

# Harvard CS 121 and CSCI E-207

## Lecture 6: Regular Languages and Countability

Harry Lewis

September 24, 2009

**Reading:** Sipser, “The Diagonalization Method,” pages 174–178 (from just before Definition 4.12 until just before Corollary 4.18).

# Converting Finite Automata to Regular Expressions

**Theorem:** For every regular language  $L$ , there is a regular expression  $R$  such that  $L(R) = L$ .

## Proof:

Define generalized NFAs (GNFAs) (of interest only for this proof)

- Transitions labelled by regular expressions (rather than symbols).
- One start state  $q_{\text{start}}$  and only one accept state  $q_{\text{accept}}$ .
- Exactly one transition from  $q_i$  to  $q_j$  for every two states  $q_i \neq q_{\text{accept}}$  and  $q_j \neq q_{\text{start}}$  (including self-loops).

## Steps toward the proof

**Lemma:** For every NFA  $N$ , there is an equivalent GNFA  $G$ .

- Add new start state, new accept state. Transitions?
- If multiple transitions between two states, combine. How?
- If no transition between two states, add one. With what transition?

**Lemma:** For every GNFA  $G$ , there is an equivalent RE  $R$ .

- By induction on the number of states  $k$  of  $G$ .
- Base case:  $k = 2$ . Set  $R$  to be the label of the transition from  $q_{\text{start}}$  to  $q_{\text{accept}}$ .

## Ripping and repairing GNFA's to reduce the number of states

- Inductive Hypothesis: Suppose every GNFA  $G$  of  $k$  or fewer states has an equivalent RE (where  $k \geq 2$ ).
- Induction Step: Given a  $(k + 1)$ -state GNFA  $G$ , we will construct an equivalent  $k$ -state GNFA  $G'$ .

*Rip*: Remove a state  $q_r$  (other than  $q_{\text{start}}, q_{\text{accept}}$ ).

*Repair*: For every two states  $q_i \notin \{q_{\text{accept}}, q_r\}$ ,  $q_j \notin \{q_{\text{start}}, q_r\}$ , let  $R_{i,j}$ ,  $R_{i,r}$ ,  $R_{r,r}$ ,  $R_{r,j}$  be REs on transitions  $q_i \rightarrow q_j$ ,  $q_i \rightarrow q_r$ ,  $q_r \rightarrow q_r$  and  $q_r \rightarrow q_j$  in  $G$ , respectively,

In  $G'$ , put RE  $R_{ij} \cup R_{i,r}R_{r,r}^*R_{r,j}$  on transition  $q_i \rightarrow q_j$ .

Argue that  $L(G') = L(G)$ , which is regular by IH.

Also constructive.

# Example conversion of an NFA to a RE

## Examples of Regular Languages

- $\{w \in \{a, b\}^* : |w| \text{ even \& every 3rd symbol is an } a\}$
- $\{w \in \{a, b\}^* : \text{There are not 7 } a\text{'s or 7 } b\text{'s in a row}\}$
- $\{w \in \{a, b\}^* : w \text{ has both an even number of } a\text{'s and an even number of } b\text{'s}\}$
- Are there non-regular languages???

## Goal: Existence of Non-Regular Languages

Intuition:

- Every regular language can be described by a finite string (namely a regular expression).
- To specify an arbitrary language requires an infinite amount of information.
  - For example, an infinite sequence of bits would suffice:
  - $\Sigma^*$  has a lexicographic ordering, and the  $i$ 'th bit of an infinite sequence specifying a language would say whether or not the  $i$ 'th string is in the language.

⇒ Some language must not be regular.

How to formalize?

## Countability

- A set  $S$  is finite if there is a bijection  $\{1, \dots, n\} \leftrightarrow S$  for some  $n \geq 0$ .

- Countably infinite if there is a bijection  $f : \mathcal{N} \leftrightarrow S$

This means that  $S$  can be “enumerated,” i.e. listed as  $\{s_0, s_1, s_2, \dots\}$  where  $s_i = f(i)$  for  $i = 0, 1, 2, 3, \dots$

So  $\mathcal{N}$  itself is countably infinite

So is  $\mathcal{Z}$  (integers) since  $\mathcal{Z} = \{0, -1, 1, -2, 2, \dots\}$

Q: What is  $f$ ?

- Countable if  $S$  is finite or countably infinite
- Uncountable if it is not countable

## Facts about Infinite Sets

- **Proposition:** The union of 2 countably infinite sets is countably infinite.

$$\text{If } A = \{a_0, a_1, \dots\}, B = \{b_0, b_1, \dots\}$$

$$\text{Then } A \cup B = C = \{c_0, c_1, \dots\}$$

$$\text{where } c_i = \begin{cases} a_{i/2} & \text{if } i \text{ is even} \\ b_{(i-1)/2} & \text{if } i \text{ is odd} \end{cases}$$

