Aligning Sentence Parts

Martin Kay

Stanford University

That is a Translation?

1. A retelling in a *Target* language of a story originally given in a *source* language
2. A text in the target language that preserves the *meaning* of the source text.

OK, so what is Meaning?!

It depends what the meaning of "is" is.

William Jefferson Clinton

That is a Translation?

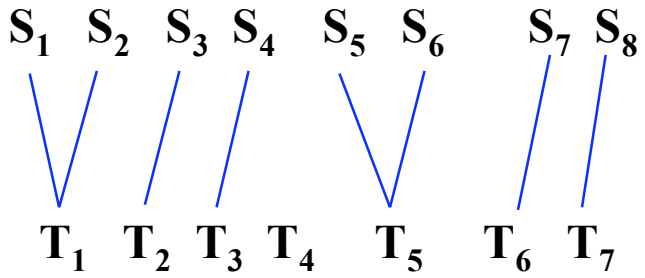
1. A retelling in a *Target* language of a story originally given in a *Source* language
2. A replacement of the words of a Target text by words with the same meaning so that the target text tells the same story as the source.

As close to 2 as possible.

A macro-translation is a better example of translation, the more micro-translations it consists of

That is a Translation?

1. *A retelling in a Target language of a story originally given in a source language*
2. **A replacement of the ~~words~~ segments of a Target text by words with the same meaning so that the target text tells the same story as the source.**



Draw as many lines as possible!

Statistical MT

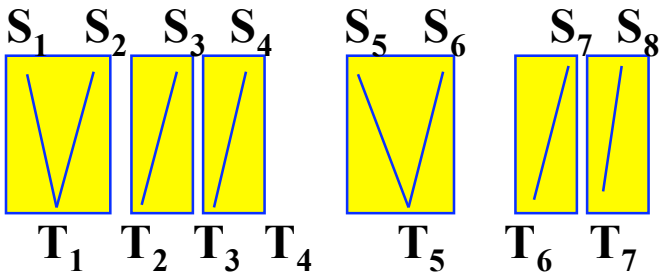
Translation model

- A probable tiling of alignments

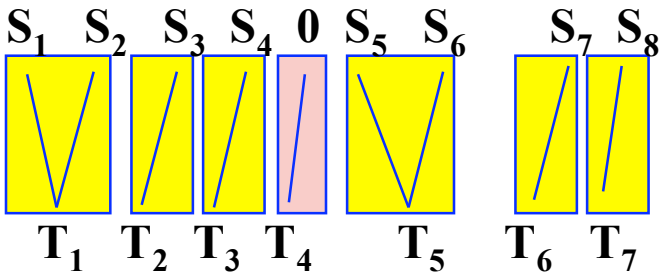
Target Language Model

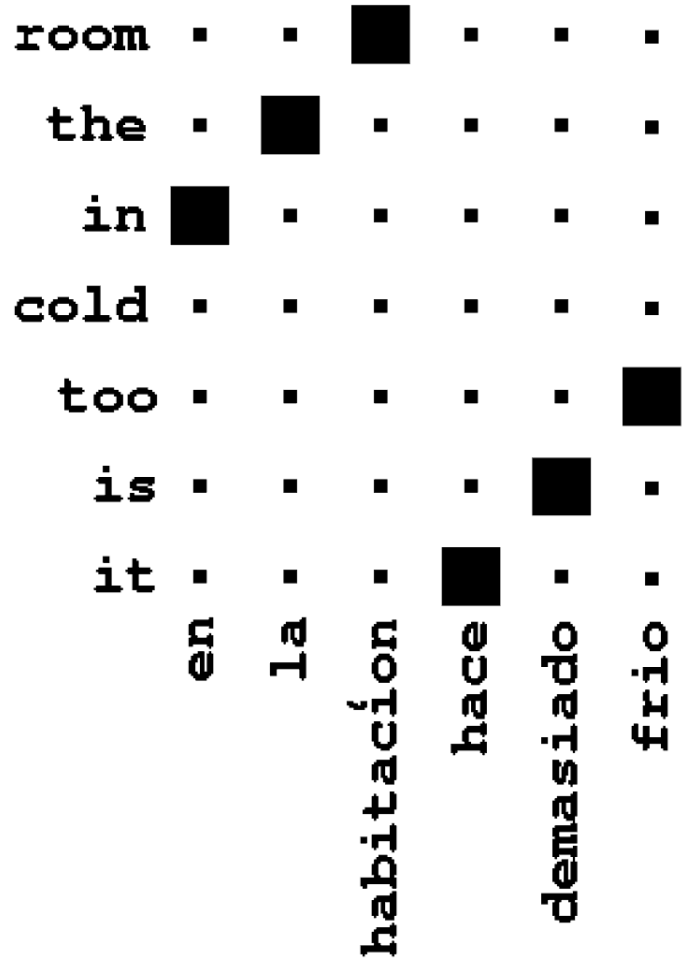
- A probable sequence of tiles

Alignments



Alignments




room
the
in
cold
too
is
it
en
la
habitación
hace
demasiado
frio

Two Approaches

1. **Use sentence alignments to induce word (segment) alignment**

   e.g. Align W_a with W_b if each tends to occur in sentences that are aligned with sentences that the other occurs in, and not elsewhere

2. **Align words independently of sentences**

   Assign to each word a vector of places in which it occurs (normalized for text length) and align words with similar vectors.

What to align?

Words

Morphemes

Lexical items

Collocations

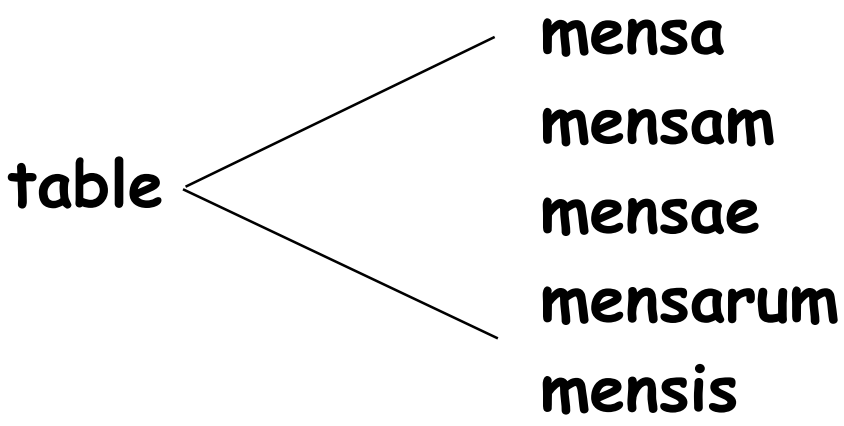
Phrases

...

Translation units

Translation units have no formal properties

Inflexions



darselo
mostrártela
estudiándola

- Align stems
- Use Lemmatized Text

Align stems

ties ~ lie

know ~ savoir ~ wissen

Idioms

without a trace

without a shadow of a doubt

in the final analysis

all things considered

see the light of day

give the finishing touch(es) to

take the liberty of

come to light

come to the help of

Idioms in lemmatized text

Collocations

conventional weapons

short story

ear, nose, and throat specialist

a little bit at a time

ignition key

spare tire

Adverbs and Conjuncts

Over and above

By and large

time and again

once and again

large and small

big and little

bread and butter

boys and girls

men and women

ladies and gentlemen

Prepositions

Composition: prep NP prep

$$\frac{\dfrac{XP\backslash XP/NP \quad NP}{XP\backslash XP} \quad XP\backslash XP/NP}{XP\backslash XP/NP}$$

Examples:

in relationship to	on account of
by virtue of	by means of
in (the) light of	in keeping with
in accordance with	with reference to
with regard to	in the presence of

Prepositions

Composition: prep NP prep

```
XP\XP/NP   NP  XP\XP/NP
    XP\XP      XP\XP/NP
          XP\XP/NP
```

Examples:

in relationship to	par rapport à
by virtue of	par le fait de
in (the) light of	en accord avec
in accordance with	*conformément à*
with regard to	*en ce qui concerne*

Subordinating conjunctions

on condition that ...

in order that ...

in the hope that ...

by the time that ...

It might be that ...

It turns out that ...

German Compounds

DIE ZEIT

Klonmeister aus Korea

Auf dem Jahrestreffen der "American Association for the Advancement of Science" konnte das Wissenschaftsland Südkorea seinen spektakulären Forschungserfolg zum therapeutischen Klonen präsentieren Von Gero von Randow für ZEIT.de

German Compounds

- Lebensversicherungsgesellschaftsangestellter
- Zweihundertsebenundfünfzig

Many Languages do not mark
word boundaries in writing

Proposal

- Any string that is interestingly long and that occurs interestingly often is a potential translation unit.
- The longer it is, the less often it has to occur to count as interesting.

How to find interesting strings?

Suffix Trees

Suffix Trees

are

- **wonderful**
- **beautiful**
- **underappreciated**
- **powerful**

Suffix Trees

are good for finding

- sequences of at least *m* characters occurring at least *n* times in the text.
- patterns in texts
- Common subsequences in two texts.
- Shortest unique subsequence.
- Frequency of substrings.

A *Really* Silly Idea

Index by all $\binom{n+1}{2}$ *substrings* of the text.

Observe: Every substring of a string is a prefix of some suffix of the string. So use a digital search tree to index the suffixes of the text.

Preprocess the Text

Suffix Trees

Patricia Arrays

Digital Acyclic Graphs

Search Time: O(p)
Prepocess Time: O(t)

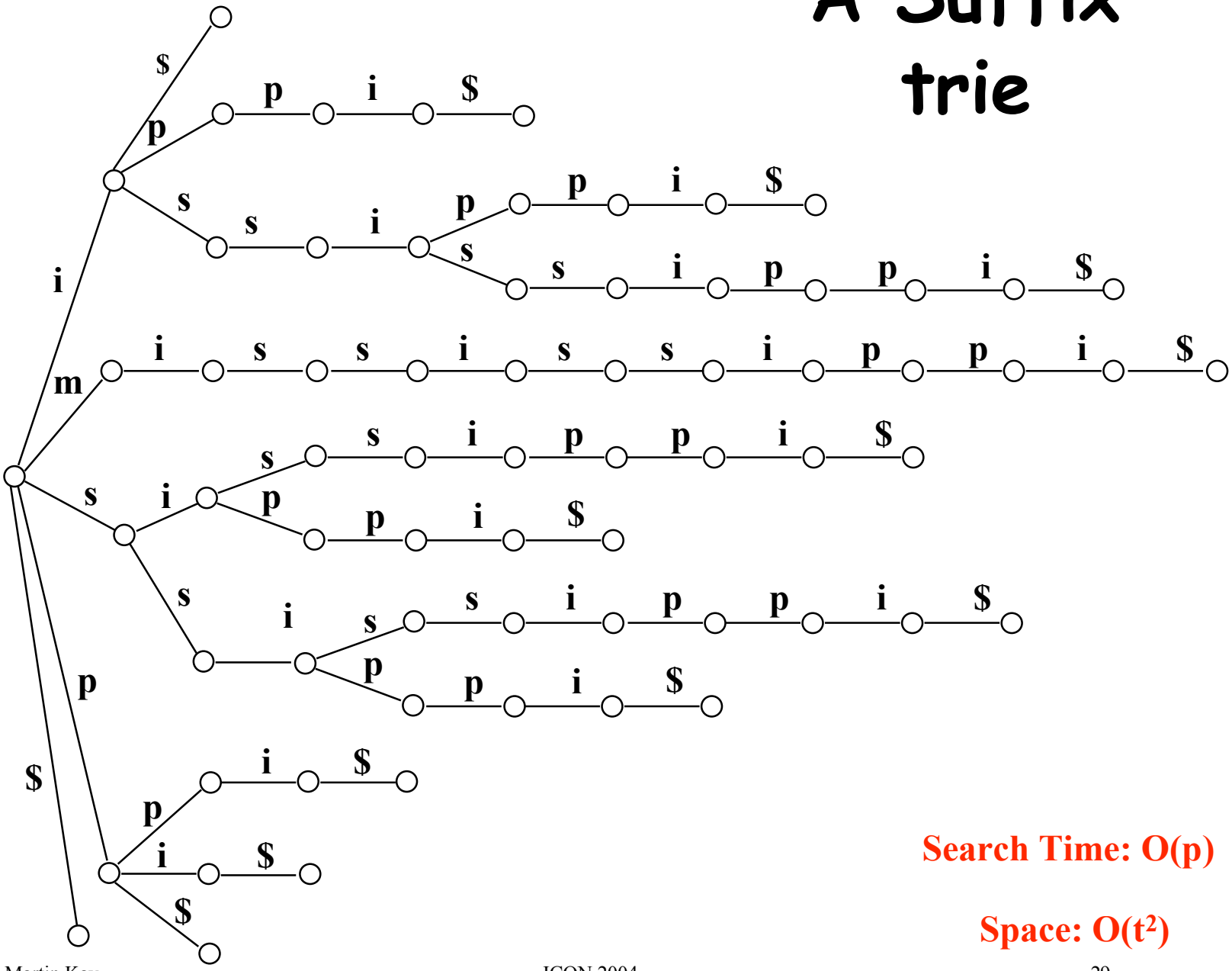
Search all occurrences:
O(p) + O(t)

