Soft Inference and Posterior Marginals

21 February 2018

The questions we answered so far

- *“What is the best path through this graph”*
- *“What is the state sequence underlying this string”*
- *“Is this string a part of this language”*
- *“How do you compose this string, with this language”*

- Decisive answers to definitive questions
- “Hard” inference

“Soft” questions

- *How probable is it for this language to produce this symbol sequence?*
- *How likely is it that the word “feed” here is a noun and not a verb?*
- *How likely is this segment to be a constituent?*
- *How probable is it that rule X→YZ has been used in composing this sentence*

- “Confidence”-type answers to questions about certainty
- “Soft” inference

Soft vs. Hard Inference

- Hard inference
  - "Give me a single solution"
  - Viterbi algorithm
  - Maximum spanning tree (Chu-Liu-Edmonds alg.)
- Soft inference
  - Task 1: Compute a distribution over outputs
  - Task 2: Compute functions on distribution
    - **marginal probabilities**, expected values, entropies, divergences

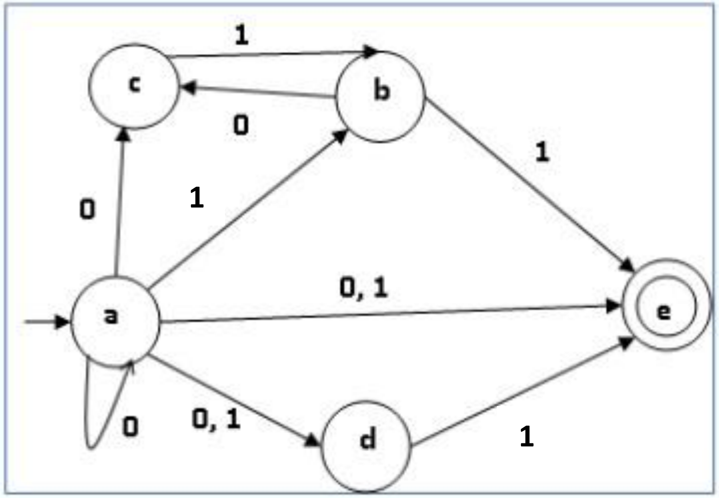
Why Soft Inference?

- Useful applications of posterior distributions
  - **Entropy**: how confused is the model?
  - **Entropy**: how confused is the model of its prediction at time *i*?
  - **Expectations**
    - What is the expected number of words in a translation of this sentence?
    - What is the expected number of times a word ending in –ed was tagged as something other than a verb?
  - **Posterior marginals**: given some input, how likely is it that some (***latent***) event of interest happened?

What we will cover

- Soft inference can be applied to any probabilistically defined model
  - Or weighted model in general
- We will specifically look at soft inference in
  - Regular grammars
    - FSGs / PFSGs
  - Context free grammars
    - HMMs / CFGs / PCFGs

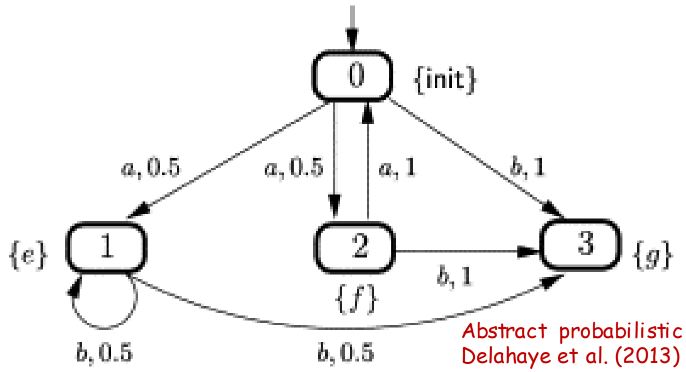
Inference in Regular Languages



- Regular languages can be recognized by a DFA or an NDFA
  - Question answered: "Does this string belong to this language"

- Can we answer : *Is the state "b" visited in recognizing "00011"*
  - DFA: Yes
  - NDFA: No
    - How about *how likely is it that the state "b" was visited in recognizing "b"*?

The probabilistic (finite) automaton



Abstract probabilistic automata
Delahaye et al. (2013)

- Probabilistic extension of NDFA
- Conventional NDFA rules:

$$s_i \overset{a}{\rightarrow} s_j$$
$$s_i \overset{a}{\rightarrow} s_k$$

  - State s_i can transition to both s_j and s_k after absorbing symbol a

- PFA rules:

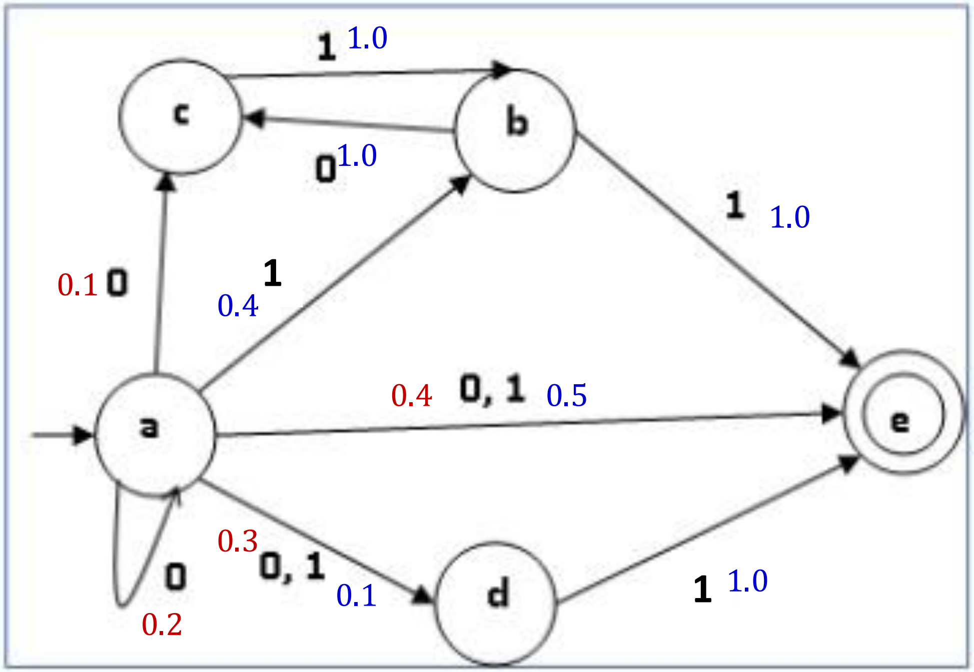
$$s_i \overset{a}{\rightarrow} s_j(0.2)$$
$$s_i \overset{a}{\rightarrow} s_k(0.8)$$

  - The different transitions have probabilities

$$\sum_k P(s_i \overset{a}{\rightarrow} s_k) = 1.0$$

  - Note: The distribution (which sums to 1.0) is specific to state-symbol combination (not just state)

Inference in Regular Languages

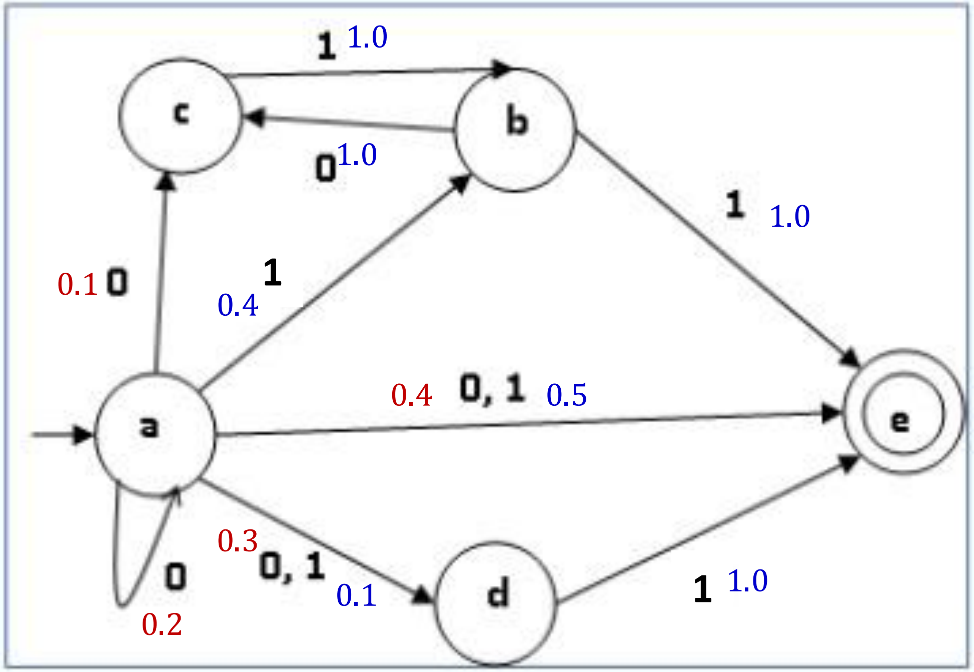


aacbe
aaabe
aaade

- *What is the probability that the state "b" is visited in recognizing "00011"*

- Can now view the recognition as a random walk *through the state sequences* that can "absorb" 00011
- What is the probability that the state "b" was visited in recognizing 00011

Inference in a PFA



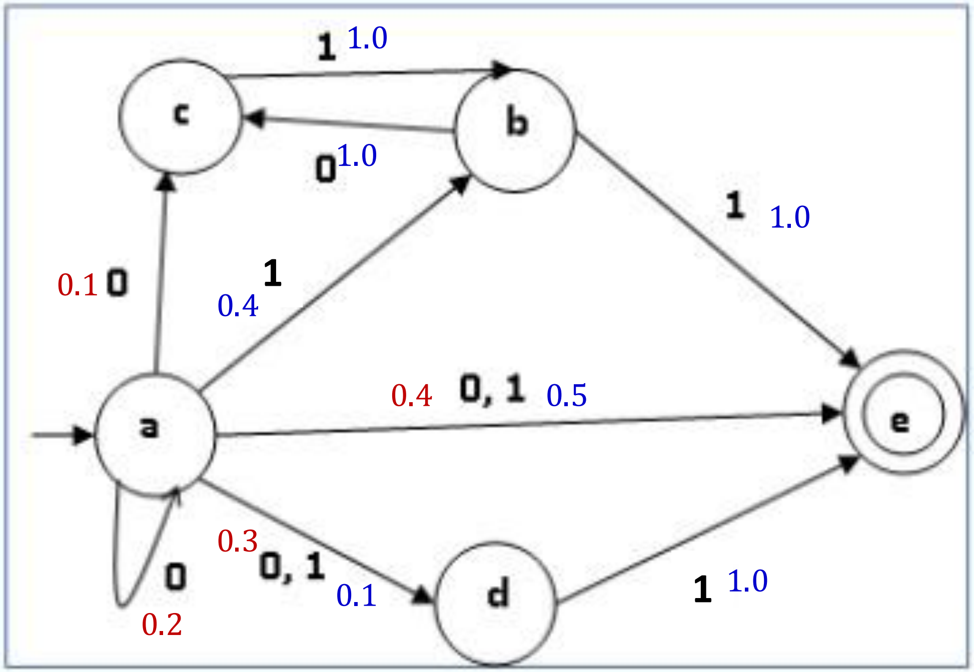
aacbe P = 0.004

aaabe P = 0.0032

aaade P = 0.0008

- *What is the probability that the state "b" is visited in recognizing "00011"*
- Can now view the recognition as a random walk *through the state sequences* that can "absorb" 00011
- What is the probability that the state "b" was visited in recognizing 00011

Inference in a PFA



aacbe P = 0.004

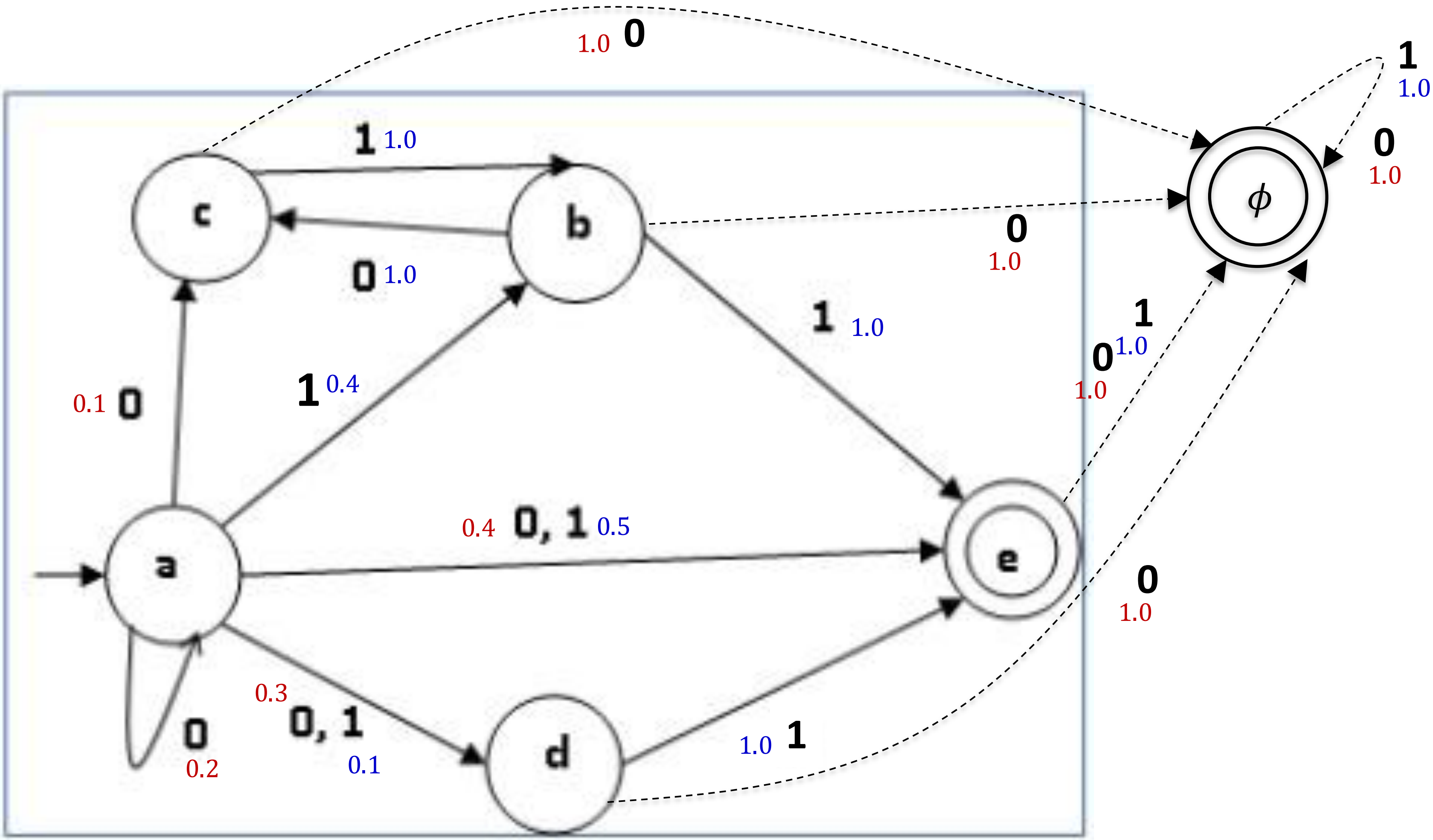
aaabe P = 0.0032

aaade P = 0.0008

Why don't these sum to 1.0?
Hint: figure is incomplete, but it doesn't affect our computation

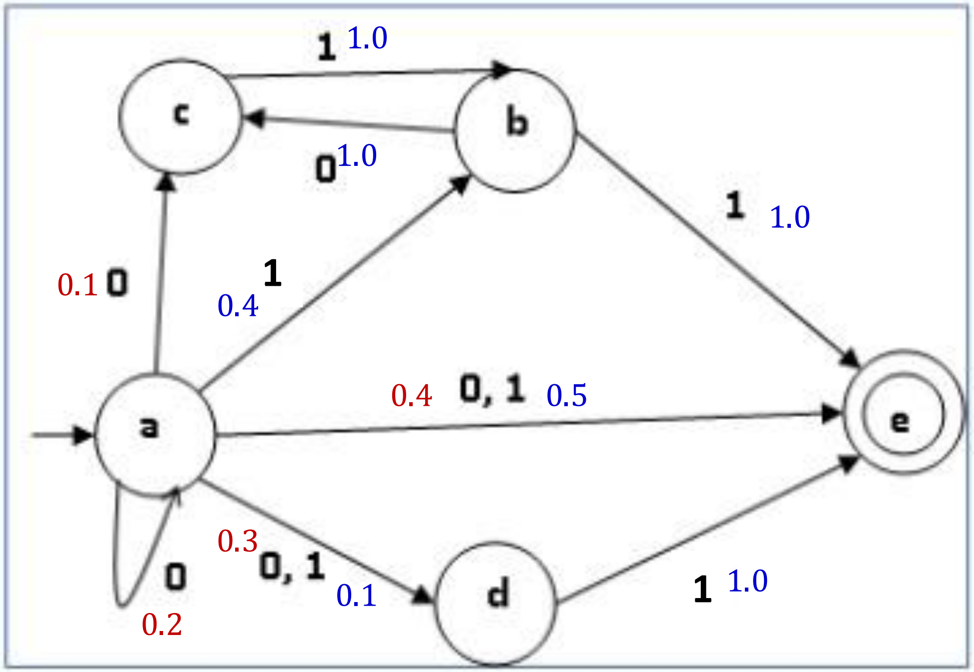
- *What is the probability that the state "b" is visited in recognizing "00011"*
- Can now view the recognition as a random walk *through the state sequences* that can "absorb" 00011
- What is the probability that the state "b" was visited in recognizing 00011

Inference in a PFA



- *We aren't interested in state sequences which end in ϕ*

Inference in a PFA



aacbe P = 0.004

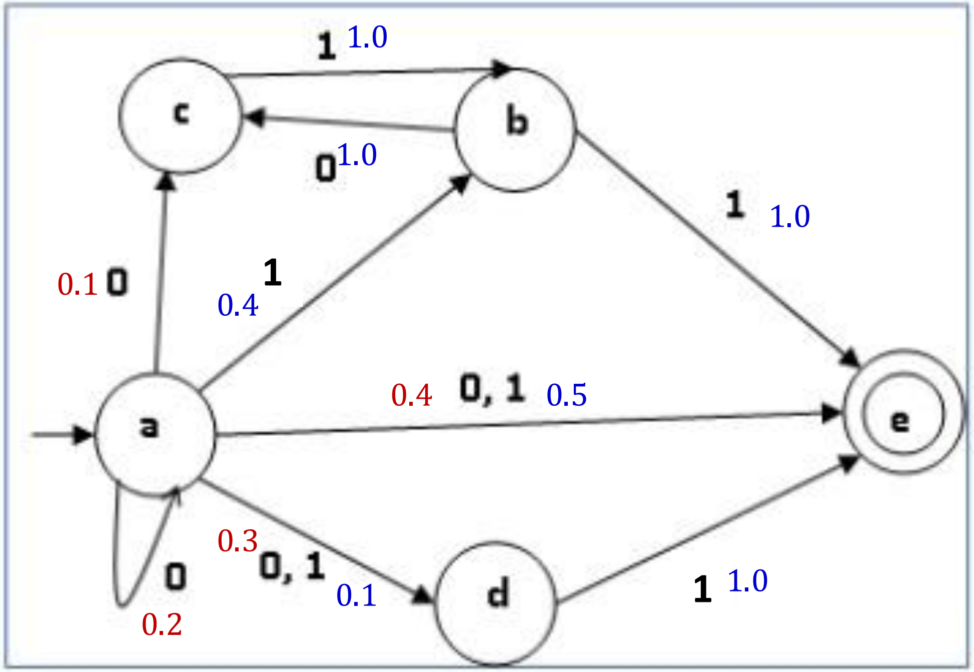
aaabe P = 0.0032

aaade P = 0.0008

P(visiting b) = 0.9

- *What is the probability that the state "b" is visited in recognizing "00011"*
  - *Given that the final state was e!*
- Can now view the recognition as a random walk *through the state sequences* that can "absorb" 00011
- What is the probability that the state "b" was visited in recognizing 00011

Inference in a PFA



aacbe P = 0.004

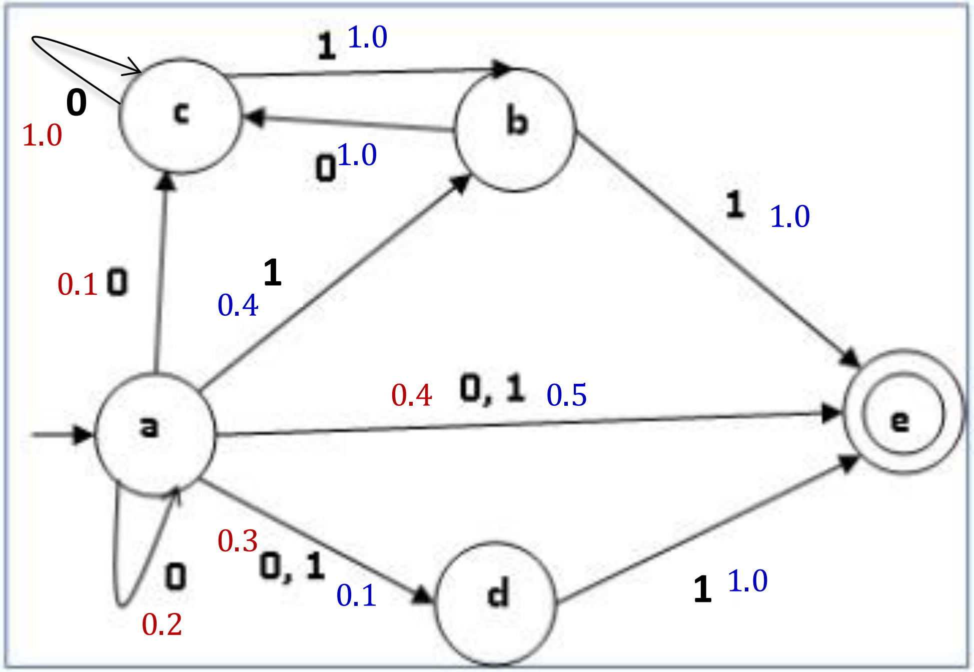
aaabe P = 0.0032

aaade P = 0.0008

P(visiting b|00011) = 0.9

- Note that we really need the probabilistic framework to make this statement
  - The PFA is actually a probability distribution over strings!!
- But the naïve computation we just performed is not scalable
  - Need an efficient algorithm!

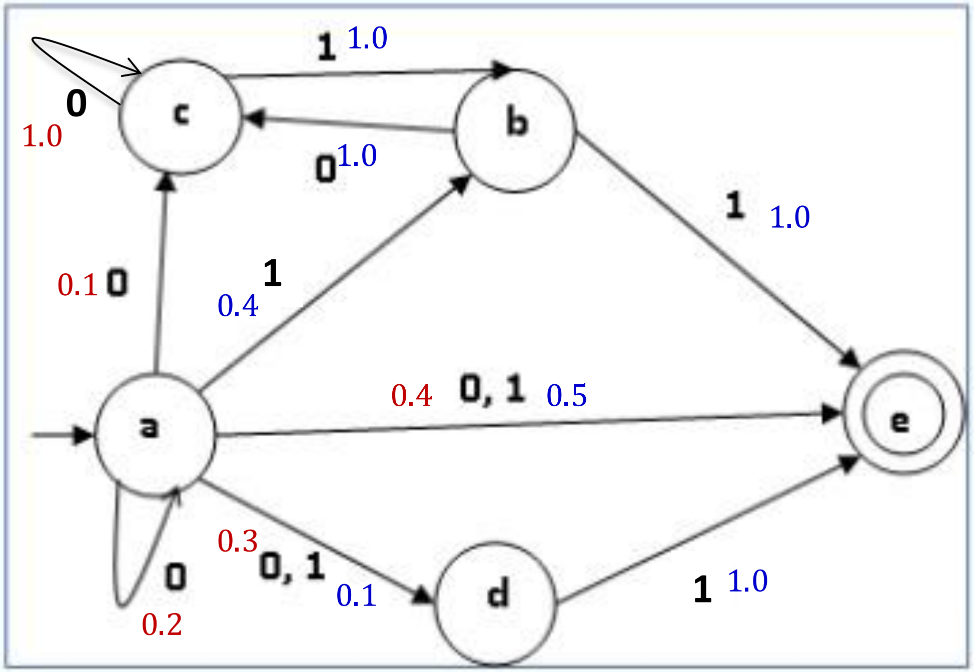
Inference in a PFA



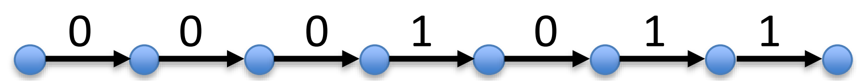
0 0 0 1 0 1 1

- Plot all the state sequences (ending in "e") that can "consume" the symbol sequence to the right

Inference in a PFA

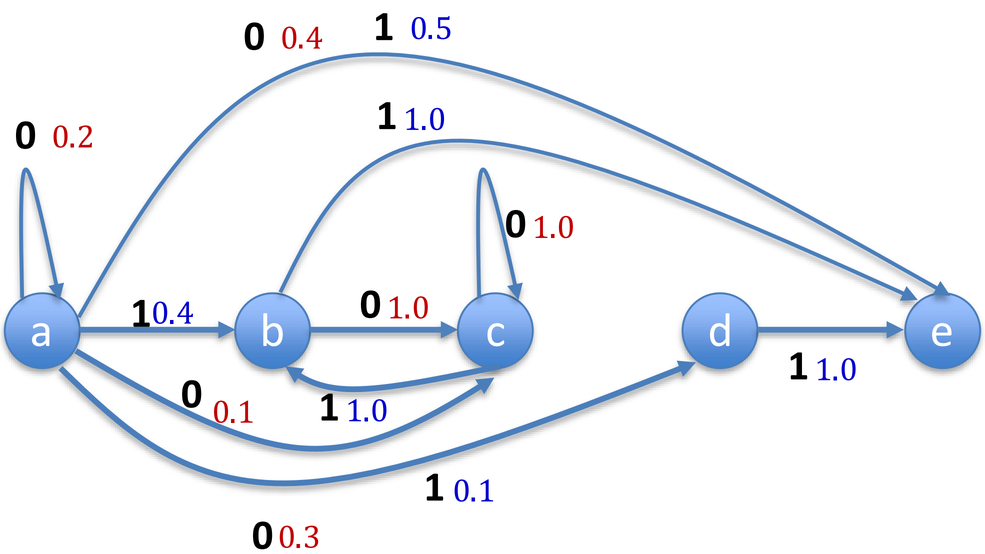


0 0 0 1 0 1 1

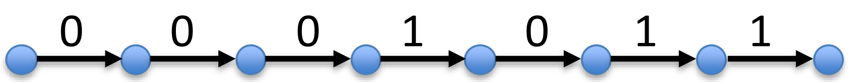


- Convert the string to an FSA
  - Note that the symbols appear on the *edges*
  - This is a *DFA* because the observed string is definitive
    - Will address what happens when we are unsure of observation later

Inference in a PFA

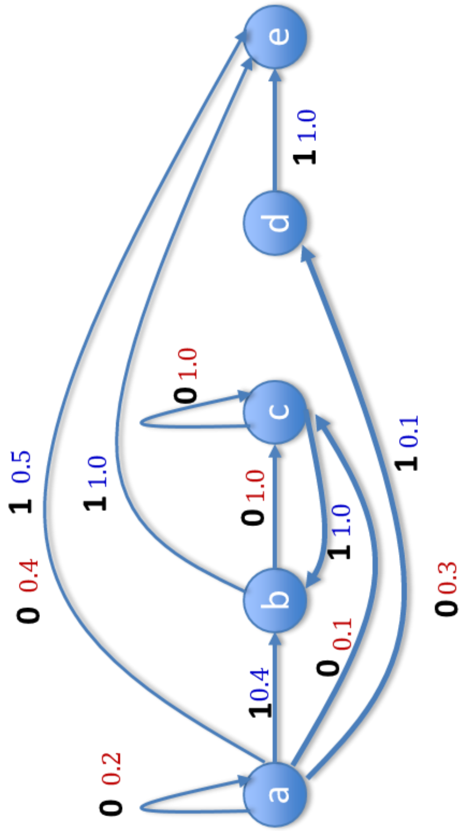


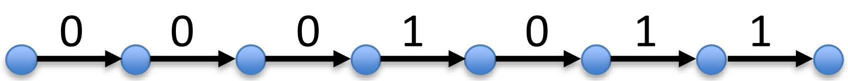
0 0 0 1 0 1 1



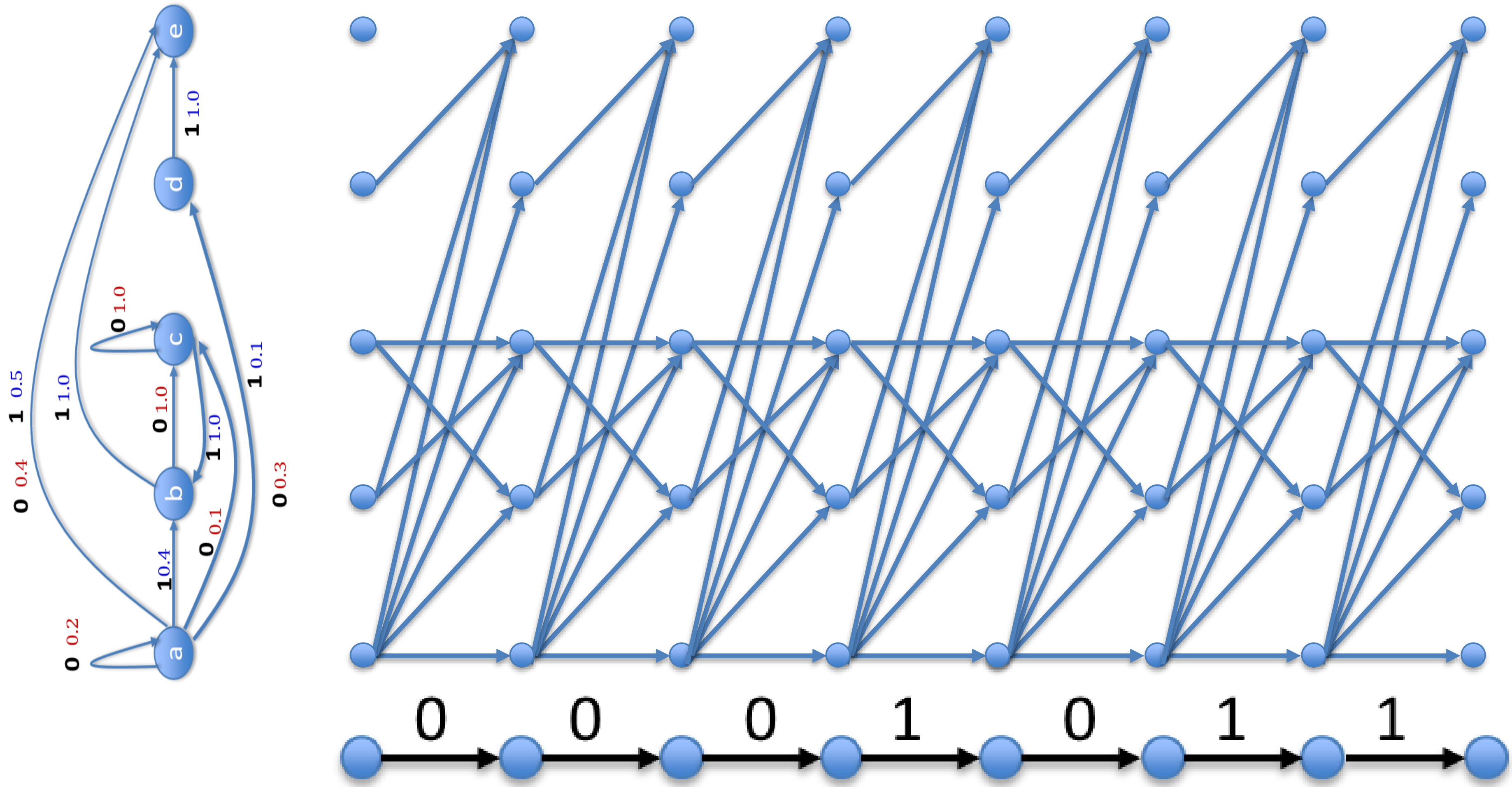
- Redrawing it linearly for illustration..

Inference in a PFA

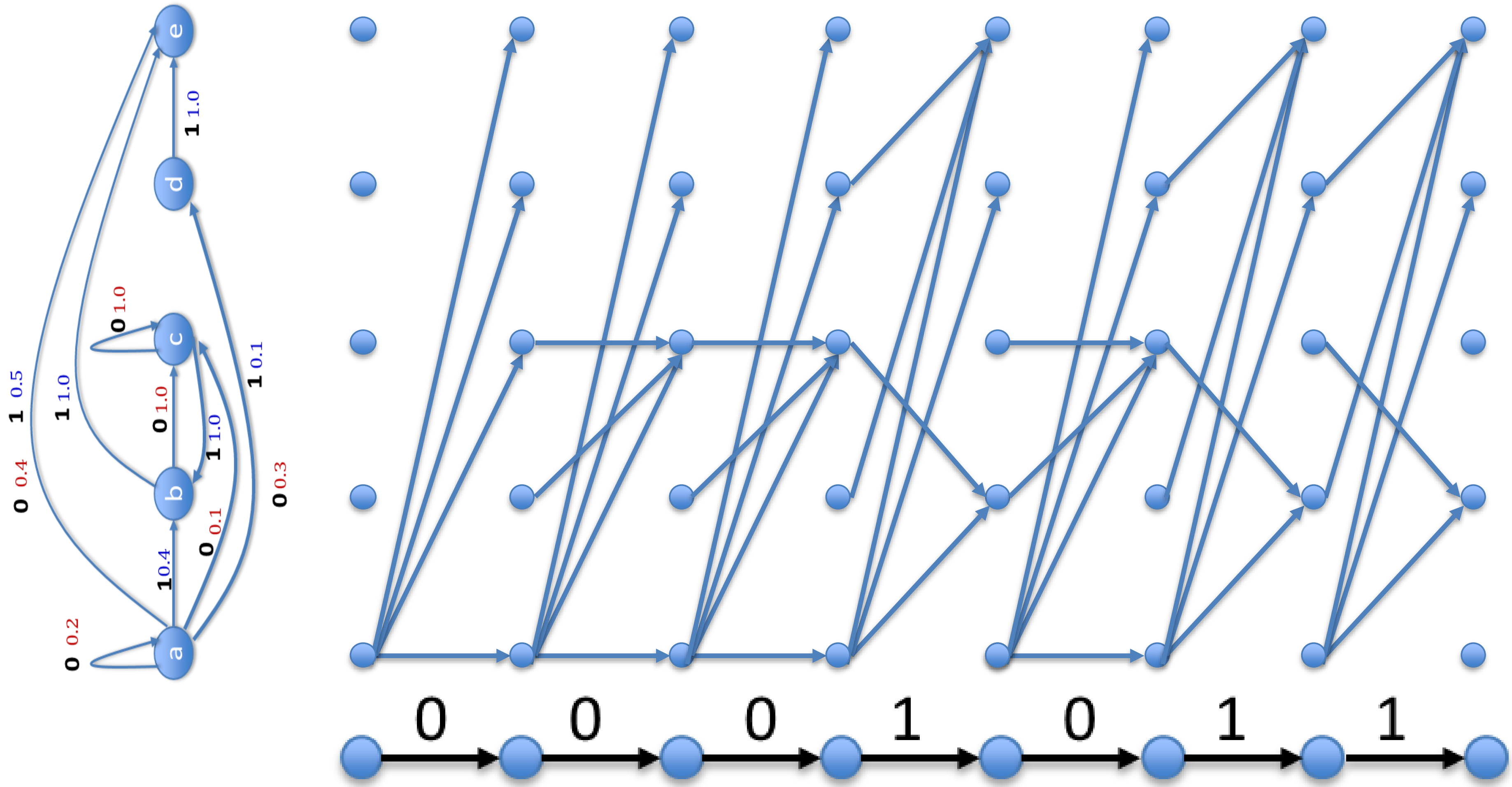




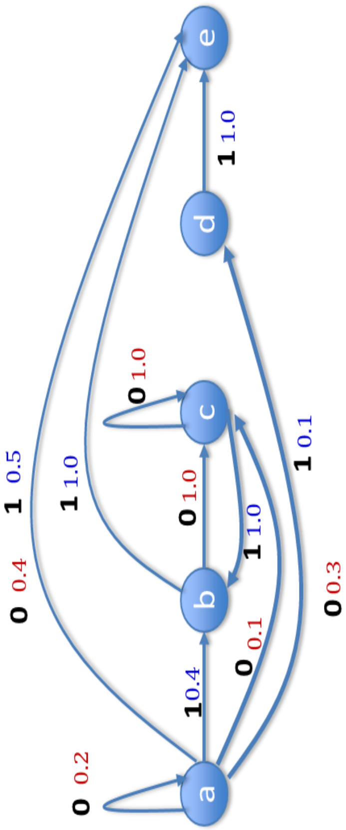
- Redrawing it linearly (and rotating it) for illustration..
  - Lets compose the two graphs!

