Topoi of automata

Ryuya Hora

Graduate School of Mathematical Sciences, University of Tokyo/ National Institute of Informatics

11th March, CSCAT2025

This talk is (mainly) based on
Topoi of automata I: Four topoi of automata and regular languages.

Table of Contents

Table of Contents

Naive question: Can we imitate Grothendieck?

- Grothendieck defined **topoi** (sites, and schemes) to introduce continuous/infinite geometric methods into number theory, which is traditionally thought of as a finite/discrete subject.

Naive question: Can we imitate Grothendieck?

- Grothendieck defined **topoi** (sites, and schemes) to introduce continuous/infinite geometric methods into number theory, which is traditionally thought of as a finite/discrete subject.
- The theory of automata and regular languages is also finite/discrete combinatorial theory. Is it possible to introduce continuous/infinite geometric methods into this setting?

Naive question: Can we imitate Grothendieck?

- Grothendieck defined **topoi** (sites, and schemes) to introduce continuous/infinite geometric methods into number theory, which is traditionally thought of as a finite/discrete subject.
- The theory of automata and regular languages is also finite/discrete combinatorial theory. Is it possible to introduce continuous/infinite geometric methods into this setting?

Question (Naive question)

Can we construct a space (= topos) that fully encodes the information of regular languages and study its topological properties and invariants?

Table of Contents

Notations and terminologies

- We fix a finite set Σ, which we will call an **alphabet**.
- Σ^* denotes the (free) monoid of finite words.
- A **language** is a subset of Σ^*.

Σ-set

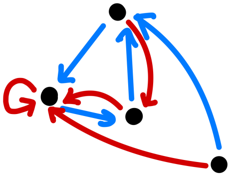
Definition (Σ-set)

A **Σ-set** is a pair (Q, δ) of

- a set Q (of states) and
- a (transition) function $\delta\colon Q \times \Sigma \to Q$.

Σ-set = right Σ^*-action = presheaf over the one-object category Σ^*.

$\Sigma = \{a, b\}$



: Σ-set

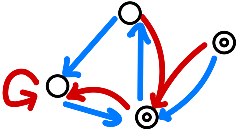
Automaton

Definition (Automaton)

A (Σ-)**automaton** is a triple (Q, δ, F) of

- a set Q (of states),
- a (transition) function $\delta \colon Q \times \Sigma \to Q$, and
- a subset $F \subset Q$ of (final states).

$\Sigma = \{a, b\}$



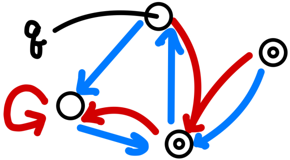
: automaton

Language recognition

Definition (Language recognition)

The language **recognised** by an automaton (Q, δ, F) and a state $q \in Q$ is

$$\{w \in \Sigma^* \mid qw \in F\} \subset \Sigma^*.$$



recognises $\{b, aa, aba, bba, aaab, \ldots\}$

$\subseteq \Sigma^*$

Table of Contents

Formal definition of topos, which we won't use.

A topos is a category satisfying a very strong condition so that we can do geometry.

Definition ((Grothendieck) topos)

A **topos** $\mathcal{E}$ is a category $\mathcal{E}$ such that

- there exsists a small category $\mathcal{C}$, and
- a fully faithful functor $\iota \colon \mathcal{E} \hookrightarrow \mathbf{PSh}(\mathcal{C})$ with a left exact left adjoint.

There are two things to emphasise:

1. A topos is just a category (without any additional structures).
2. Topoi are rare!
   ($\mathbf{Vect}$, $\mathbf{Top}$, $\mathbf{Group}$, $\mathbf{Ab}$, $\mathbf{Rings}$, $\mathbf{Mnfd}$, $\mathbf{Monoids}$, $\cdots$ are not topos.)

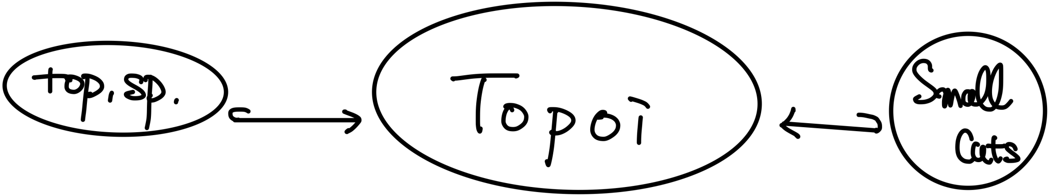
Typical Examples of topoi

Example (Topological example, which we won't use.)