t=occurrences of pattern

A Corpus

mississippi

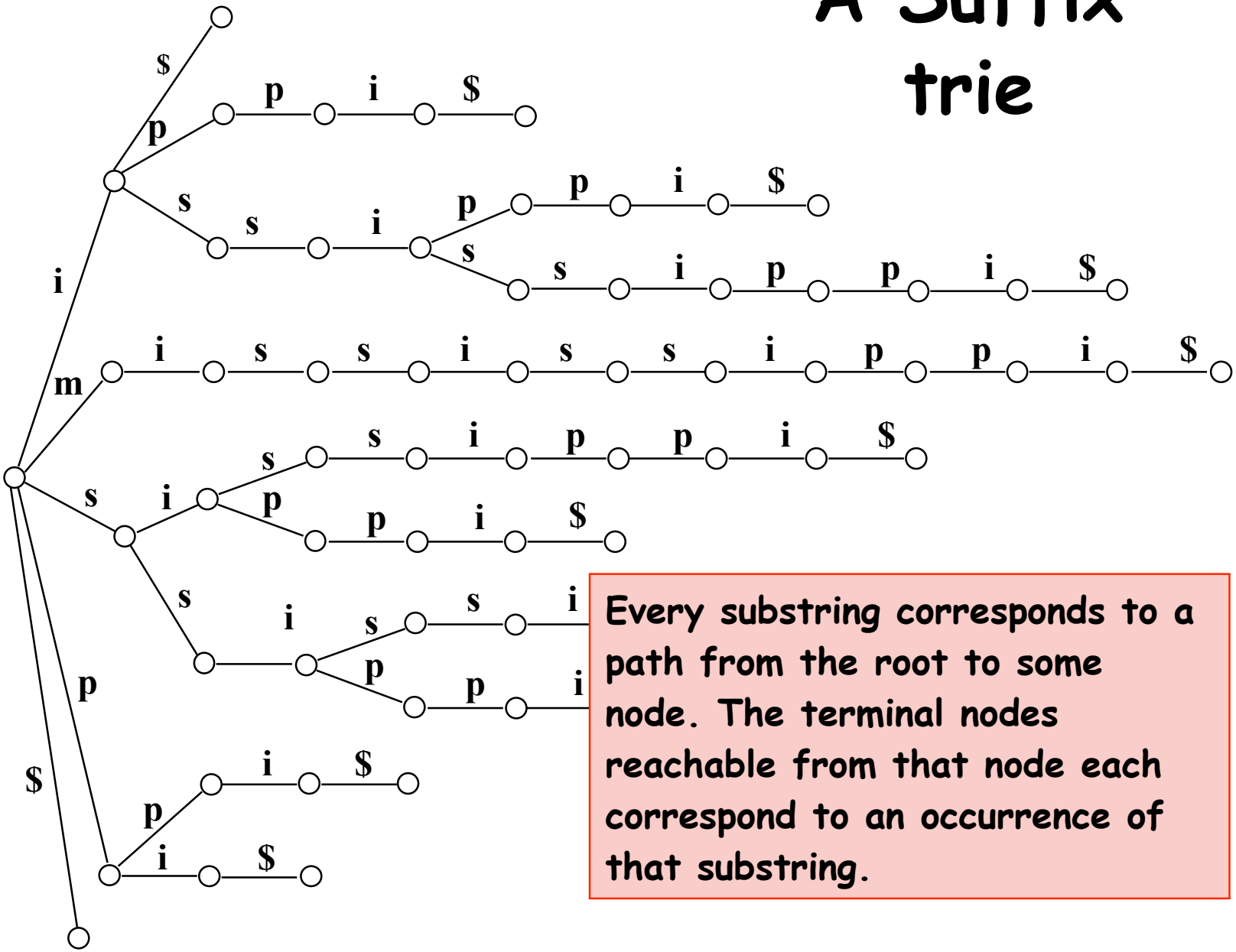
A Suffix trie



Search Time: O(p)

Space: $O(t^2)$

A Suffix trie



For a text of *n* words

The tree contains (at most)