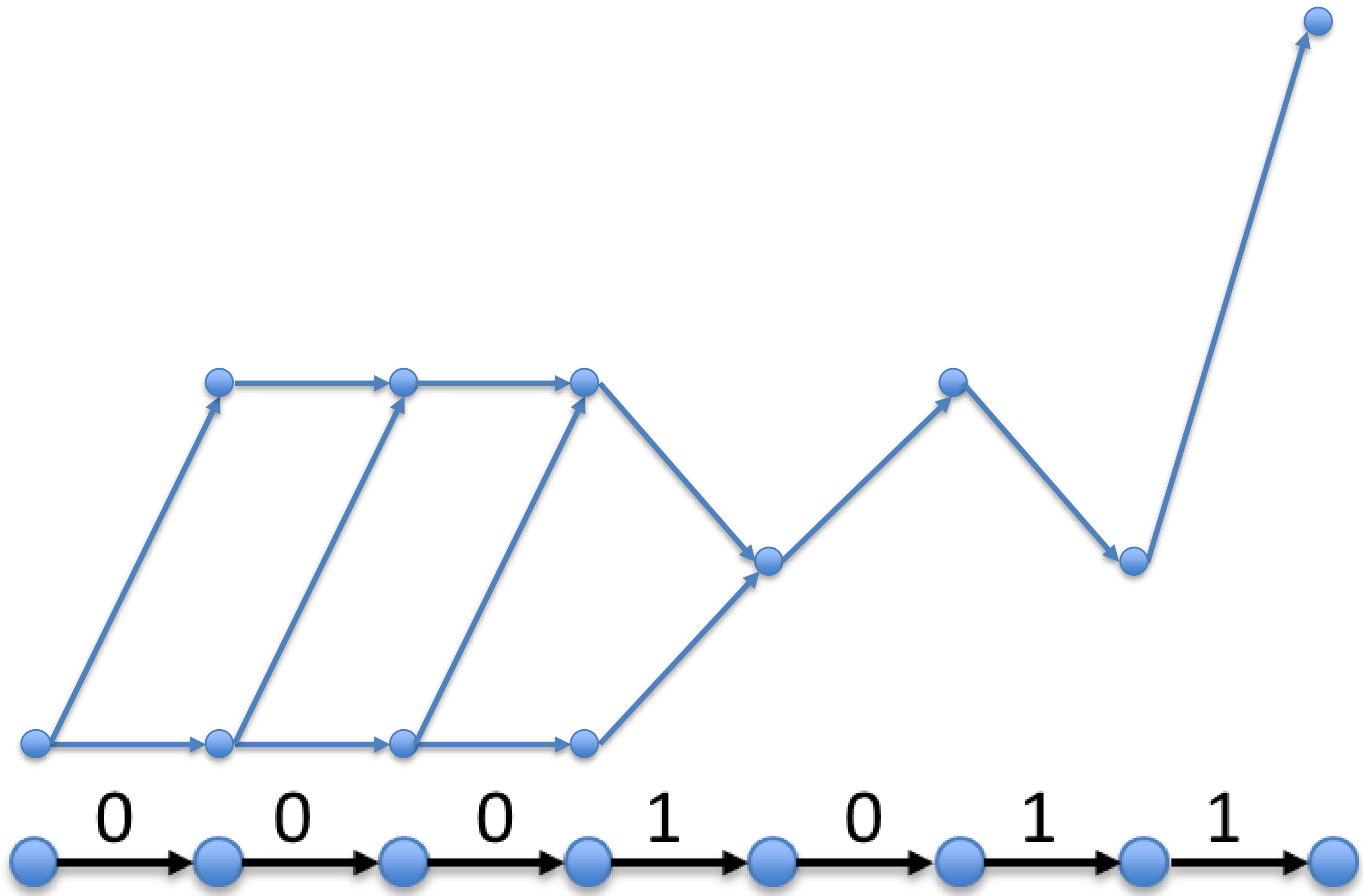


- This graph shows all paths that can consume *any* sequence of seven symbols
- But we are only interested in the paths that consume the actual observation

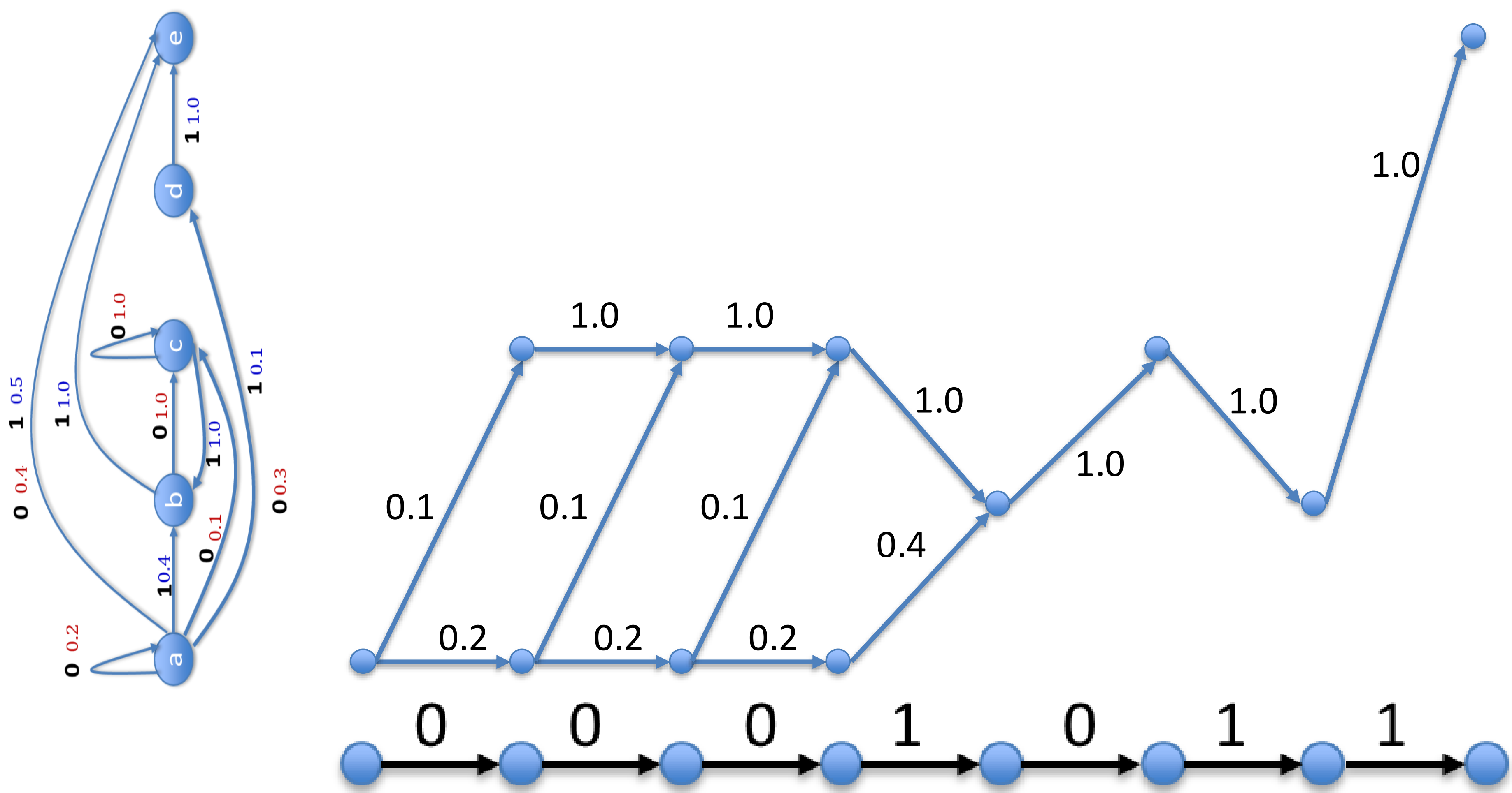


- Cleanup: Eliminate all nodes without incoming edges, and all nodes (except in the last column) without outgoing edges

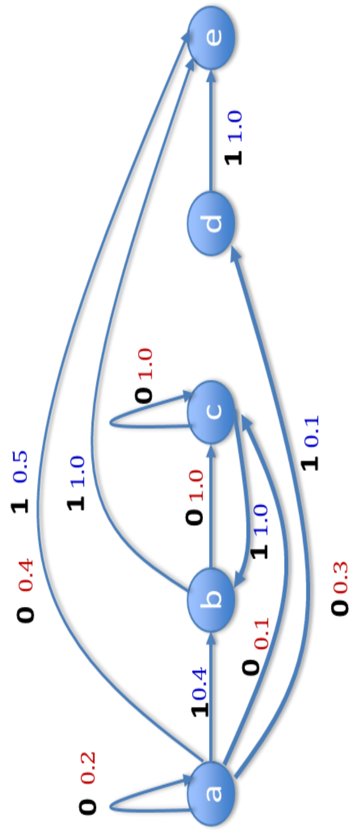


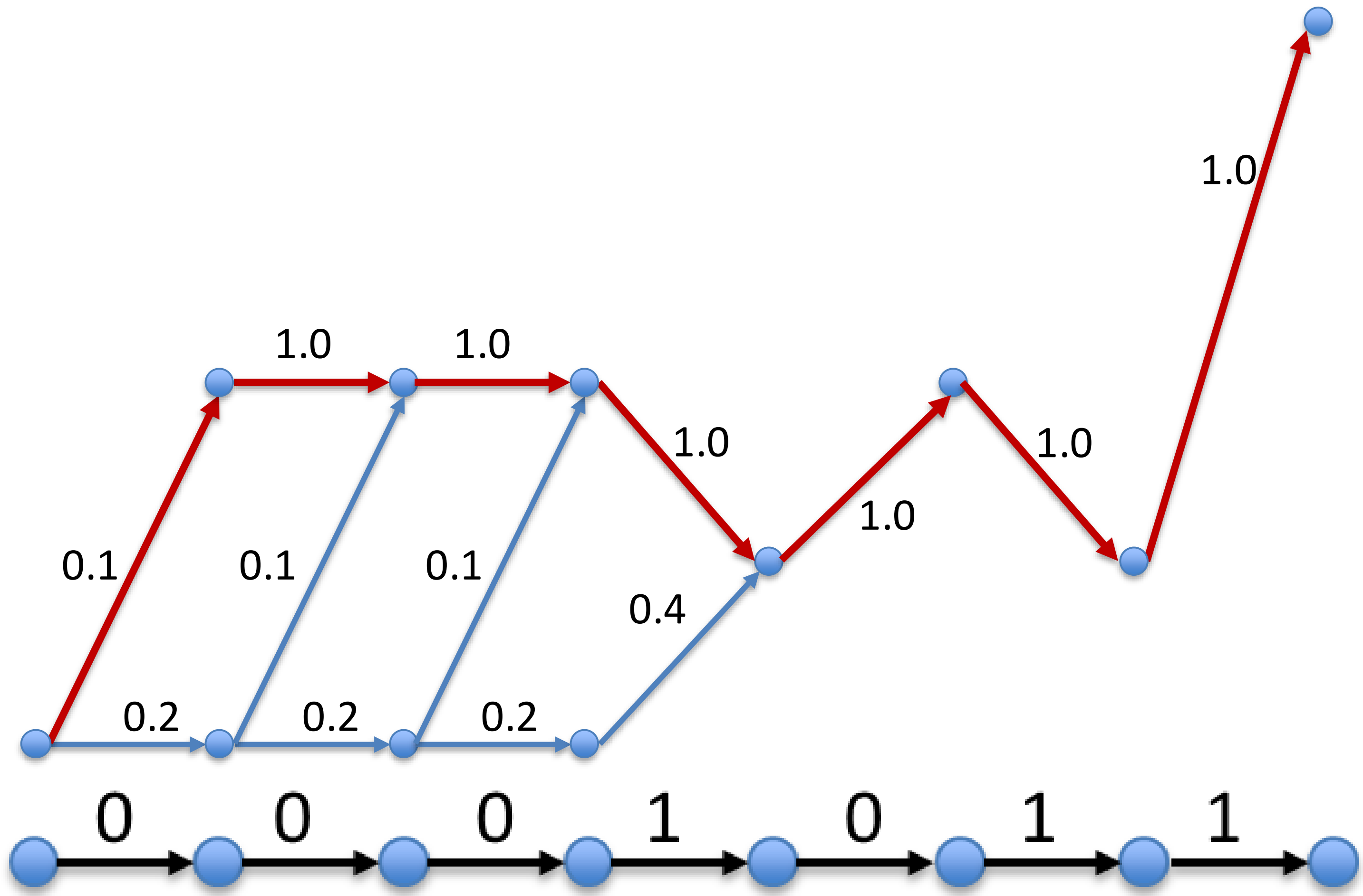


- The complete set of all paths that can absorb the observed sequence
  - But what weights do the edges carry?

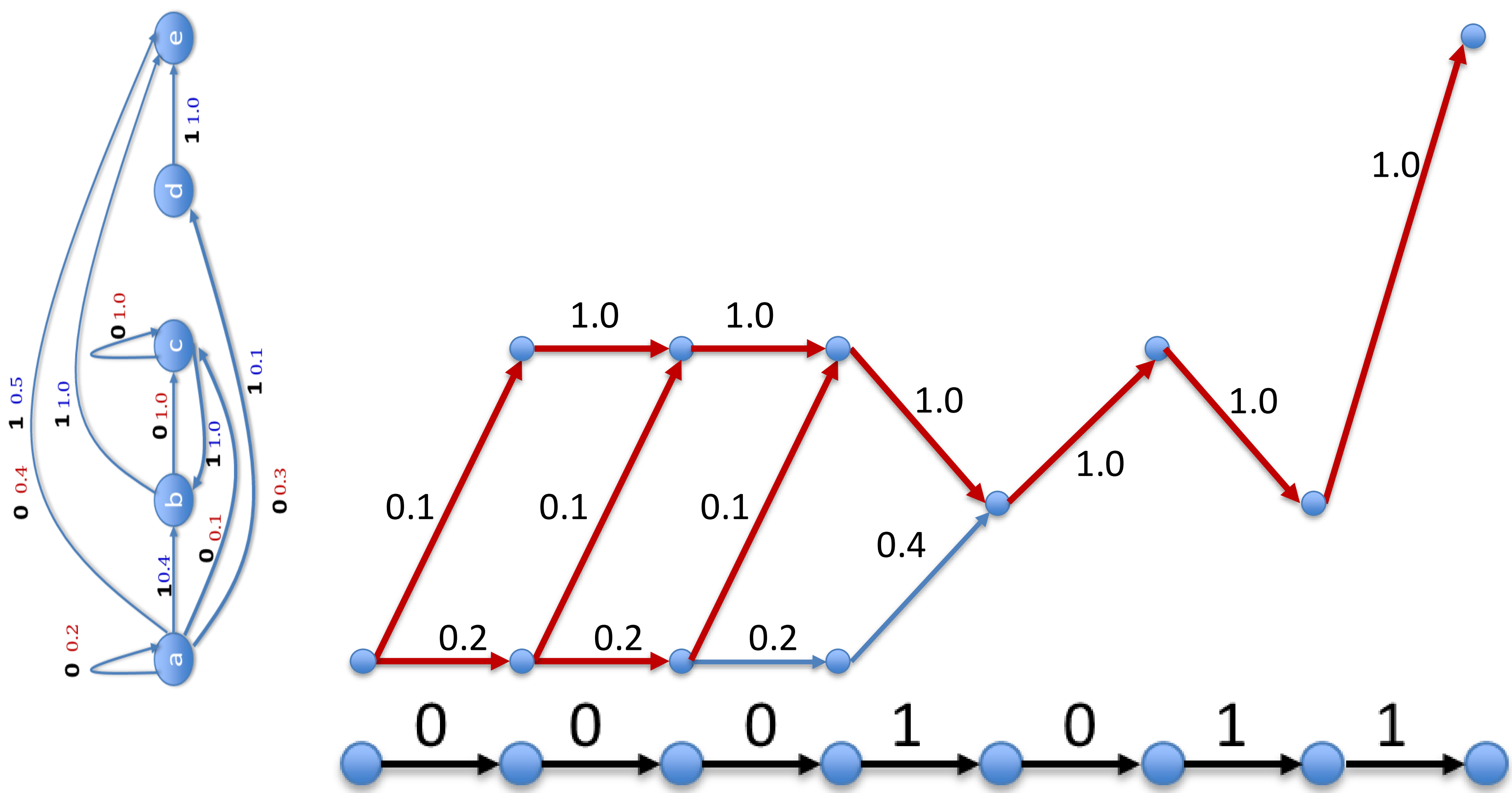


- The complete set of all paths that can absorb the observed sequence
  - Edges carry the probability of the particular symbol absorbed

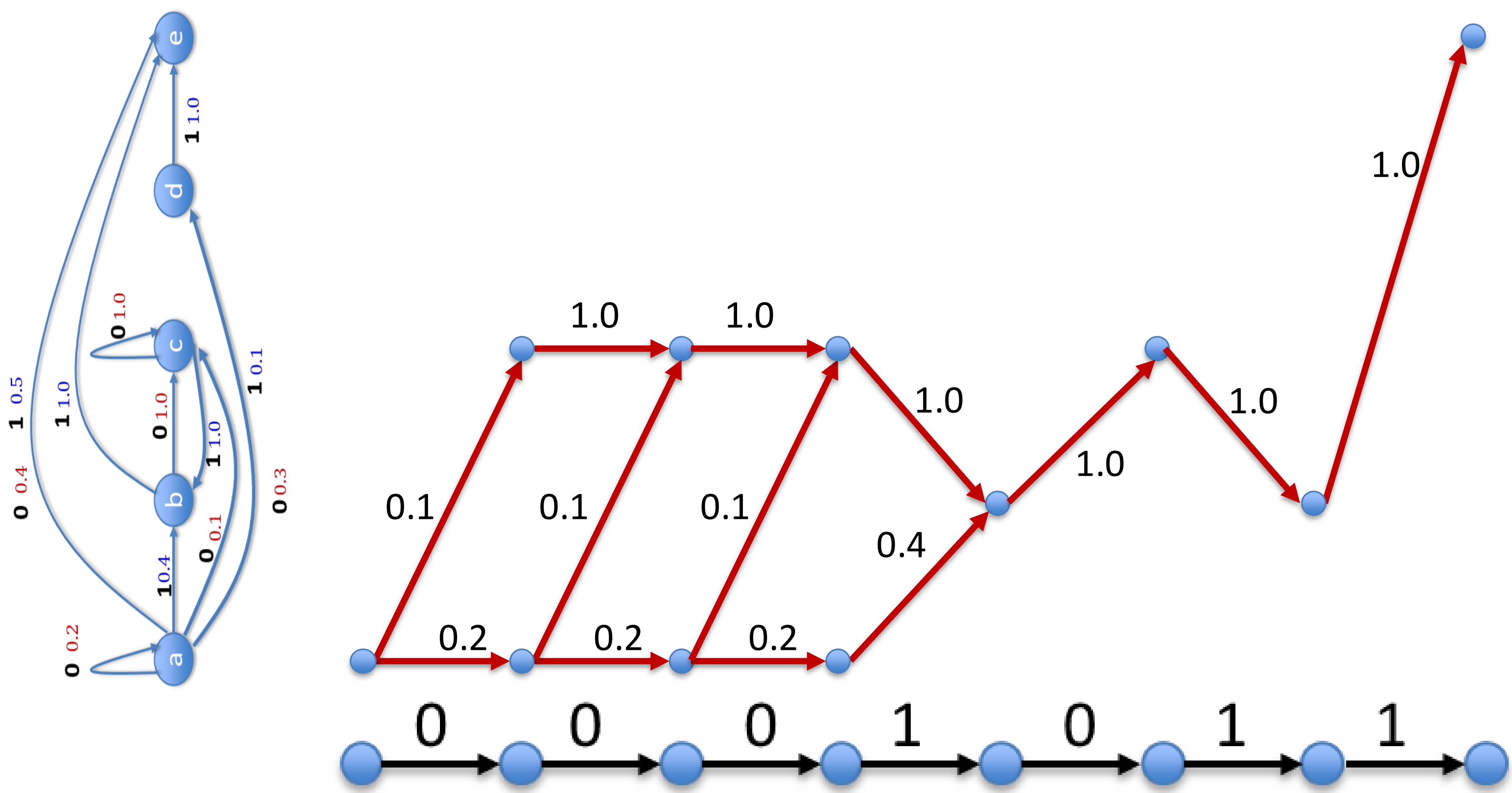




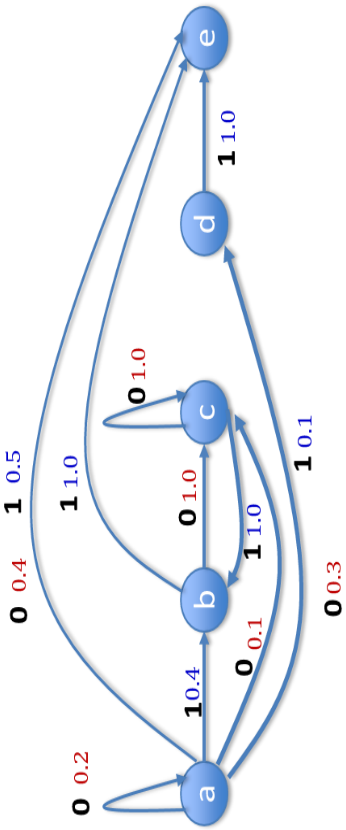
- The probability of any given state sequence is the product of the probabilities on all the edges representing the state sequence
  - The probability of *a c c c b c b e* is 0.1



- The total probability of visiting “c” is the total probability of all paths that go through “c” and end at “e”

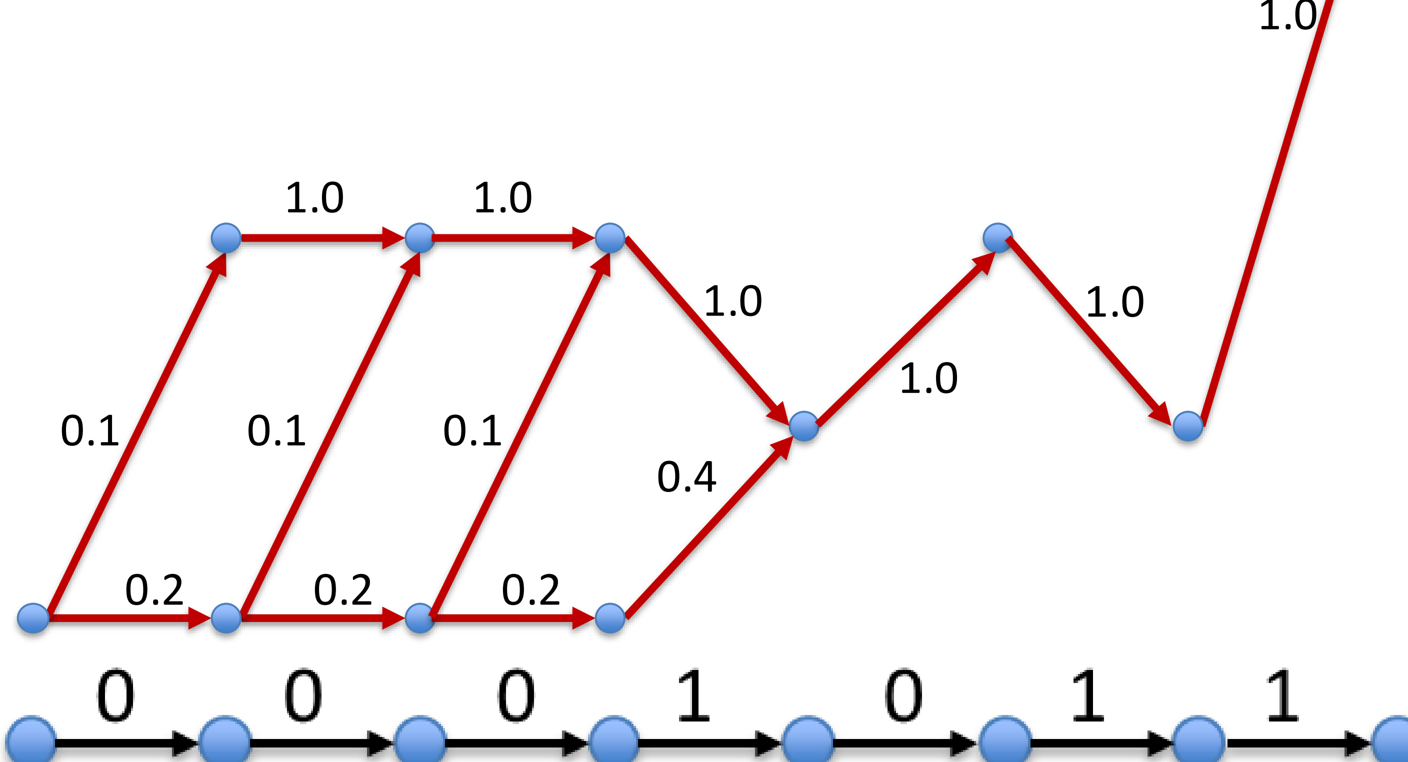


- The total probability of all paths that get to the final state "e" is the probability of the entire graph



The actual algorithm to compute these probabilities is the forward-backward algorithm

We will get to that shortly

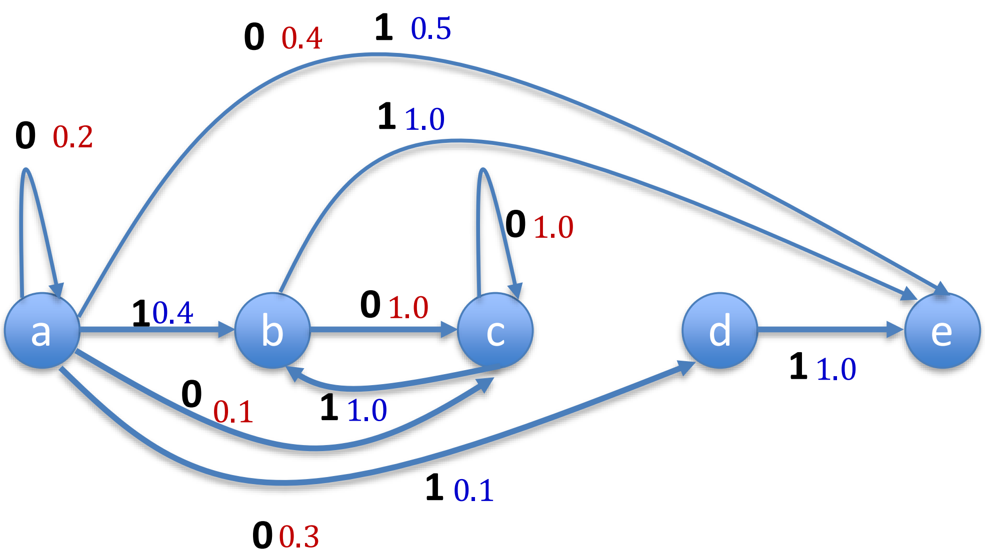


- The total probability of all paths that get to the final state "e" is the probability of the entire graph

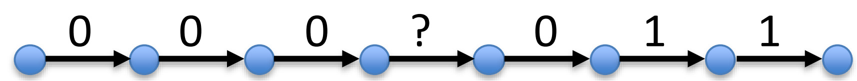
Composition and computation

- Composition and computation can be done dynamically as one processes the input string
- Alternately, one may use any of the FSA composition algorithms in the literature (and tools available on the web)
  - These can be highly efficient

Dealing with uncertainty..

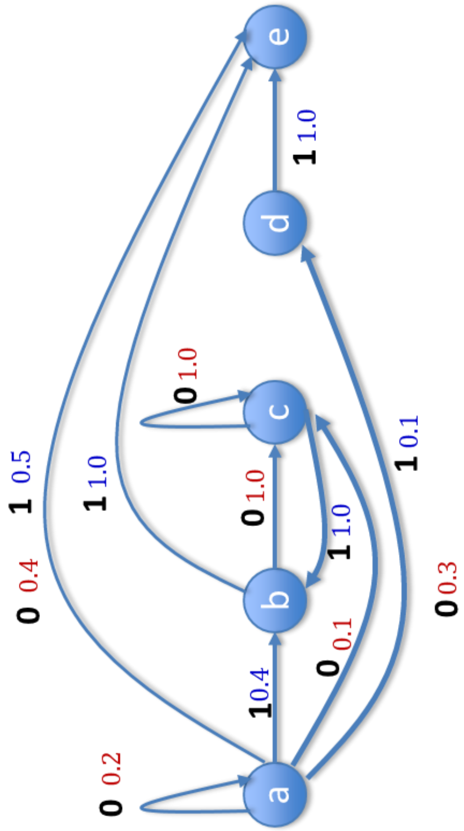


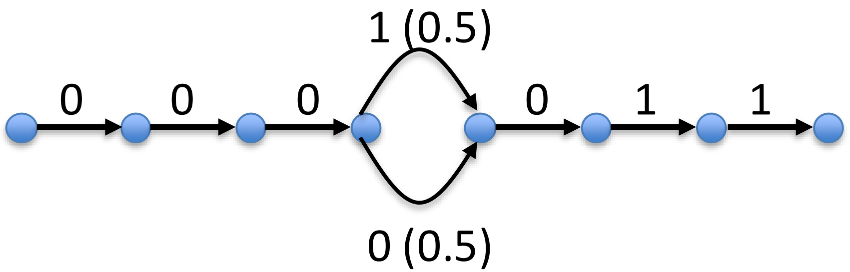
0 0 0 ? 0 1 1



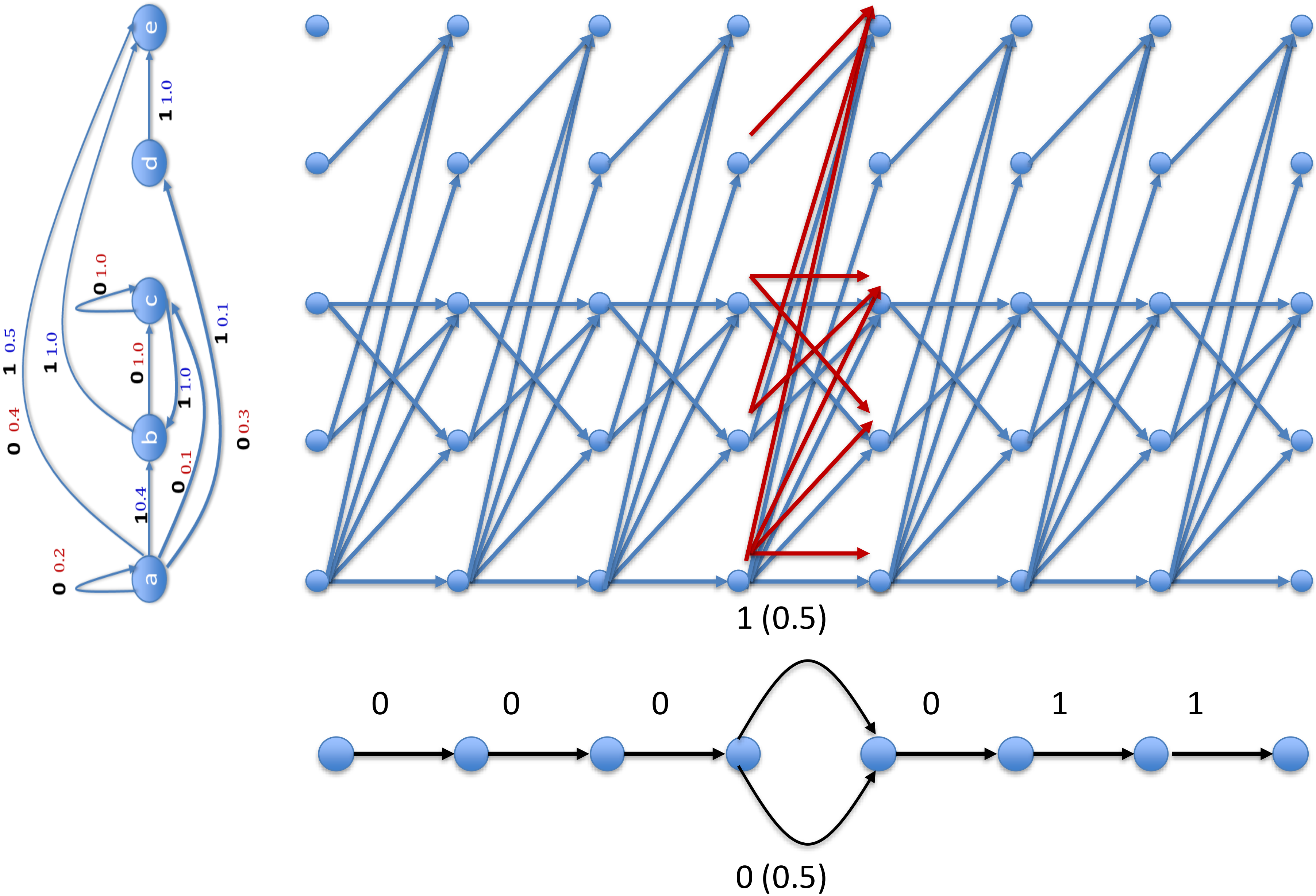
- Easily adapted to deal with uncertainty..

