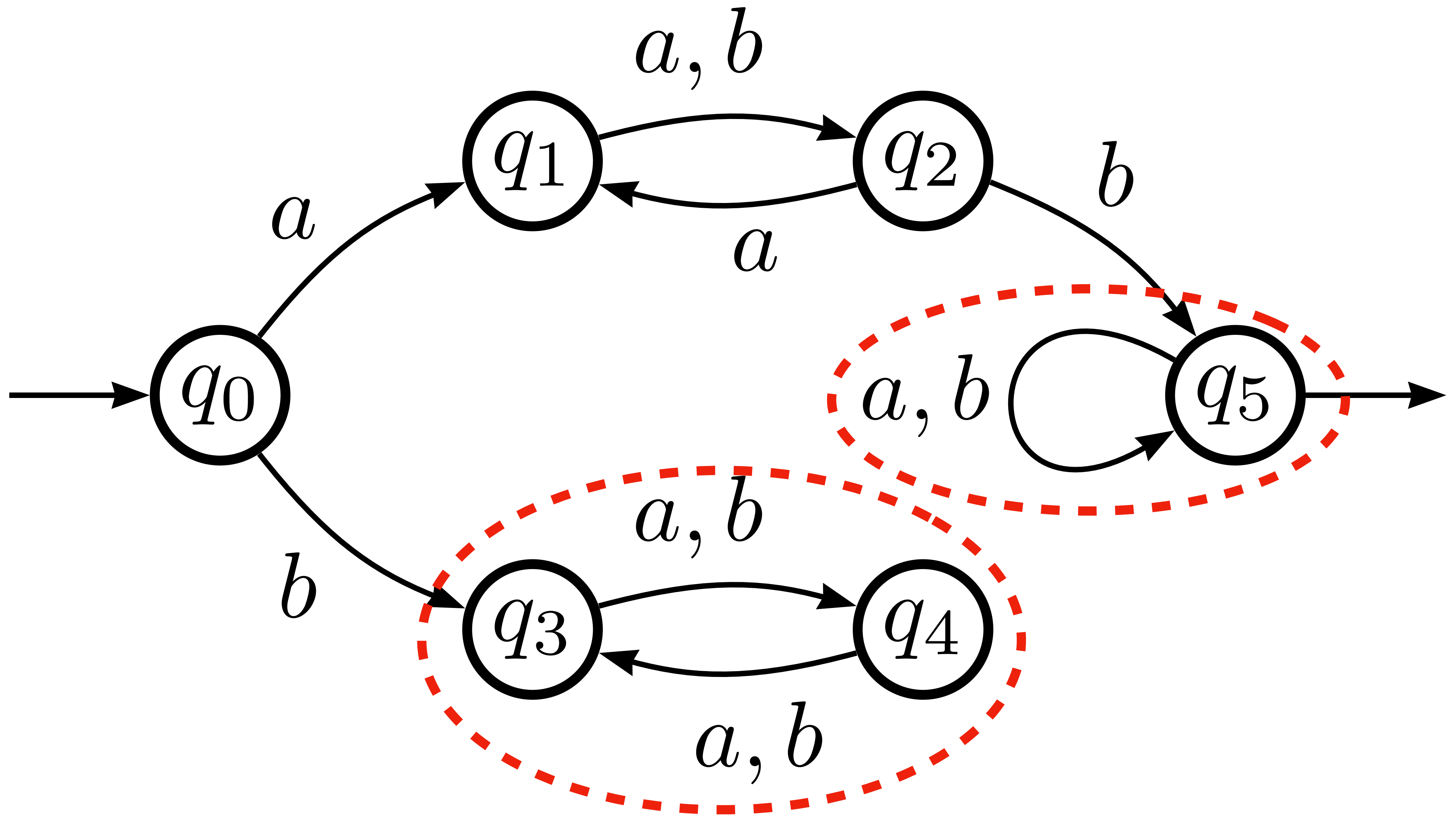
Sink components can be considered as a **minimal** SCC with respect to the reachability relation.