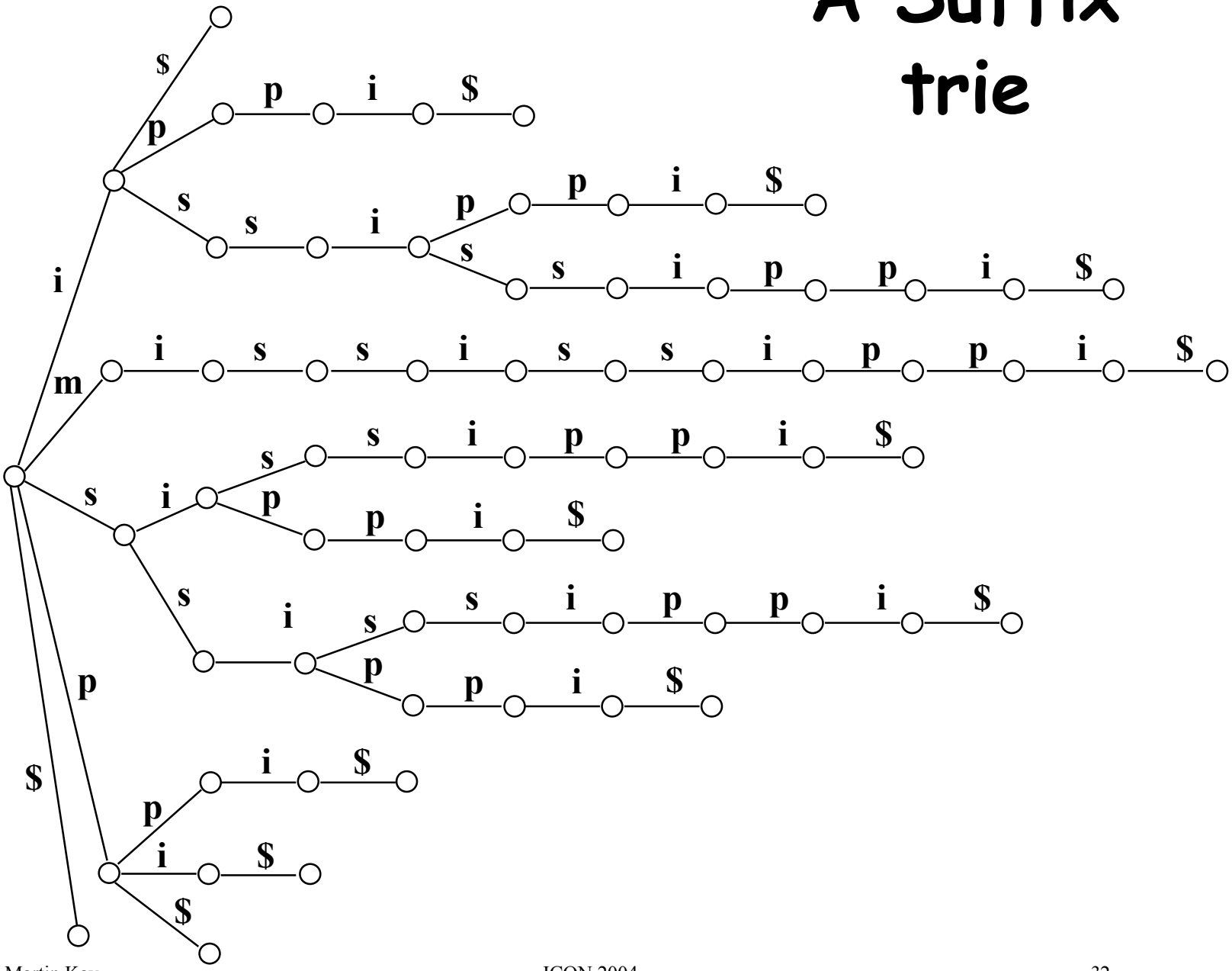
$\binom{n+1}{2}$ edges

$\binom{n+1}{2}+1$ nodes

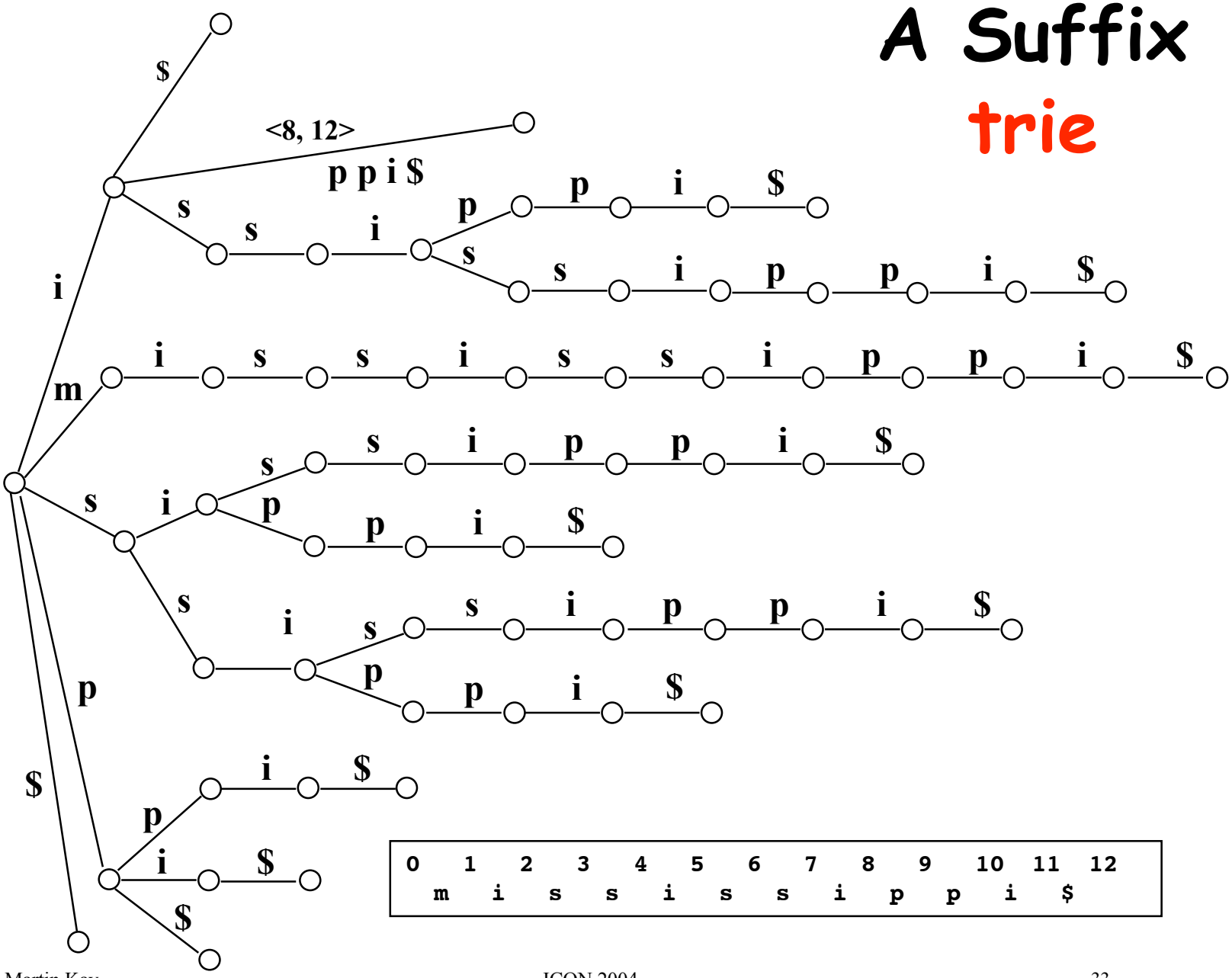
n terminal nodes

n-1 branching nodes

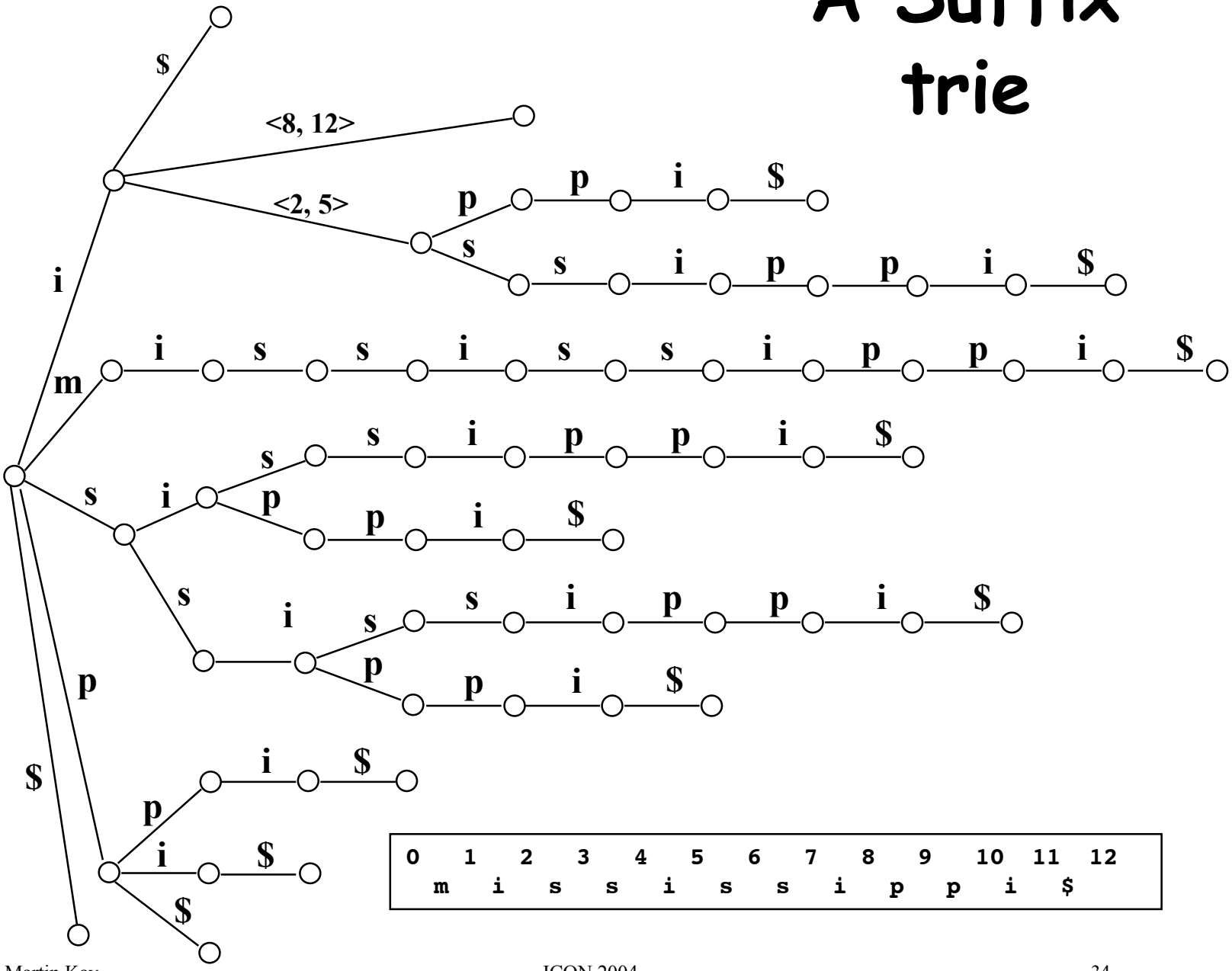
A Suffix trie



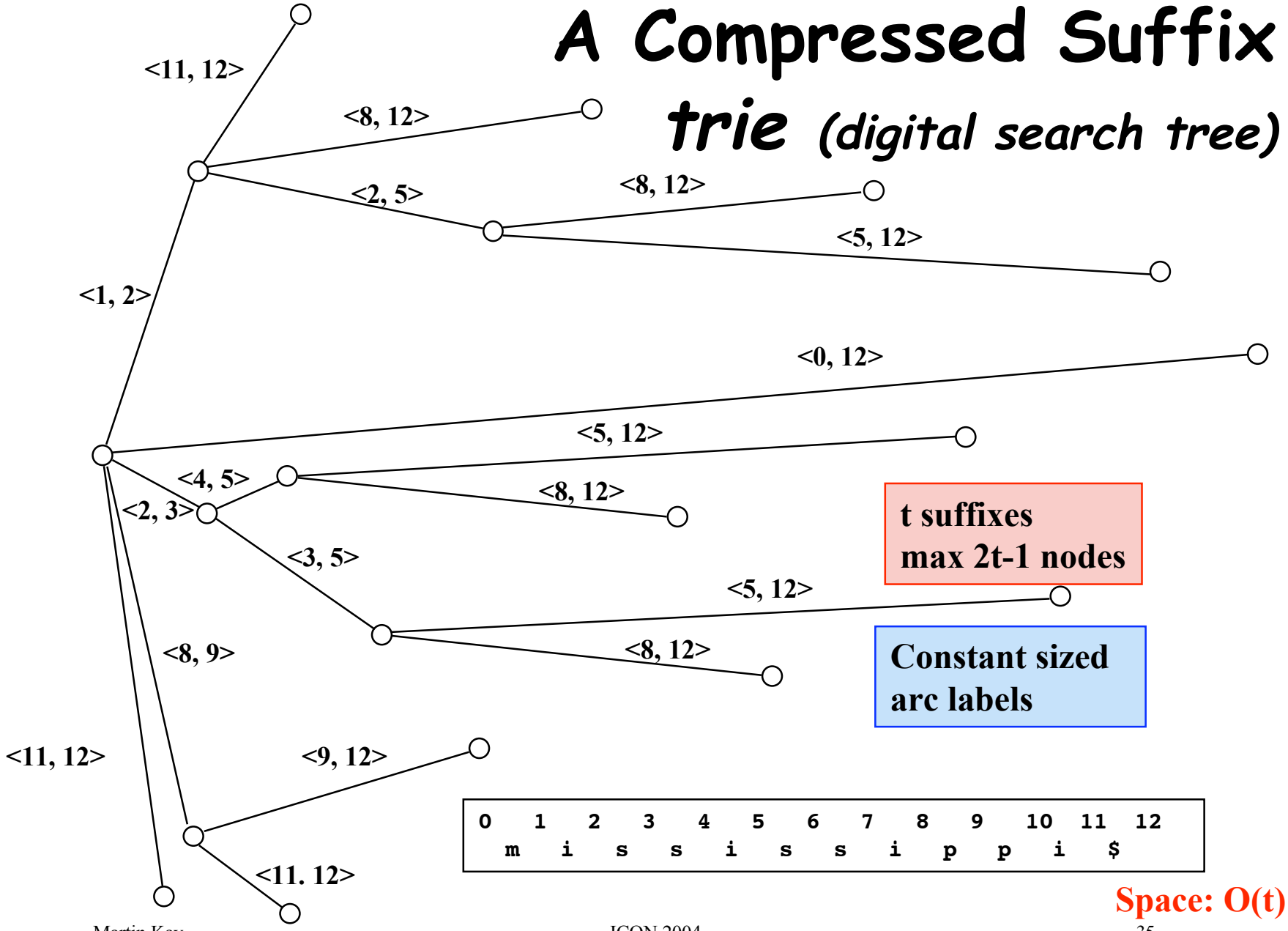
A Suffix trie



A Suffix trie



A Compressed Suffix *trie* (digital search tree)