Inference in a PFA





- Uncertainty is reflected in the input string
- The rest of the process remains largely unchanged


e
d
c
b
a
1 1.0
0 1.0
0 1.0
1 1.0
1 0.4
0 0.1
1 0.1
0 0.3
1 0.5
0 0.4
1 1.0
0 0.2
1 (0.5)
0
0
0
0
1
1
0 (0.5)

Moving on: Generative models

- Hidden Markov Models
  - “Stochastic functions of Markov Chains”

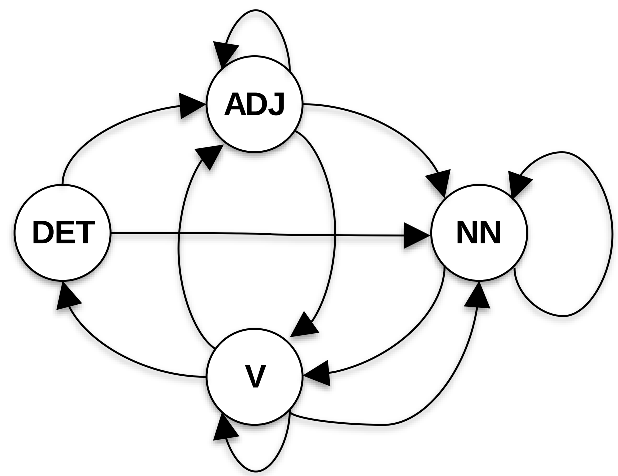
- E.g. a finite-state automaton over tags, that can generate word sequences

Initial Probabilities:

DET	ADJ	NN	V
0.5	0.1	0.3	0.1

η Transition Probabilities:

	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
	0.0	0.0	0.2	0.1



γ Emission Probabilities:

DET		ADJ		NN		V	
the	0.7	green	0.1	book	0.3	might	0.2
a	0.3	big	0.4	plants	0.2	watch	0.3
		old	0.4	people	0.2	watches	0.2
		might	0.1	person	0.1	loves	0.1
				John	0.1	reads	0.19
				watch	0.1	books	0.01

Examples:

John	might	watch
NN	**V**	**V**

the	old	person	loves	big	books
DET	**ADJ**	**NN**	**V**	**ADJ**	**NN**

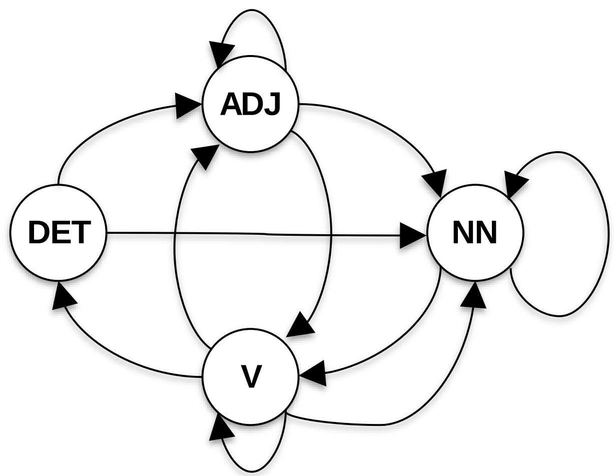
String Marginals

- Inference question for HMMs
  - What is the probability of a string **w**?
    Answer: generate all possible tag sequences and explicitly *marginalize*

$$O(|\Omega|^{|\mathbf{w}|}) \quad \text{time}$$

Initial Probabilities:

DET	ADJ	NN	V
0.5	0.1	0.3	0.1



η **Transition Probabilities:**

	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
	0.0	0.0	0.2	0.1

γ **Emission Probabilities:**

DET		**ADJ**		**NN**		**V**	
the	0.7	green	0.1	book	0.3	might	0.2
a	0.3	big	0.4	plants	0.2	watch	0.3
		old	0.4	people	0.2	watches	0.2
		might	0.1	person	0.1	loves	0.1
				John	0.1	reads	0.19
				watch	0.1	books	0.01

Examples:

John	might	watch
NN	**V**	**V**

the	old	person	loves	big	books
DET	**ADJ**	**NN**	**V**	**ADJ**	**NN**

John	**might**	**watch**	$\Pr(x, y)$
DET	DET	DET	0.0
DET	DET	ADJ	0.0
DET	DET	NN	0.0
DET	DET	V	0.0
DET	ADJ	DET	0.0
DET	ADJ	ADJ	0.0
DET	ADJ	NN	0.0
DET	ADJ	V	0.0
DET	NN	DET	0.0
DET	NN	ADJ	0.0
DET	NN	NN	0.0
DET	NN	V	0.0
DET	V	DET	0.0
DET	V	ADJ	0.0
DET	V	NN	0.0
DET	V	V	0.0

John	**might**	**watch**	$\Pr(x, y)$
ADJ	DET	DET	0.0
ADJ	DET	ADJ	0.0
ADJ	DET	NN	0.0
ADJ	DET	V	0.0
ADJ	ADJ	DET	0.0
ADJ	ADJ	ADJ	0.0
ADJ	ADJ	NN	0.0
ADJ	ADJ	V	0.0
ADJ	NN	DET	0.0
ADJ	NN	ADJ	0.0
ADJ	NN	NN	0.0
ADJ	NN	V	0.0
ADJ	V	DET	0.0
ADJ	V	ADJ	0.0
ADJ	V	NN	0.0
ADJ	V	V	0.0

John	**might**	**watch**	$\Pr(x, y)$
NN	DET	DET	0.0
NN	DET	ADJ	0.0
NN	DET	NN	0.0
NN	DET	V	0.0
NN	ADJ	DET	0.0
NN	ADJ	ADJ	0.0
NN	**ADJ**	**NN**	**0.0000042**
NN	**ADJ**	**V**	**0.0000009**
NN	NN	DET	0.0
NN	NN	ADJ	0.0
NN	NN	NN	0.0
NN	NN	V	0.0
NN	V	DET	0.0
NN	V	ADJ	0.0
NN	***V***	***NN***	***0.0000096***
NN	**V**	**V**	**0.0000072**

John	**might**	**watch**	$\Pr(x, y)$
V	DET	DET	0.0
V	DET	ADJ	0.0
V	DET	NN	0.0
V	DET	V	0.0
V	ADJ	DET	0.0
V	ADJ	ADJ	0.0
V	ADJ	NN	0.0
V	ADJ	V	0.0
V	NN	DET	0.0
V	NN	ADJ	0.0
V	NN	NN	0.0
V	NN	V	0.0
V	V	DET	0.0
V	V	ADJ	0.0
V	V	NN	0.0
V	V	V	0.0

John	**might**	**watch**	$\mathrm{Pr}(x, y)$	**John**	**might**	**watch**	$\mathrm{Pr}(x, y)$	**John**	**might**	**watch**	$\mathrm{Pr}(x, y)$	**John**	**might**	**watch**	$\mathrm{Pr}(x, y)$
DET	DET	DET	0.0	ADJ	DET	DET	0.0	NN	DET	DET	0.0	V	DET	DET	0.0
DET	DET	ADJ	0.0	ADJ	DET	ADJ	0.0	NN	DET	ADJ	0.0	V	DET	ADJ	0.0
DET	DET	NN	0.0	ADJ	DET	NN	0.0	NN	DET	NN	0.0	V	DET	NN	0.0
DET	DET	V	0.0	ADJ	DET	V	0.0	NN	DET	V	0.0	V	DET	V	0.0
DET	ADJ	DET	0.0	ADJ	ADJ	DET	0.0	NN	ADJ	DET	0.0	V	ADJ	DET	0.0
DET	ADJ	ADJ	0.0	ADJ	ADJ	ADJ	0.0	NN	ADJ	ADJ	0.0	V	ADJ	ADJ	0.0
DET	ADJ	NN	0.0	ADJ	ADJ	NN	0.0	**NN**	**ADJ**	**NN**	**0.0000042**	V	ADJ	NN	0.0
DET	ADJ	V	0.0	ADJ	ADJ	V	0.0	**NN**	**ADJ**	**V**	**0.0000009**	V	ADJ	V	0.0
DET	NN	DET	0.0	ADJ	NN	DET	0.0	NN	NN	DET	0.0	V	NN	DET	0.0
DET	NN	ADJ	0.0	ADJ	NN	ADJ	0.0	NN	NN	ADJ	0.0	V	NN	ADJ	0.0
DET	NN	NN	0.0	ADJ	NN	NN	0.0	NN	NN	NN	0.0	V	NN	NN	0.0
DET	NN	V	0.0	ADJ	NN	V	0.0	NN	NN	V	0.0	V	NN	V	0.0
DET	V	DET	0.0	ADJ	V	DET	0.0	NN	V	DET	0.0	V	V	DET	0.0
DET	V	ADJ	0.0	ADJ	V	ADJ	0.0	NN	V	ADJ	0.0	V	V	ADJ	0.0
DET	V	NN	0.0	ADJ	V	NN	0.0	***NN***	***V***	***NN***	***0.0000096***	V	V	NN	0.0
DET	V	V	0.0	ADJ	V	V	0.0	**NN**	**V**	**V**	**0.0000072**	V	V	V	0.0

$$p = 0.0000219$$

John	might	watch	$\Pr(x,y)$
DET	DET	DET	0.0
DET	DET	ADJ	0.0
DET	DET	NN	0.0
DET	DET	V	0.0
DET	ADJ	DET	0.0
DET	ADJ	ADJ	0.0
DET	ADJ	NN	0.0
DET	ADJ	V	0.0
DET	NN	DET	0.0
DET	NN	ADJ	0.0
DET	NN	NN	0.0
DET	NN	V	0.0
DET	V	DET	0.0
DET	V	ADJ	0.0
DET	V	NN	0.0
DET	V	V	0.0

John	might	watch	$\Pr(x,y)$
ADJ	DET	DET	0.0
ADJ	DET	ADJ	0.0
ADJ	DET	NN	0.0
ADJ	DET	V	0.0
ADJ	ADJ	DET	0.0
ADJ	ADJ	ADJ	0.0
ADJ	ADJ	NN	0.0
ADJ	ADJ	V	0.0
ADJ	NN	DET	0.0
ADJ	NN	ADJ	0.0
ADJ	NN	NN	0.0
ADJ	NN	V	0.0
ADJ	V	DET	0.0
ADJ	V	ADJ	0.0
ADJ	V	NN	0.0
ADJ	V	V	0.0

John	might	watch	$\Pr(x,y)$
NN	DET	DET	0.0
NN	DET	ADJ	0.0
NN	DET	NN	0.0
NN	DET	V	0.0
NN	ADJ	DET	0.0
NN	ADJ	ADJ	0.0
NN	**ADJ**	**NN**	**0.0000042**
NN	**ADJ**	**V**	**0.0000009**
NN	NN	DET	0.0
NN	NN	ADJ	0.0
NN	NN	NN	0.0
NN	NN	V	0.0
NN	V	DET	0.0
NN	V	ADJ	0.0
NN	***V***	***NN***	***0.0000096***
NN	**V**	**V**	**0.0000072**

John	might	watch	$\Pr(x,y)$
V	DET	DET	0.0
V	DET	ADJ	0.0
V	DET	NN	0.0
V	DET	V	0.0
V	ADJ	DET	0.0
V	ADJ	ADJ	0.0
V	ADJ	NN	0.0
V	ADJ	V	0.0
V	NN	DET	0.0
V	NN	ADJ	0.0
V	NN	NN	0.0
V	NN	V	0.0
V	V	DET	0.0
V	V	ADJ	0.0
V	V	NN	0.0
V	V	V	0.0

Exponential computation, if done naïvely.

$$p = 0.0000219$$

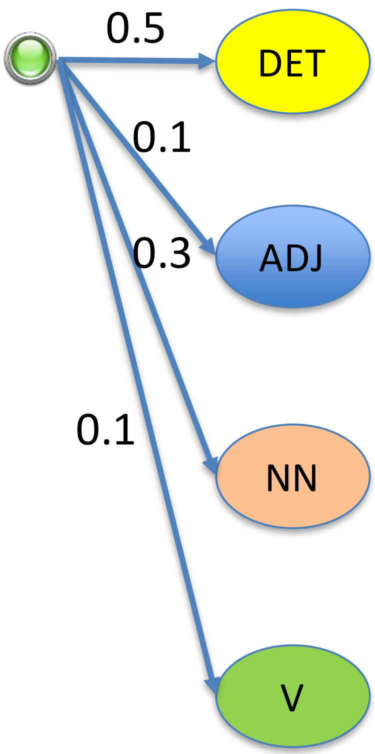
A different perspective

- “Graphical” view of the generative process..

Initial Probabilities:

DET	ADJ	NN	V
0.5	0.1	0.3	0.1

JOHN MIGHT WATCH



Initial Probabilities:

DET	ADJ	NN	V
0.5	0.1	0.3	0.1

JOHN MIGHT WATCH

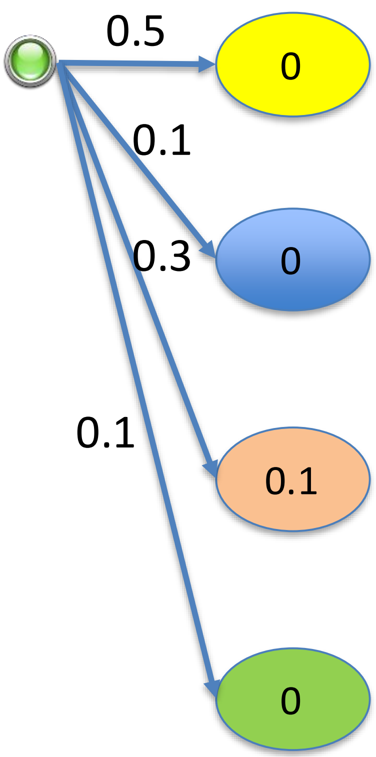
γ **Emission Probabilities:**

DET	
the	0.7
a	0.3

ADJ	
green	0.1
big	0.4
old	0.4
might	0.1

NN	
book	0.3
plants	0.2
people	0.2
person	0.1
John	0.1
watch	0.1

V	
might	0.2
watch	0.3
watches	0.2
loves	0.1
reads	0.19
books	0.01

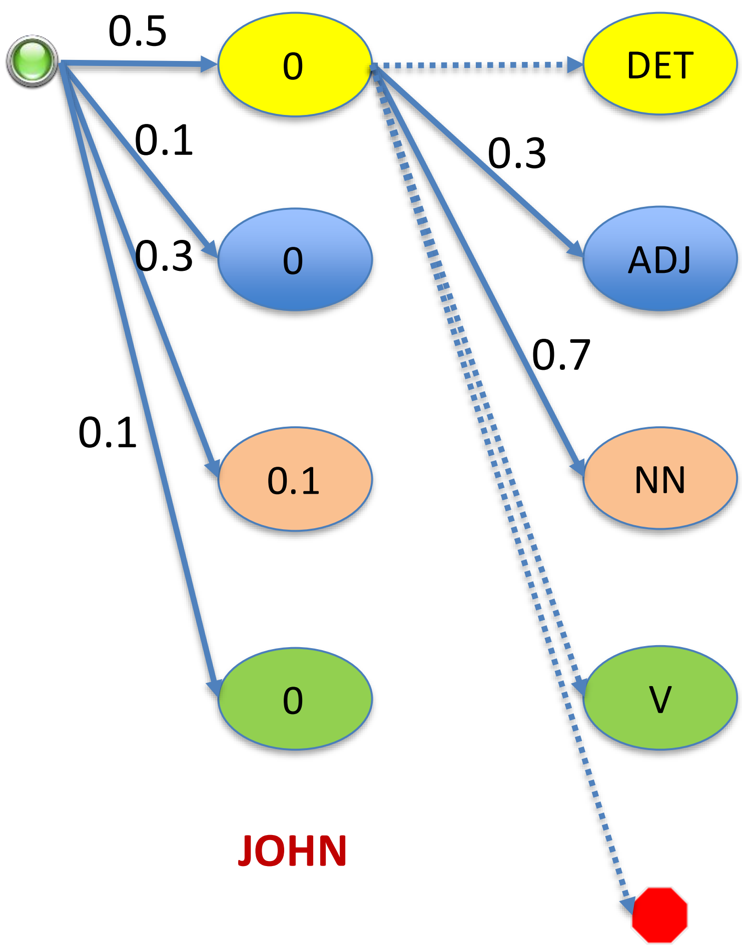


JOHN

Initial Probabilities:

DET	ADJ	NN	V
0.5	0.1	0.3	0.1

JOHN MIGHT WATCH



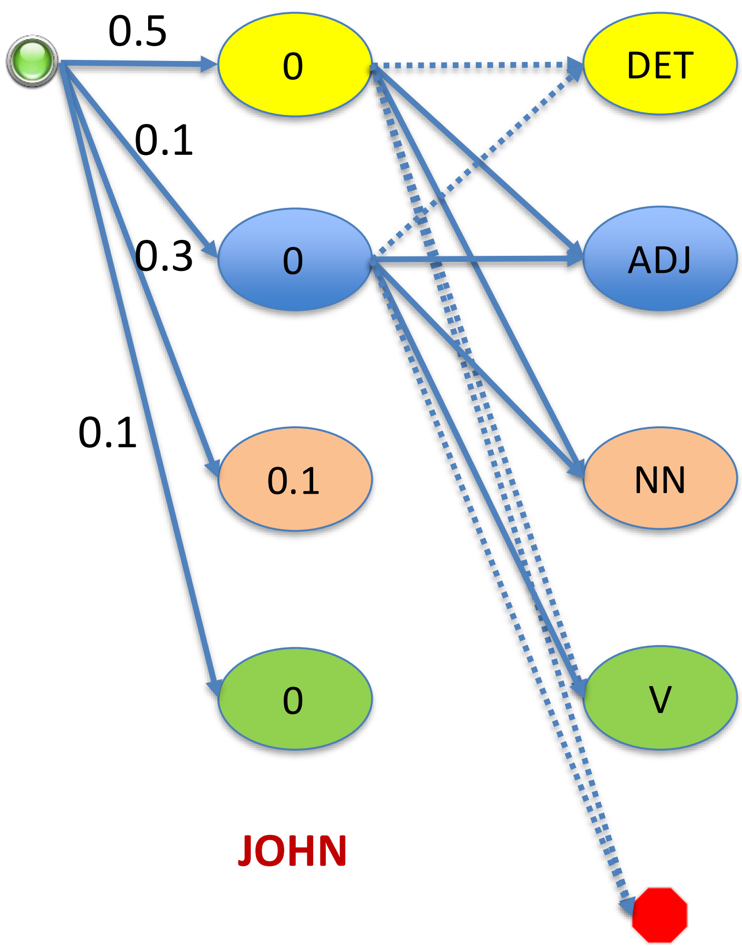
η **Transition Probabilities:**

	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
	0.0	0.0	0.2	0.1

Initial Probabilities:

DET	ADJ	NN	V
0.5	0.1	0.3	0.1

JOHN MIGHT WATCH



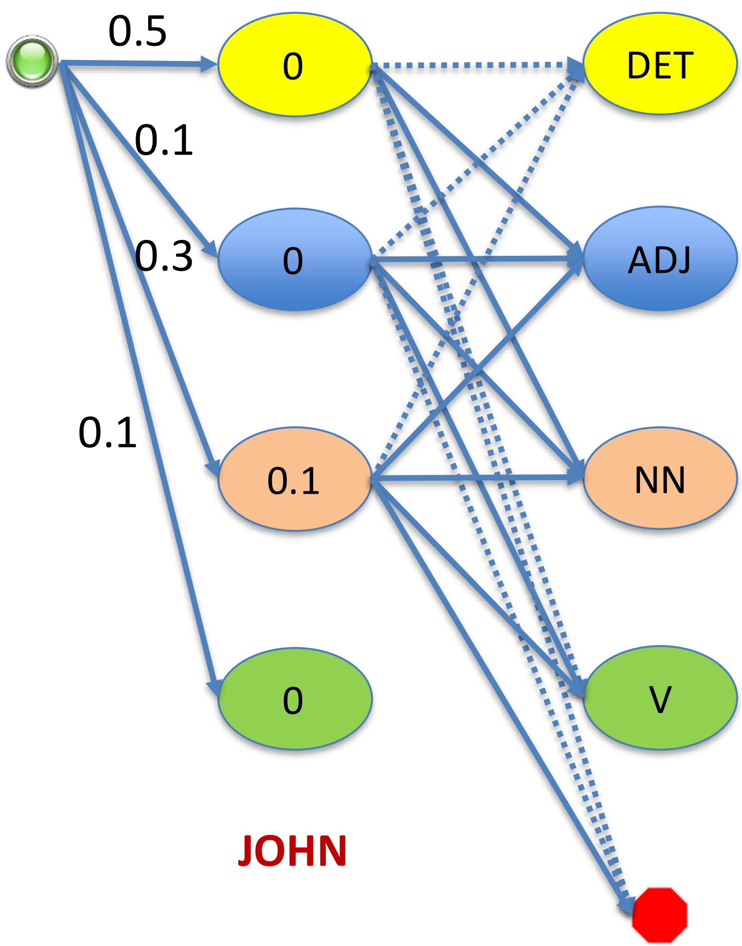
η **Transition Probabilities:**

	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
	0.0	0.0	0.2	0.1

Initial Probabilities:

DET	ADJ	NN	V
0.5	0.1	0.3	0.1

JOHN MIGHT WATCH



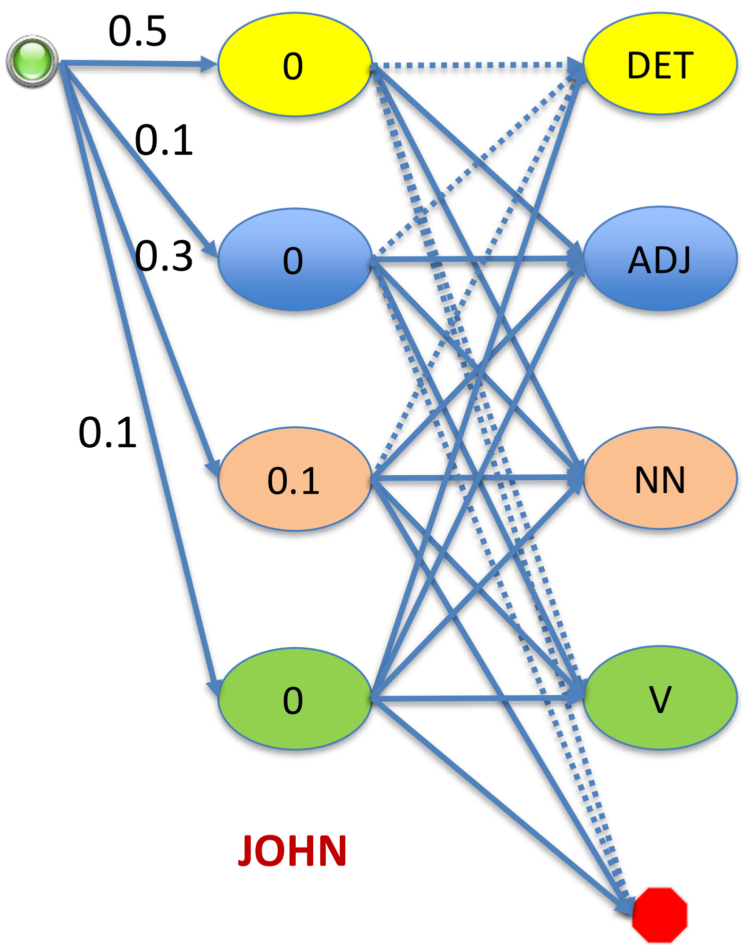
η **Transition Probabilities:**

	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
	0.0	0.0	0.2	0.1

Initial Probabilities:

DET	ADJ	NN	V
0.5	0.1	0.3	0.1

JOHN MIGHT WATCH



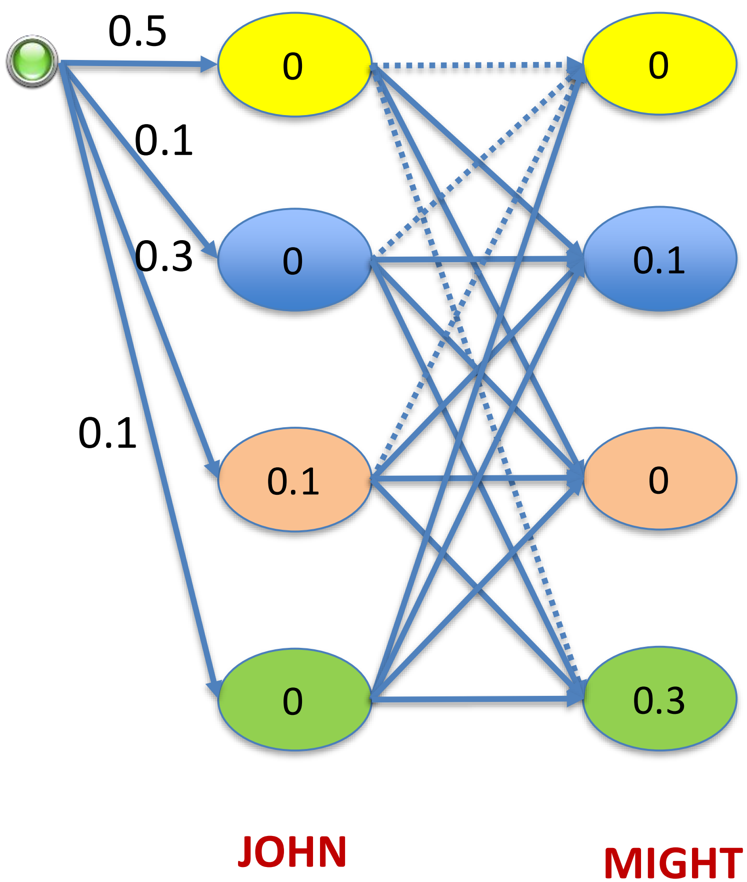
η **Transition Probabilities:**

	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
	0.0	0.0	0.2	0.1

Initial Probabilities:

DET	ADJ	NN	V
0.5	0.1	0.3	0.1

JOHN MIGHT WATCH



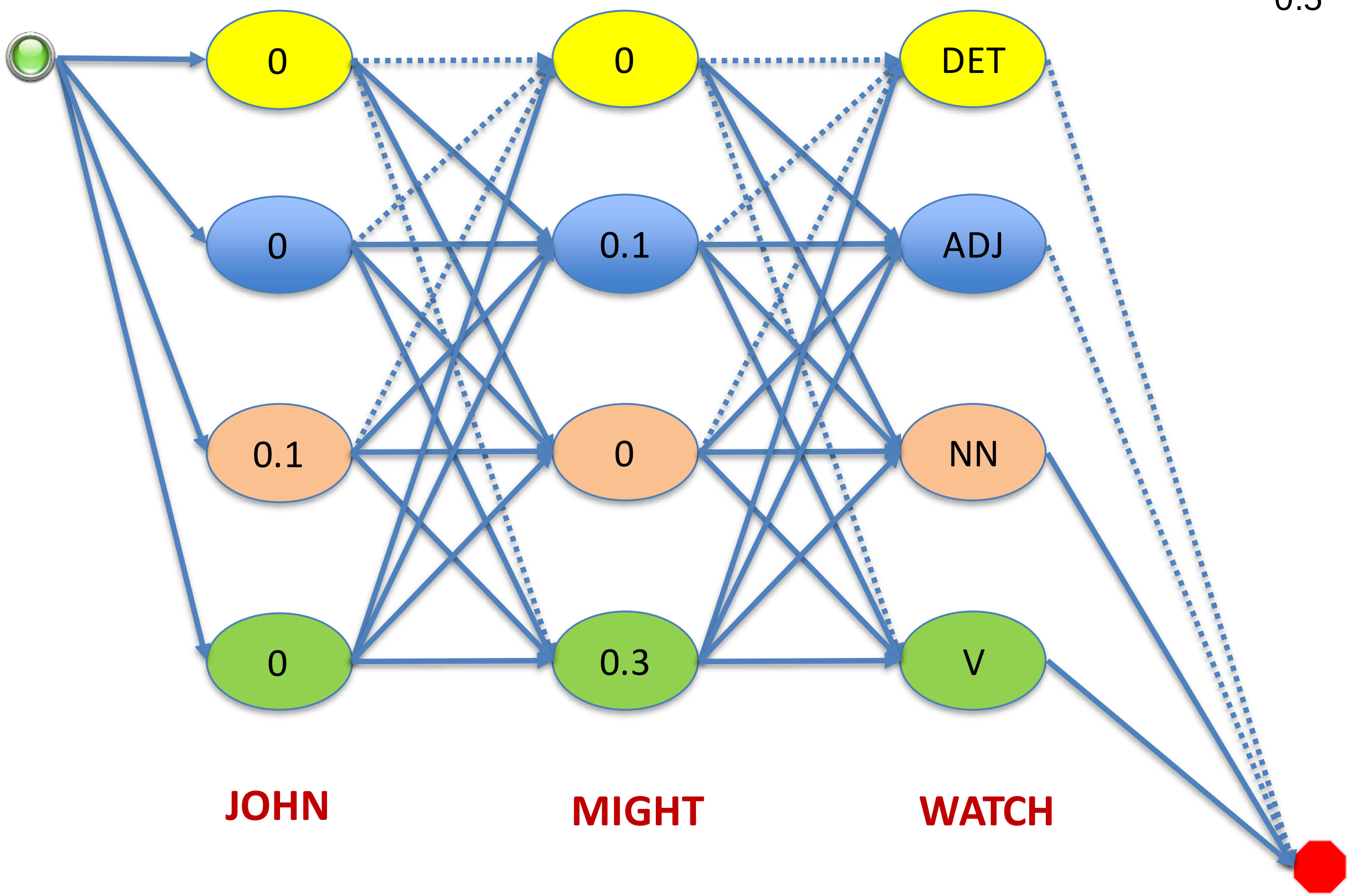
ADJ	
green	0.1
big	0.4
old	0.4
might	0.1

V	
might	0.2
watch	0.3
watches	0.2
loves	0.1
reads	0.19
books	0.01

This is the “trellis” that shows all possible ways of generating the word sequence

Initial Probabilities:

DET	ADJ	NN	V
0.5	0.1	0.3	0.1

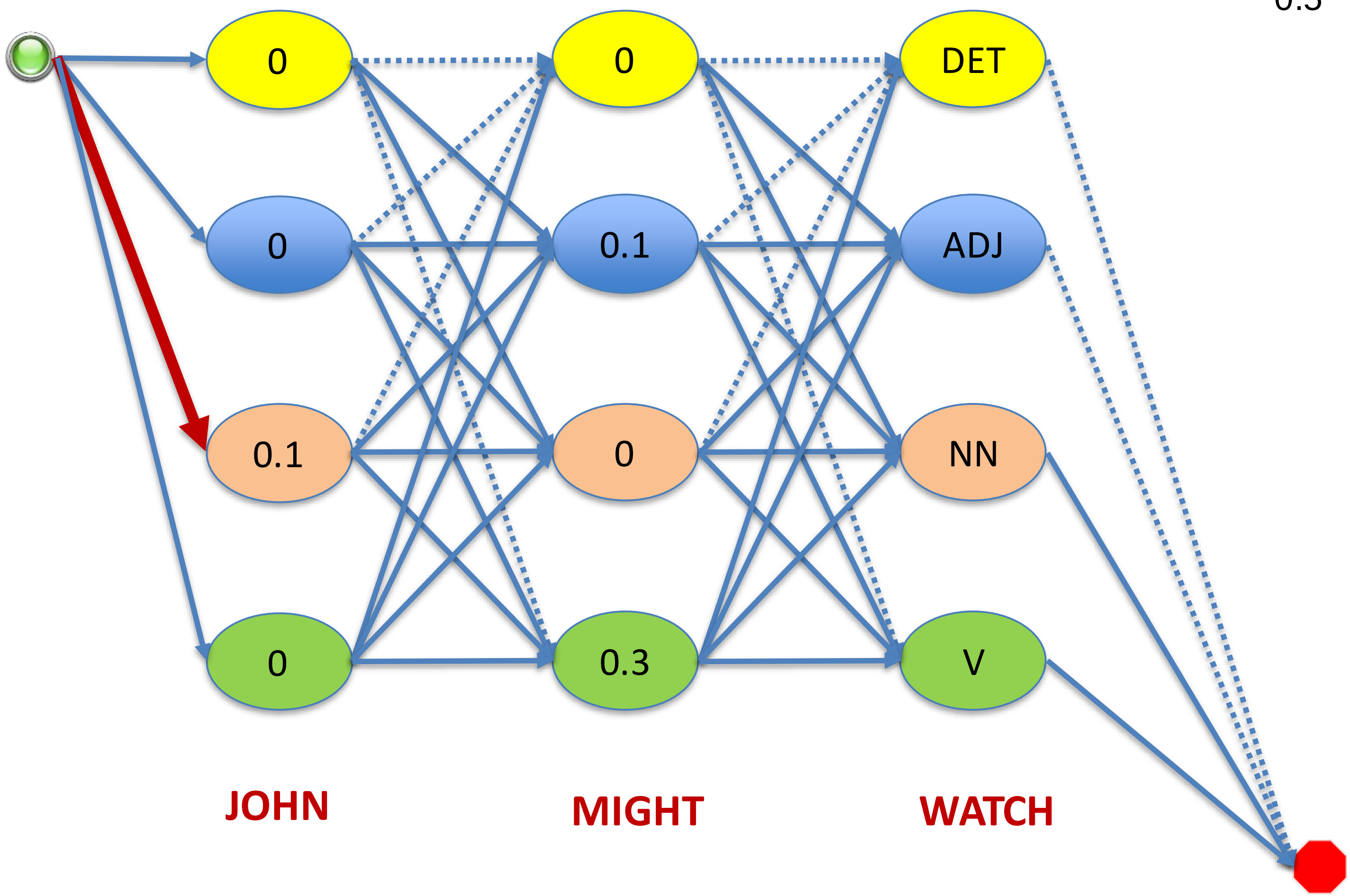


This is the “trellis” that shows all possible ways of generating the word sequence

$P(s_1 = NN) = 0.3$

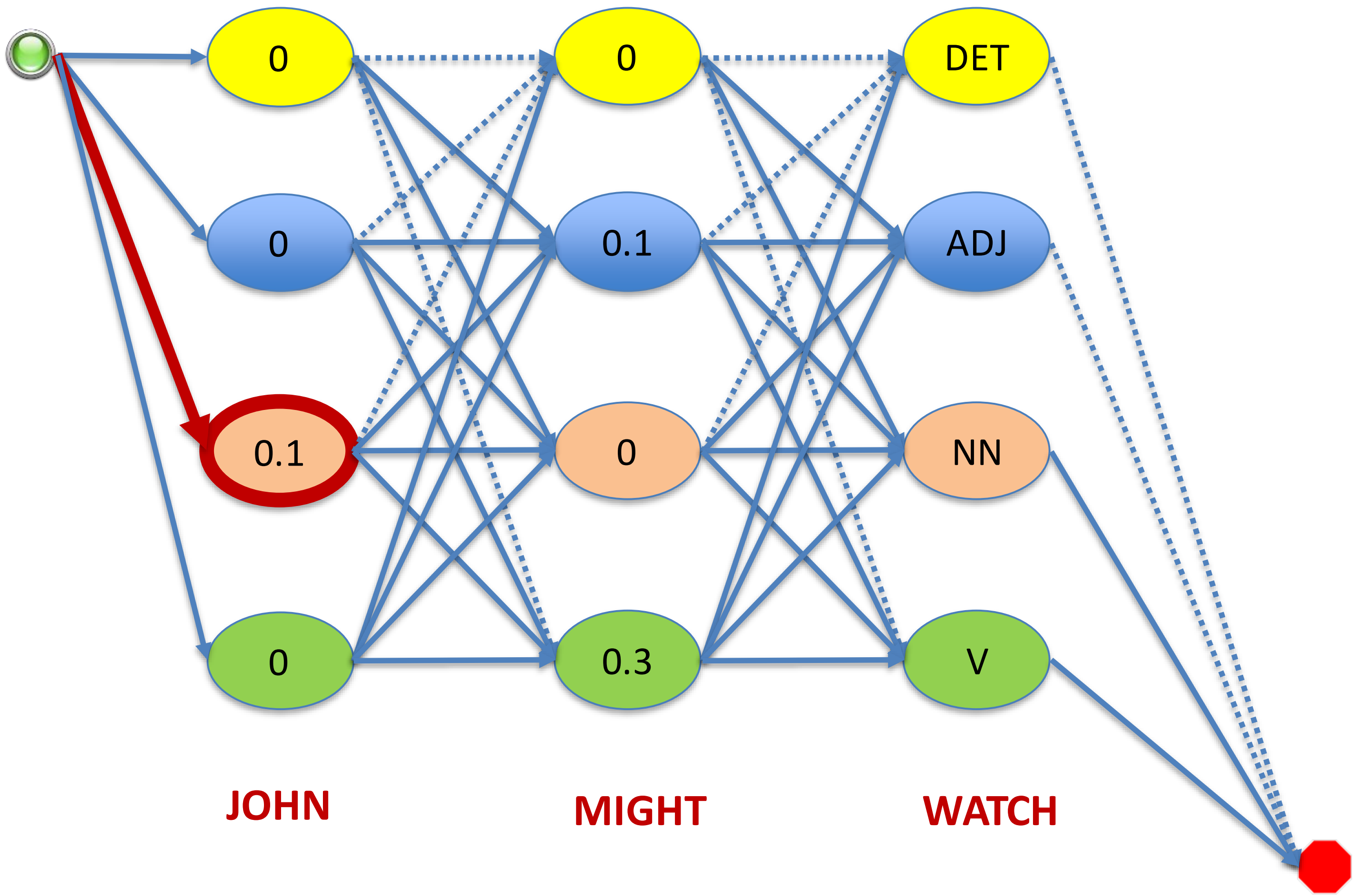
Initial Probabilities:

DET	ADJ	NN	V
0.5	0.1	0.3	0.1



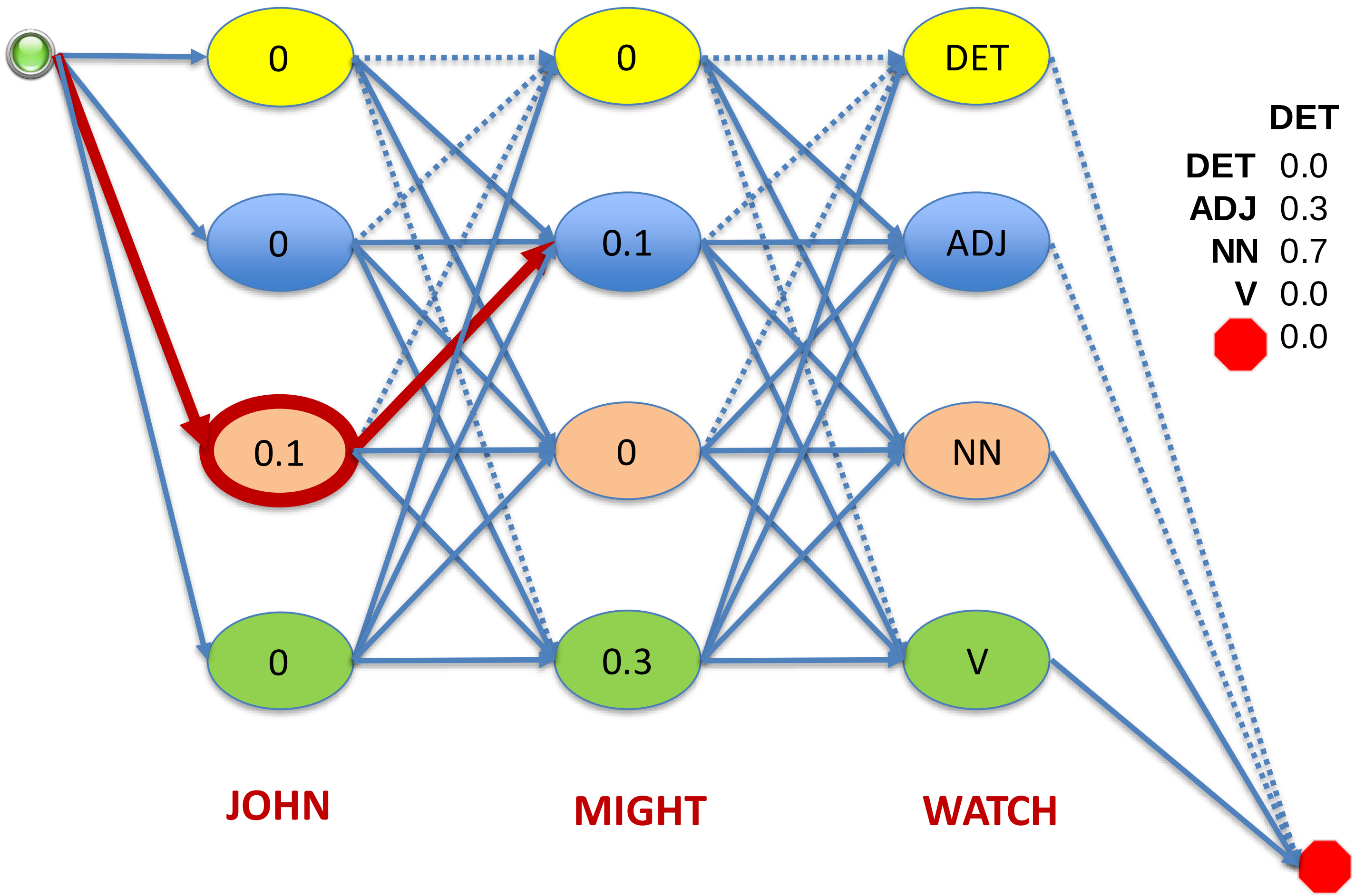
This is the "trellis" that shows all possible ways of generating the word sequence

$$P(s_1 = NN, x_1 = JOHN) = 0.3 \times 0.1$$



This is the “trellis” that shows all possible ways of generating the word sequence