For a topological space X, we can construct its **sheaf topos** $\mathbf{Sh}(X)$.

Example (Presheaf topos, which we will use.)

For a small category $\mathcal{C}$, its **presheaf category** $\mathbf{PSh}(\mathcal{C})$ is a topos.



What to remember

- We fix an **alphabet** Σ.
- A Σ-**set** is a right Σ^*-action $(Q, \delta\colon Q \times A \to Q)$.
- An **automaton** $(Q, \delta\colon Q \times \Sigma \to Q, F \subset Q)$ together with a(n initial) state $q \in Q$ recognises a **language** $L \subset \Sigma^*$.
- A **topos** is a very nice category (so that we can do geometry).

Table of Contents

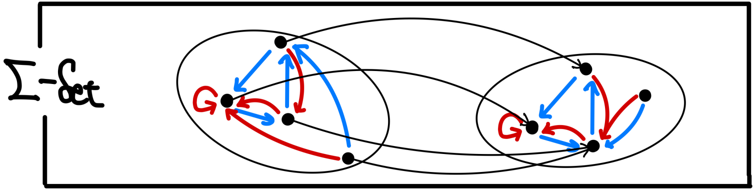
Topos of Σ-sets

Definition

Σ-**Set** is the category of Σ-sets defined by Σ-**Set** := **PSh**(Σ*).

Σ-**Set** is a (presheaf) topos.

$\Sigma = \{a, b\}$



Two objects of Σ-sets

The topos Σ-**Set** has two important objects:

- **The Σ-set of words Σ^***
  - The Σ^*-action is given by $v * w = vw$.
  - Universality: Yoneda $\Sigma\text{-}\mathbf{Set}(\Sigma^*, (Q, \delta)) \cong Q$.
- **The Σ-set of languages $\mathcal{P}(\Sigma^*)$**
  - The Σ^*-action is given by *derivative*:

$$L * w := \{v \in \Sigma^* \mid wv \in L\}.$$

  - Universality: We will see soon!

Σ-**Set** as a space

Even for this very simple topos, there remain so many "geometric" things that we can calculate! (Classifying its *quotients* solves an open problem!)

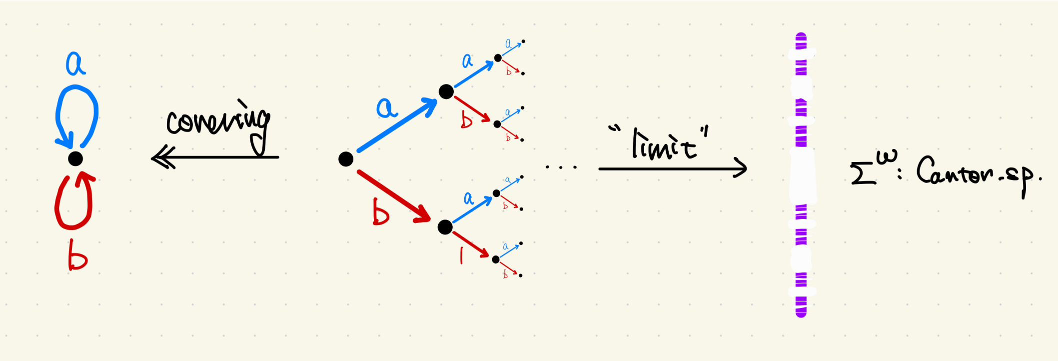


Table of Contents

Points of a topos

Definition (Points)

For a Grothendieck topos $\mathcal{E}$, its **point** p is an adjunction

$$\begin{tikzcd} \mathbf{Set} \arrow[r, shift right=2, "p_*"'] \arrow[r, phantom, "\perp"] & \mathcal{E} \arrow[l, shift right=2, "p^*\colon\ \text{lex}"'] \end{tikzcd}$$

whose left adjoint part p^* preserves finite limits.