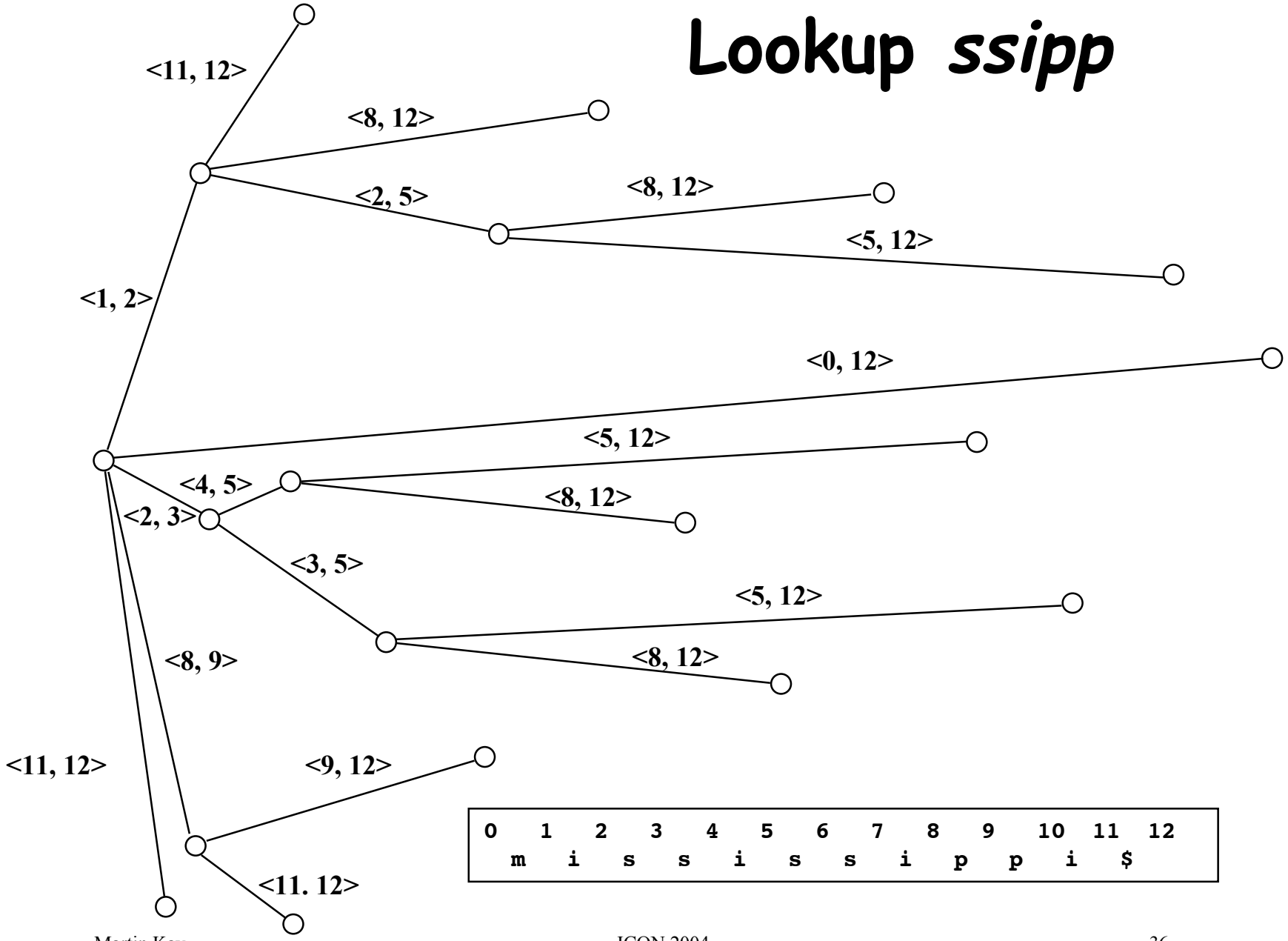


t suffixes
max 2t-1 nodes

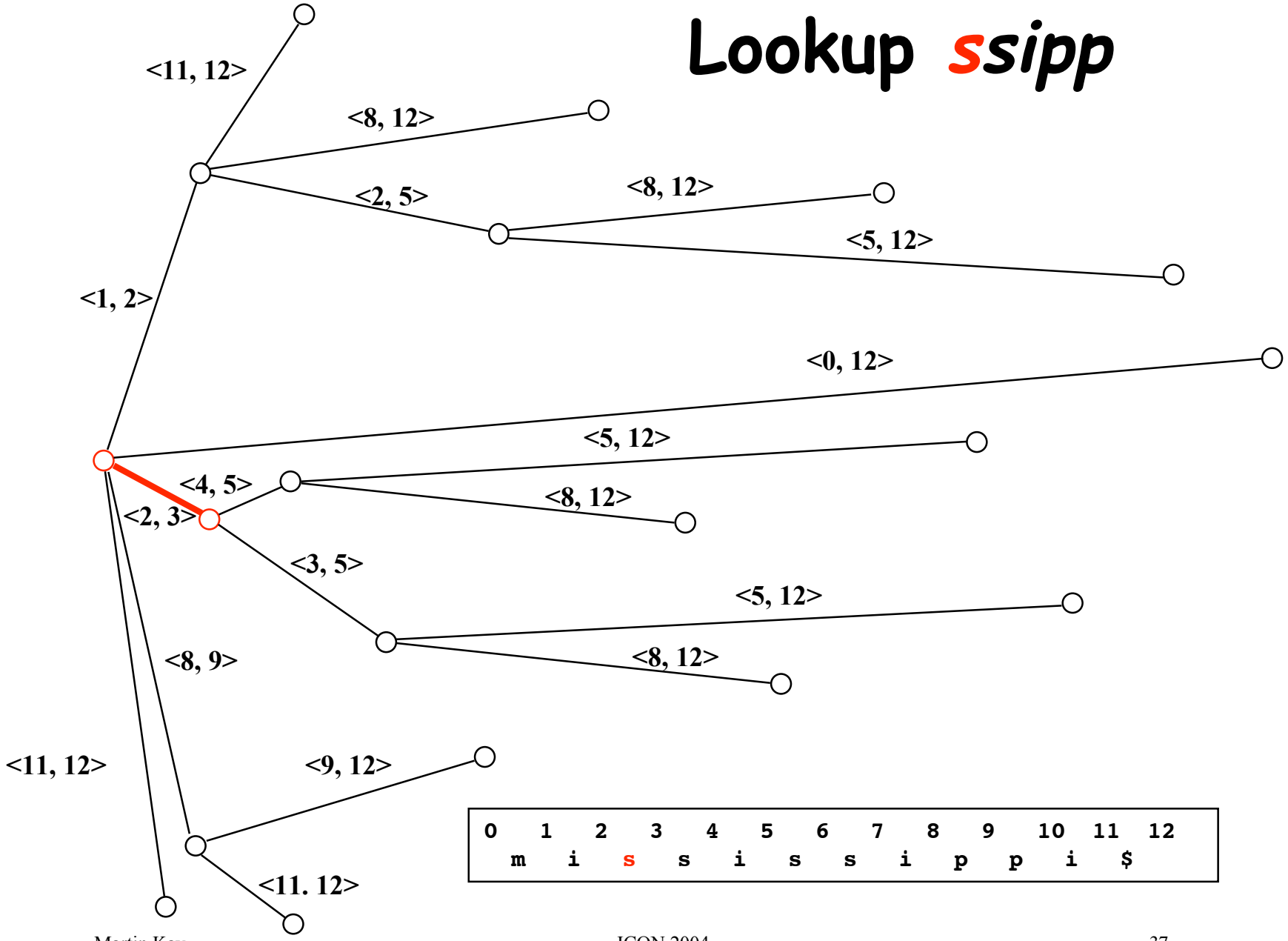
Constant sized
arc labels

Space: O(t)