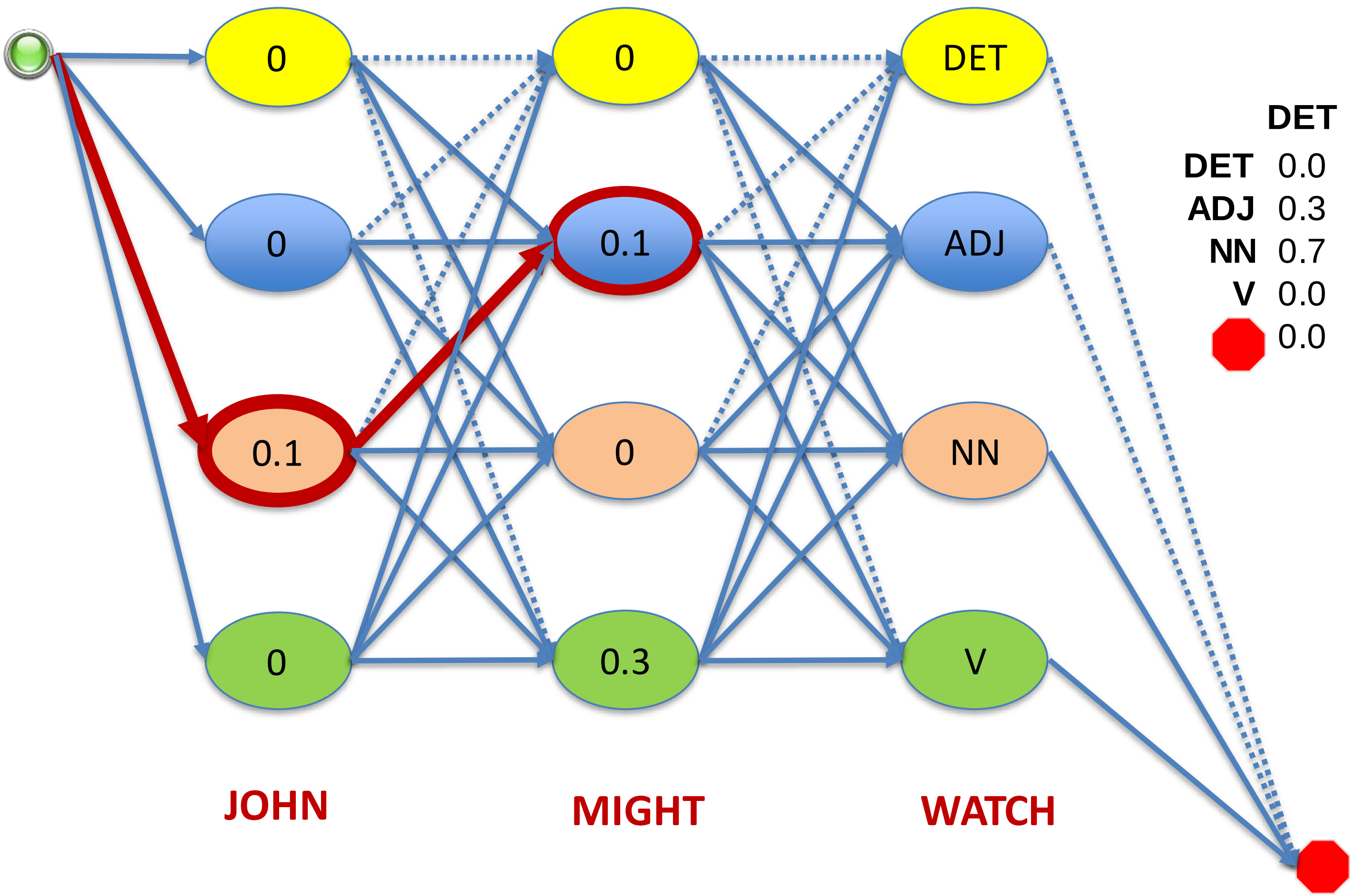
$$P(s_1 = NN, x_1 = JOHN, s_2 = ADJ) = 0.3 \times 0.1 \times 0.1$$



	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
	0.0	0.0	0.2	0.1

This is the “trellis” that shows all possible ways of generating the word sequence

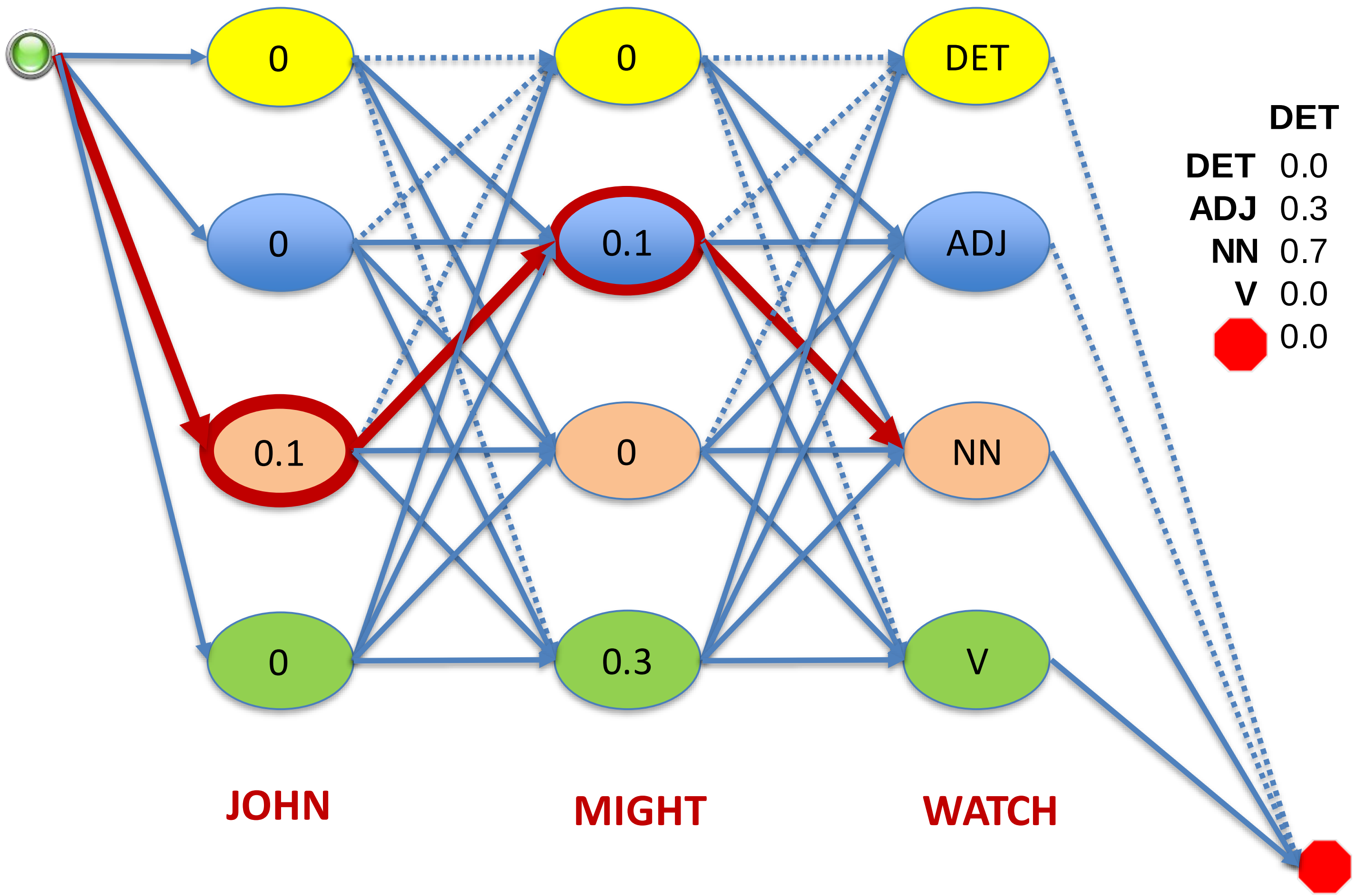
$$P(s_1 = NN, x_1 = JOHN, s_2 = ADJ, x_2 = MIGHT) = 0.3 \times 0.1 \times 0.1 \times 0.1$$



	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
	0.0	0.0	0.2	0.1

This is the "trellis" that shows all possible ways of generating the word sequence

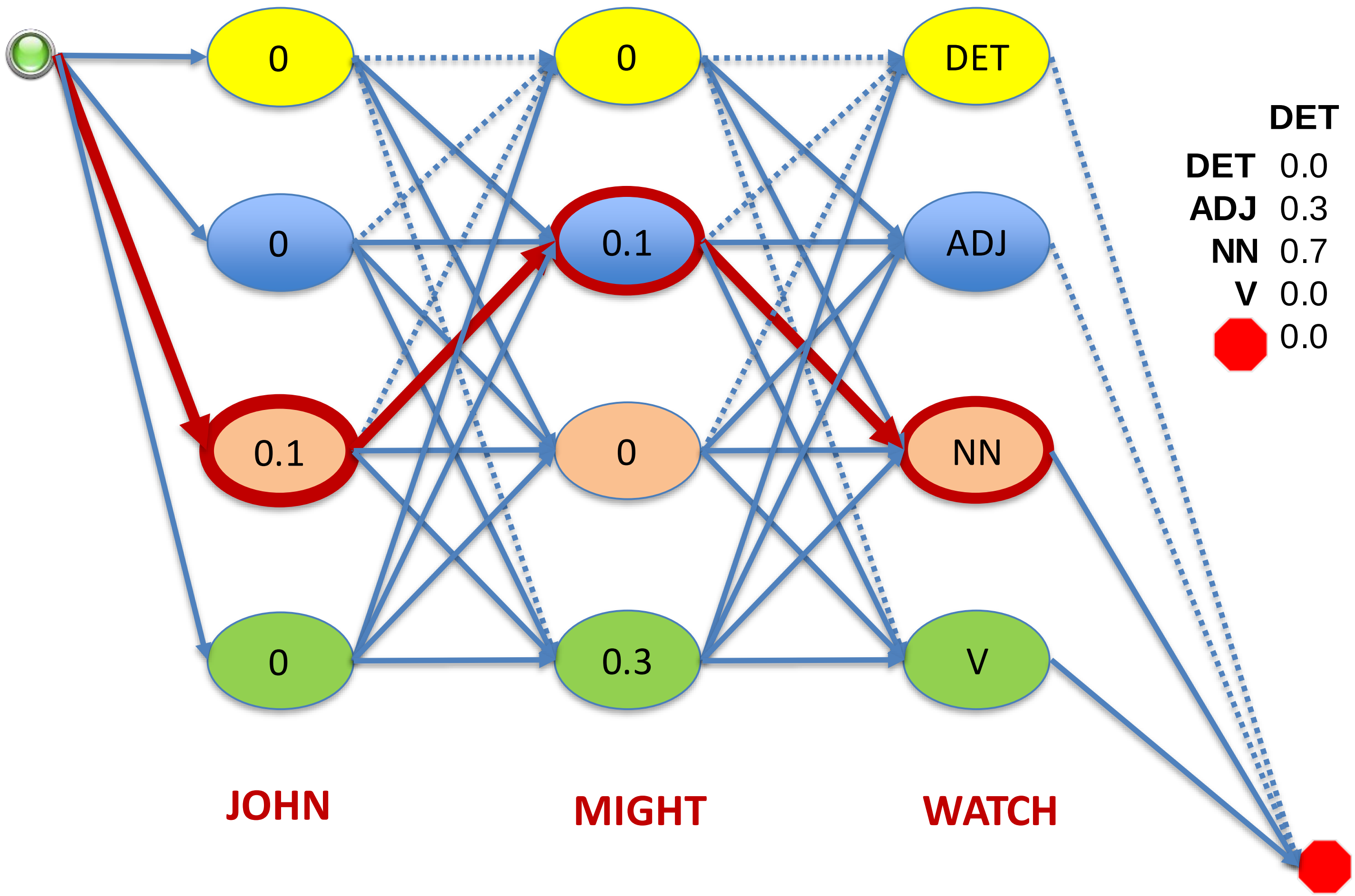
$$P(s_1 = NN, x_1 = JOHN, s_2 = ADJ, x_2 = MIGHT, s_3 = NN)$$
$$= 0.3 \times 0.1 \times 0.1 \times 0.1 \times 0.7$$



	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
	0.0	0.0	0.2	0.1

This is the “trellis” that shows all possible ways of generating the word sequence

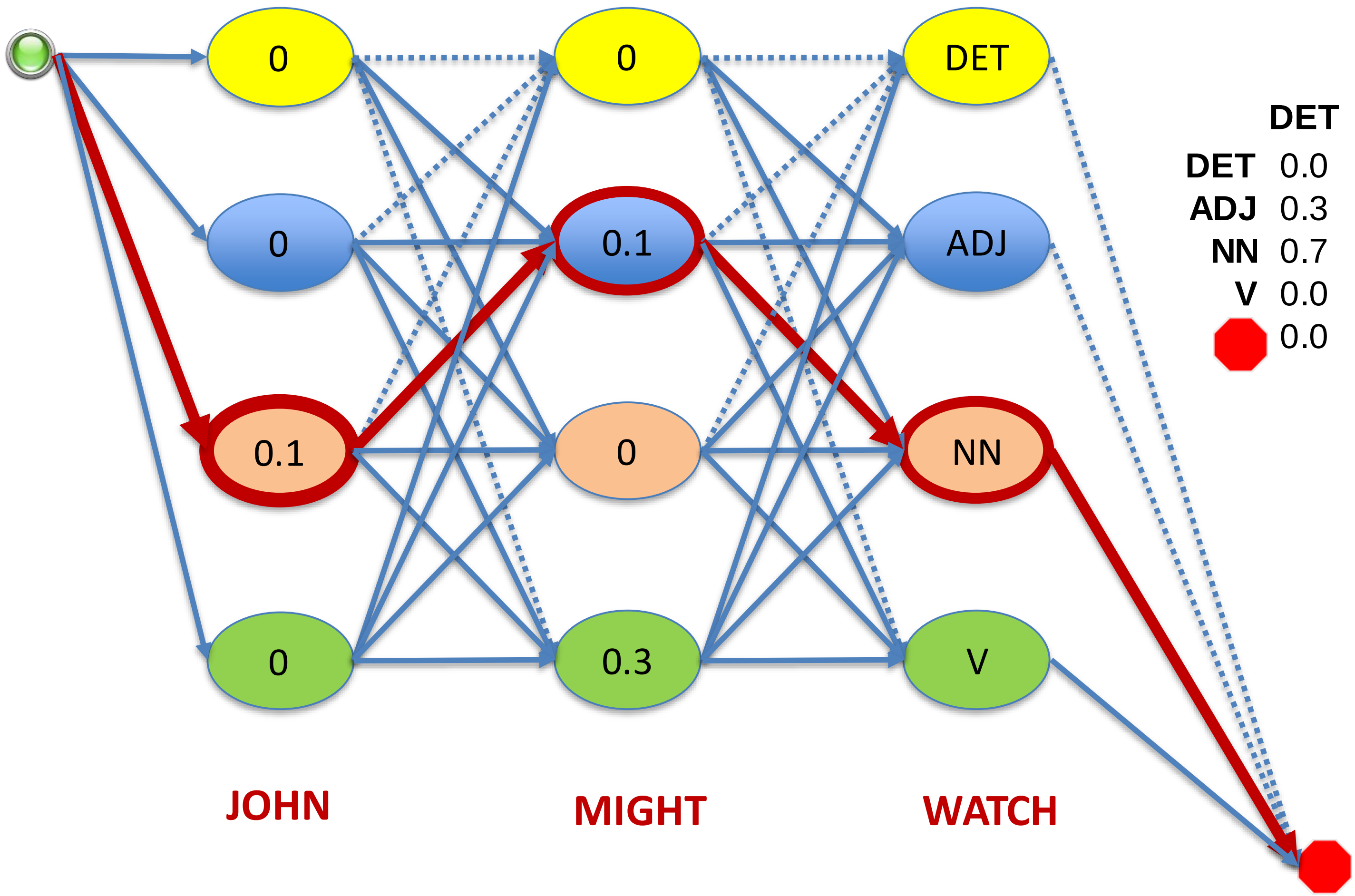
$$P(s_1 = NN, x_1 = JOHN, s_2 = ADJ, x_2 = MIGHT, s_3 = NN, x_3 = WATCH)$$
$$= 0.3 \times 0.1 \times 0.1 \times 0.1 \times 0.7$$



	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
	0.0	0.0	0.2	0.1

This is the “trellis” that shows all possible ways of generating the word sequence

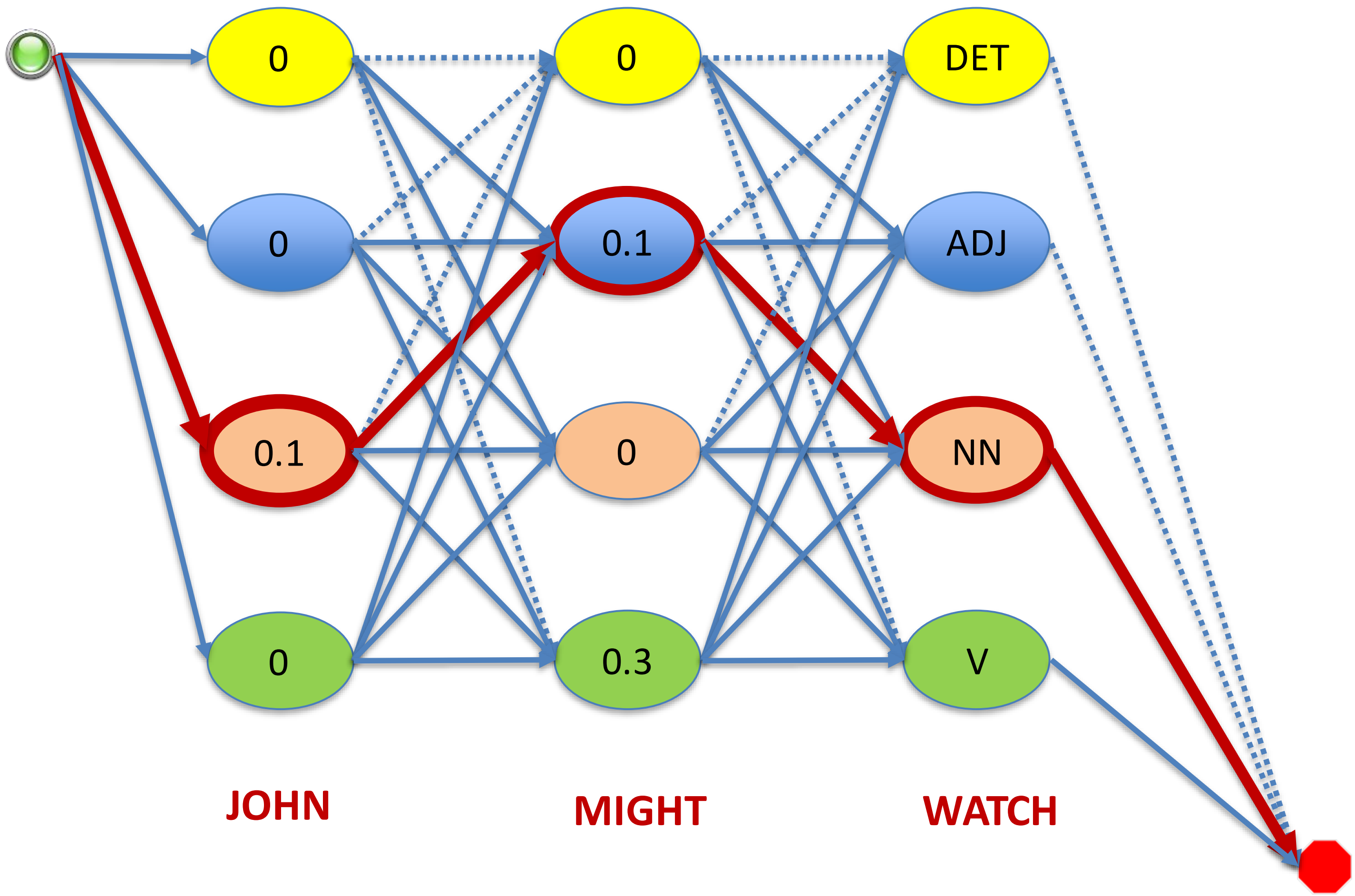
$$P(s_1 = NN, x_1 = JOHN, s_2 = ADJ, x_2 = MIGHT, s_3 = NN, x_3 = WATCH, s_4 = STOP)$$
$$= 0.3 \times 0.1 \times 0.1 \times 0.1 \times 0.7 \times 0.2$$



	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
	0.0	0.0	0.2	0.1

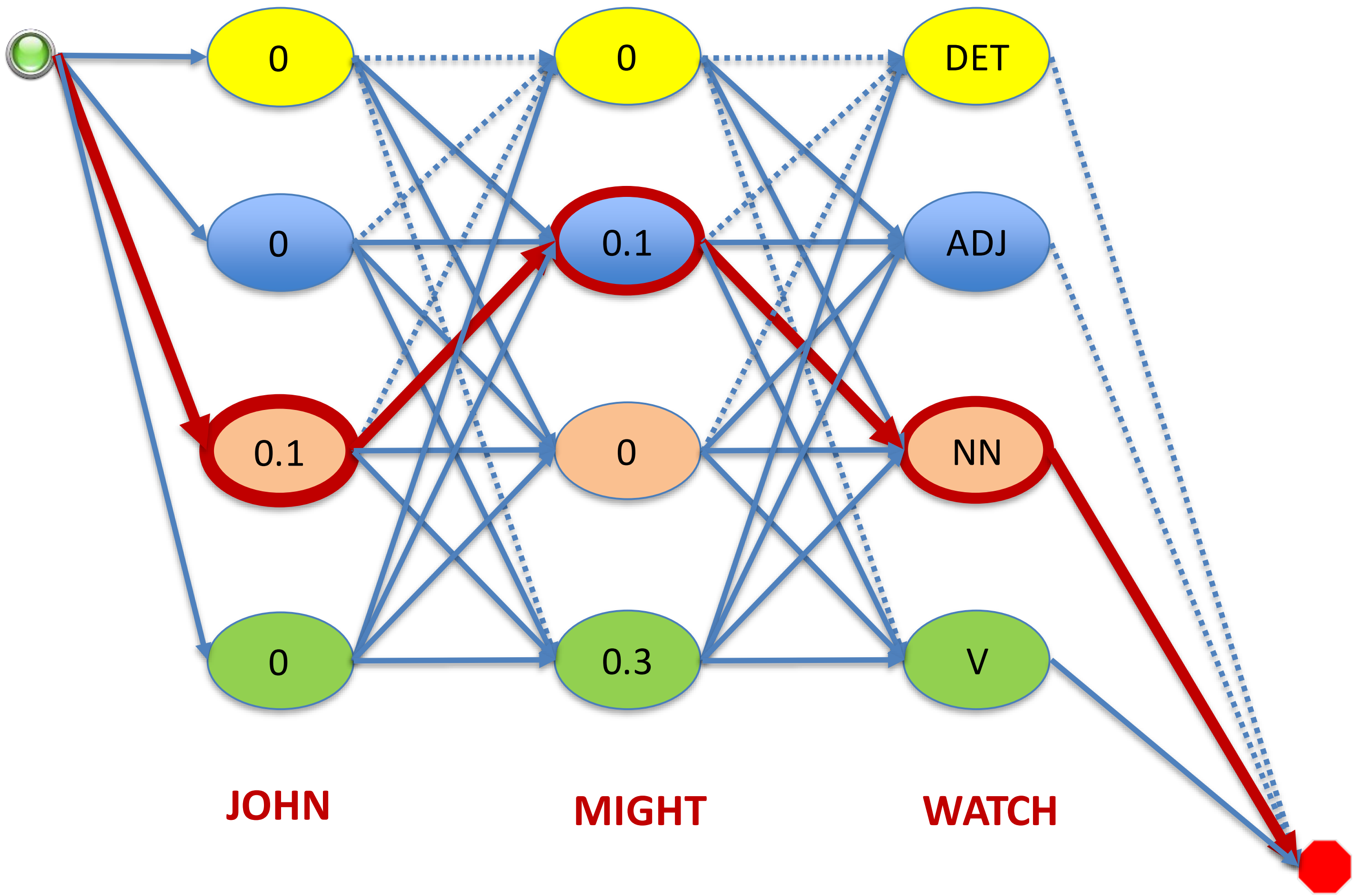
This is the “trellis” that shows all possible ways of generating the word sequence

$$P(s_1 = NN, x_1 = JOHN, s_2 = ADJ, x_2 = MIGHT, s_3 = NN, x_3 = WATCH, s_4 = STOP)$$
$$= 0.3 \times 0.1 \times 0.1 \times 0.1 \times 0.7 \times 0.2$$



$$P(s_1, x_1, s_2, x_2, \ldots, s_T, x_T, \mathrm{STOP}) = P_{in}(s_1)\gamma(s_1 \downarrow x_1)\eta(s_1 \rightarrow s_2)\,\gamma(s_2 \downarrow x_2)\ldots\eta(s_T \rightarrow STOP)$$

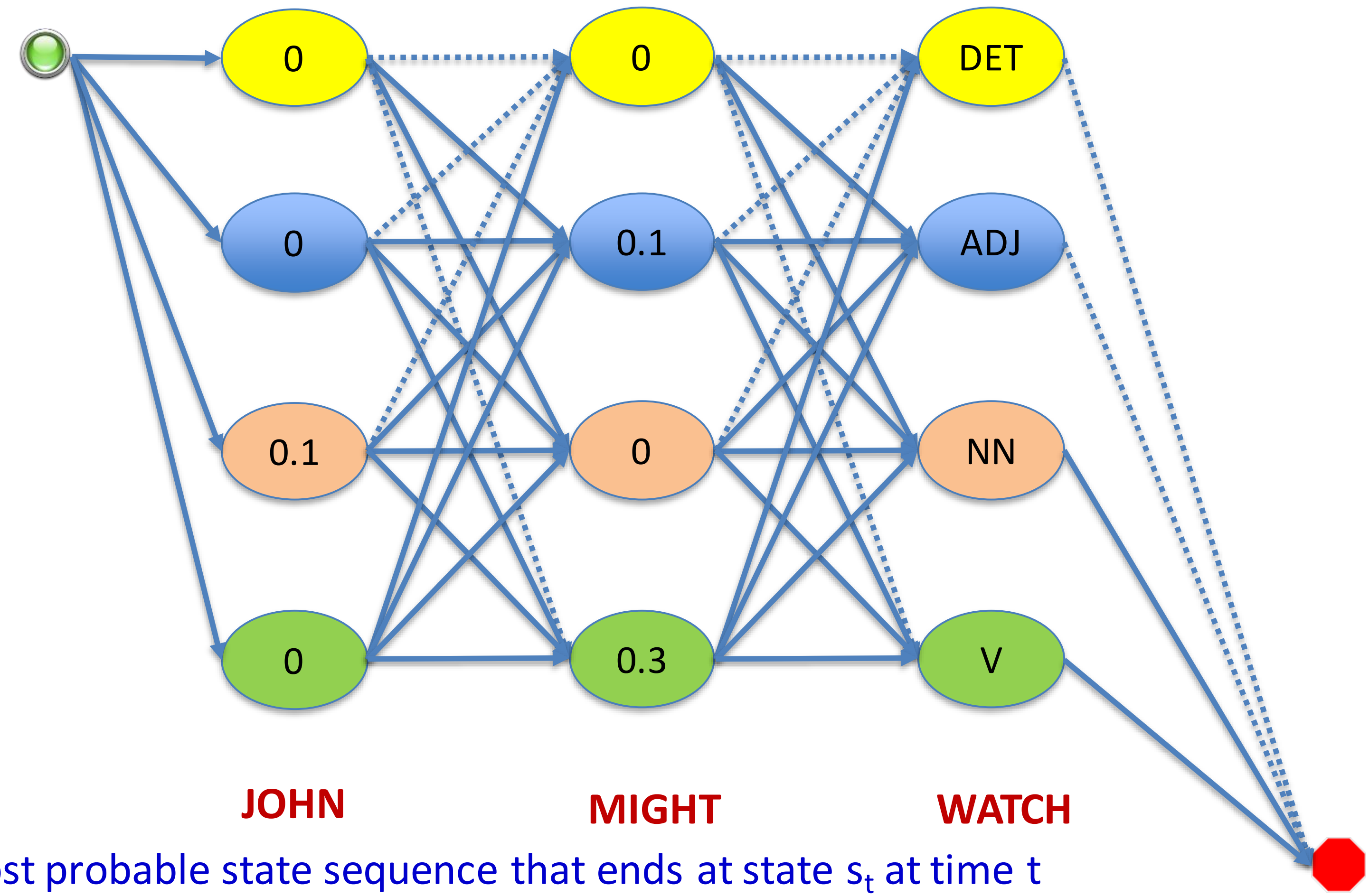
Viterbi : Find the best path (most probable)



$$P(s_1, x_1, s_2, x_2, \ldots, s_T, x_T, \text{STOP}) = P_{in}(s_1)\gamma(s_1 \downarrow x_1)\eta(s_1 \to s_2)\,\gamma(s_2 \downarrow x_2) \ldots \eta(s_T \to STOP)$$

Viterbi : Find the best path (most probable)

DP Argument: For a Markov process, the best N-length path to any state *must* be an extension of a best N-1 length path to some state



The most probable state sequence that ends at state s_t at time t

$$max_{s_1 s_2,\ldots,s_{t-1}} P(s_1, x_1, \ldots, s_t, x_t) = max_{s_1 s_2,\ldots,s_{t-1}} P(s_1, x_1, \ldots, s_{t-1}, x_{t-1}) \, \eta(s_{t-1} \rightarrow s_t) \, \gamma(s_t \downarrow x_t)$$

The Viterbi Algorithm

Probability of the most likely state sequence that ends at state s_t at t and produces $x_1 \ldots x_t$

$$max_{s_1 s_2,\ldots,s_{t-1}} P(s_1, x_1, \ldots, s_t, x_t)$$

$$= max_{s_1 s_2,\ldots,s_{t-1}} P(s_1, x_1, \ldots, s_{t-1}, x_{t-1}) \eta(s_{t-1} \rightarrow s_t)\, \gamma(s_t \downarrow x_t)$$

$$= max_{s_{t-1}} max_{s_1 s_2,\ldots,s_{t-2}} P(s_1, x_1, \ldots, s_{t-1}, x_{t-1}) \eta(s_{t-1} \rightarrow s_t)\, \gamma(s_t \downarrow x_t)$$

Probability of the most likely state sequence that ends at state s_{t-1} at $t-1$ and produces $x_1 \ldots x_{t-1}$

- The probabilities are decomposed in a manner suited to DP

Viterbi

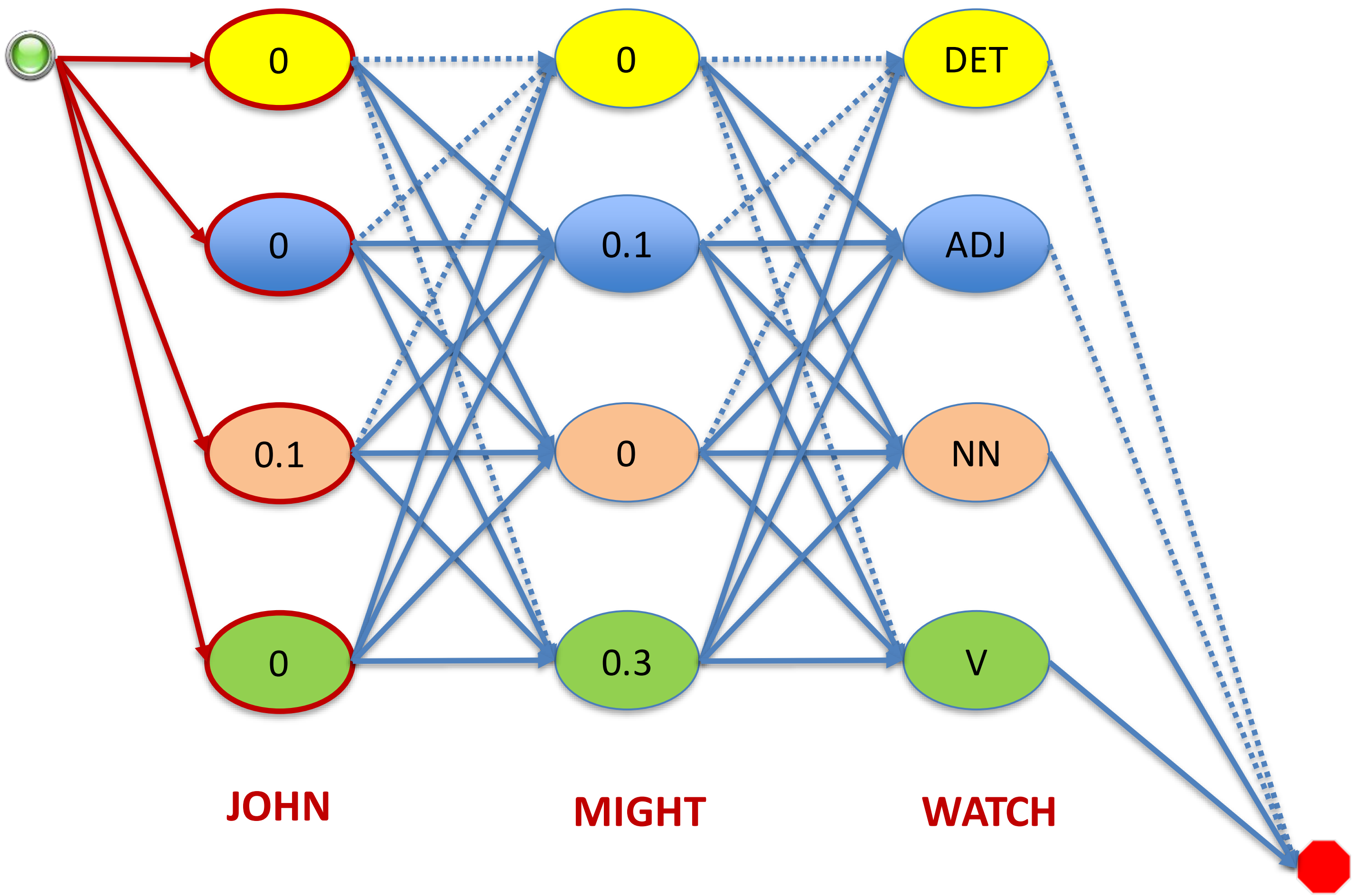
$$max_{s_1 s_2,\dots,s_{t-1}} P(s_1, x_1, \dots, s_t, x_t)$$
$$= max_{s_{t-1}} max_{s_1 s_2,\dots,s_{t-2}} P(s_1, x_1, \dots, s_{t-1}, x_{t-1}) \eta(s_{t-1} \to s_t)\, \gamma(s_t \downarrow x_t)$$

- Let $max_{s_1 s_2,\dots,s_{t-1}} P(s_1, x_1, \dots, s_t, x_t) = R(s_t, t)$
- for $t \ = 1{:}T$

$$h(s,t) = \underset{s'}{\operatorname{argmax}} R(s', t-1)\eta(s' \to s)$$
$$R(s,t) = R(h(s,t), t-1)$$
$$\eta(h(s,t) \to s_t)\gamma(s \downarrow x_t)$$