Example (Geometric example)

For a "sober" topological space X, we have

$$\text{Points of } X \overset{1:1}{\longleftrightarrow} \text{Points of } \mathbf{Sh}(X).$$

The canonical point of Σ-**Set**

Example (The canonical point)

The forgetful-cofree adjunction defines a point of Σ-**Set**.

$$p_* : \mathbf{Set} \underset{S \mapsto S^{\Sigma^*}}{\overset{Q \leftarrow\!\shortmid (Q,\delta)}{\rightleftarrows}} \Sigma\text{-}\mathbf{Set} : p^*$$

We have $p_*(S) = S^{\Sigma^*}$. In particular, we have $p_*(\{\top, \bot\}) = \mathcal{P}(\Sigma^*)$

Lemma (Universality of the Σ-set of languages)

Σ-$\mathbf{Set}((Q, \delta), \mathcal{P}(\Sigma^*)) \cong \mathbf{Set}(Q, \{\bot\top\}) \cong \mathcal{P}(Q).$

The canonical point, recognition, and minimalization (1/2)

This adjunction (= point) subsumes the language recognition!
For a given Σ-set (Q, δ), we have the following correspondence

$$\frac{\dfrac{F \qquad \subset \qquad Q}{Q \xrightarrow{\chi_F} \{\top, \bot\} \text{ in } \mathbf{Set}}}{(Q,\delta) \xrightarrow{\chi_F^\sharp} \mathcal{P}(\Sigma^*) \text{ in } \Sigma\text{-}\mathbf{Set}}$$

The canonical point, recognition, and minimalization (2/2)

The data $(Q, \delta, F \subset Q, q \in Q)$ corresponds to

$$\Sigma^* \xrightarrow{\text{pick } q} (Q, \delta) \xrightarrow{\chi_F^\sharp} \mathcal{P}(\Sigma^*).$$

The canonical point, recognition, and minimalization (2/2)

The data $(Q, \delta, F \subset Q, q \in Q)$ corresponds to

$$\Sigma^* \xrightarrow{\text{pick } q} (Q, \delta) \xrightarrow{\chi_F^\sharp} \mathcal{P}(\Sigma^*).$$

We have

$$\Sigma^* \xrightarrow{\text{the recognised language}} \mathcal{P}(\Sigma^*),$$

The canonical point, recognition, and minimalization (2/2)

The data $(Q, \delta, F \subset Q, q \in Q)$ corresponds to

$$\Sigma^* \xrightarrow{\text{pick } q} (Q, \delta) \xrightarrow{\chi_F^\sharp} \mathcal{P}(\Sigma^*).$$

We have

$$\Sigma^* \xrightarrow{\text{the recognised language}} \mathcal{P}(\Sigma^*),$$

and

$$\Sigma^* \twoheadrightarrow (\text{min. atmt.}) \rightarrowtail \mathcal{P}(\Sigma^*).$$

(cf. Functorial point of view [Colcombet and Petrișan 2020])

The canonical point and topos of automata

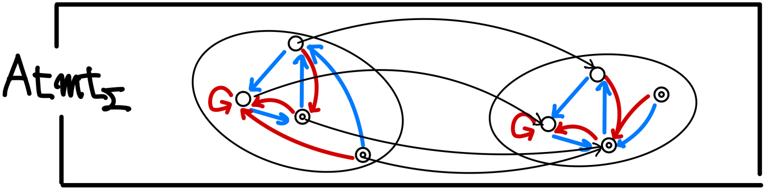
Proposition

The category of automata $\mathbf{Atmt}_\Sigma = \mathbf{Coalg}_{2\times\Sigma}$ *is a presheaf topos.*

Proof.

$\mathbf{Atmt}_\Sigma \simeq \Sigma\text{-}\mathbf{Set}/\mathcal{P}(\Sigma^*) \simeq \mathbf{PSh}\left(\int \mathcal{P}(\Sigma^*)\right)$ $\square$

$\Sigma = \{a, b\}$



Digression: Other points capture infinite words recognition!?

We can apply the same argument to other points as well.