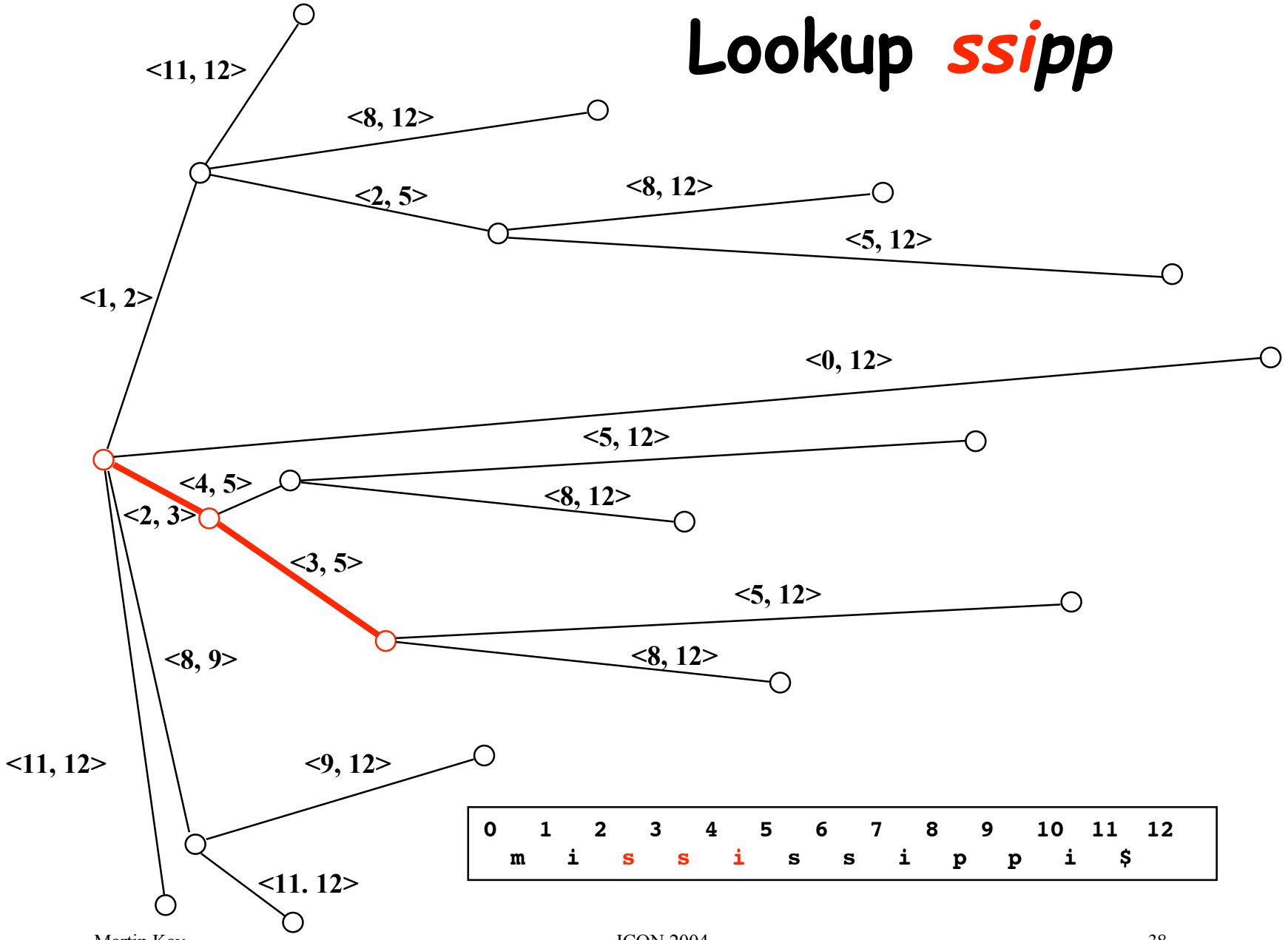
Lookup *ssipp*



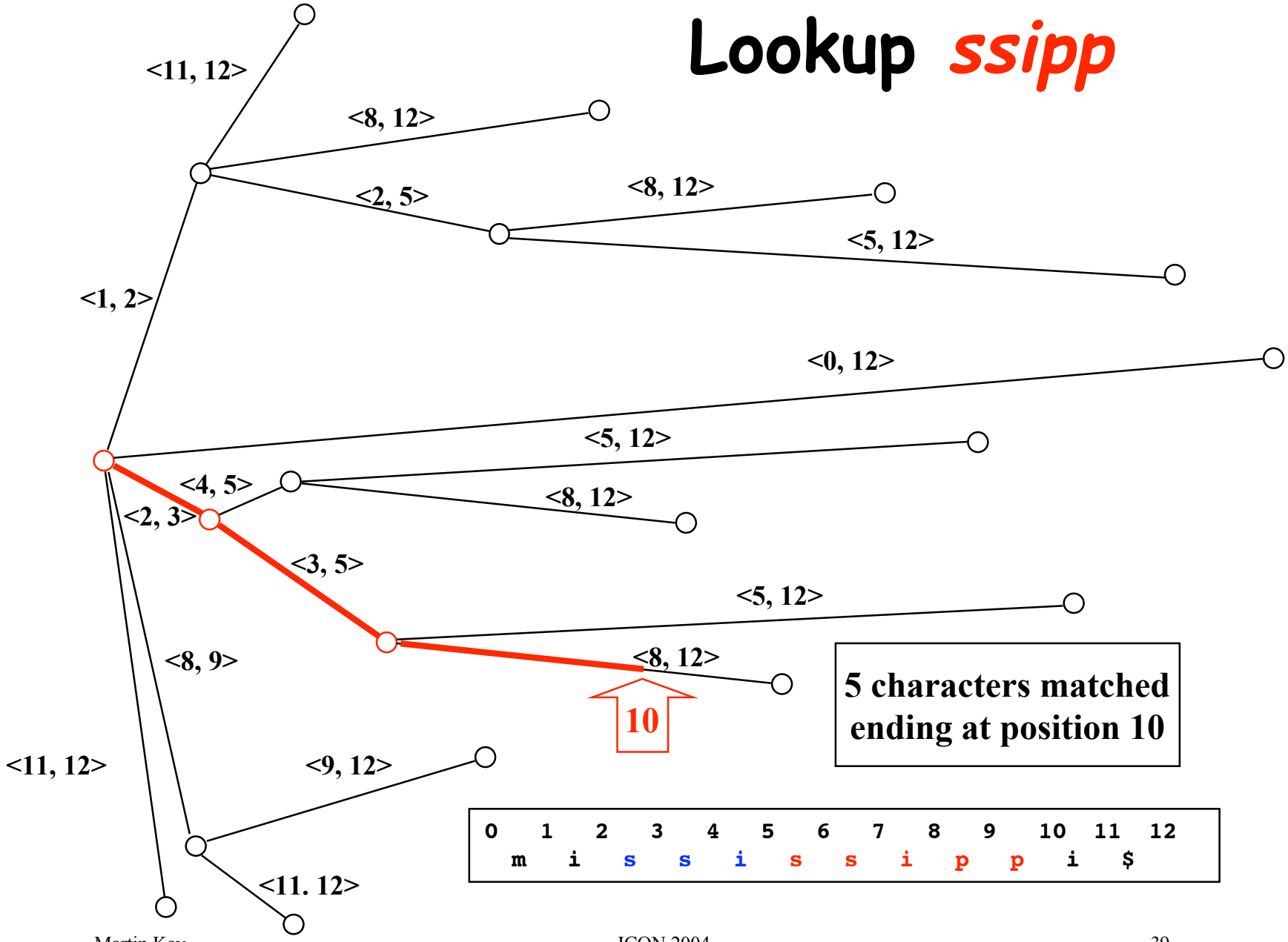
Lookup *ssipp*



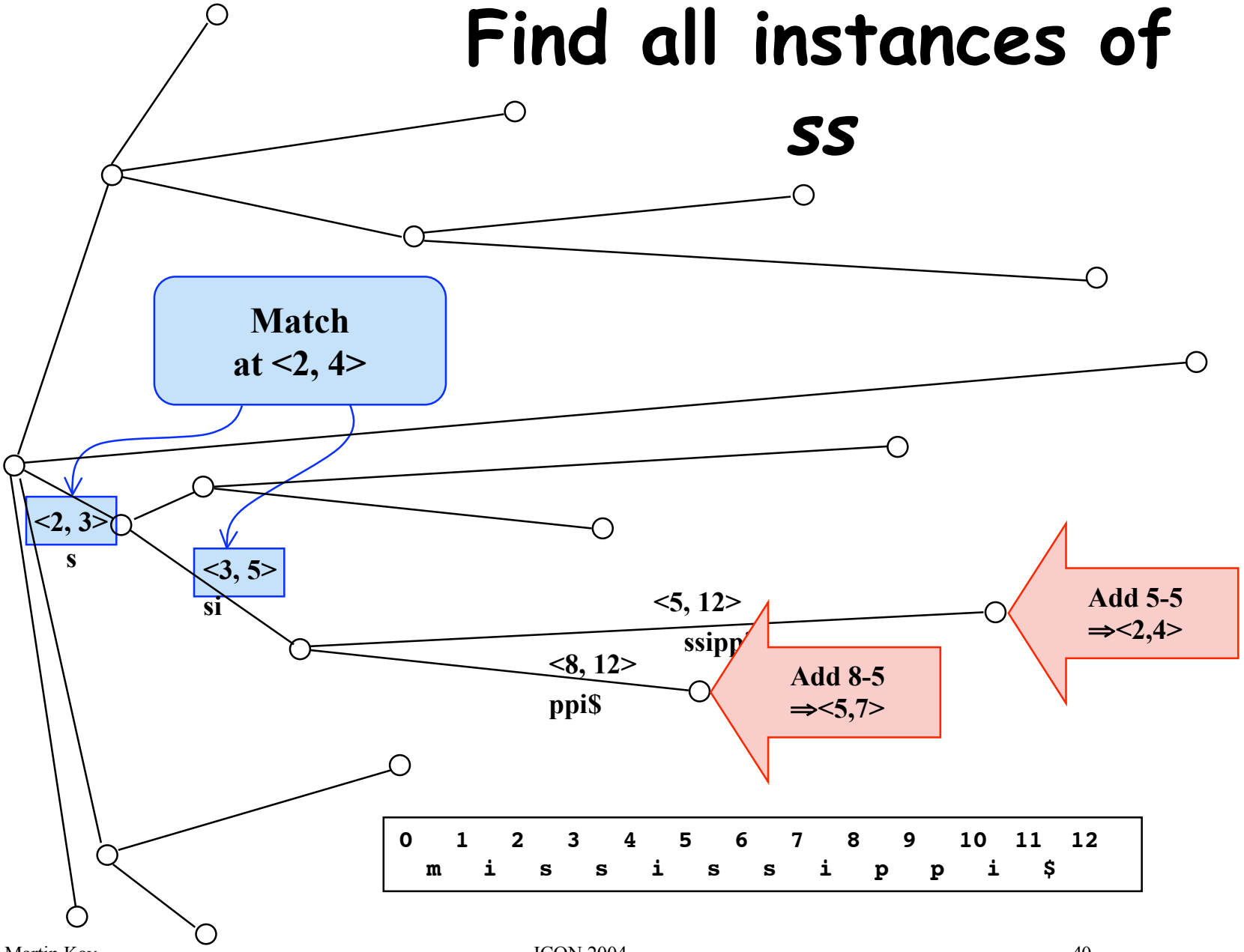
Lookup *ssipp*



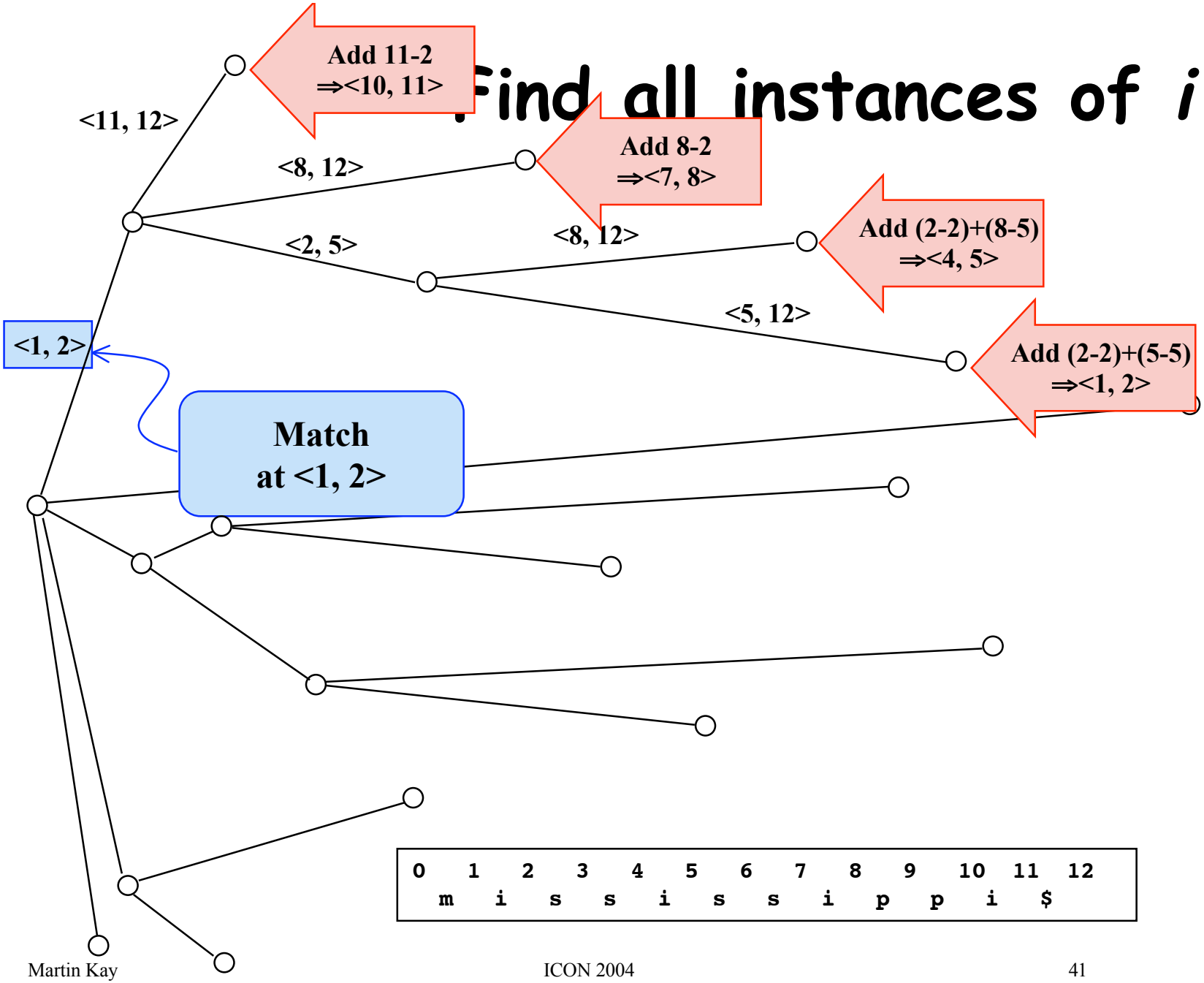
Lookup *ssipp*



Find all instances of *ss*



Find all instances of *i*

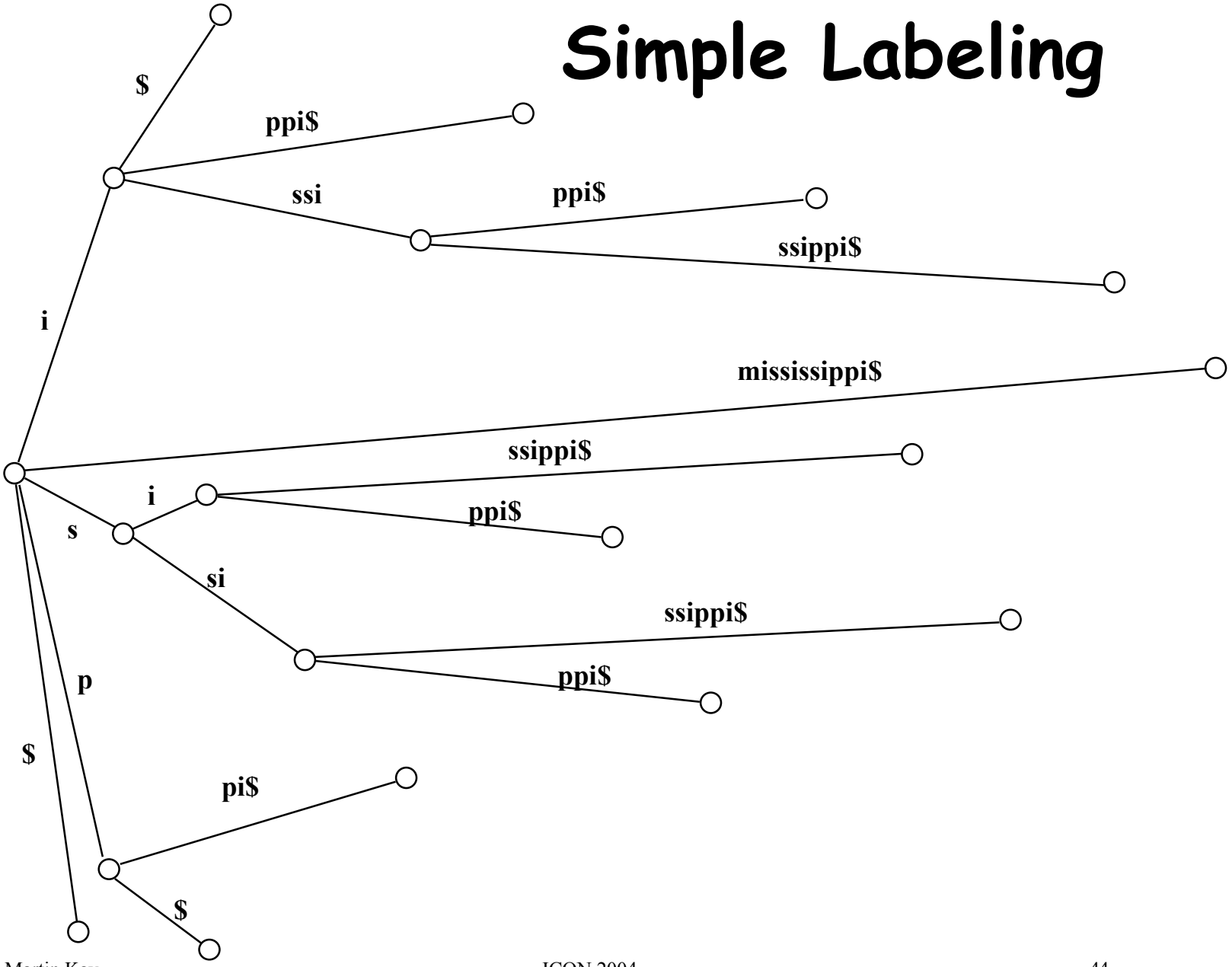


Observe:

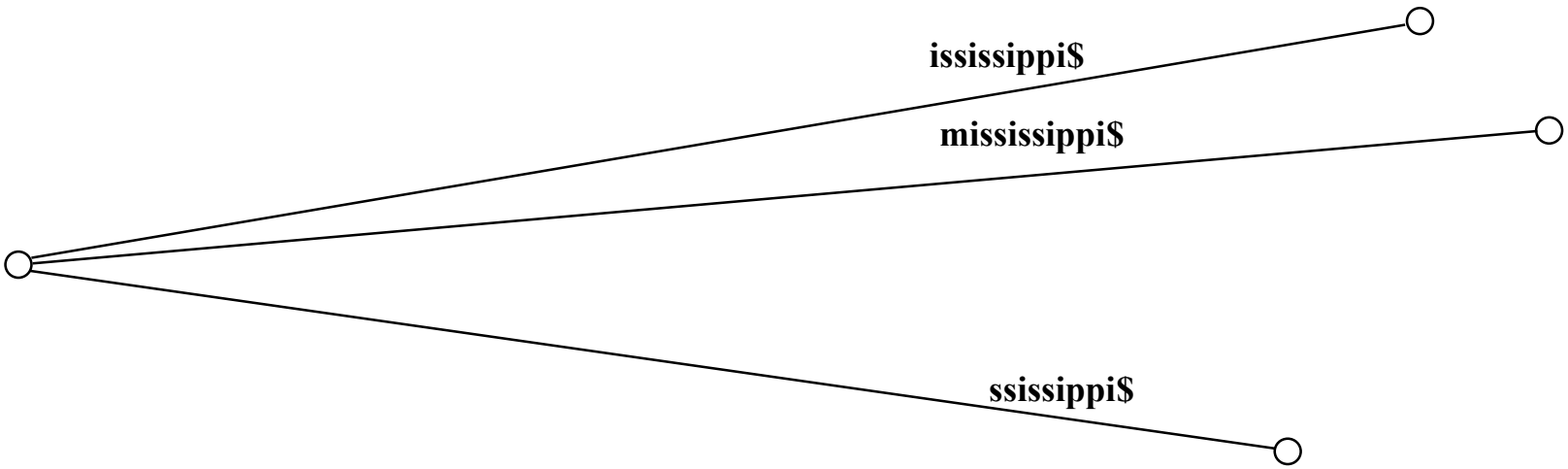
A (sub)tree with *n* terminals contains a total of at most 2*n*-1 nodes. Therefore finding all occurrences, once the first has been found requires, at most, that 2 nodes be visited for each hit.