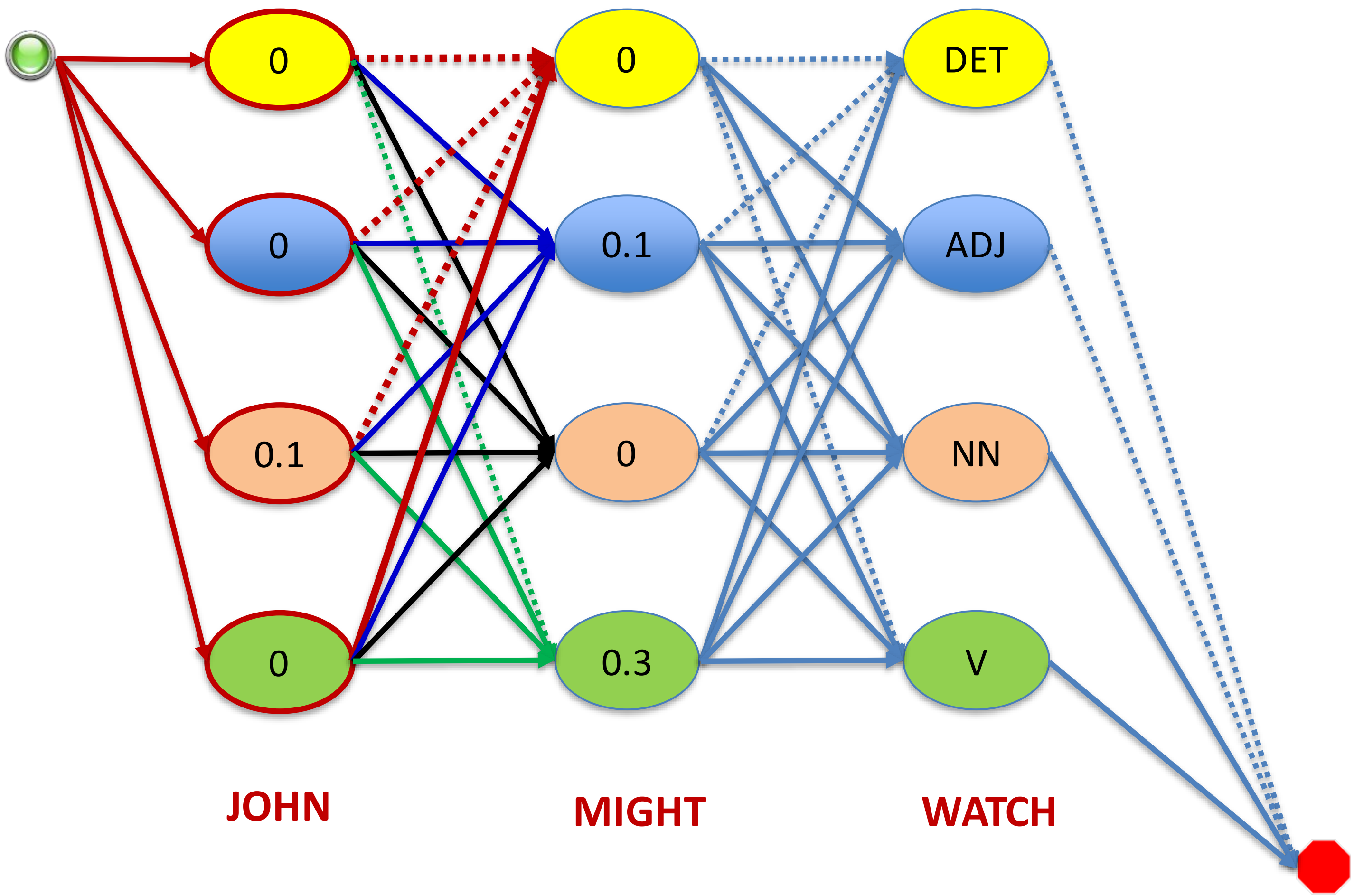
Viterbi Algorithm

$$R(s,1) = P_{in}(s)\gamma(s \downarrow x_1)$$



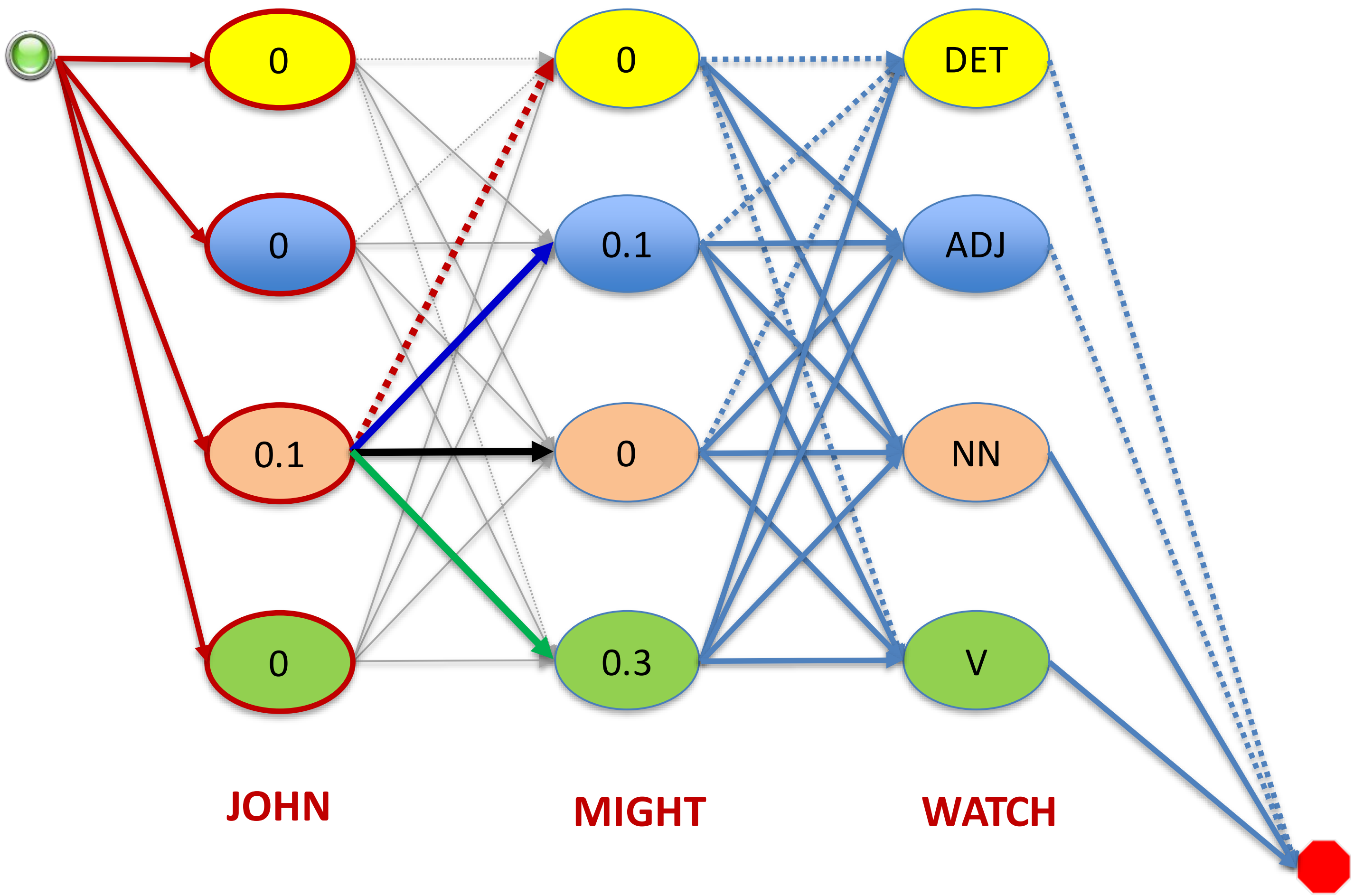
Viterbi Algorithm

$$h(s,t) = \underset{s'}{\operatorname{argmax}}\, R(s', t-1)\eta(s' \to s)$$



Viterbi Algorithm

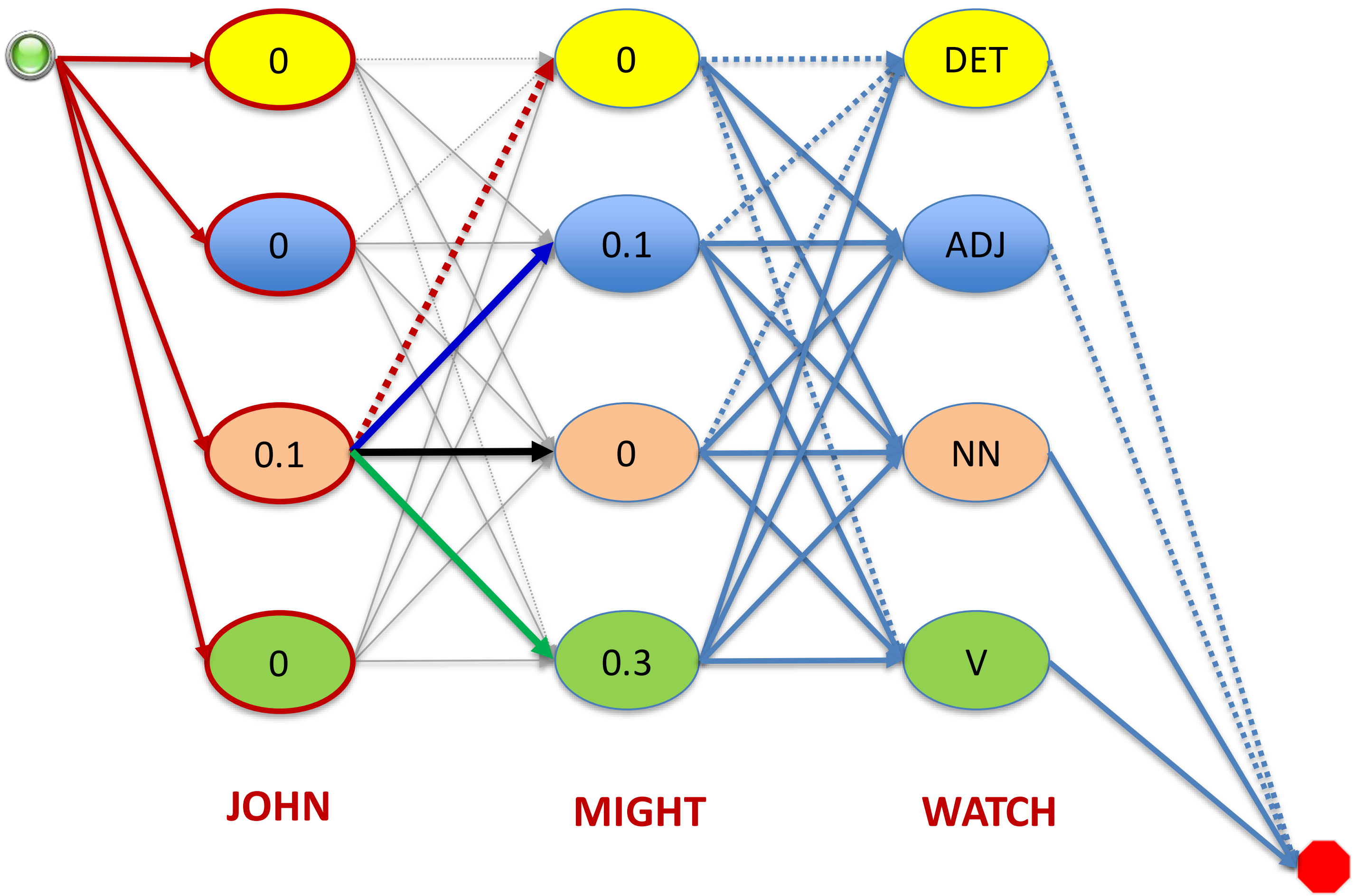
$$h(s,t) = \underset{s'}{\operatorname{argmax}}\, R(s', t-1)\eta(s' \to s)$$



Viterbi Algorithm

$$h(s,t) = \underset{s'}{\operatorname{argmax}}\, R(s',t-1)\eta(s' \rightarrow s)$$

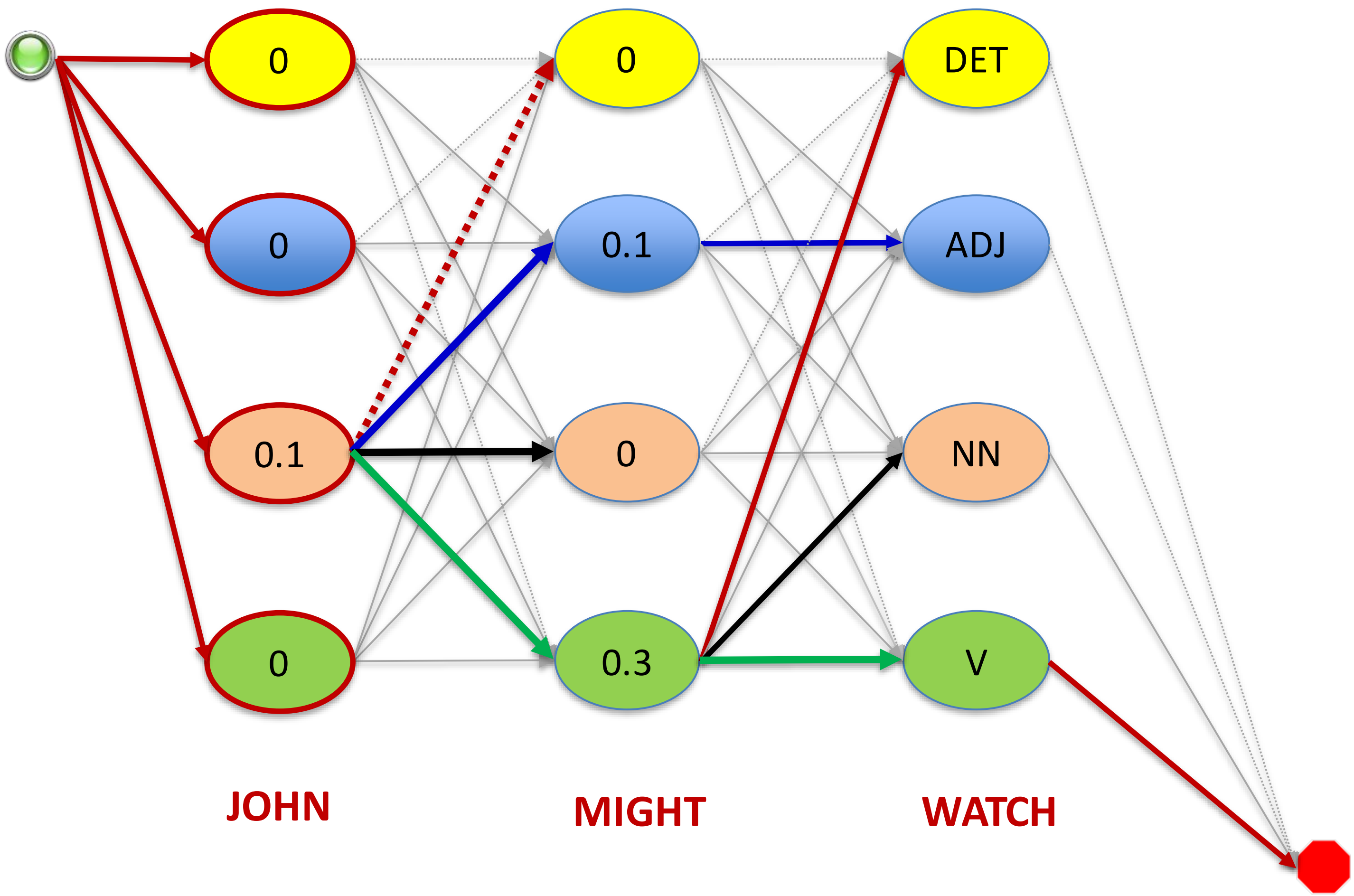
$$R(s,t) = R(h(s,t),t-1)\eta(h(s,t) \rightarrow s_t)\gamma(s \downarrow x_t)$$



Viterbi Algorithm

$$h(s,t) = \operatorname*{argmax}_{s'} R(s', t-1)\eta(s' \to s)$$

$$R(s,t) = R(h(s,t), t-1)\eta(h(s,t) \to s_t)\gamma(s \downarrow x_t)$$



String Marginals

- Inference question for HMMs
  - What is the probability of a string **w**?

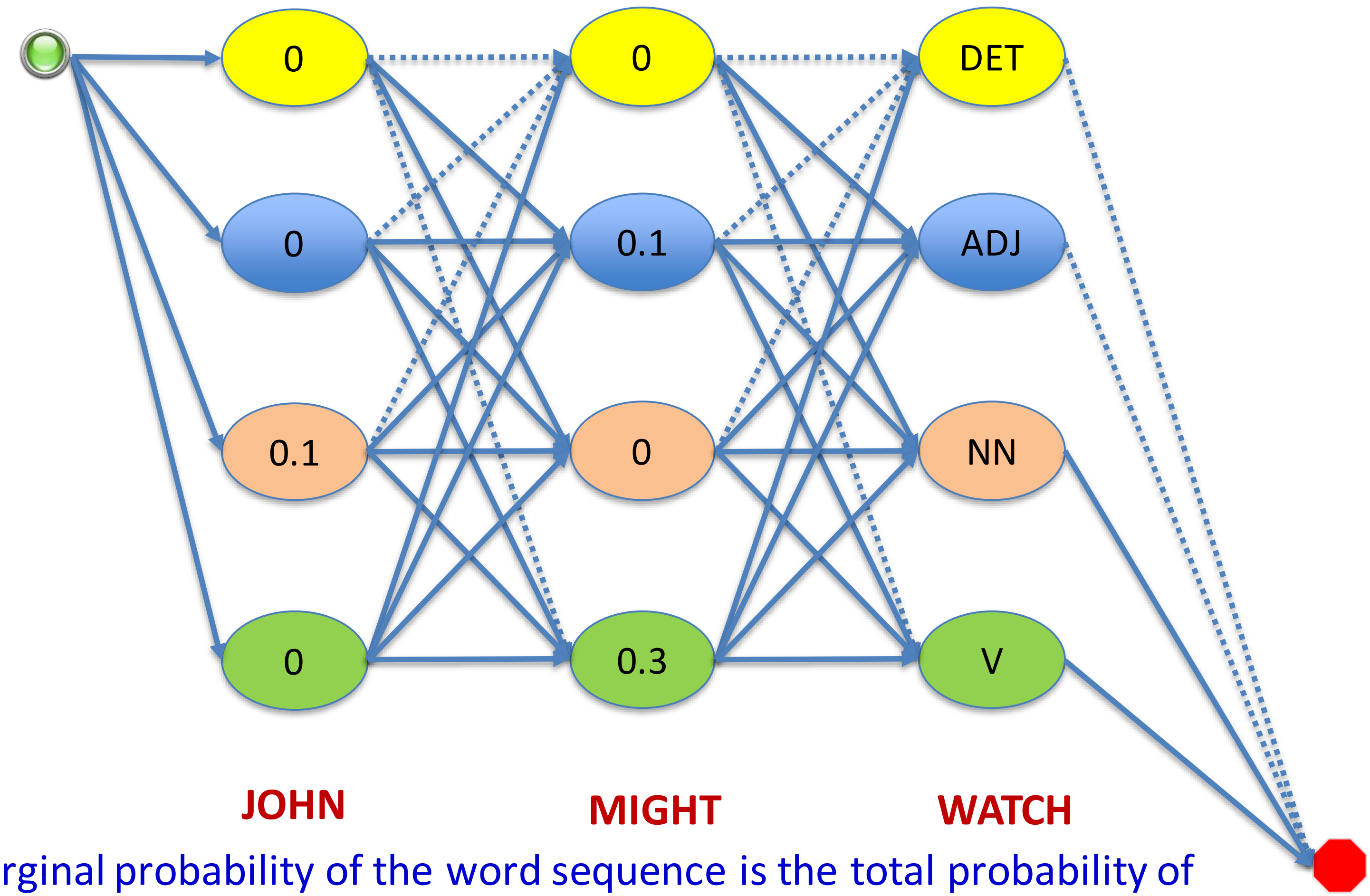
    Answer: generate all possible tag sequences and explicitly *marginalize*

    $$O(|\Omega|^{|\mathbf{w}|}) \text{ time}$$

    Can we do this efficiently?

Viterbi : Find the best path (most probable)

String Marginals



The marginal probability of the word sequence is the total probability of all paths through the trellis

$$P(x_1, \ldots, x_t) = \sum_{s_1 s_2, \ldots, s_{t-1}} P(s_1, x_1, \ldots, s_t, x_t)$$

Forward Algorithm

- How to compute: use the **forward algorithm**

- Analogous to Viterbi
  - Instead of computing a **max** of inputs at each node, use **addition**
- Same run-time, same space requirements

$$O(|\Omega|^2 \times |\mathbf{w}|) \quad \text{time}$$

$$O(|\Omega|) \quad \text{space}$$

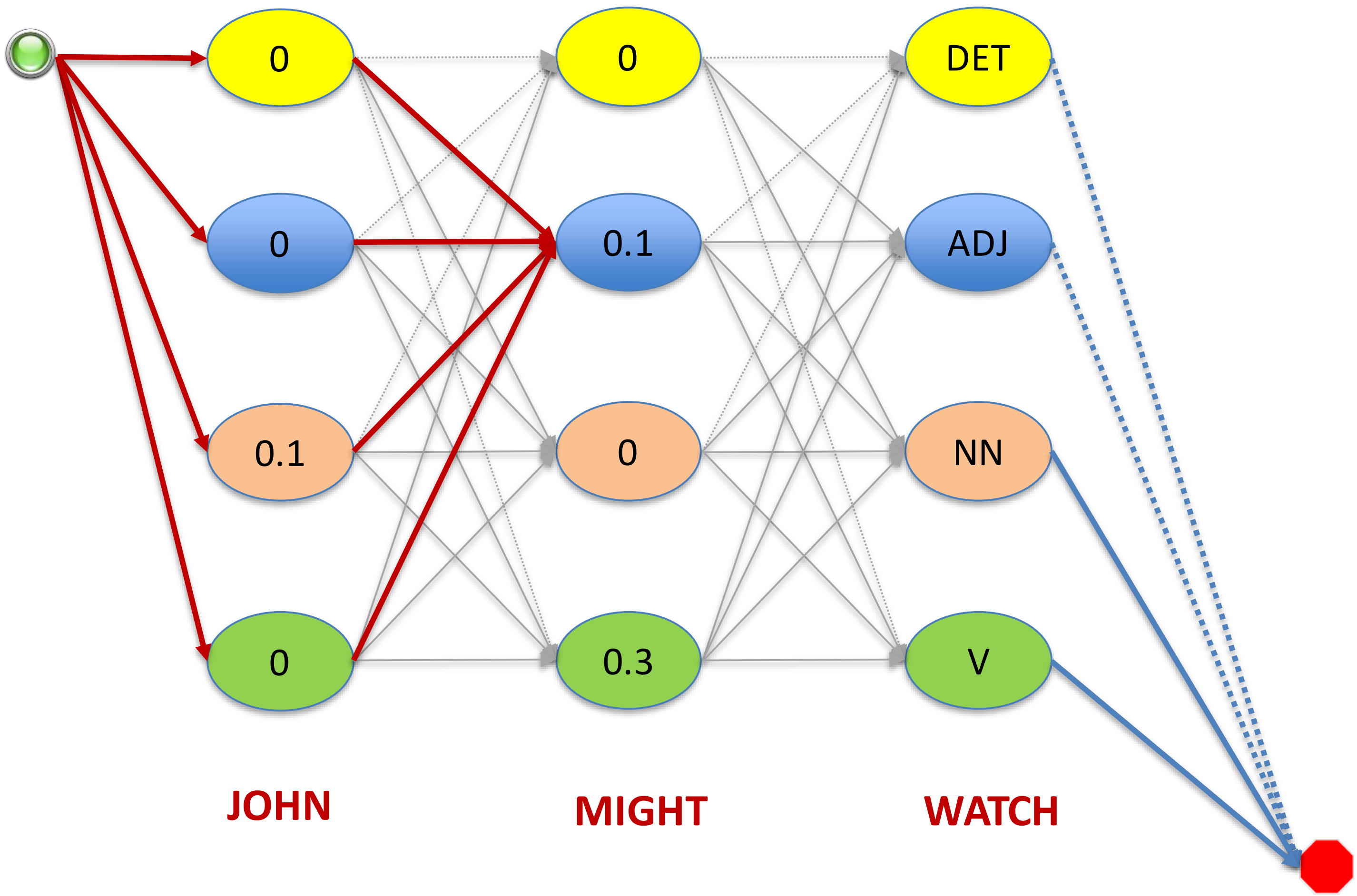
Define

$$\alpha_t(s) = P(x_1 \ldots x_t, s_t = s)$$

- The probability of generating $x_1 \ldots x_t$ such that the process is in state s at time t

Viterbi : Find the best path (most probable)

String Marginals



$\alpha_2(ADJ)$ is the probability of producing JOHN MIGHT such that the second word is an adjective

This is the total probability of all paths leading to ADJ at t=2, while producing JOHN MIGHT

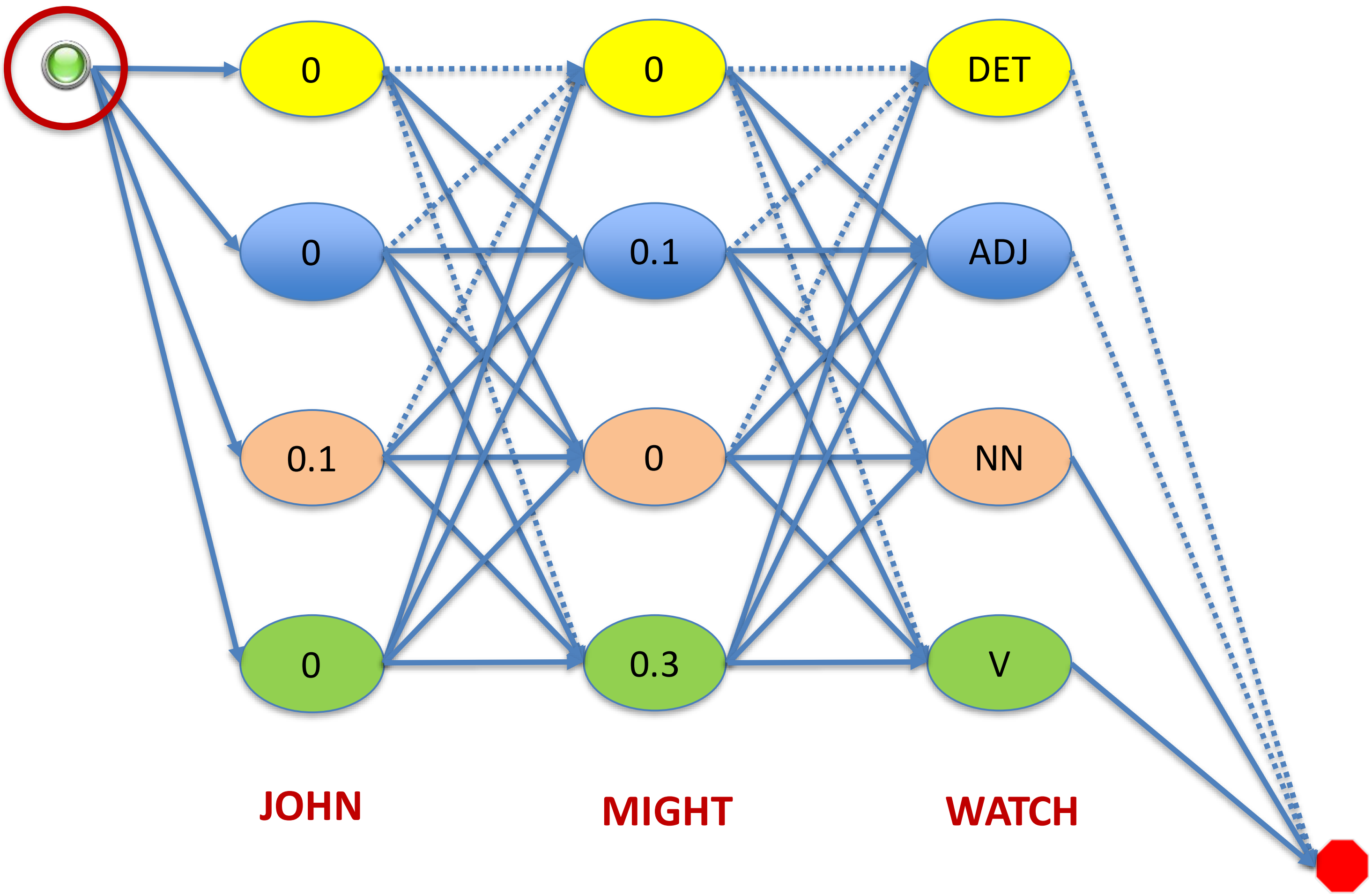
Forward Algorithm Recurrence

$$\alpha_0(\mathrm{START}) = 1$$

$$\alpha_t(r) = \sum_{q \in \Omega} \alpha_{t-1}(q)\eta(q \rightarrow r)\,\gamma(r \downarrow x_t)$$

Viterbi : Find the best path (most probable)

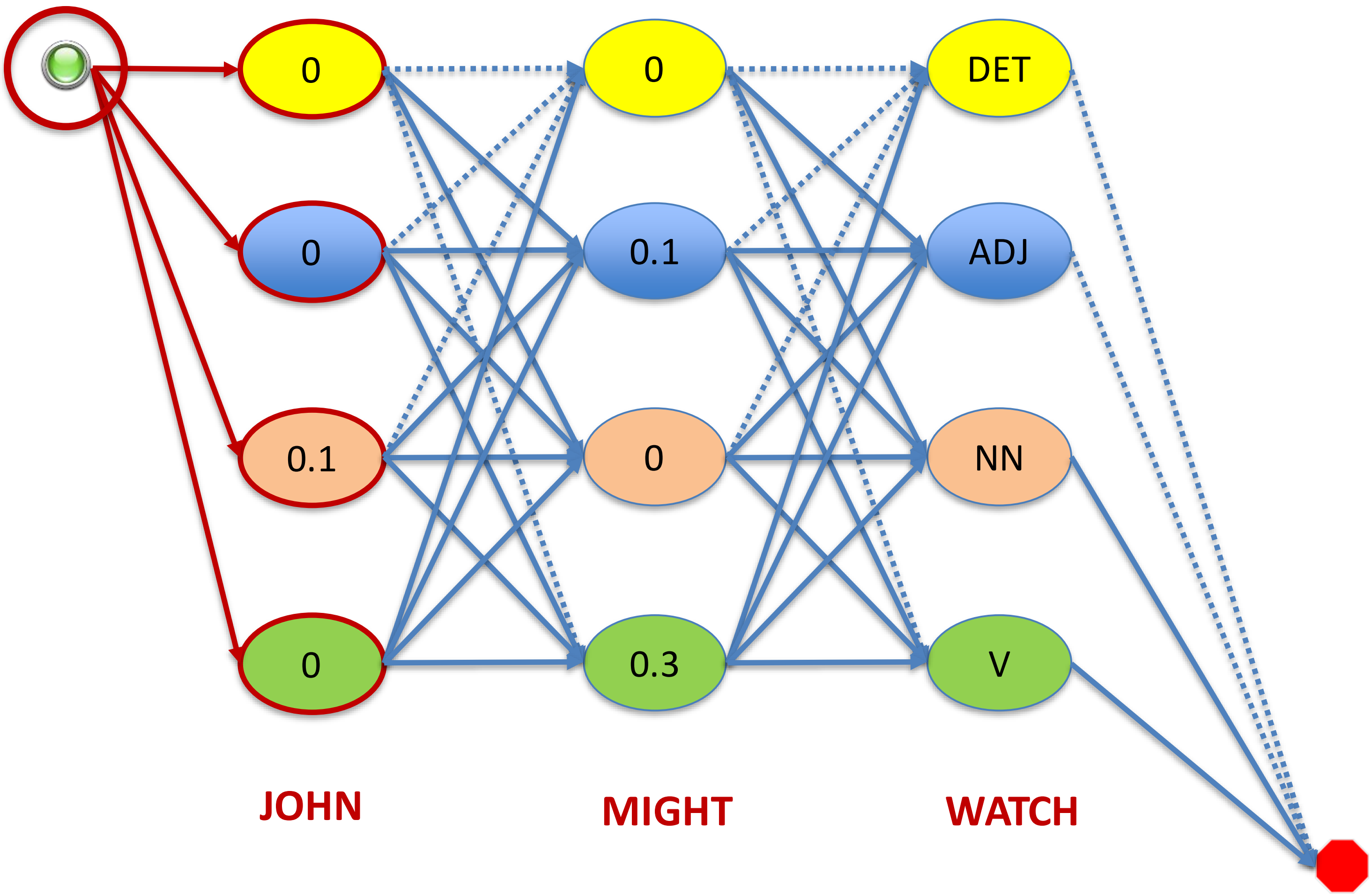
Forward Algorithm



$\alpha_0(START)$

Viterbi : Find the best path (most probable)

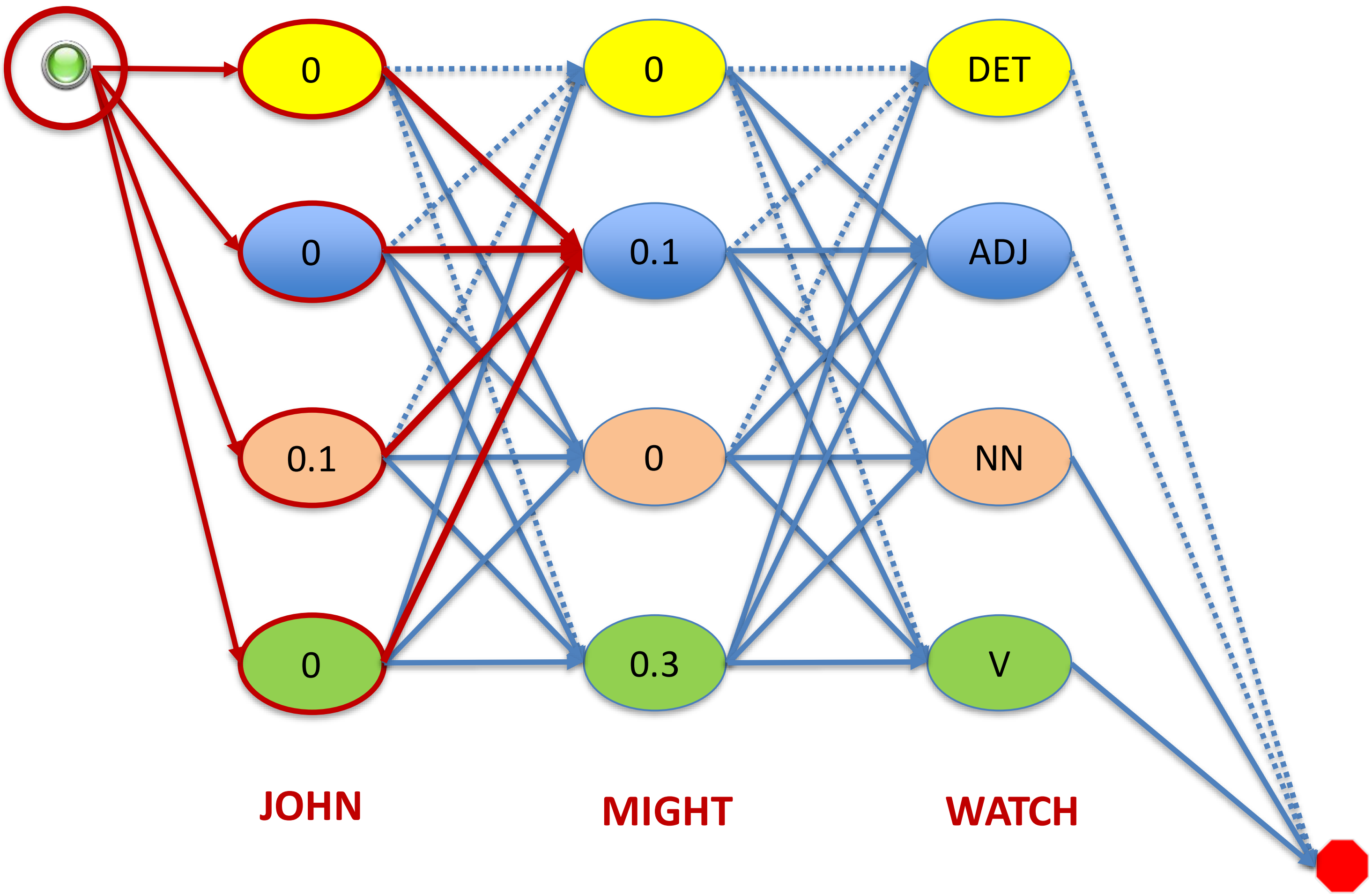
Forward Algorithm



$$\alpha_1(s) = \alpha_0(START)\eta(START \rightarrow s)\,\gamma(s \downarrow x_1)$$

Viterbi : Find the best path (most probable)

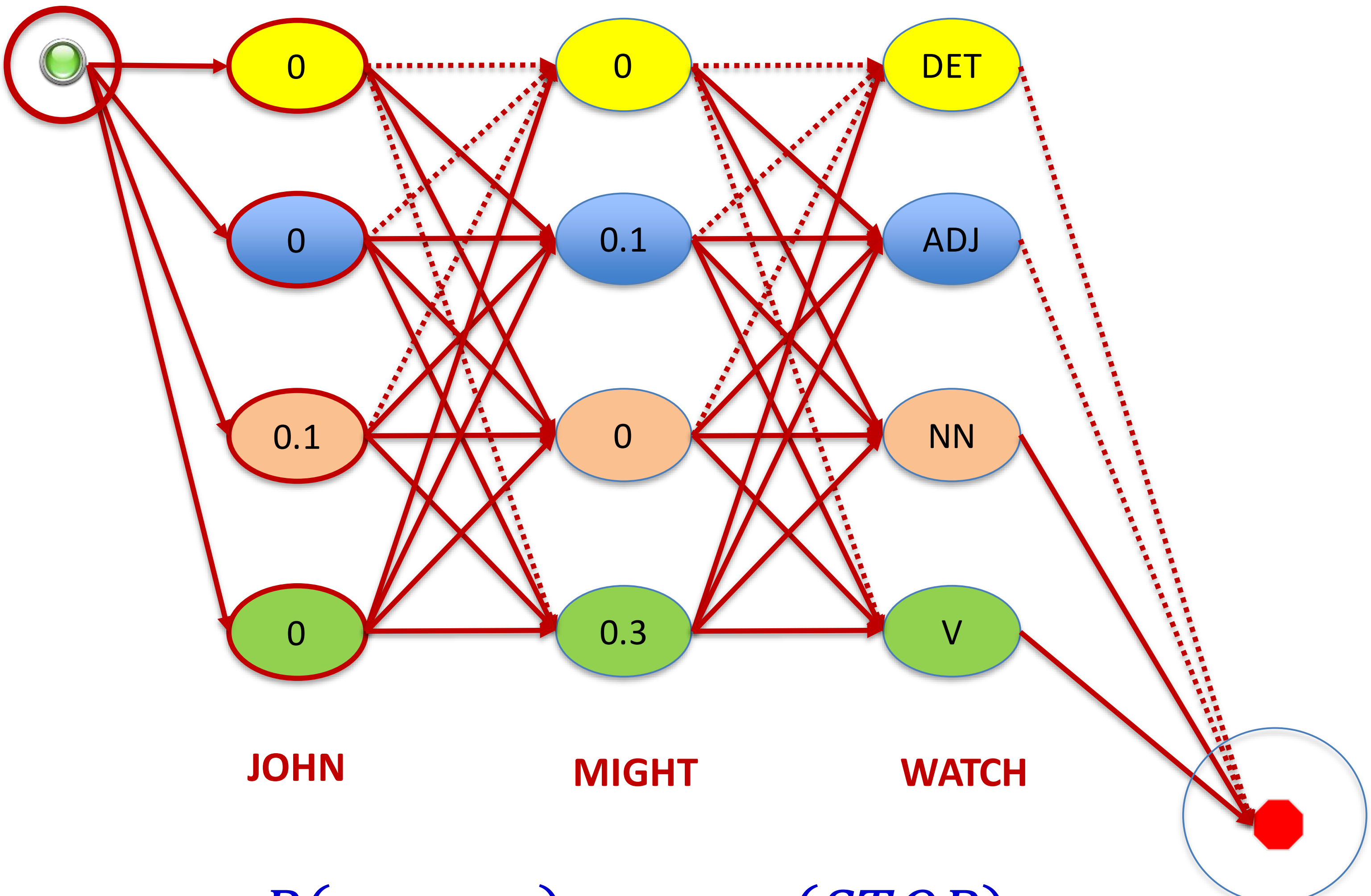
Forward Algorithm



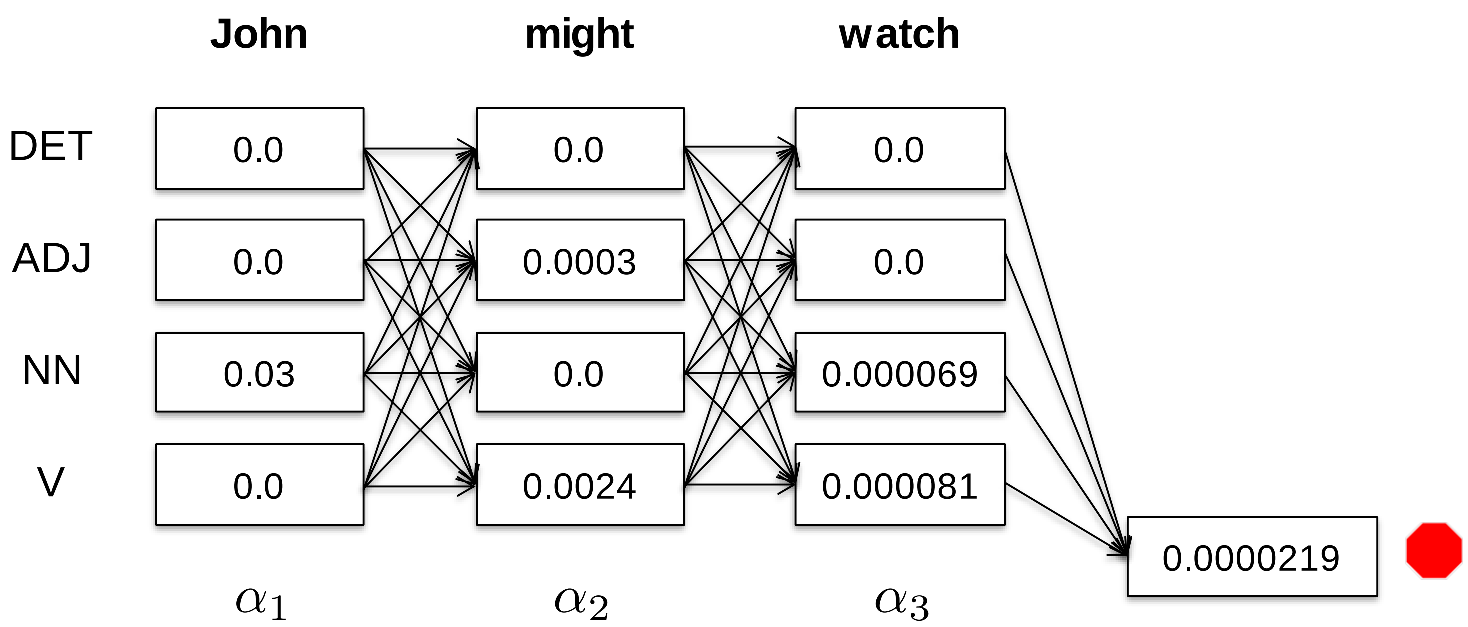
$$\alpha_t(r) = \sum_{q \in \Omega} \alpha_{t-1}(q)\eta(q \to r)\,\gamma(r \downarrow x_t)$$