Characterisation of RD-measurable REGs

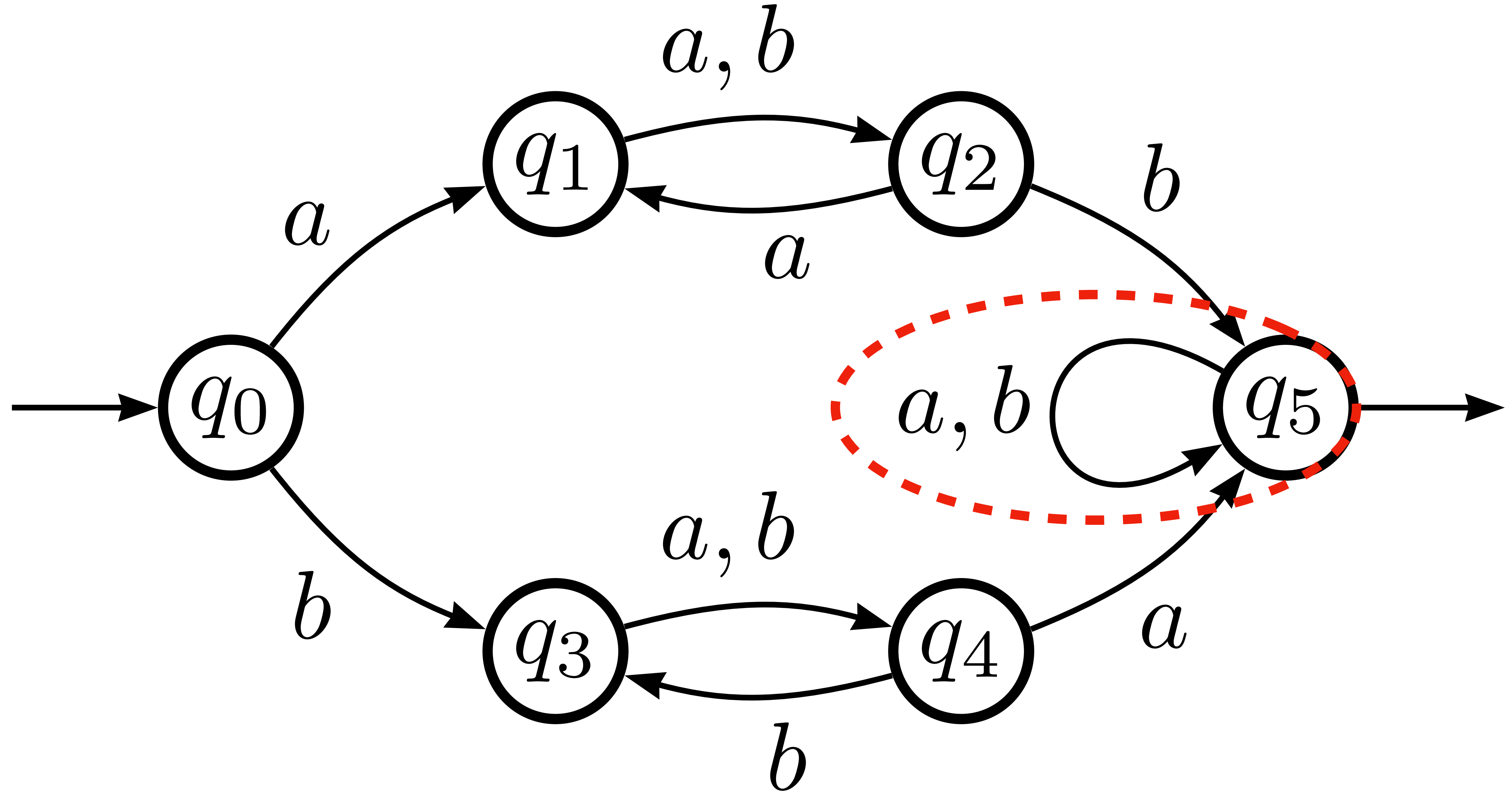
Theorem: Let $\mathscr{A} = (Q, \cdot, q_0, F)$ be a minimal deterministic automaton. TFAE:

(1) $L(\mathscr{A})$ is RD-measurable.

(2) Every sink component of $\mathscr{A}$ is a singleton (contains only one state).



RD-**im**measurable



RD-measurable

Characterisation of RD-measurable REGs

Theorem: Let $\mathscr{A} = (Q, \cdot\,, q_0, F)$ be a minimal deterministic automaton. TFAE:

(1) $L(\mathscr{A})$ is RD-measurable.

(2) Every sink component of $\mathscr{A}$ is a singleton (contains only one state).