Building the tree

Simple Labeling

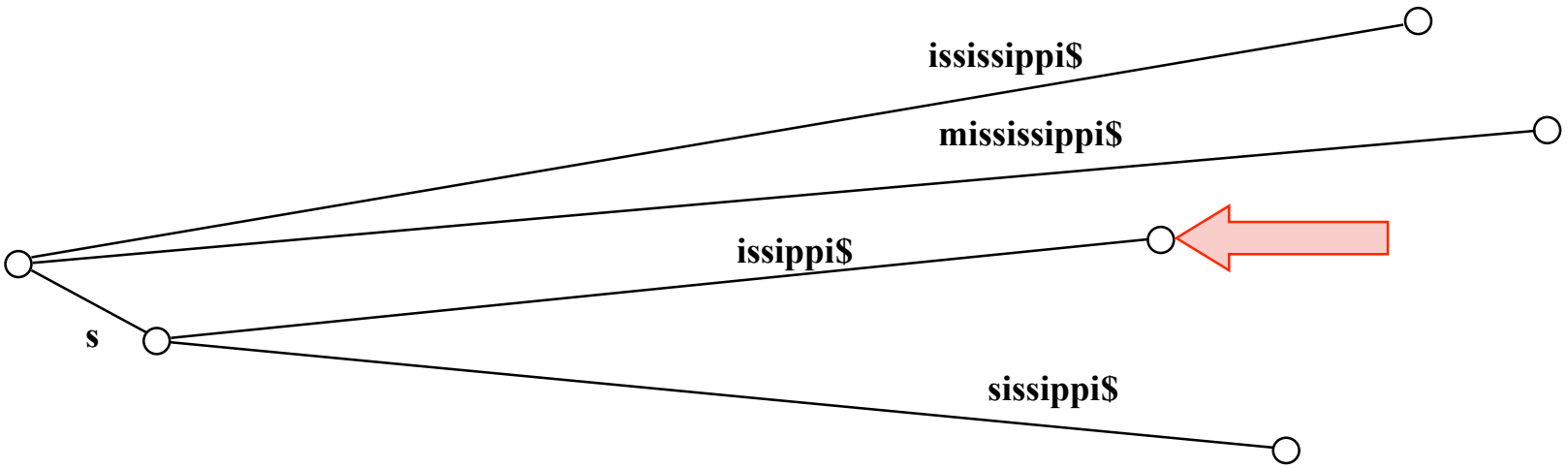


Building



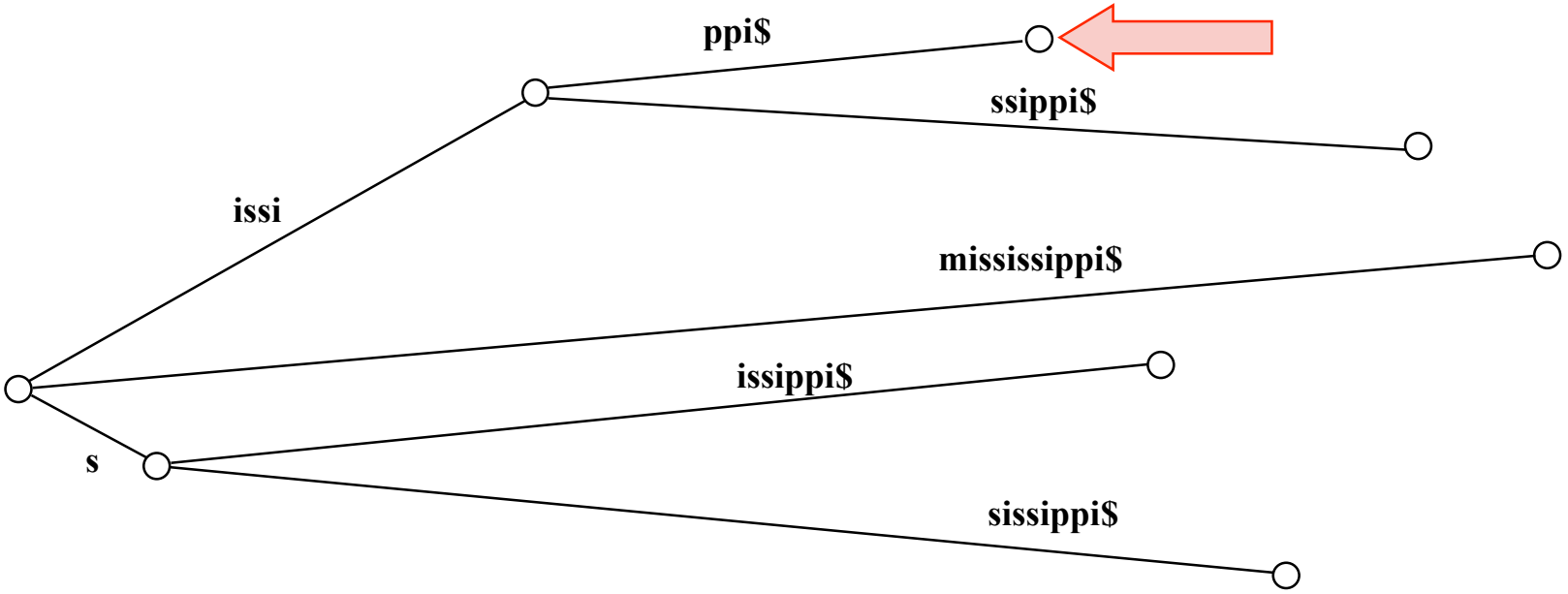
m	i	s	s	i	s	s	i	p	p	i	$

Building



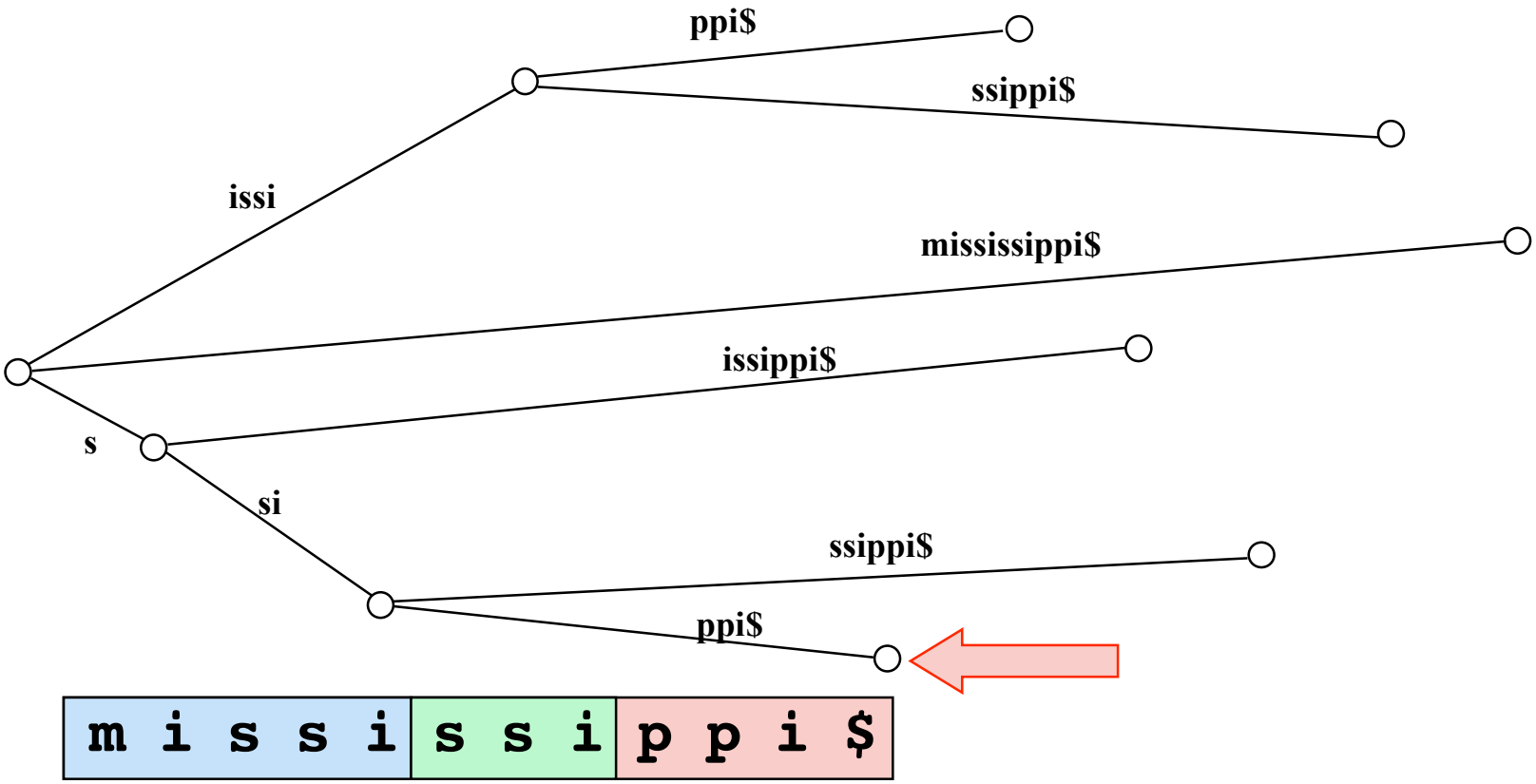
m	i	s	s	i	s	s	i	p	p	i	$

Building

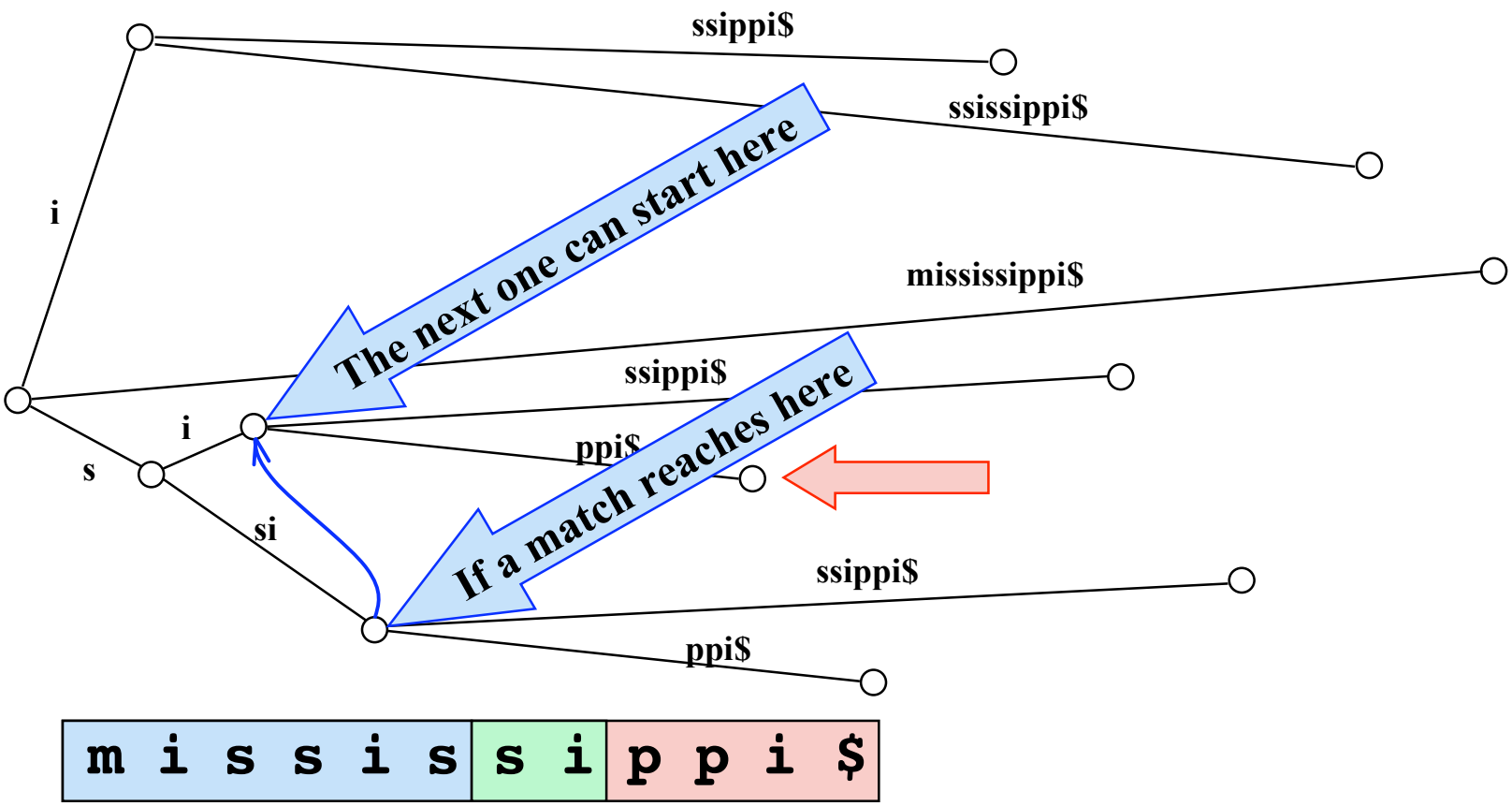


m	i	s	s	i	s	s	i	p	p	i	$

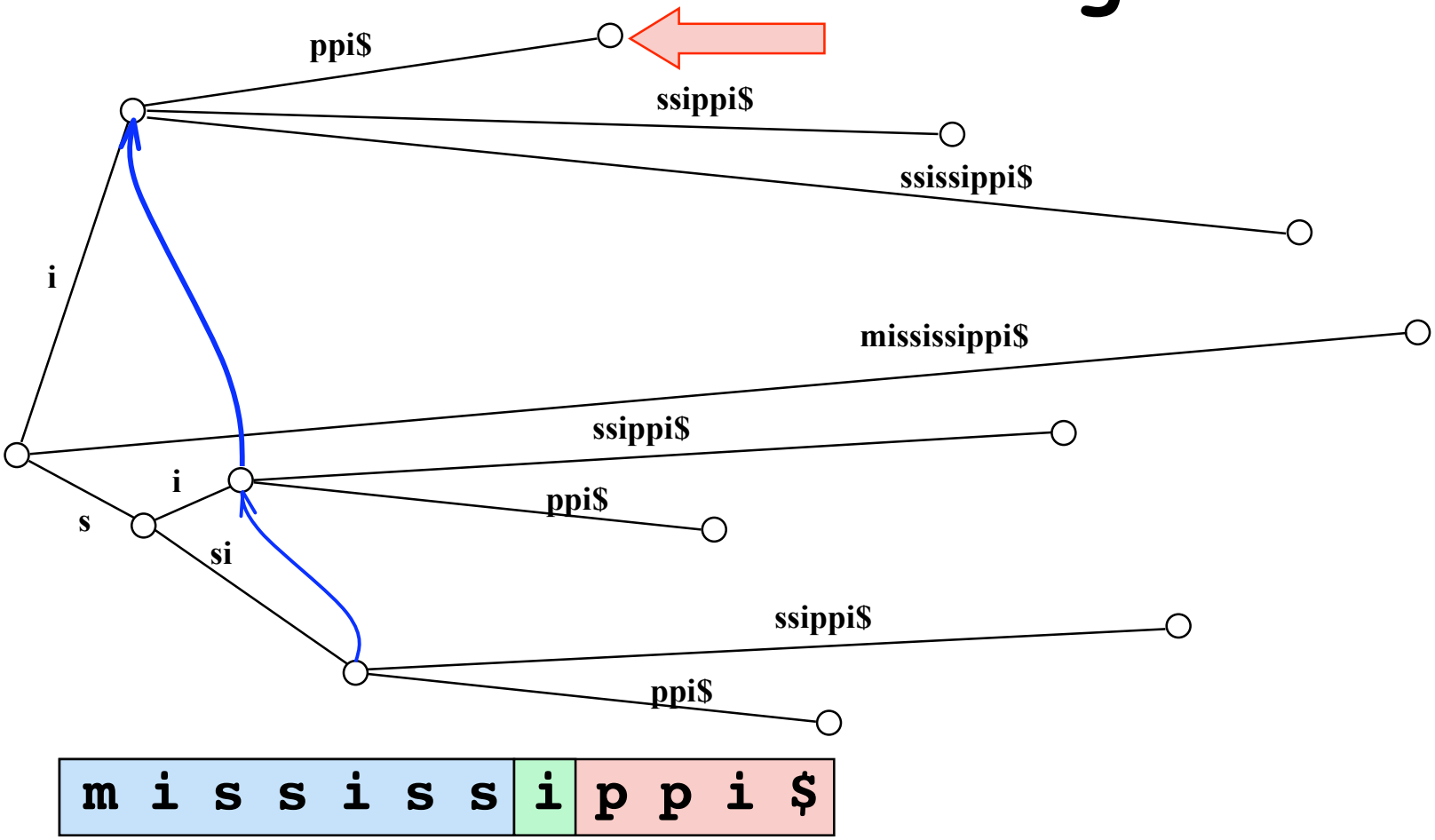
Building


ppi$
ssippi$
issi
mississippi$
issippi$
s
si
ssippi$
ppi$
m i s s i s s i p p i $

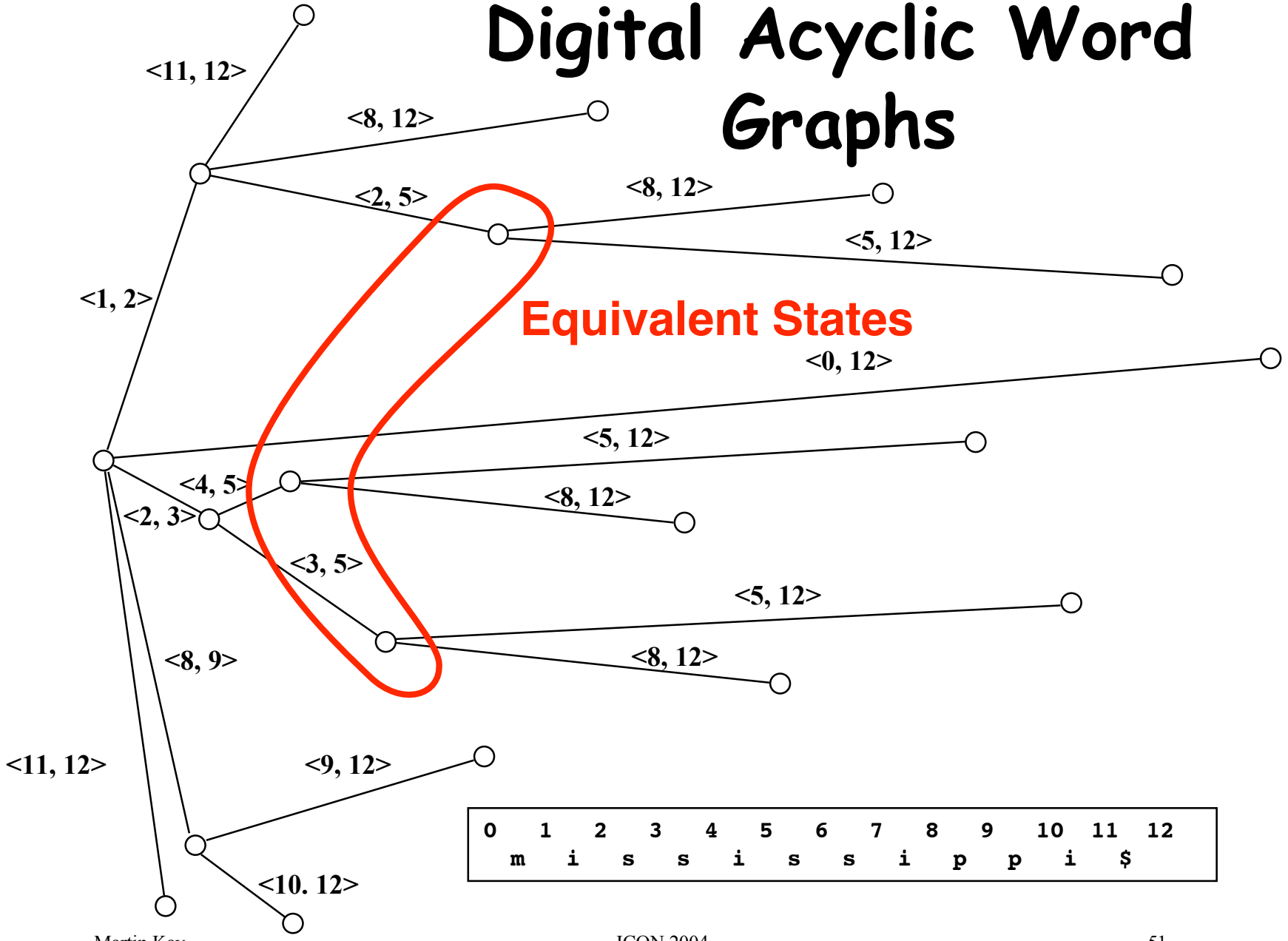
Building


ssippi$
ssissippi$
i
mississippi$
The next one can start here
ssippi$
i
ppi$
s
If a match reaches here
si
ssippi$
ppi$
m i s s i s s i p p i $