Viterbi : Find the best path (most probable)

Forward Algorithm



$$P(x_1 \ldots x_T) = \alpha_{T+1}(STOP)$$


John
might
watch
DET
ADJ
NN
V
0.0
0.0
0.0
0.0
0.0003
0.0
0.03
0.0
0.000069
0.0
0.0024
0.000081
0.0000219
α_1
α_2
α_3

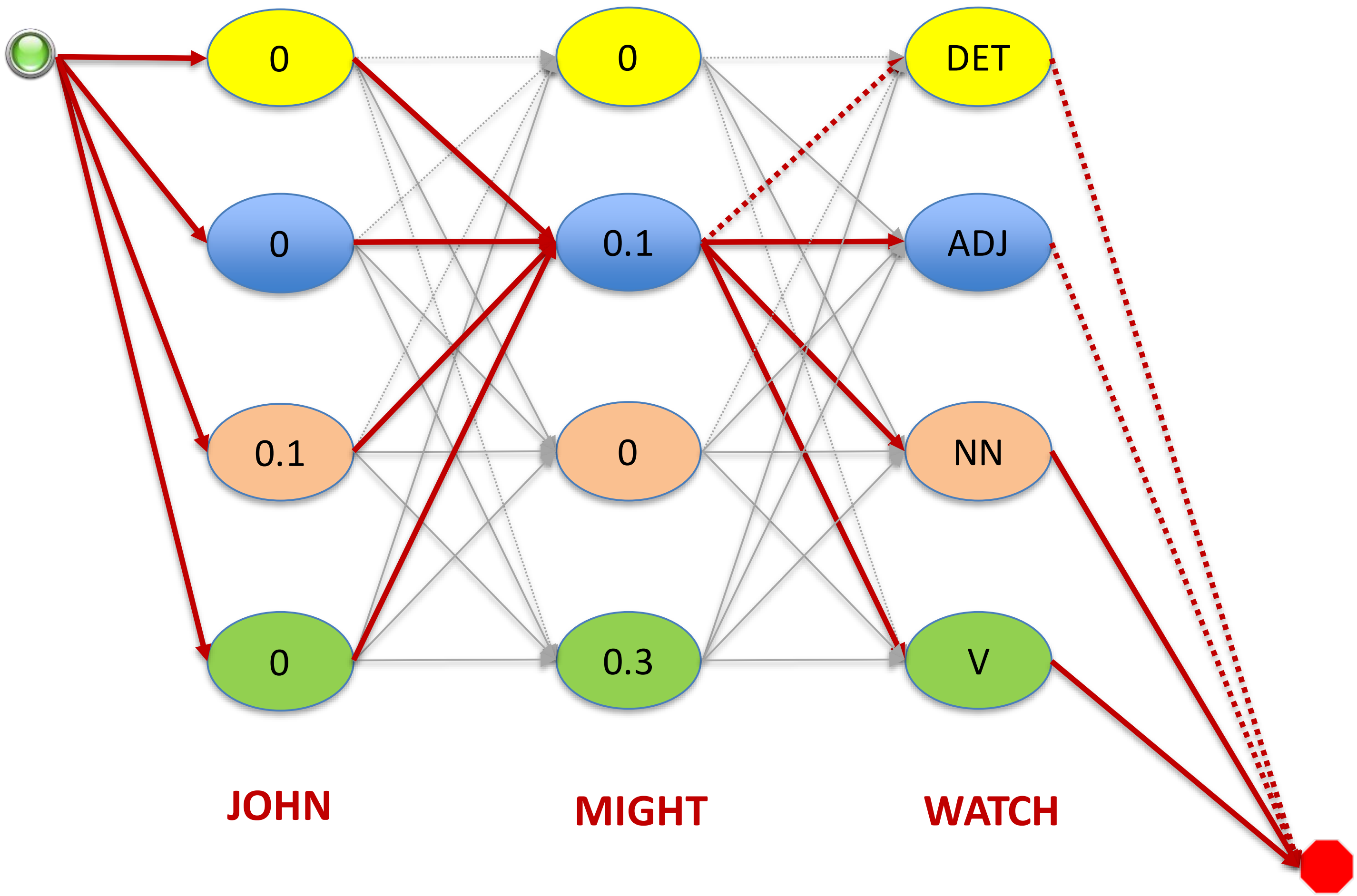
$$p = 0.0000219$$

Posterior Marginals

- Marginal inference question for HMMs
  - *Posterior Marginal*: Given **x**, what is the probability of being in a state *q* at time *t*?
  - *Marginal*: What is the probability of ***x*** *and* being in state *q* at time *t*?

$$
\begin{aligned}
p(s_t = q | x_1, \dots, x_T) &\propto p(x_1, \dots, x_T, s_t = q) \\
&= p(x_1, \dots, x_t, s_t = q, x_{t+1}, \dots, x_T)
\end{aligned}
$$

Posterior Marginals



The marginal of ADJ at t=2 is the total probability of all paths that go through ADJ at t=2

Posterior Marginals

- Marginal inference question for HMMs
  - State: Given **x**, what is the probability of being in a state q at time t?

$$p(s_t = q | x_1, \dots, x_T) \propto p(x_1, \dots, x_T, s_t = q)$$
$$= p(x_1, \dots, x_t, s_t = q, x_{t+1}, \dots, x_T)$$

  - Transition: Given **x**, what is the probability of transitioning from state q to r at time t?

$$p(s_t = q, s_{t+1} = r | x_1, \dots, x_T) \propto p(x_1, \dots, x_T, s_t = q, s_{t+1} = r)$$
$$= p(x_1, \dots, x_t, s_t = q, s_{t+1} = r, x_{t+1}, \dots, x_T)$$

Posterior Marginals

- Marginal inference question for HMMs
  - State: What is the probability of **x** and being in a state *q* at time *t*?

$$p(x_1, \ldots, x_T, s_t = q) = p(x_1, \ldots, x_t, s_t = q)p(x_{t+1}, \ldots, x_T | s_t = q)$$

  - Transition: What is the probability of **x** and of transitioning from state *q* to *r* at time *t*?

$$p(x_1, \ldots, x_T, s_t = q, s_{t+1} = r)$$
$$= p(x_1, \ldots, x_t, s_t = q\ )\eta(q \to r)\gamma(r \downarrow x_{t+1})p(x_{t+1}, \ldots, x_T | s_{t+1} = r)$$

Posterior Marginals

- Marginal inference question for HMMs
  - State: What is the probability of **x** and being in a state *q* at time *t*?

$$p(x_1, \dots, x_T, s_t = q) = \boxed{p(x_1, \dots, x_t, s_t = q)} p(x_{t+1}, \dots, x_T | s_t = q)$$

  - Transition: What is the probability of **x** and of transitioning from state *q* to *r* at time *t*?

$$p(x_1, \dots, x_T, s_t = q, s_{t+1} = r)$$
$$= \boxed{p(x_1, \dots, x_t, s_t = q\ )} \eta(q \to r) \gamma(r \downarrow x_{t+1}) p(x_{t+1}, \dots, x_T | s_{t+1} = r)$$

Posterior Marginals

- Marginal inference question for HMMs
  - State: What is the probability of **x** and being in a state *q* at time *t*?

$$p(x_1, \dots, x_T, s_t = q) = p(x_1, \dots, x_t, s_t = q)\boxed{p(x_{t+1}, \dots, x_T | s_t = q)}$$

  - Transition: What is the probability of **x** and of transitioning from state *q* to *r* at time *t*?

$$p(x_1, \dots, x_T, s_t = q, s_{t+1} = r)$$
$$= p(x_1, \dots, x_t, s_t = q\ )\eta(q \rightarrow r)\gamma(r \downarrow x_{t+1})\boxed{p(x_{t+1}, \dots, x_T | s_{t+1} = r)}$$

The Backward Probability

$$P(x_{t+1}, \ldots, x_T | s_t = q) = \sum_s P(x_{t+1}, \ldots, x_T, s_{t+1} = r | s_t = q)$$

$$P(x_{t+1}, \ldots, x_T, s_{t+1} = r | s_t = q) =$$
$$\eta(q \to s)\gamma(s \downarrow x_{t+1})P(x_{t+2}, \ldots, x_T | s_{t+1} = r)$$

$$P(x_{t+1}, \ldots, x_T | s_t = q) =$$
$$\sum_s \eta(q \to s)\gamma(s \downarrow x_{t+1})\, P(x_{t+2}, \ldots, x_T | s_{t+1} = r)$$

Backward Algorithm

$$P(x_{t+1},\dots,x_T|s_t=q) = \sum_r \eta(q \to r)\gamma(r \downarrow x_{t+1})\, P(x_{t+2},\dots,x_T|s_{t+1}=r)$$

- Define $\beta_t(q) = P(x_{t+1},\dots,x_T|s_t=q)$

- Recursion

$$\beta_t(q) = \sum_r \eta(q \to r)\gamma(r \downarrow x_{i+1})\, \beta_{t+1}(r)$$

Backward Algorithm

- Start at the goal node(s) and work **backwards** through the trellis

Backward Recurrence

$$\beta_{|\mathbf{x}|+1}(\text{STOP}) = 1$$

$$\beta_t(q) = \sum_{r \in \Omega} \eta(q \to r)\, \gamma(r \downarrow x_{t+1}) \beta_{t+1}(r)$$

Backward Chart

$\cdots$

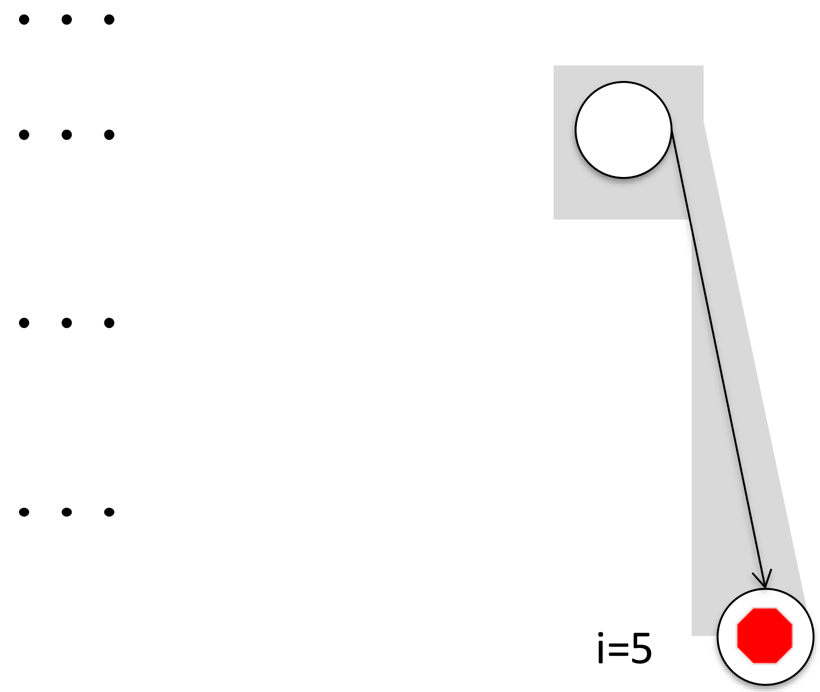
$\cdots$

$\cdots$

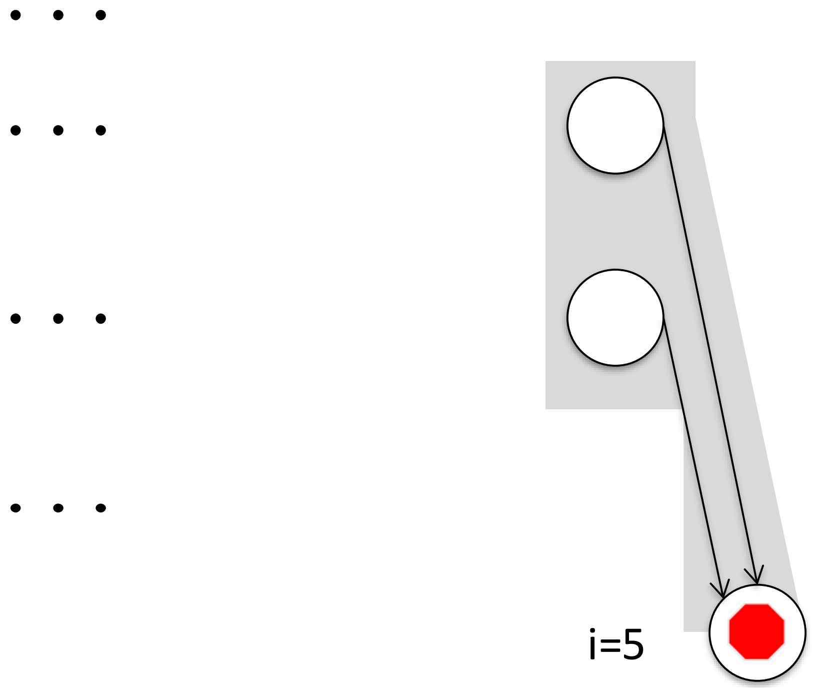
$\cdots$

$$\beta_{|\mathbf{x}|+1}(\textsc{stop}) = 1$$

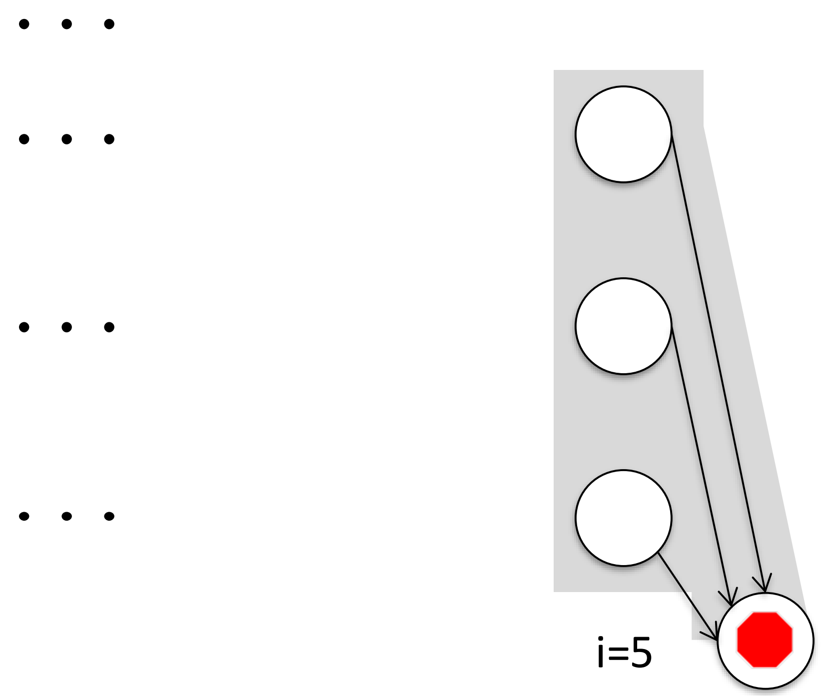
Backward Chart



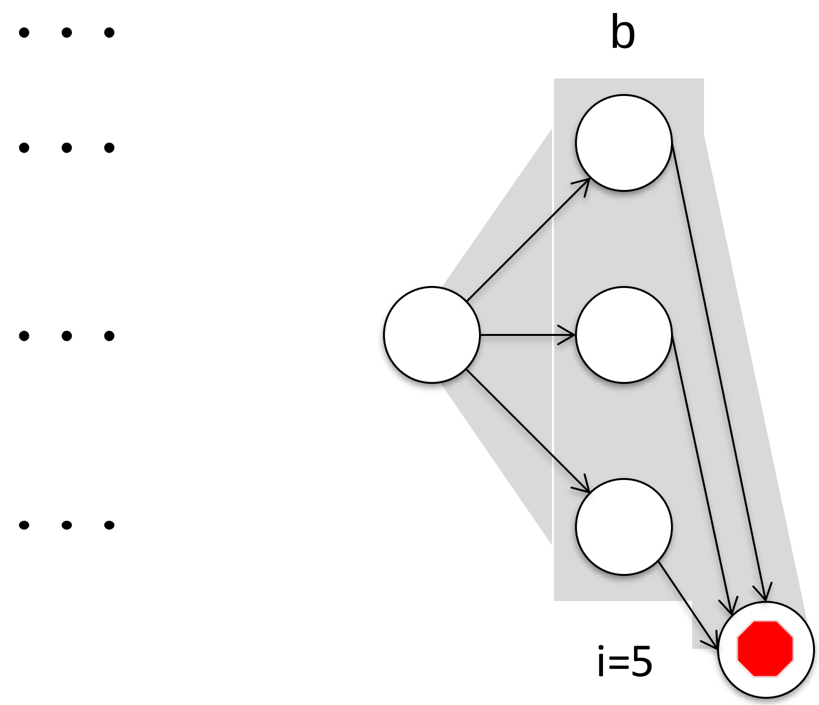
Backward Chart



Backward Chart

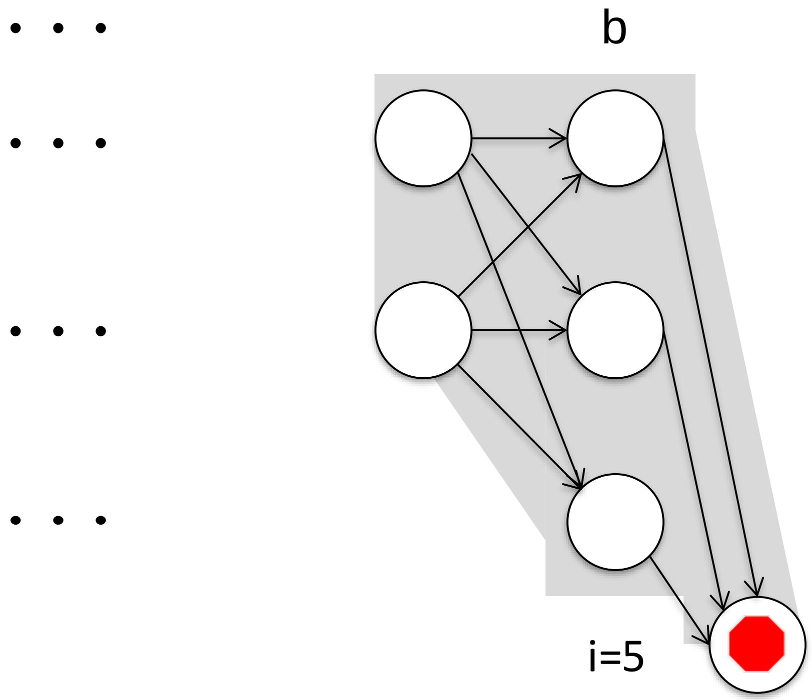


Backward Chart



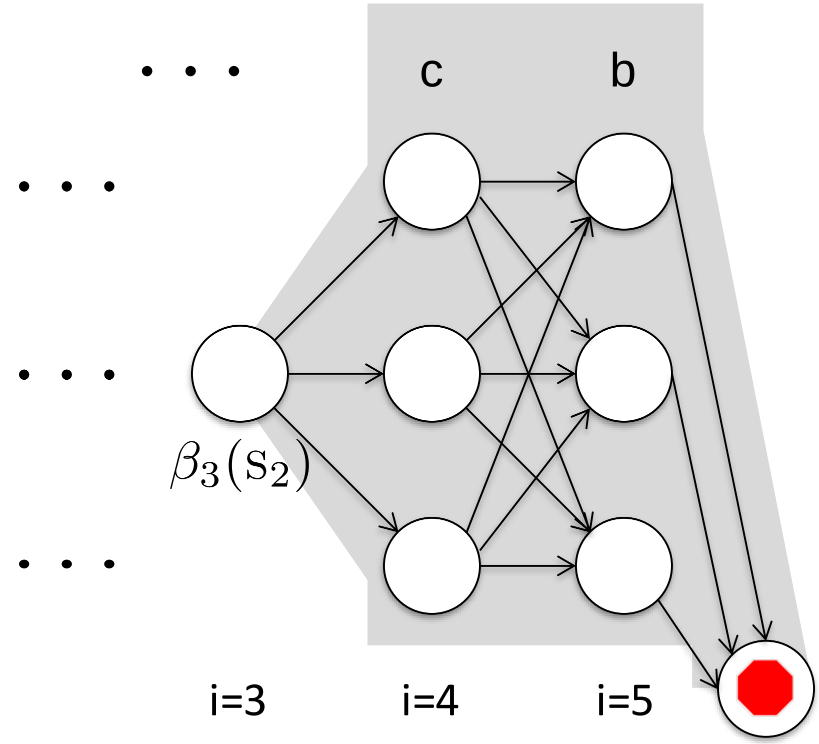
$$\beta_t(q) = \sum_{r \in \Omega} \eta(q \rightarrow r)\,\gamma(r \downarrow x_{t+1})\beta_{t+1}(r)$$

Backward Chart



Backward Chart

$$\beta_t(q) = p(x_{t+1}, \ldots, x_{|\mathbf{x}|} \mid y_t = q)$$

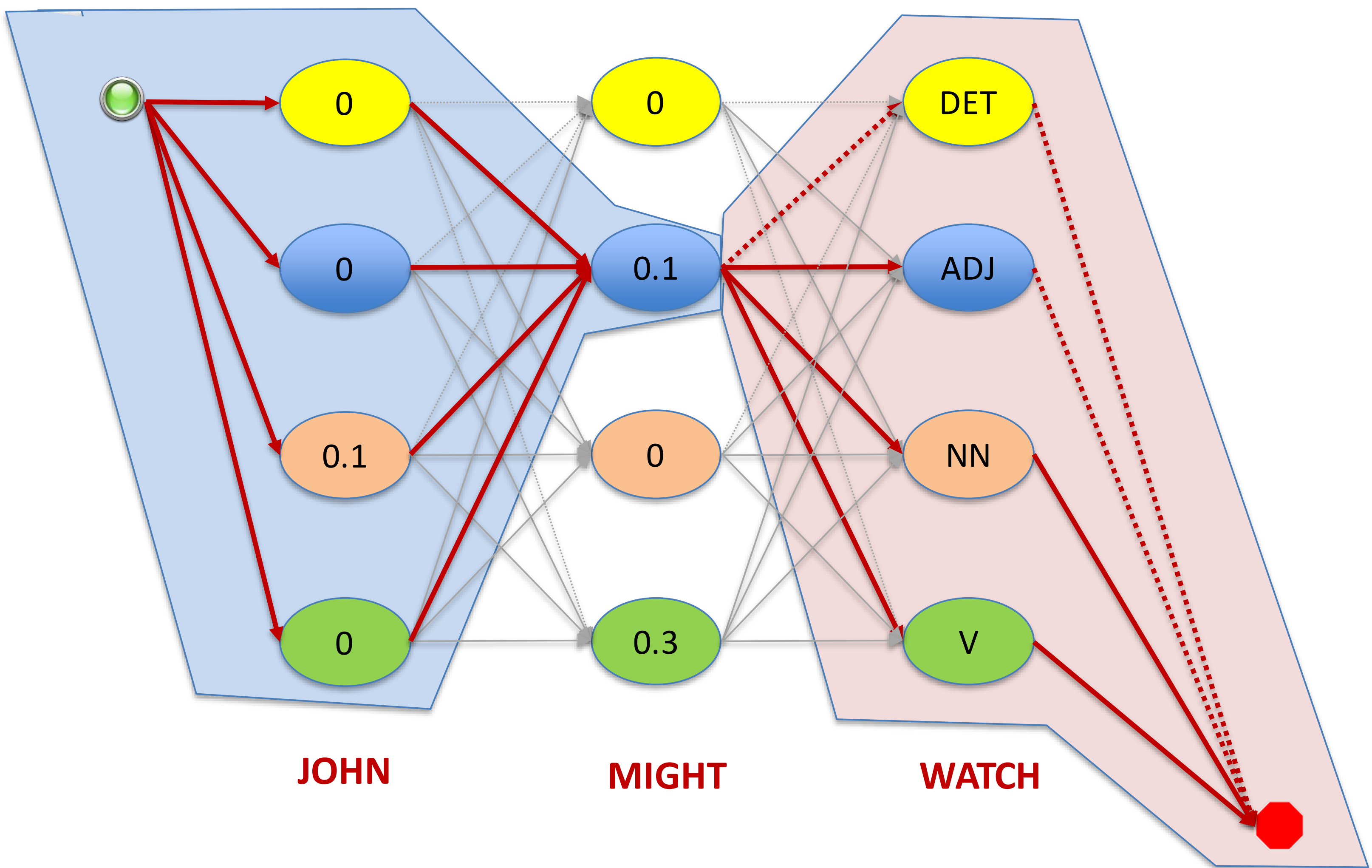


$$\beta_t(q) = \sum_{r \in \Omega} \eta(q \to r)\,\gamma(r \downarrow x_{t+1})\beta_{t+1}(r)$$

Forward-Backward

- Compute forward chart
$$\alpha_t(q) = p(\text{START}, x_1, \ldots, x_t, y_t = q)$$
- Compute backward chart
$$\beta_t(q) = p(x_{t+1}, \ldots, x_{|\mathbf{x}|}, \text{STOP} \mid y_t = q)$$

Forward Backward



The forward and backwards probabilities

Forward-Backward

- Compute forward chart

$$\alpha_t(q) = p(\text{START}, x_1, \ldots, x_t, y_t = q)$$

- Compute backward chart

$$\beta_t(q) = p(x_{t+1}, \ldots, x_{|\mathbf{x}|}, \text{STOP} \mid y_t = q)$$

$$\alpha_t(q) \times \beta_t(q)$$

$$= p(START, x_1, \ldots, x_t | y_t = q) p(x_{t+1}, \ldots, x_T, STOP | y_t = q)$$

$$p(\mathbf{x}, y_t = q) = \alpha_t(q) \times \beta_t(q)$$

Edge probability

- What is the probability that **x** was generated and $q \rightarrow r$ happened at time t?

$$p(x_1, \ldots, x_T, s_t = q, s_{t+1} = r) =$$

$$p(x_1, \ldots, x_t, s_t = q\ )$$

$$\eta(q \rightarrow r)\gamma(r \downarrow x_{t+1})$$

$$p(x_{t+1}, \ldots, x_T | s_{t+1} = r)$$

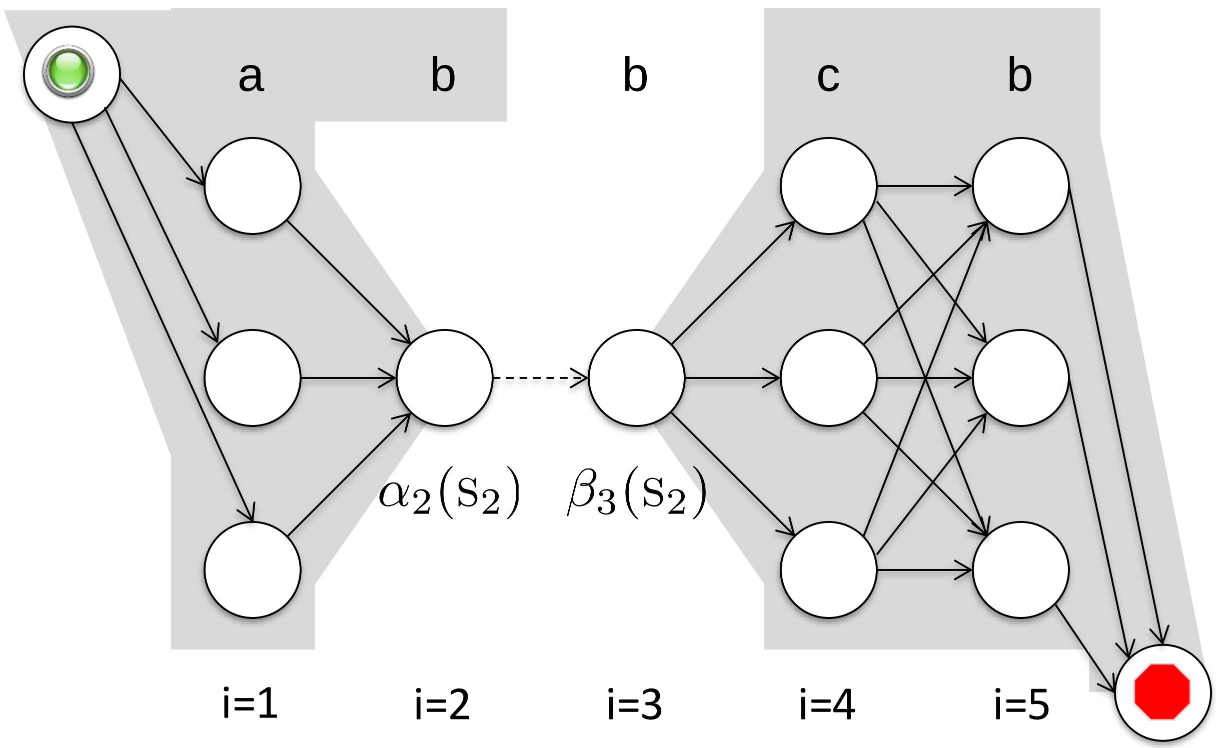
Edge probability

- What is the probability that **x** was generated and $q \rightarrow r$ happened at time t?

$$p(x_1, \ldots, x_T, s_t = q, s_{t+1} = r) =$$
$$p(x_1, \ldots, x_t, s_t = q\ )$$
$$\eta(q \rightarrow r)\gamma(r \downarrow x_{t+1})$$
$$p(x_{t+1}, \ldots, x_T | s_{t+1} = r)$$

$$\alpha_t(q)\eta(q \rightarrow r)\gamma(r \downarrow x_{t+1})\beta_{t+1}(r)$$

Forward-Backward



$$p(x_1, \ldots, x_T, s_t = q, s_{t+1} = r) = \alpha_t(q)\eta(q \rightarrow r)\gamma(r \rightarrow x_{t+1})\beta_{t+1}(r)$$

Actual Marginals

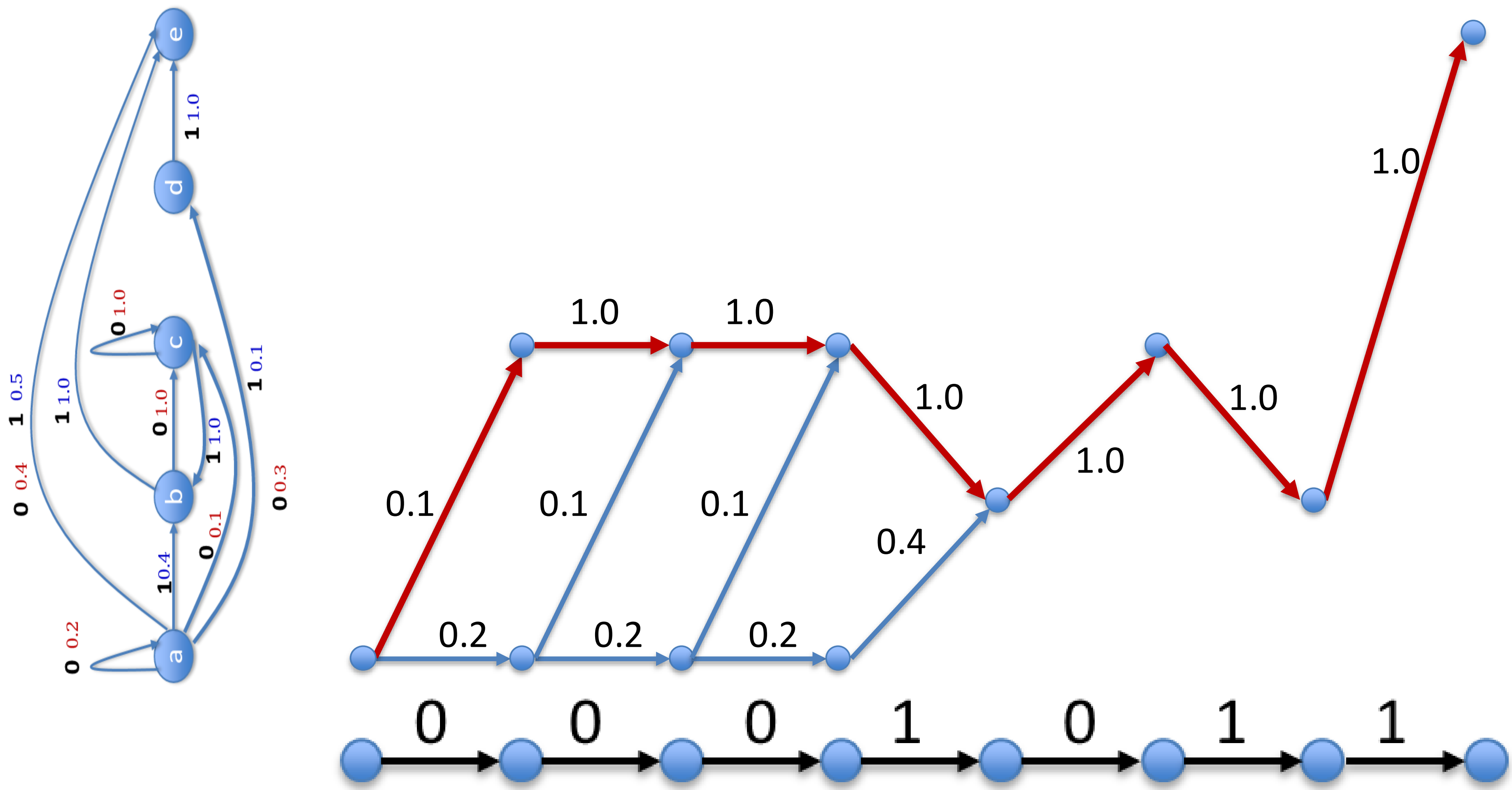
- Posterior Marginal

$$p(s_i = q | x_1, \ldots, x_T) = \frac{p(x_1, \ldots, x_T, s_i = q)}{p(x_1, \ldots, x_T\ )} = \frac{\alpha_t(q)\beta_t(q)}{\alpha_{T+1}(STOP)}$$

- Edge Marginal

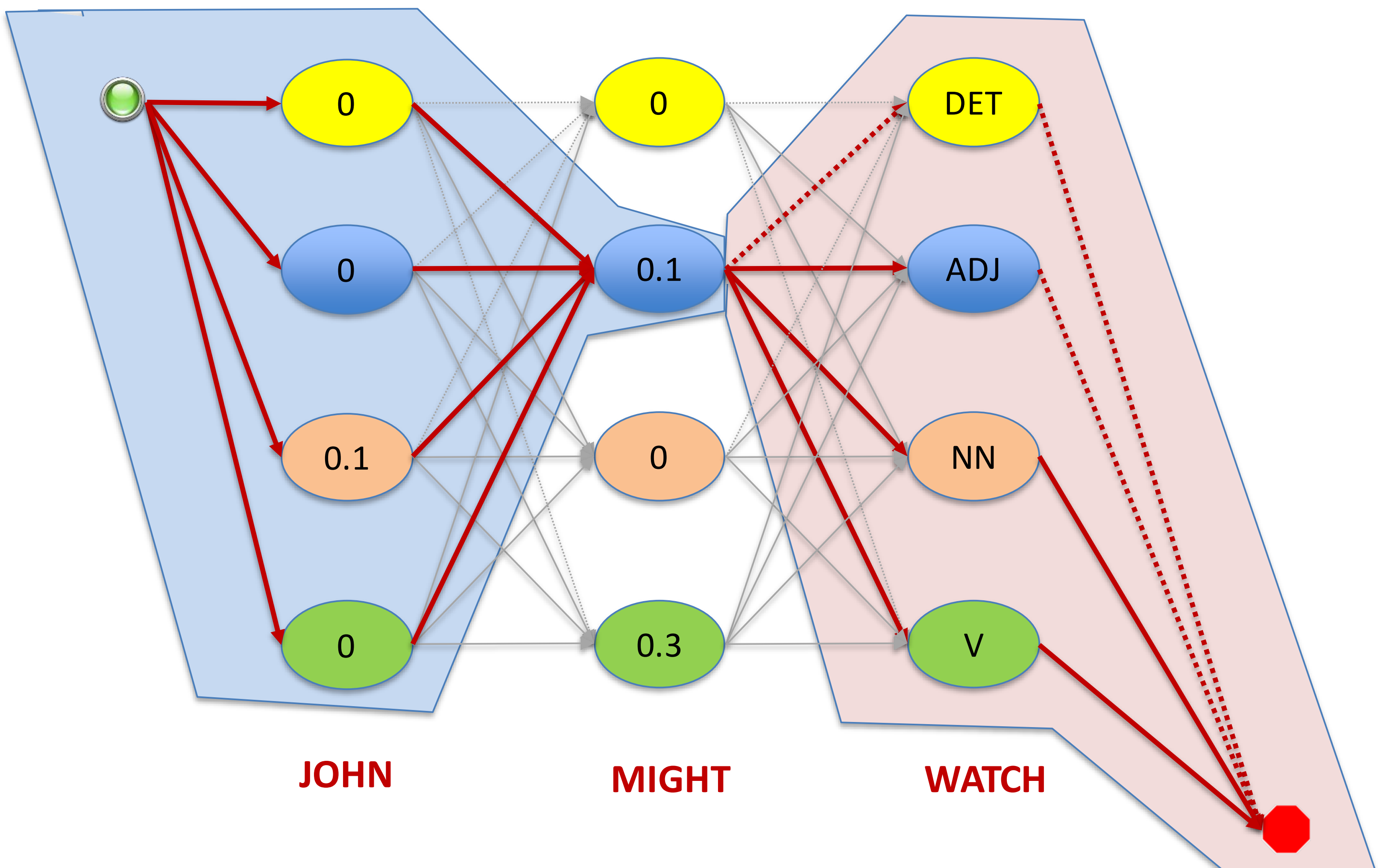
$$p(s_i = q, s_{i+1} = r | x_1, \ldots, x_T) = \frac{\alpha_t(q)\eta(q \to r)\gamma(r \to x_{i+1})\beta_{t+1}(r)}{\alpha_{T+1}(STOP)}$$

RECAP: Inference from PFA



- The probability of any given state sequence is the product of the probabilities on all the edges representing the state sequence
  - The probability of *a c c c b c b e* is 0.1

Recap: Inference on HMMs



The forward and backwards probabilities

What we've done

- Able to answer questions about marginal distributions of components of finite-state models of language
  - Finite state grammars
  - HMMs

- E.g.
  - How probable is state q at time t, given $\mathbf{x}$
    - E.g. how likely is it that the second word in "John might watch" is an adjective
  - How probable is the state transition $q \rightarrow r$ at time t, given $\mathbf{x}$
    - E.g. how likely is it that "might watch" consists of an adjective followed by a verb

A different problem

- We’ve answered the following questions:
  - How probable is state q at time t, given **x**
  - How probable is the state transition $q \rightarrow r$ at time t, given **x**

- More generic question:
  - How probable is it that the process visited state q, given **x**
    - Is there an adjective in “John might watch”?
  - How probable is it that the transition $q \rightarrow r$ occurred, given **x**
    - Does the sentence have an adjective followed by a verb?