**Q:** If we are being fussy, there is a small problem with this argument. What is it?

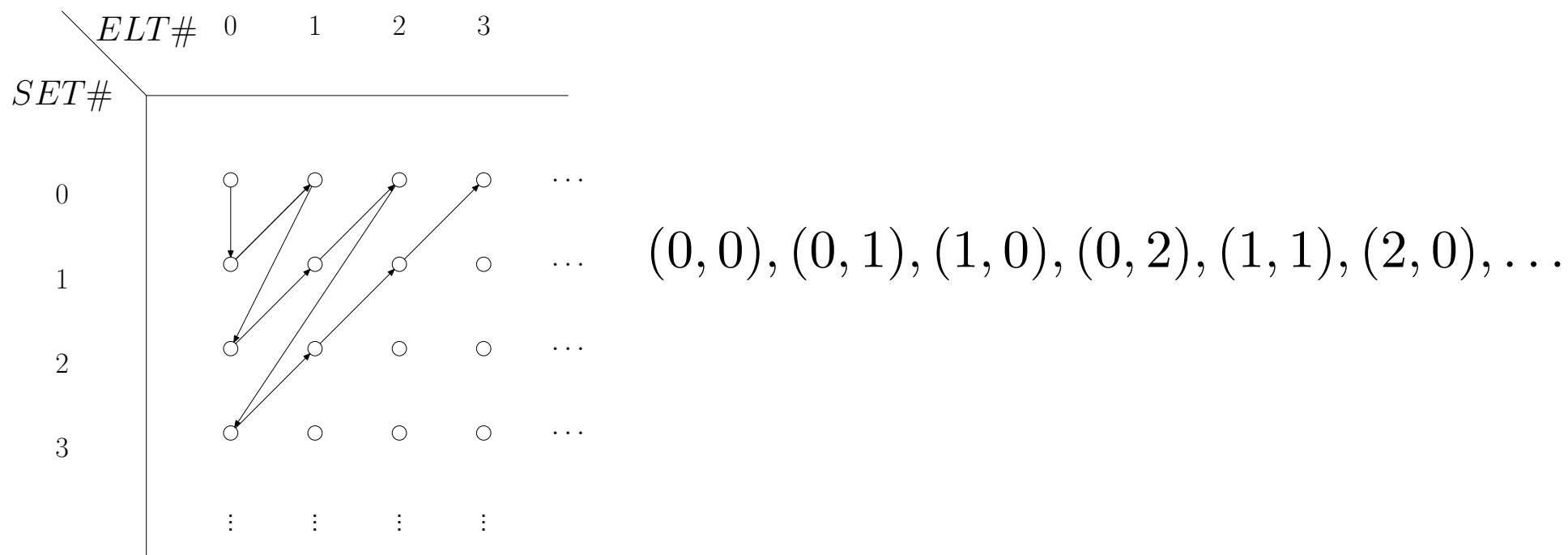
- **Proposition:** If there is a function  $f : \mathcal{N} \rightarrow S$  that is onto  $S$  then  $S$  is countable.

## Countable Unions of Countable Sets

- **Proposition:** The union of countably many countably infinite sets is countably infinite

## Countable Unions of Countable Sets

- **Proposition:** The union of countably many countably infinite sets is countably infinite



Each element is “reached” eventually in this ordering

**Q:** What is the bijection  $\mathcal{N} \leftrightarrow \mathcal{N} \times \mathcal{N}$ ?

# Are there uncountable sets? (Infinite but not countably infinite)

**Theorem:**  $P(\mathcal{N})$  is uncountable  
(The set of all sets of natural numbers)

## Proof by contradiction:

(i.e. assume that  $P(\mathcal{N})$  is countable and show that this results in a contradiction)

- Suppose that  $P(\mathcal{N})$  were countable.
- Then there is an enumeration of all subsets of  $\mathcal{N}$  say  $P(\mathcal{N}) = \{S_0, S_1, \dots\}$

# Diagonalization

$S_i$	$j =$	0	1	2	3	4	
$S_0$	Y	N	N	Y	N	...	
$S_1$	N	N	N	N	N	...	
$S_2$	Y	Y	N	Y	Y	...	
$S_3$	N	N	N	Y	N	...	
⋮				$D$			

“Y” in row  $i$ , column  $j$  means  $j \in S_i$

- Let  $D = \{i \in \mathcal{N} : i \in S_i\}$  be the diagonal.
  - $D = YNNY \dots = \{0, 3, \dots\}$
  - Let  $\bar{D} = \mathcal{N} - D$  be its complement.
  - $\bar{D} = NYYN \dots = \{1, 2, \dots\}$
  - **Claim:**  $\bar{D}$  is omitted from the enumeration, contradicting the assumption that every set of natural numbers is one of the  $S_i$ s.
- Pf:**  $\bar{D}$  is different from each row because they differ at the diagonal.

## Cardinality of Languages

- An alphabet  $\Sigma$  is finite by definition
- **Proposition:**  $\Sigma^*$  is countably infinite
- So every language is either finite or countably infinite
- $P(\Sigma^*)$  is uncountable, being the set of subsets of a countable infinite set.

i.e. There are uncountably many languages over any alphabet

**Q:** Even if  $|\Sigma| = 1$ ?

## Existence of Non-regular Languages

**Theorem:** For every alphabet  $\Sigma$ , there exists a non-regular language over  $\Sigma$ .

**Proof:**

- There are only countably many regular expressions over  $\Sigma$ .  
 $\Rightarrow$  There are only countably many regular languages over  $\Sigma$ .
- There are uncountably many languages over  $\Sigma$ .
- Thus at least one language must be non-regular.

$\Rightarrow$  In fact, “almost all” languages must be non-regular.

**Q:** Could we do this proof using DFAs instead?

**Q:** Can we get our hands on an *explicit* non-regular language?