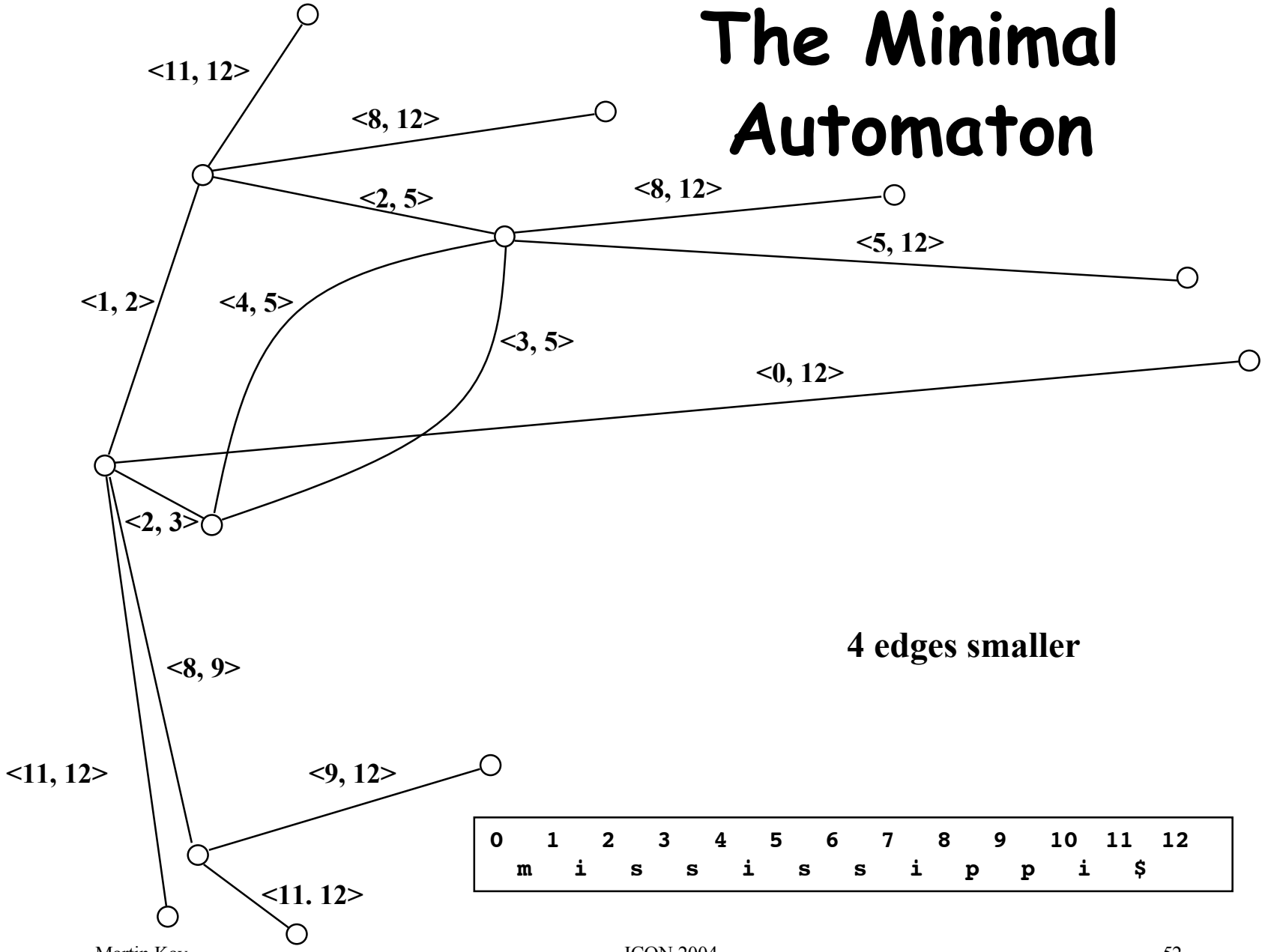
Building


ppi$
ssippi$
ssissippi$
i
mississippi$
ssippi$
i
ppi$
s
si
ssippi$
ppi$
m i s s i s s i p p i $

Digital Acyclic Word Graphs



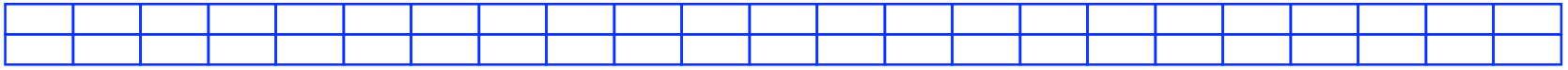
The Minimal Automaton



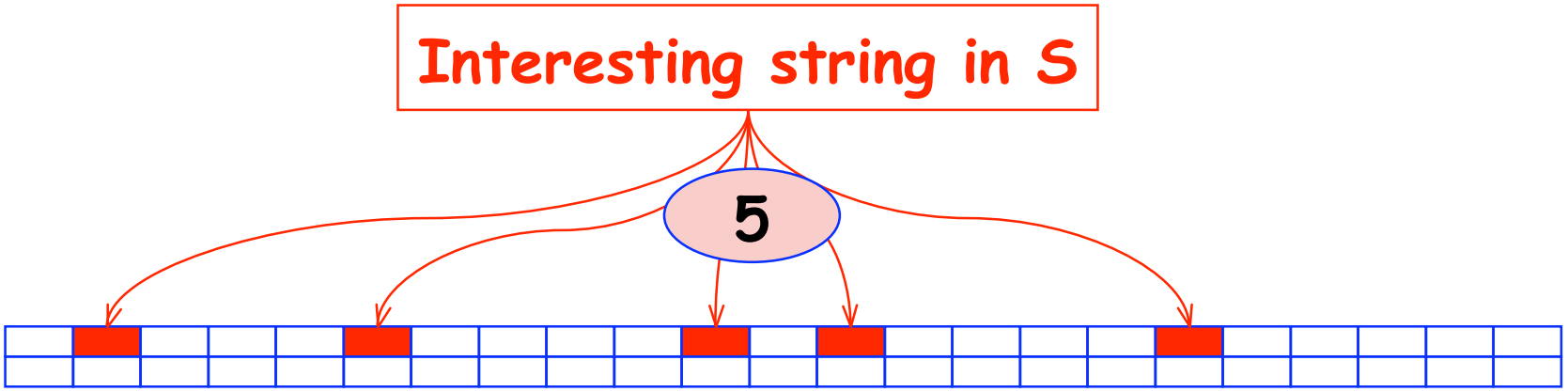
4 edges smaller

0	1	2	3	4	5	6	7	8	9	10	11	12
m	i	s	s	i	s	s	i	p	p	i	$	

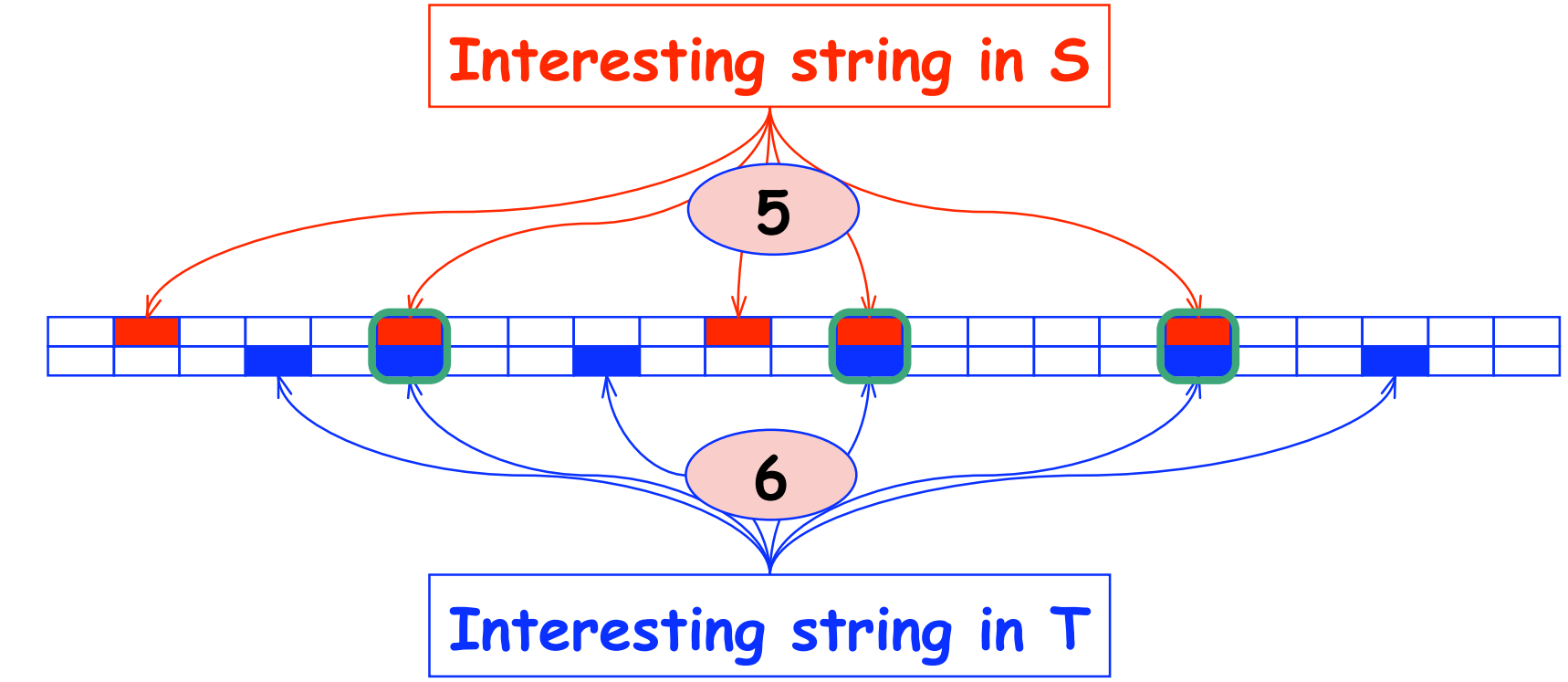
Finding Alignments



Finding Alignments

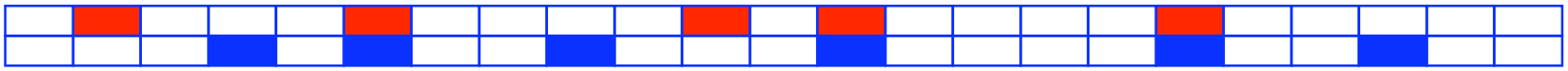


Finding Alignments



4 Common segments

Finding Alignments



$$\frac{3}{8} \qquad \frac{|\text{Intersection}|}{|\text{Union}|}$$

Variants

- Align *b* with *a* if it is the most similar (interesting) string to *a*.
- Align a and b if each is the most similar (interesting) string to the other.
- Lemmatize
- Tag

A Few Alignments

-ly	-ment
-s	les
anti-	contraire
clockwise	dans le sense des aiguilles
	une montre
conditioning	climatisation
lighter	cigar

A Few More

anticlockwise	gegen den Uhrzeigersinn
Hand break	Feststellbremse
break	-bremse
turned to "AUX"	in die "AUX" Stellung gedreht
all-wheel drive	der Allradantrieb
ignition key	Zündschlüssel
the engine and	abstellen und
safety	Sicherheits-

Discontinuous items ???

Phrasal verbs

as far as ... is concerned

...

(Untried) Solution

- Use the suffix-tree to align each text with itself