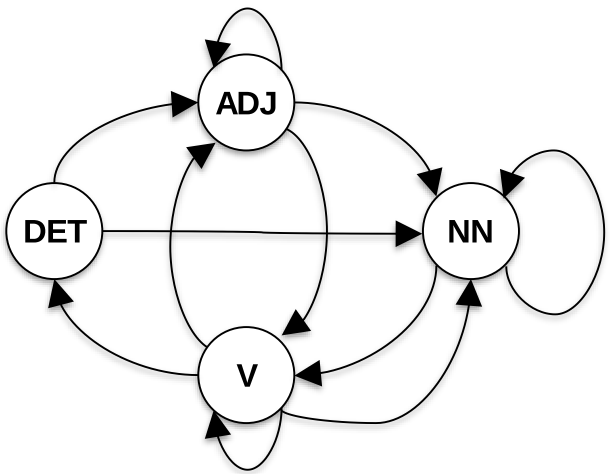
HMMs are PCFGs too

Initial Probabilities:

	DET	ADJ	NN	V
start	0.5	0.1	0.3	0.1

η **Transition Probabilities:**

	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
stop	0.0	0.0	0.2	0.1



γ **Emission Probabilities:**

DET		ADJ		NN		V	
the	0.7	green	0.1	book	0.3	might	0.2
a	0.3	big	0.4	plants	0.2	watch	0.3
		old	0.4	people	0.2	watches	0.2
		might	0.1	person	0.1	loves	0.1
				John	0.1	reads	0.19
				watch	0.1	books	0.01

EXERCISE: Convert this HMM to a PCFG

HMM→PCFG

- Split a state into State-transition NT and State-emission NT

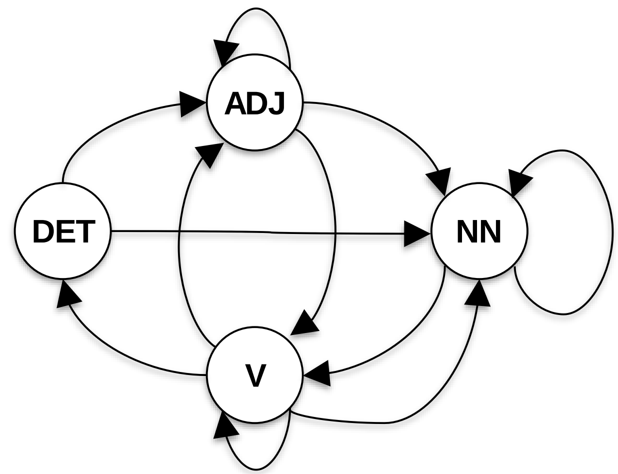
$$S \rightarrow Q_i \; \{Q_i = \; ADJ, NN, DET, V\} \;\; P_{in}(Q_i)$$
$$Q_i \rightarrow E_i Q_j \; \{Q_j = \; ADJ, NN, DET, V, STOP\} \; P(Q_j | Q_i)$$
$$E_i \rightarrow word \;\; P(word \, | E_i)$$

- Note: We do not define the second rule for the $STOP$ state

HMMs can be cast as PFAs too

Initial Probabilities:

DET	ADJ	NN	V
0.5	0.1	0.3	0.1



η **Transition Probabilities:**

	DET	ADJ	NN	V
DET	0.0	0.0	0.0	0.5
ADJ	0.3	0.2	0.1	0.1
NN	0.7	0.7	0.3	0.2
V	0.0	0.1	0.4	0.1
	0.0	0.0	0.2	0.1

γ **Emission Probabilities:**

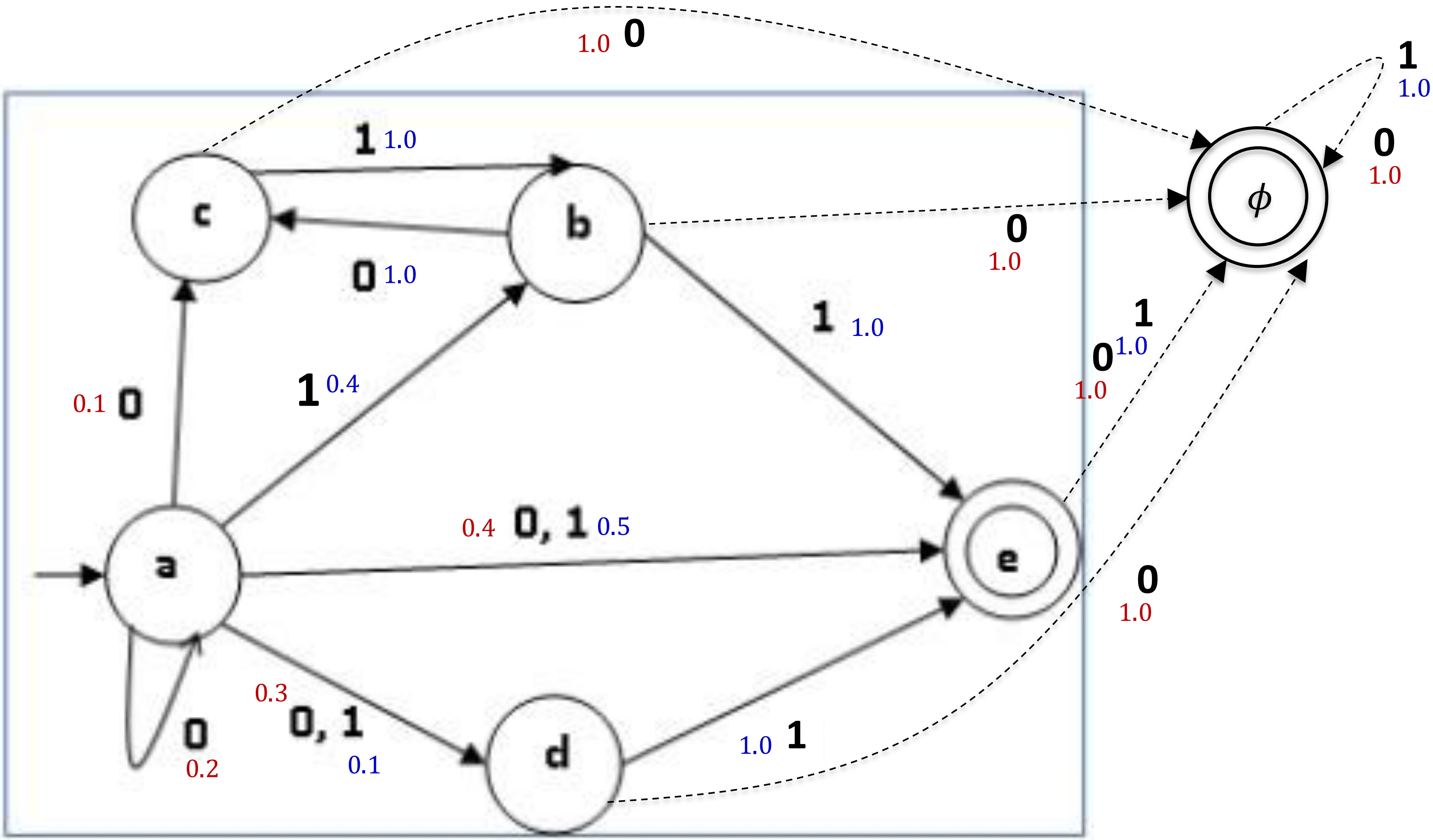
DET		ADJ		NN		V	
the	0.7	green	0.1	book	0.3	might	0.2
a	0.3	big	0.4	plants	0.2	watch	0.3
		old	0.4	people	0.2	watches	0.2
		might	0.1	person	0.1	loves	0.1
				John	0.1	reads	0.19
				watch	0.1	books	0.01

EXERCISE: Convert this HMM to a PFA

HMM→PFA

- First, recall an earlier grammar

Inference in a PFA



- *We aren't interested in state sequences which end in ϕ.*
  - So no need to explicitly represent it

HMM→PFA: Back to our grammar

- Let $Q = \{ADJ, NN, V, DET, STOP\}$
- Let $W = \{w_1, w_2, \ldots, w_N, \blacksquare\}$
  - $\blacksquare$ is a termination symbol to signify end of sentence
- PFA:

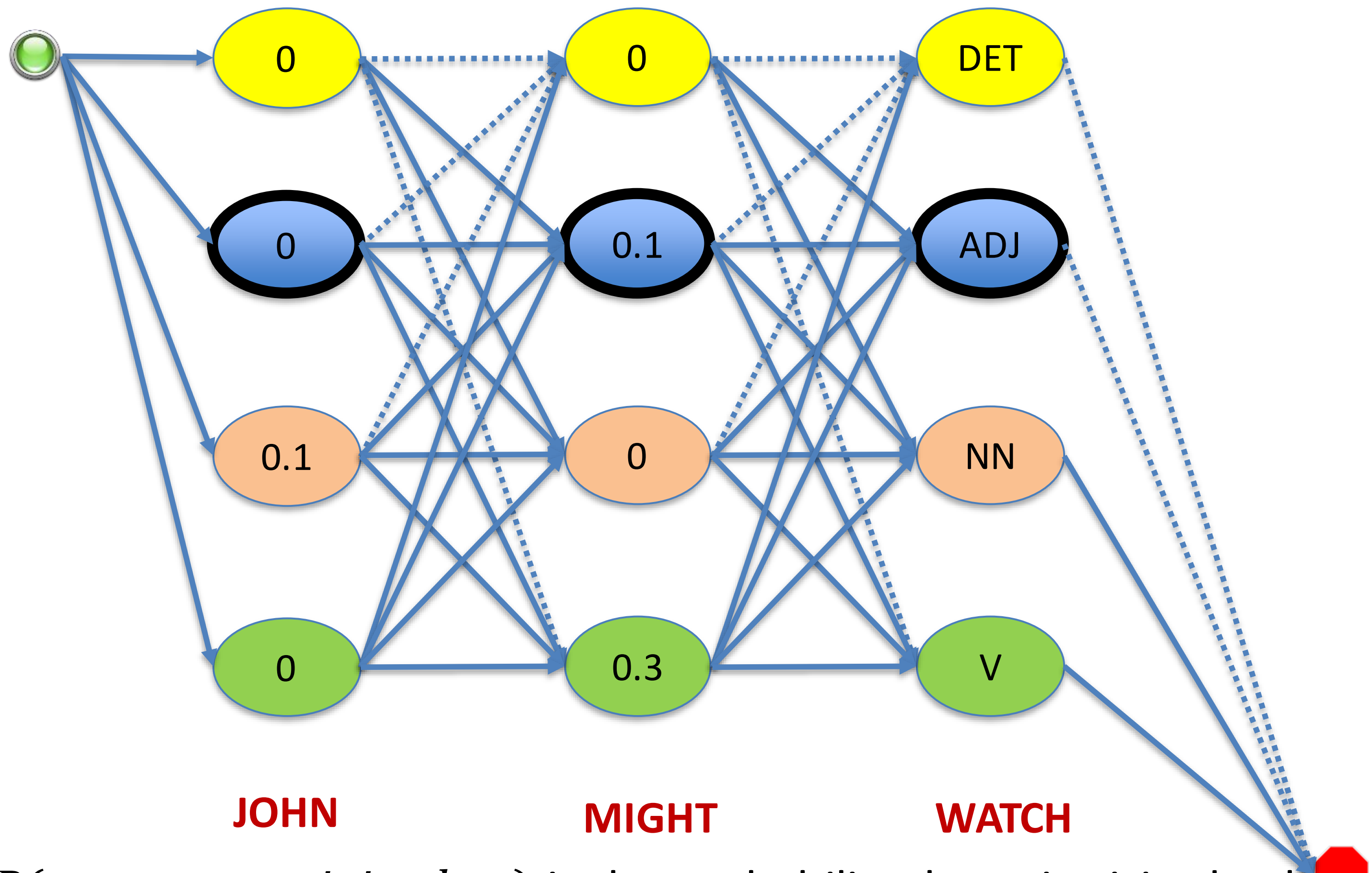
$$\pi(Q_i) = P_{in}(Q_i)$$

$$Q_i \xrightarrow{w} Q_j \; : \; P = P\left(Q_j \middle| Q_i\right) P(w|Q_i) \quad \forall i, j, w$$

A different problem

- We've answered the following questions:
  - How probable is state q at time t, given $\mathbf{x}$
  - How probable is the state transition $q \rightarrow r$ at time t, given $\mathbf{x}$
- More generic question:
  - How probable is it that the process visited state q, given $\mathbf{x}$
    - E.g. does the sentence have an adjective

Visiting a state



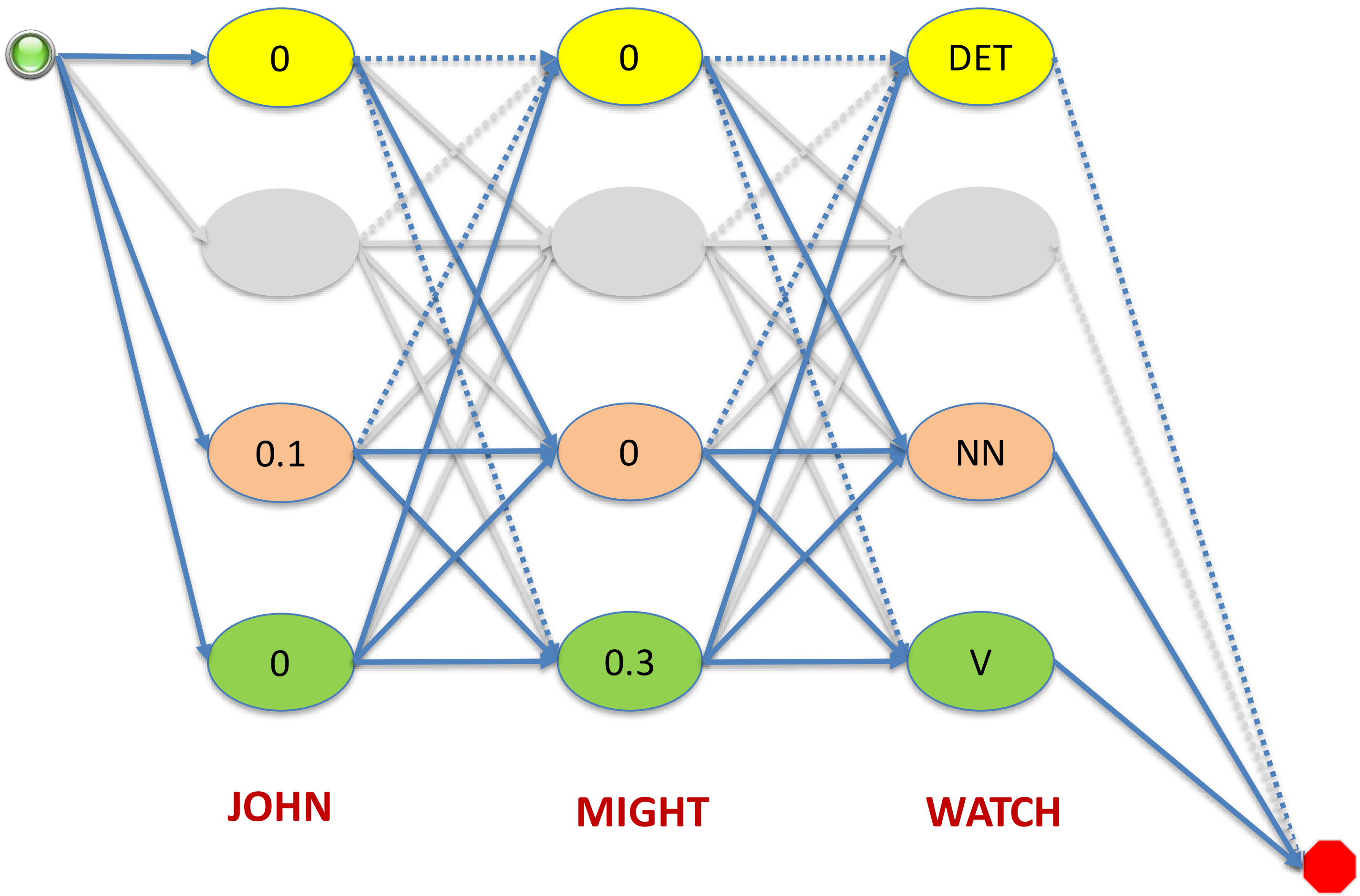
- $P(state\ s\ =\ visited,\ \mathbf{x})$ is the probability that s is visited at least once.
  - E.g. "What is the probability that at least one of the words is an adjective
- This is the total probability of the subset of the trellis where every path from start to end visits state s (e.g. ADJ) at least once
  - ***Its generally difficult or impossible to isolate this portion of the trellis***

Derivation by ablation

- $p(\mathbf{x}) = p(s, \mathbf{x}) + p(\bar{s}, \mathbf{x})$
  - s = state s is visited at least once
  - $\bar{s}$ = state s is never visited

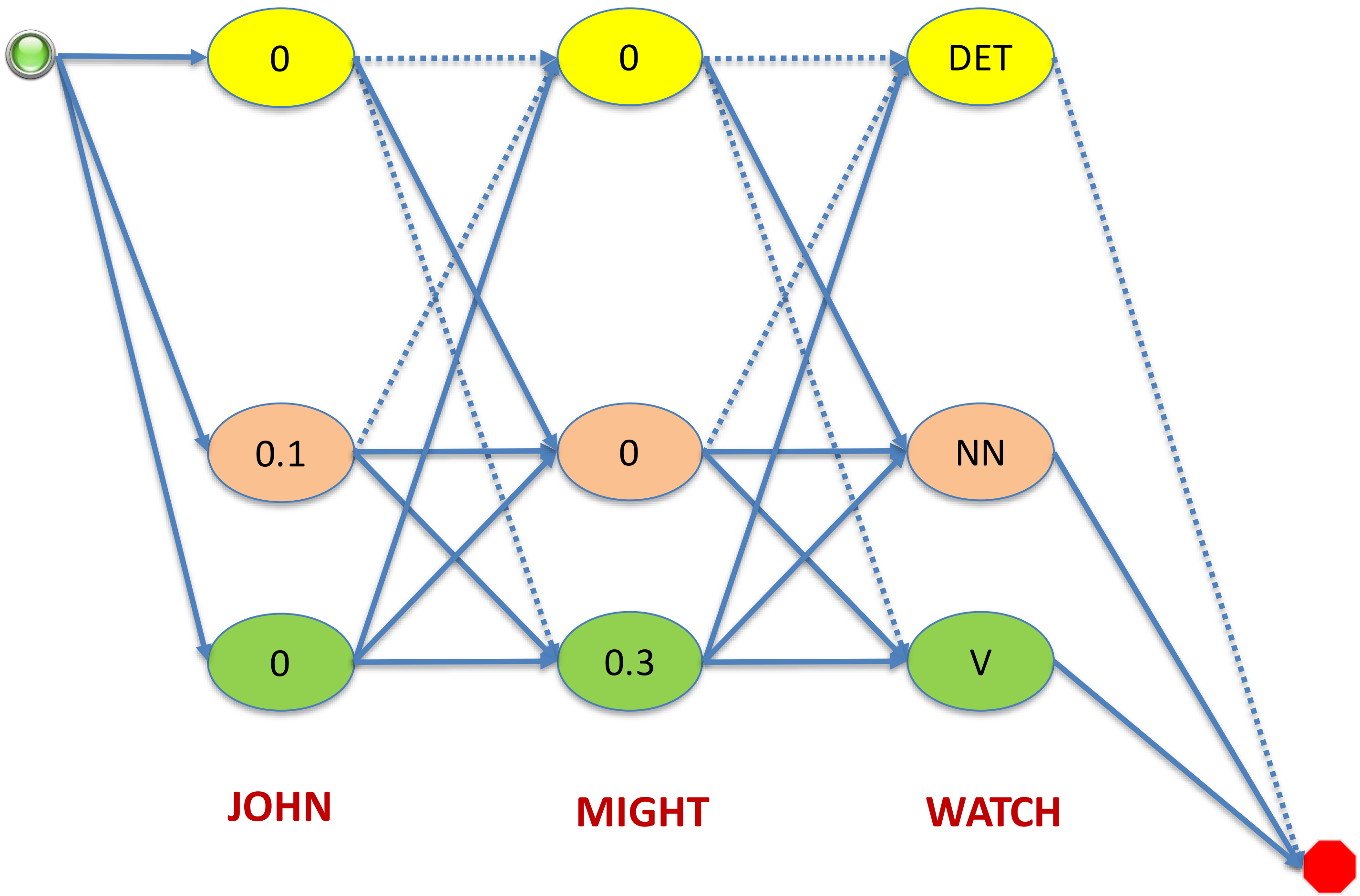
- $p(s, \mathbf{x}) = p(\mathbf{x}) - p(\bar{s}, \mathbf{x})$
- $p(s|\mathbf{x}) = 1 - \frac{p(\bar{s}, \mathbf{x})}{p(\mathbf{x})}$

Not visiting a state



- The portion of the trellis where *no* path visits state *s*
  - This is complete; there are no other paths that do not visit *s*

Not visiting a state



- The portion of the trellis where *no* path visits state s
  - This is complete; there are no other paths that do not visit s
  - The total probability of this trellis is $p(\bar{s}, \mathbf{x})$
    - Can be computed using the forward algorithm on this trellis

Derivation by ablation

- $p(\mathbf{x}) = p(s, \mathbf{x}) + p(\bar{s}, \mathbf{x})$
  - s = state s is visited at least once
  - $\bar{s}$ = state s is never visited

- $p(s, \mathbf{x}) = p(\mathbf{x}) - p(\bar{s}, \mathbf{x})$
- $p(s|\mathbf{x}) = 1 - \frac{p(\bar{s},\mathbf{x})}{p(\mathbf{x})}$

Computed by the forward algorithm on the *reduced* trellis

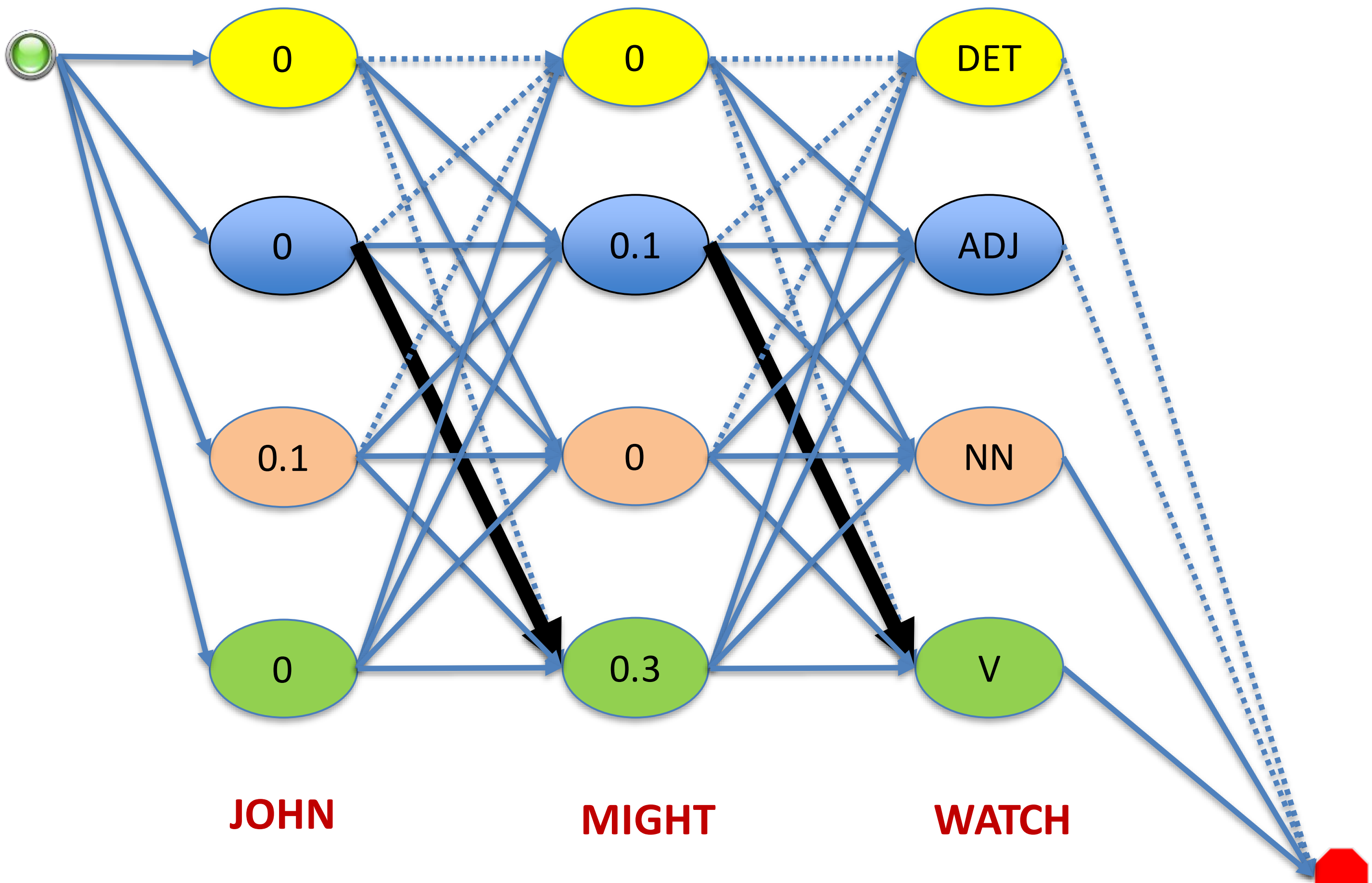
Computed by the forward algorithm on the complete trellis

A different problem

- We've answered the following questions:
  - How probable is state q at time t, given **x**
  - How probable is the state transition $q \rightarrow r$ at time t, given **x**

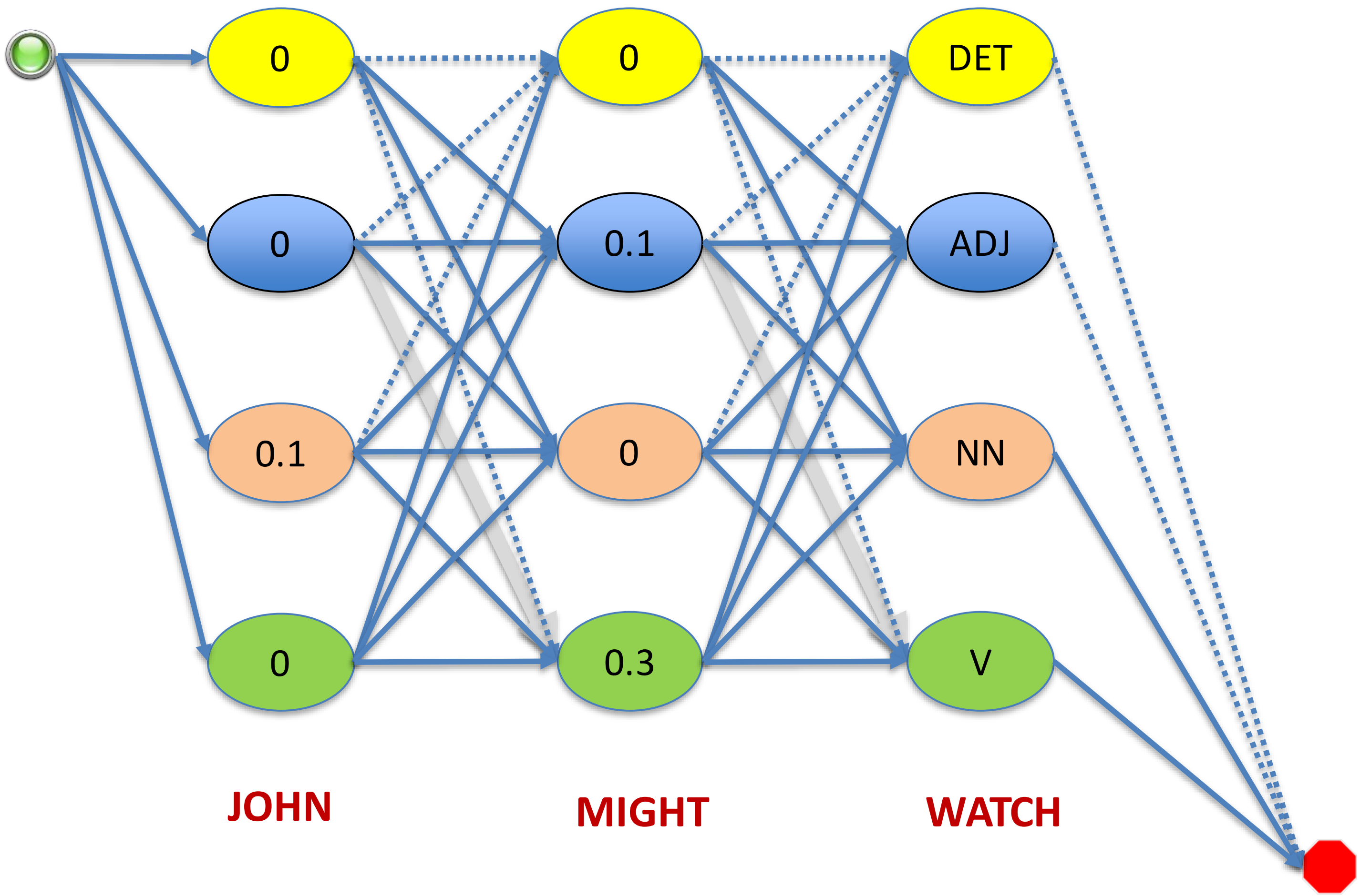
- More generic question:
  - How probable is it that the process visited state q, given **x**
  - How probable is it that the transition $q \rightarrow r$ occurred, given **x**
    - E.g. Is an adjective followed by a verb in this sentence?

Visiting a transition



- This is the total probability of all paths that use the shown transition
  - No possible to isolate this portion of the trellis

Not visiting a transition

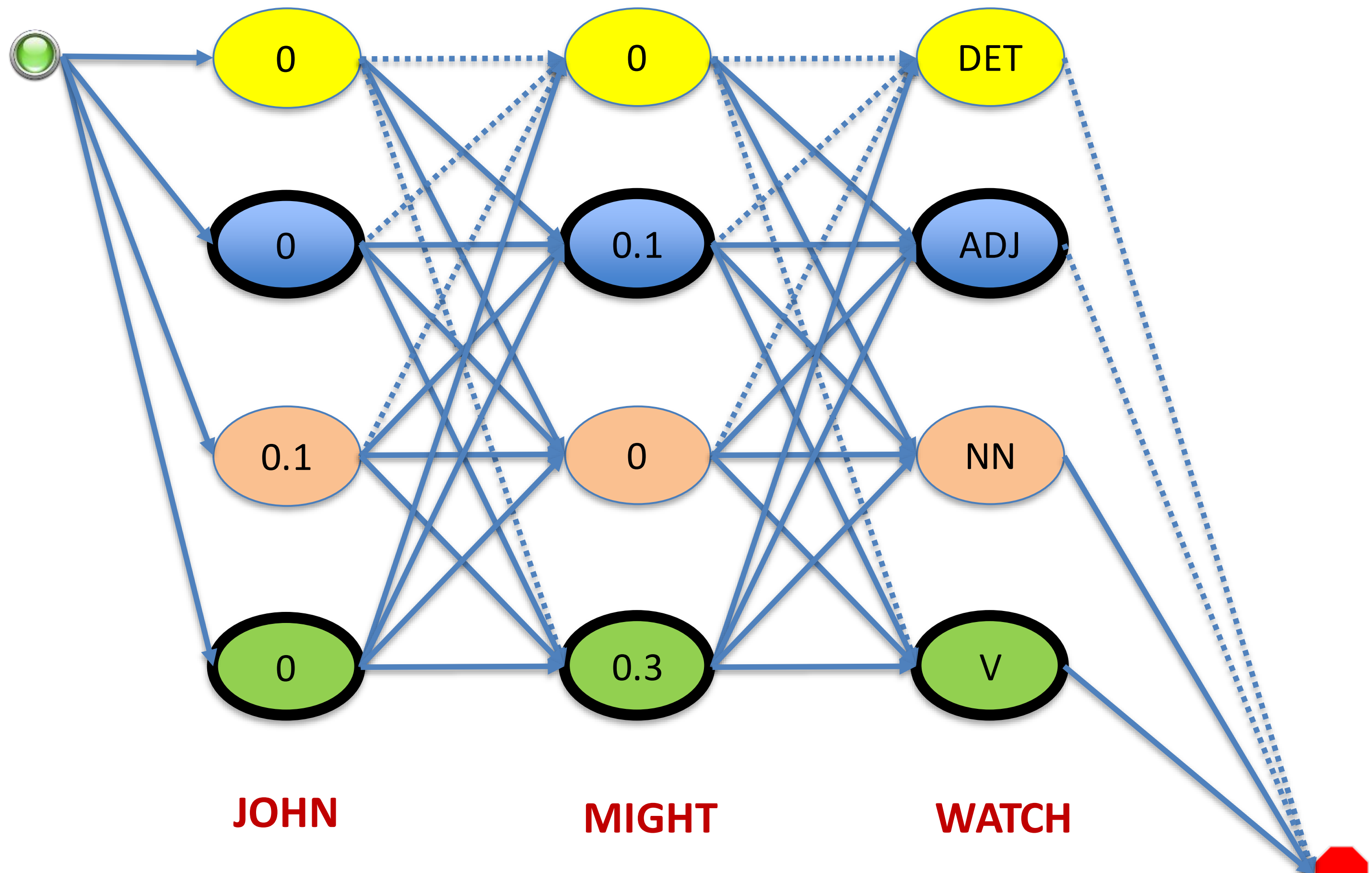


- Delete the transition and compute the total probability of the remaining trellis
  - $p(\mathbf{x}, state\ q\ is\ never\ followed\ by\ state\ r)$

More complex inferences

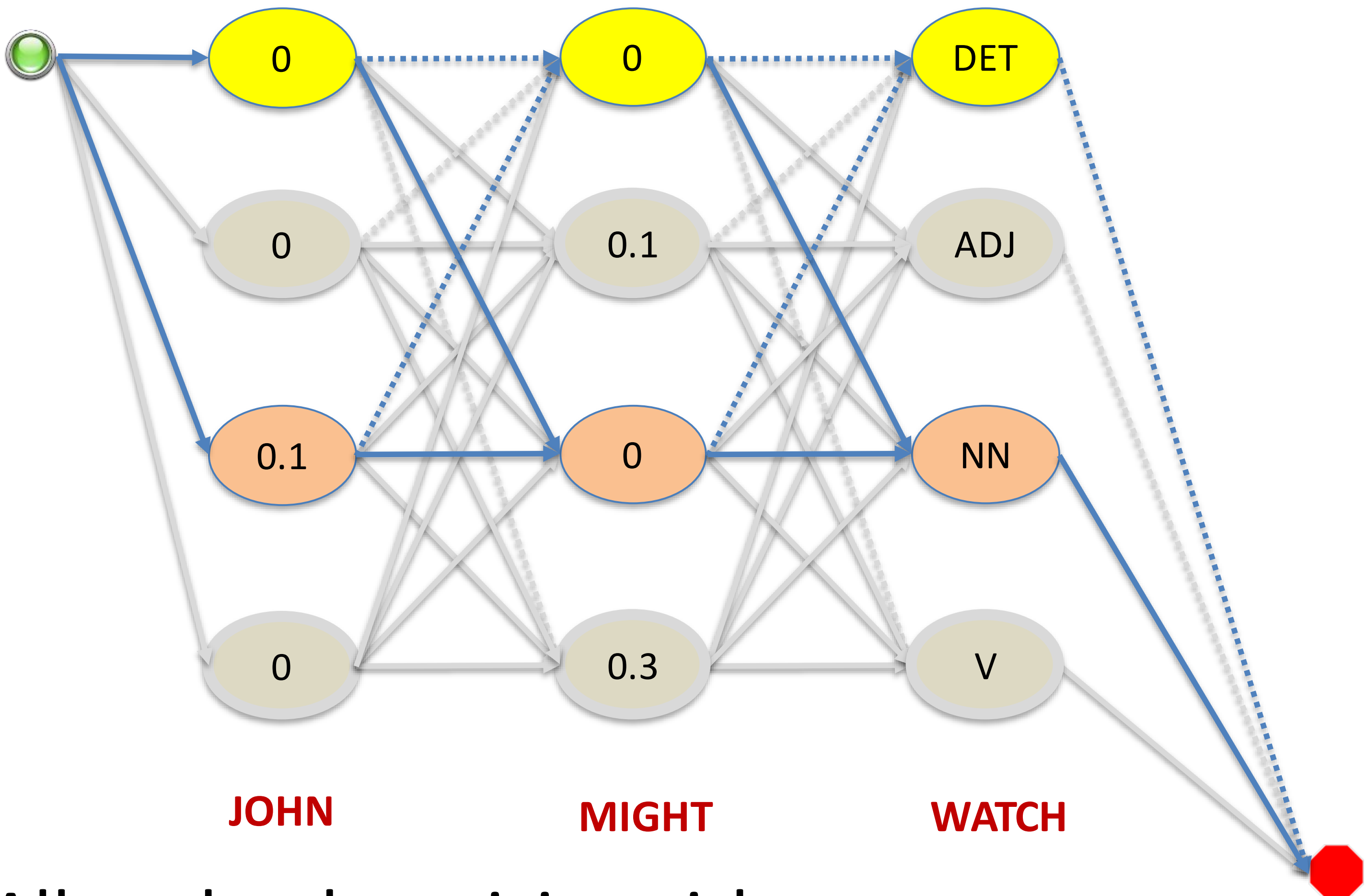
- What is the probability that *both* state *s* AND state *r* were visited?
  - What is the probability that “John might watch” includes both a verb and an adjective?