Proposition

*Non-canonical points of Σ-***Set** *are in bijective correspondence with* **infinite words** *up to finite edit distance.*

(cf. Büchi automata)

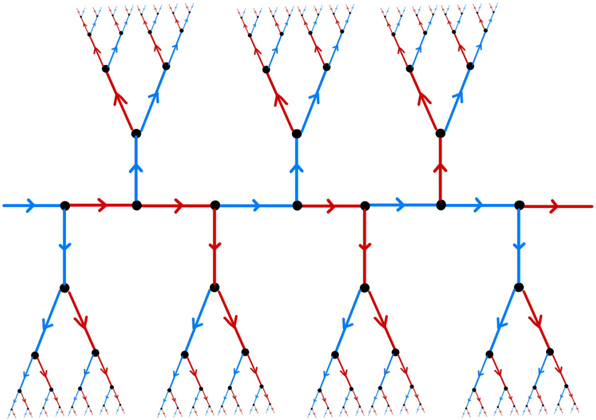


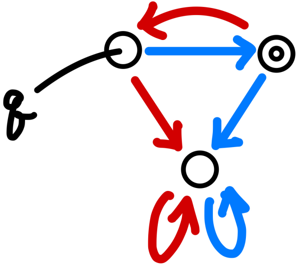
Table of Contents

Regular languages

Definition

A language $L \subset \Sigma^*$ is **regular** if L is recognised by an automaton (Q, δ, F) and a state $q \in Q$ such that $|Q| < \infty$.

- $\{a, aba, ababa, abababa, \dots\}$ is regular.
- $\{ab, aabb, aaabbb, aaaabbbb, \dots\}$ is not regular.



Can we imitate Grothendieck?

Let us return to the original "naive question."

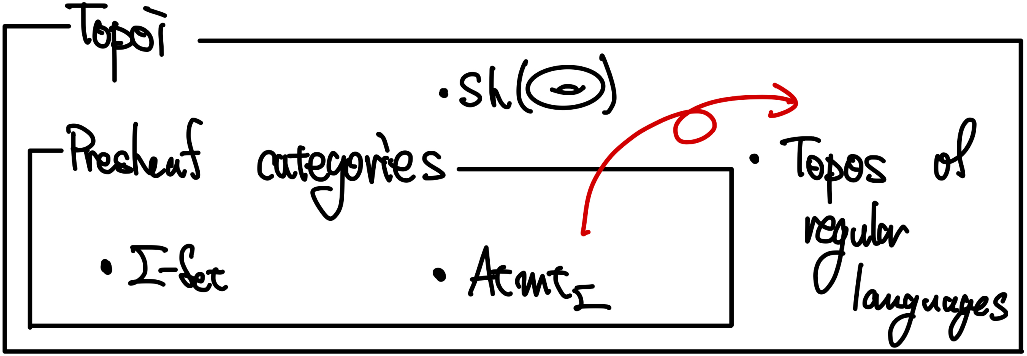
Question (Naive question)

Can we construct a space (= topos) that fully encodes the information of regular languages and study its topological properties and invariants?

We need finiteness, we need topoi!

So far,

- we did not care finiteness, and
- we only consider presheaf categories ($\subsetneq$ topoi).



Orbit-finiteness

Can we capture the **finiteness** related to regular languages?

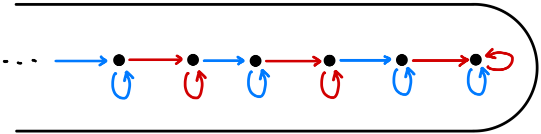
Fact

Finite Σ-sets do not form a topos.

Instead of considering finiteness, we can consider orbit-finiteness.

Definition

A Σ-set $(Q, \delta\colon Q \times \Sigma \to Q)$ is said to be **orbit-finite** if, for any state $q \in Q$, its orbit $\{qw \mid w \in \Sigma^*\}$ is a finite set.



The third topos: Σ-**Set**$_{\text{o.f.}}$

We write Σ-**Set**$_{\text{o.f.}}$ for the category of orbit-finite Σ-sets.