Corollary: RD-measurability for minimal DFAs are decidable in **linear time**.

Corollary: D-measurability for DFAs are decidable.

Characterisation of GD-measurable REGs

Theorem: Let $\mathscr{A} = (Q, \cdot, q_0, F)$ be a deterministic automaton, and $Q_1, \ldots, Q_k$ be its all sink components.

Define $P_i = \{w \in A^* \mid q_0 \cdot w \in Q_i\}$,

$S_i = \{w \in A^* \mid Q_i \cdot w \subseteq F\}$ and $S_i' = \{w \in A^* \mid Q_i \cdot w \cap F = \varnothing\}$.

Let $$M = \bigcup_{i=1}^{k} P_i S_i \quad \text{and} \quad M' = \bigcup_{i=1}^{k} P_i S_i'.$$

Then $L(\mathscr{A})$ is GD-measurable if and only if $\delta_A(M) + \delta_A(M') = 1$.

Characterisation of GD-measurable REGs

Intuition: M is a largest subset of $L(\mathscr{A})$ that can be represented as a (possibly infinite) union of languages of the form uA^*w.

That is, M is a largest GD-measurable subset of $L(\mathscr{A})$.

Also, M' is a largest GD-measurable subset of $\overline{L(\mathscr{A})}$.

Let $M = \bigcup_{i=1}^{k} P_i S_i$ and $M' = \bigcup_{i=1}^{k} P_i S_i'$.

Then $L(\mathscr{A})$ is GD-measurable if and only if $\delta_A(M) + \delta_A(M') = 1$.

This condition means $\delta_A(L(\mathscr{A})) = \delta_A(M)$ and $\delta_A(\overline{L(\mathscr{A})}) = \delta_A(M')$.

Characterisation of GD-measurable REGs

Theorem: Let $\mathscr{A} = (Q, \cdot, q_0, F)$ be a deterministic automaton, and $Q_1, \ldots, Q_k$ be its all sink components.
Define $P_i = \{w \in A^* \mid q_0 \cdot w \in Q_i\}$,
$S_i = \{w \in A^* \mid Q_i \cdot w \subseteq F\}$ and $S_i' = \{w \in A^* \mid Q_i \cdot w \cap F = \varnothing\}$.

Let $M = \bigcup_{i=1}^{k} P_i S_i$ and $M' = \bigcup_{i=1}^{k} P_i S_i'$.

Then $L(\mathscr{A})$ is GD-measurable if and only if $\delta_A(M) + \delta_A(M') = 1$.

Corollary: GD-measurability for DFAs is **decidable** (in PSPACE).
(because the density of a regular language is computable and M, M' are regular by the construction)

Outline

1. Background I: density and measurability (7 min.)
2. Background II: known properties (5 min.)
3. Complexity results (8 min.)
4. Conclusion (3 min.)

Summary

- PT-measurability and RD-measurability for DFAs is decidable in linear time, while AT-measurability for DFA is coNP-complete.

- GD-measurability is decidable for DFAs (in PSPACE).

Progress: we (S., Y. Nakamura and Y. Yamaguchi) found that it is in PTIME.

Open problem

Is the measuring power of GD and SF (the class of all **star-free** languages) equivalent or not?

Definition: L is ***star-free*** if and only if it can be represented as a finite Boolean combination and concatenation finite languages.

[S. DLT'22]: $\mathscr{M}_A(\mathrm{AT}) \subsetneq \mathscr{M}_A(\mathrm{PT}) \subsetneq \underline{\mathscr{M}_A(\mathrm{GD}) \subseteq \mathscr{M}_A(\mathrm{SF})}$.

Is this inclusion strict?

How much GD-measurability and SF are weaker than regular measurability?

Application?

The decidable characterisation of GD-measurability gives us the following **approximation scheme**:

Input : an automaton $\mathscr{A}$ and an admissible error ratio $\epsilon > 0$.
Output: an automaton $\mathscr{B}$ (if exists) such that
(1) $L(\mathscr{A}) \subseteq L(\mathscr{B})$, (2) $L(\mathscr{B})$ is generalised definite,
and (3) $|\delta_A(L(\mathscr{A})) - \delta_A(L(\mathscr{A}))| \leq \epsilon$.

Can we apply this scheme to, say,
obtain an efficient regular expression matching algorithm?
(or other decision problems, e.g., inclusion ?)

(Akita-Inu)