Theorem

*The category of orbit-finite Σ-sets Σ-**Set**$_{\text{o.f.}}$ is a Grothendieck topos.*

(Rough proof: We prove that Σ-**Set**$_{\text{o.f.}}$ is a quotient topos Σ-**Set** $\twoheadrightarrow$ Σ-**Set**$_{\text{o.f.}}$.)
This topos Σ-**Set**$_{\text{o.f.}}$ is not a presheaf topos!

The third topos: Σ-**Set**$_{\text{o.f.}}$

We write Σ-**Set**$_{\text{o.f.}}$ for the category of orbit-finite Σ-sets.

Theorem

*The category of orbit-finite Σ-sets Σ-**Set**$_{\text{o.f.}}$ is a Grothendieck topos.*

(Rough proof: We prove that Σ-**Set**$_{\text{o.f.}}$ is a quotient topos Σ-**Set** $\twoheadrightarrow$ Σ-**Set**$_{\text{o.f.}}$.)
This topos Σ-**Set**$_{\text{o.f.}}$ is not a presheaf topos!

Question

Does this deserve to be called 'the topos of regular languages'?

Let me mention two (of some) reasons.

Reason 1: This space has the information of regular languages.

Σ-**Set**$_{\mathrm{o.f.}}$ has the canonical point

$$p_{\mathrm{o.f.}} : \mathbf{Set} \to \Sigma\text{-}\mathbf{Set} \to \Sigma\text{-}\mathbf{Set}_{\mathrm{o.f.}}.$$

This induces the canonical (internal Boolean algebra) object $(p_{\mathrm{o.f.}})_*(\{\top, \bot\})$.

Proposition (Hora 2024 preprint)

The canonical (internal Boolean algebra) object $(p_{\mathrm{o.f.}})_*(\{\top, \bot\})$ coincides with the Boolean algebra of regular languages (equipped with the orbit-finite action).

(Digression: This follows from a topos-theoretic variant of Myhill-Nerode theorem.)

Reason 2: It has 4 other definitions!

Theorem (Hora 2024 preprint (and my ongoing paper))

All of the following (Boolean algebra-ed) topoi are equivalent to $(\Sigma\text{-}\mathbf{Set}_{\mathrm{o.f.}}, \mathrm{Reg}_\Sigma)$.

- $(\mathbf{Sh}(\Sigma\text{-}\mathbf{FinSet}, J), \mathrm{DFA})$: *The sheaf topos over the (Boolean algebra-ed) site of* **DFA**.
- $(\mathbf{Sh}(\Sigma\text{-}\mathbf{FinMon}, J), \mathcal{P})$: *The sheaf topos over the (Boolean algebra-ed) site of* **finite (Σ-) monoids.**
- $(\mathbf{Cont}(\widehat{\Sigma^*}), \mathrm{Clopen}(\widehat{\Sigma^*}))$: *The continuous action topos of the profinite monoid* $\widehat{\Sigma^*}$ *of* **profinite words** *with the Boolean algebra of clopen subsets.*
- *The hyperconnected quotient topos of* Σ-**Set** *corresponding to finite orbit right congruences. (***Myhill-Nerode theorem***)*

What I've been working on (Topoi of automata II, III, ...)

- Lacal variety theorem. (Topoi of automata II.)
  - We can generalize the notion of syntactic monoids.
  - The complete lattice of hyperconnected quotients of Σ-**Set**$_{\mathrm{o.f.}}$ is isomorphisc to ...
  - The Σ-set of congruences Ξ is the local state classifier of Σ^*.
  - We obtain the topos of star-free languages, the topos of group languages, ...
- Subspaces of Σ-**Set** (j.w.w. Morgan Rogers).
  - They are in bijection with *self-similar sublocale* of the Cantor space Σ^ω.
- Cohomology
  - There is the notions of cohomology for any topoi.
  - The topos of group languages for $\Sigma = \{1\}$ is the topos **Cont**$(\hat{\mathbb{Z}})$, for which case the cohomology is well-known in number theory.
  - (I'm motivated by the **star-height** problem.)

References

- Colcombet, Thomas and Daniela Petrişan (2020). "Automata minimization: a functorial approach". In: *Logical Methods in Computer Science* 16.
- Hora, Ryuya (2024). "Topoi of automata I: Four topoi of automata and regular languages". In: *arXiv:2411.06358*.