Visiting multiple states



- Total probability of all paths that visit both the blue and green states
  - Again, not possible to isolate the corresponding portion of the trellis

NOT visiting multiple states



- All paths that visit neither state

$$\overline{(s \cup r)}$$

More complex inferences

- What is the probability that *both* state *s* AND state *r* were visited?

$$p(s \cap r) = 1 - (p(\bar{s}) + p(\bar{r}) - p(\overline{s \cup r}))$$

$$P(s\ , r, \mathbf{x})$$
$$= P(\mathbf{x}) - (P(\bar{s}, \mathbf{x}) + P(\bar{r}, \mathbf{x}) - P(\overline{s \cup r}, \mathbf{x})\ )$$

$$P(s\ , r|\mathbf{x}) = 1 - \frac{P(\bar{s}, \mathbf{x}) + P(\bar{r}, \mathbf{x}) - P(\overline{s \cup r}, \mathbf{x})}{P(\mathbf{x})}$$

More complex inferences

- What is the probability that *both* state *s* AND state *r* were visited?

$$p(s \cap r) = 1 - (p(\bar{s}) + p(\bar{r}) - p(\overline{s \cup r}))$$

Computed from the Trellis with the *r* row removed

$$P(s, r, \mathbf{x}) = P(\mathbf{x}) - (P(\bar{s}, \mathbf{x}) + P(\bar{r}, \mathbf{x}) - P(\overline{s \cup r}, \mathbf{x}))$$

Computed from the Trellis with the *s* row removed

Computed from the Trellis with both *s* and *r* rows removed

$$P(s, r|\mathbf{x}) = 1 - \frac{P(\bar{s}, \mathbf{x}) + P(\bar{r}, \mathbf{x}) - P(\overline{s \cup r}, \mathbf{x})}{P(\mathbf{x})}$$

Computed from the full Trellis

More complex inferences

- Other more complex inferences can be similarly obtained
- Becomes increasingly more computationally expensive as the order of the inference increases

Higher-level grammars

- We have derived probabilistic inferences from PFAs and HMMs
- More generally we want similar inferences from CFGs and PCFGs
  - Given word sequence ***w***, what is the probability of having a constituent of type Z from *i* to *j*?
    - '*A person who trusts no one can't be trusted*' : what is the probability that the "A person who trusts no one" is a noun phrase?
  - Given **w**, what is the probability of having a constituent of ***any*** type from *i* to *j*?
    - What is the probability that `*A person…trusted'* is *any type* of phrase
  - Given **w**, what is the probability of using rule Z -> XY to derive the span from *i* to *j*?
    - What is the probability that VP → V N generated *"can't be trusted"*
- That will require a generalization of the algorithms we just saw..

Generalizing Forward-Backward

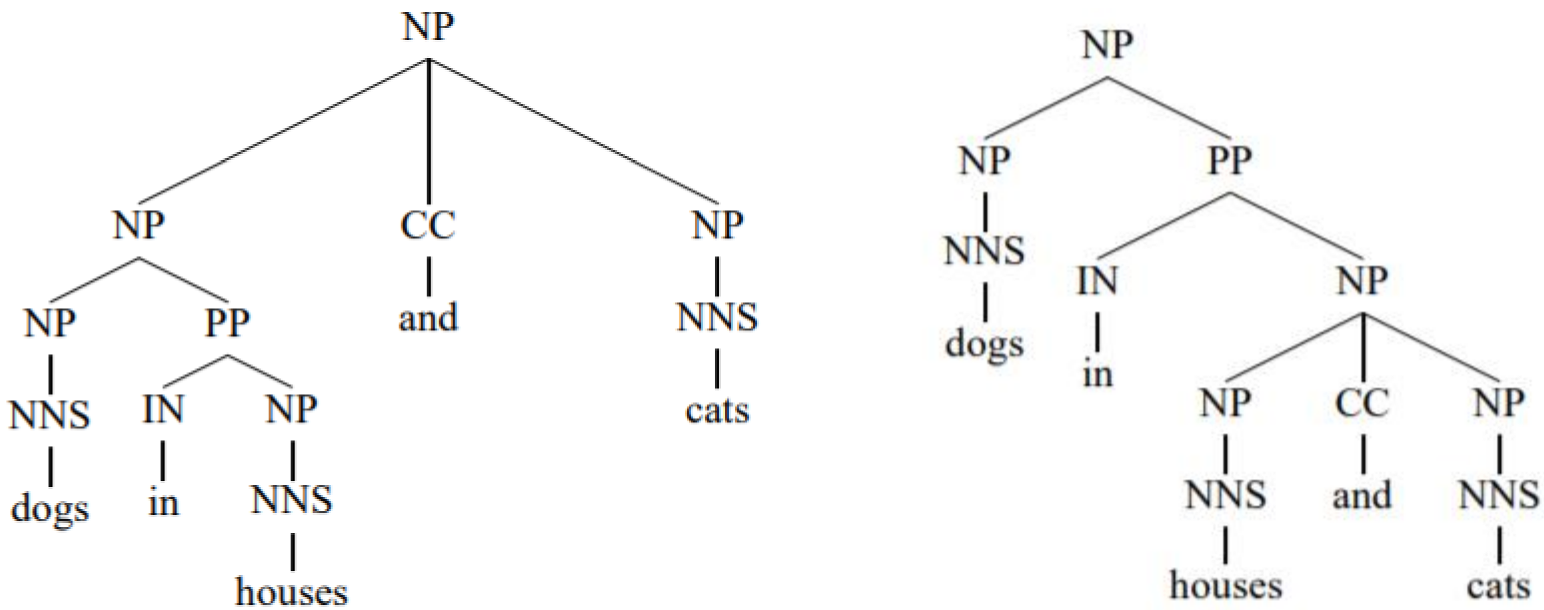
- Inference in HMMs was performed using the forward-backward algorithm
  - Recall that HMMs are instances of PCFGs
- For more general PCFGs we will use the inside-outside algorithm
  - A generalization of the forward backward algorithm
  - Builds upon the CKY algorithm

Inside/Outside Algorithm

Have you seen this man somewhere?

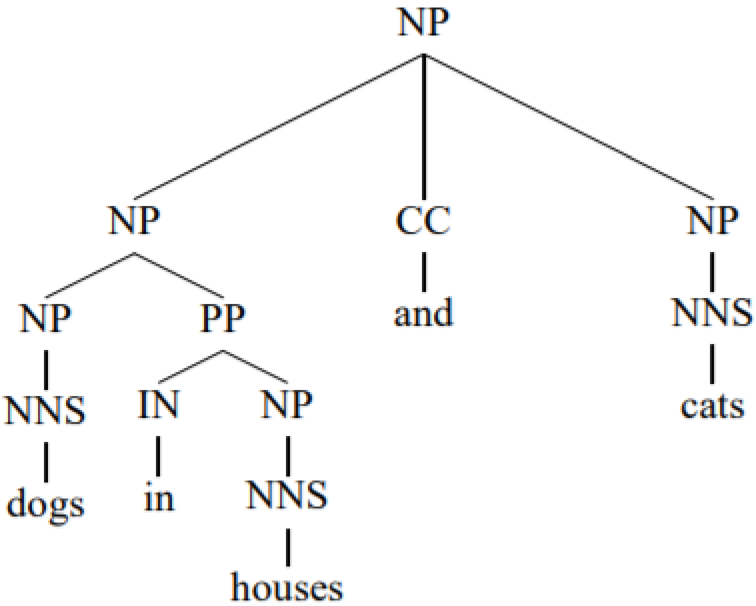
- "Trainable grammars for speech recognition," J. K. Baker, 1979

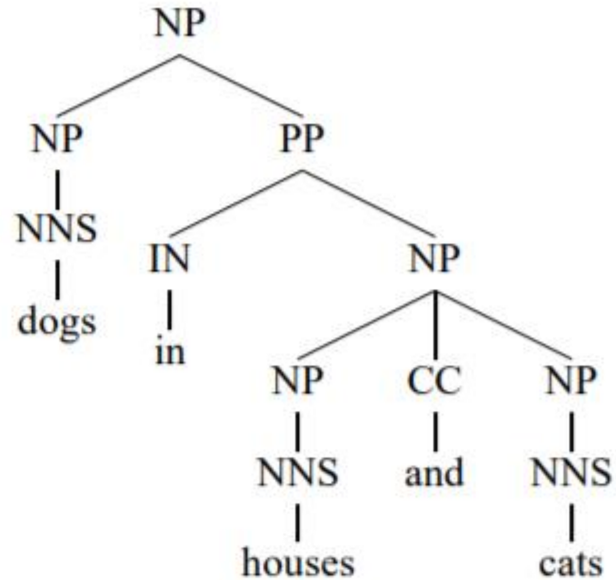
Inferences we would like to make..



- What is the probability of “dogs in houses and cats”?
- What is the probability that “houses and cats” is a clause by itself?
  - What is the probability that its an *NP?*
- Is there a *PP* in the sentence?

Inferences we would like to make..

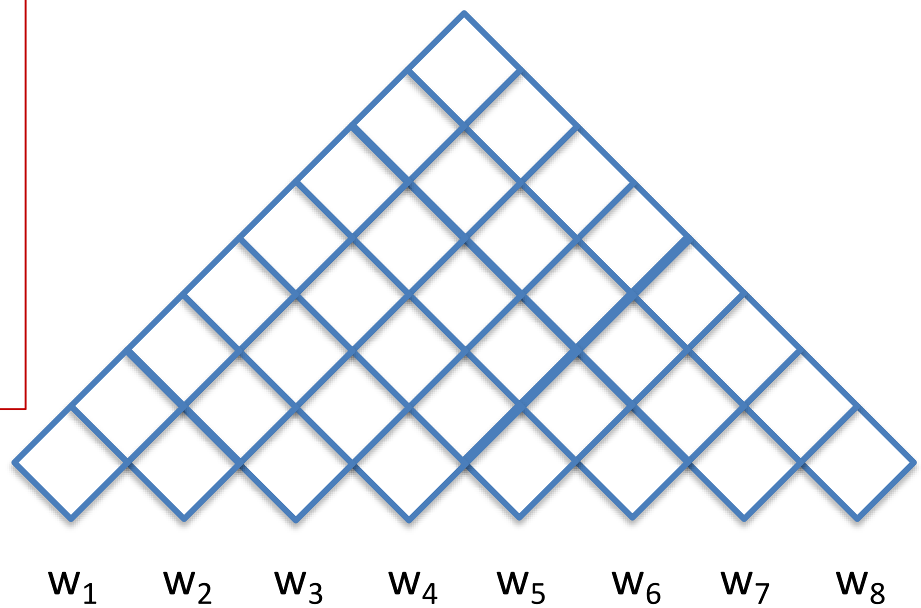




- Which of the probability of "dogs in houses and cats"
  - P("dogs in houses and cats")
- What is the probability that "houses and cats" is a clause by itself?
  - P("houses and cats" = clause | "dogs in houses and cats")
- What is the probability that its an *NP?*
  - P("houses and cats" = NP | "dogs in houses and cats")
- Is there a *PP* in the sentence?
  - P(PP | "dogs in houses and cats")

Recall the CKY algorithm

$$R_0: \quad S \rightarrow NT_1\, NT_2 \quad [P(R_0)]$$
$$R_1: NT_1 \rightarrow NT_3\, NT_4 \quad [P(R_1)]$$
$$R_2: NT_2 \rightarrow NT_5\, NT_6 \quad [P(R_2)]$$
$$..$$
$$R_k: NT_k \rightarrow w1 \quad [P(R_k)]$$
$$R_l: NT_l \rightarrow w2 \quad [P(R_l)]$$
$$..$$



- Given: A PCFG in CNF, and a word sequence
- Build a skeleton that can hold every possible tree

Recall the CKY algorithm

$R_0 : \quad S \rightarrow NT_1 \, NT_2 \quad [P(R_0)]$

$R_1 : NT_1 \rightarrow NT_3 \, NT_4 \quad [P(R_1)]$

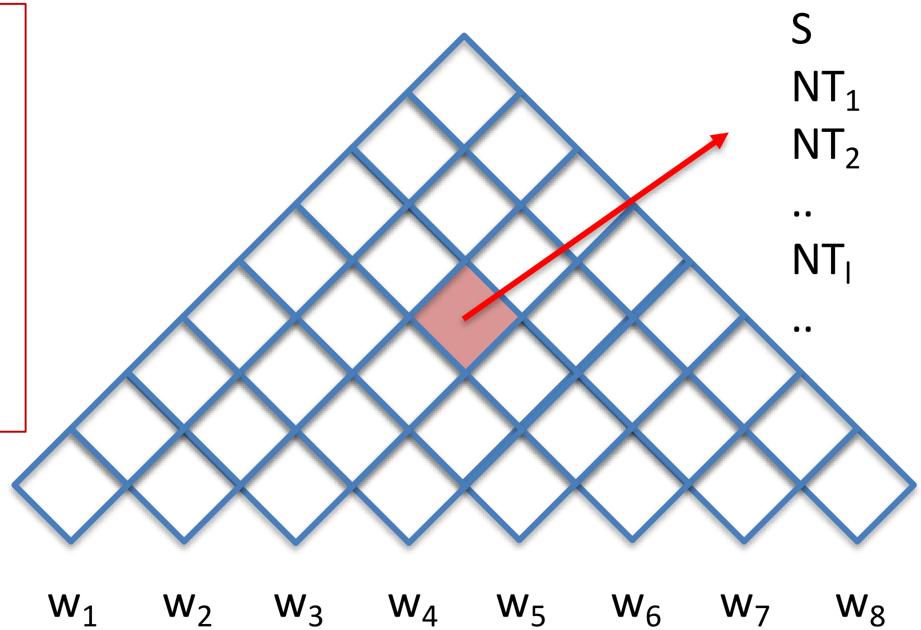
$R_2 : NT_2 \rightarrow NT_5 \, NT_6 \quad [P(R_2)]$

..

$R_k : NT_k \rightarrow w1 \quad [P(R_k)]$

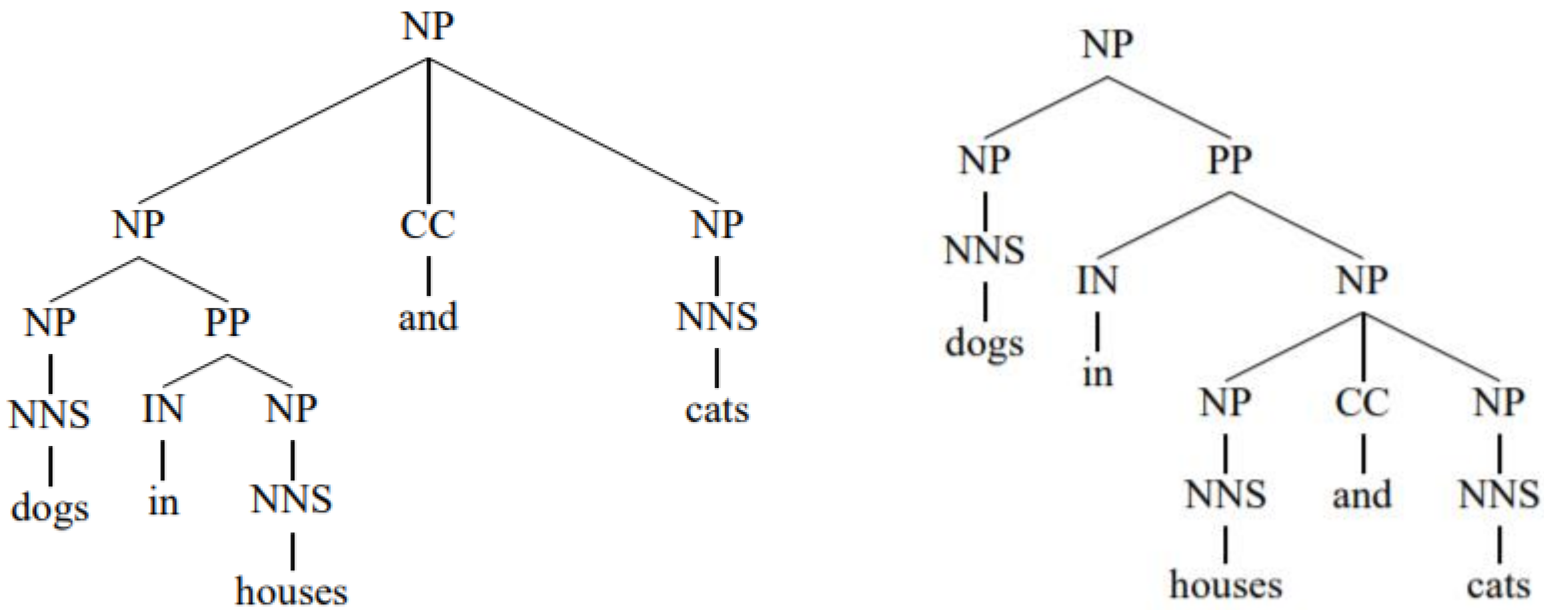
$R_l : NT_l \rightarrow w2 \quad [P(R_l)]$

..



- *Each* box in the grid (potentially) holds *every* non-terminal

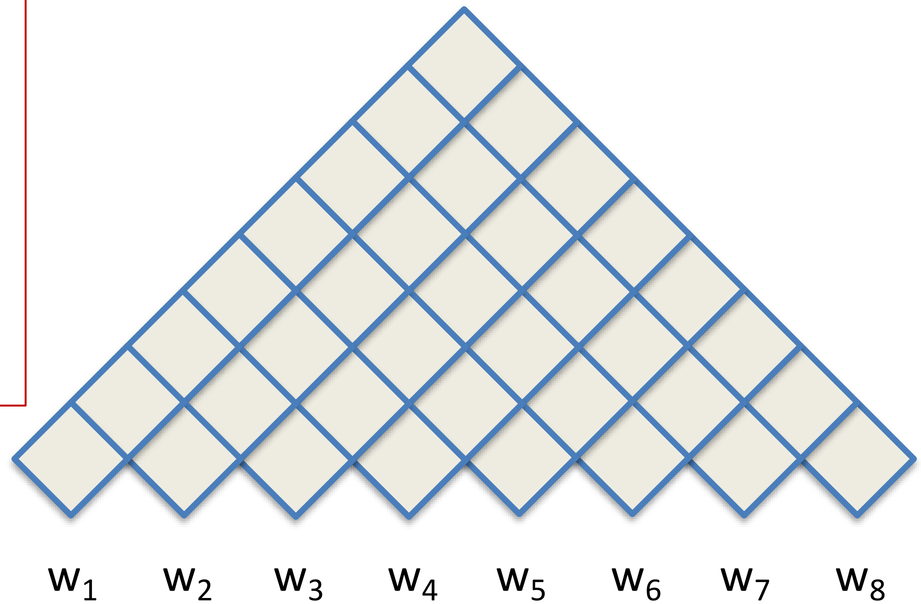
Inferences we would like to make..



- Which of the probability of “dogs in houses and cats”
  - P(“dogs in houses and cats”)
- What is the probability that “houses and cats” is a clause by itself?
  - P(“houses and cats” = clause | “dogs in houses and cats”)
- What is the probability that its an *NP?*
  - P(“houses and cats” = NP | “dogs in houses and cats”)
- Is there a *PP* in the sentence?
  - P(PP | “dogs in houses and cats”)

Probability computation using CKY

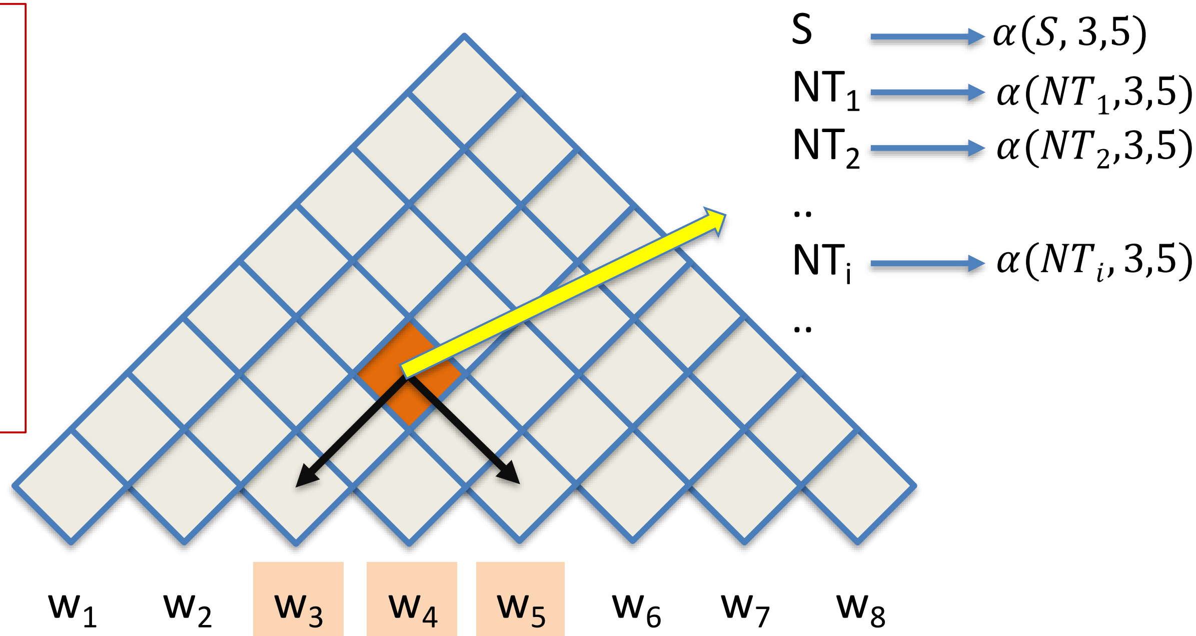
$R_0 : \quad S \rightarrow NT_1 \, NT_2 \quad [P(R_0)]$
$R_1 : NT_1 \rightarrow NT_3 \, NT_4 \quad [P(R_1)]$
$R_2 : NT_2 \rightarrow NT_5 \, NT_6 \quad [P(R_2)]$
..
$R_k : NT_k \rightarrow w1 \quad [P(R_k)]$
$R_l : NT_l \rightarrow w2 \quad [P(R_l)]$
..



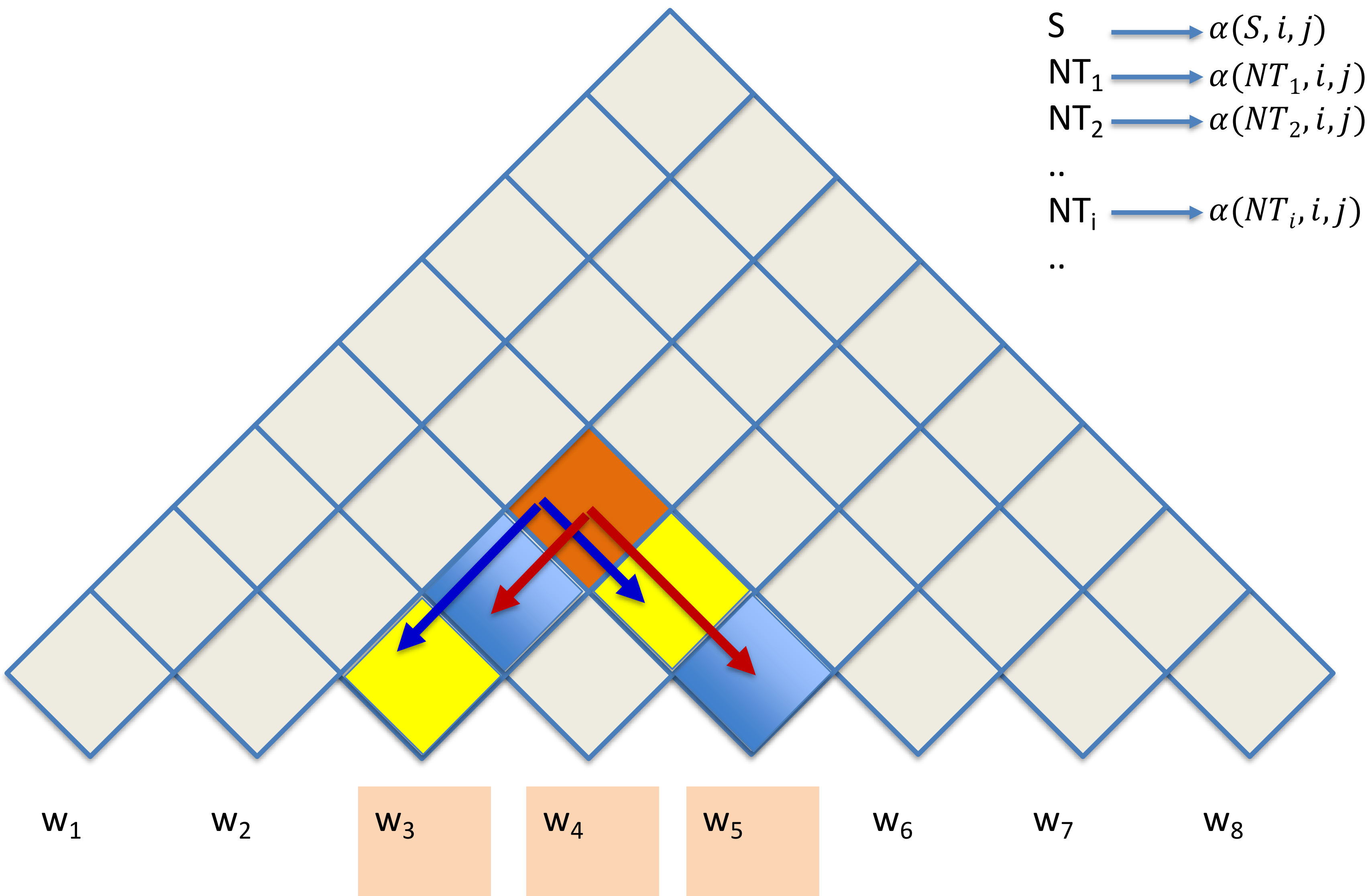
- What we desire to compute:
  - $P(w_1, \ldots, w_N)$: Probability of producing the word sequence
    - Total possibility of every possible tree that could produce the word sequence

The Inside Algorithm

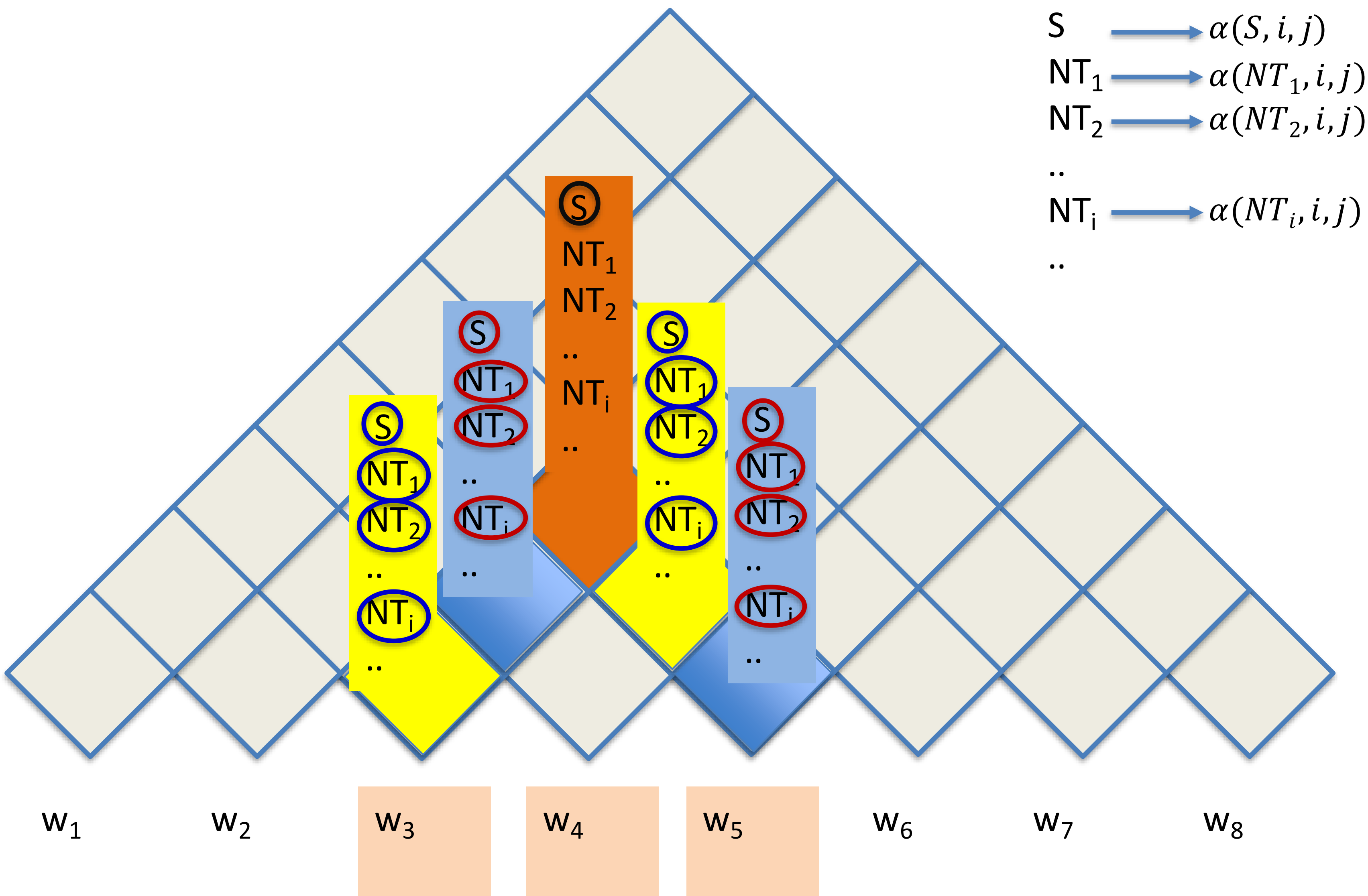
$$R_0: \quad S \rightarrow NT_1\, NT_2 \quad [P(R_0)]$$
$$R_1: NT_1 \rightarrow NT_3\, NT_4 \quad [P(R_1)]$$
$$R_2: NT_2 \rightarrow NT_5\, NT_6 \quad [P(R_2)]$$
$$..$$
$$R_k: NT_k \rightarrow w1 \quad [P(R_k)]$$
$$R_l: NT_l \rightarrow w2 \quad [P(R_l)]$$
$$..$$



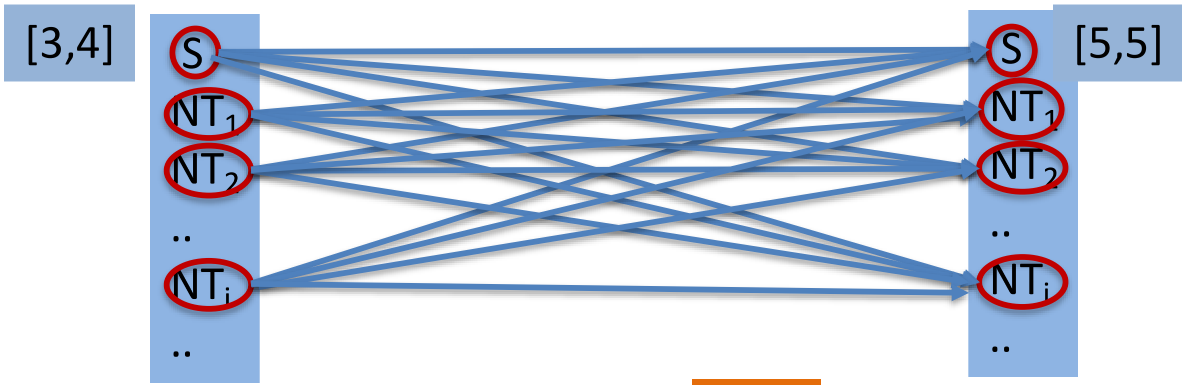
- Let $\alpha(NT, i, j)$ be the probability that the non-terminal NT produced words $w_i \dots w_j$ (at the word positions $i \dots j$ within the sentence)
  - $\alpha(NT, i, j) = p(NT \rightarrow w_i \dots w_j) = p(w_i \dots w_j | c(i, j) = NT)$



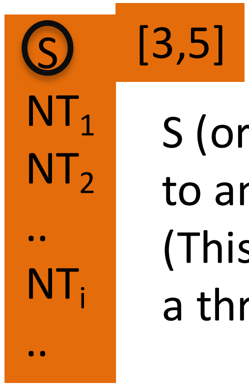
- Ways in which NT could occur at (i, j)



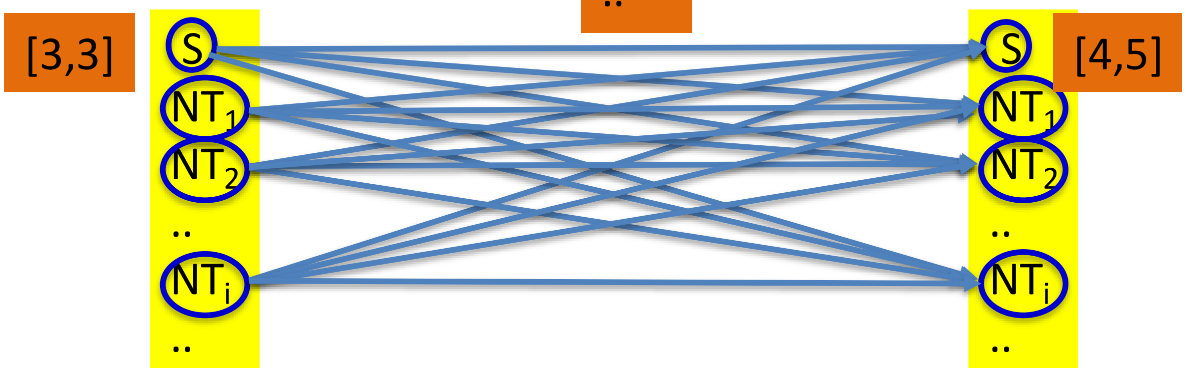
- Ways in which NT could occur at (i, j)



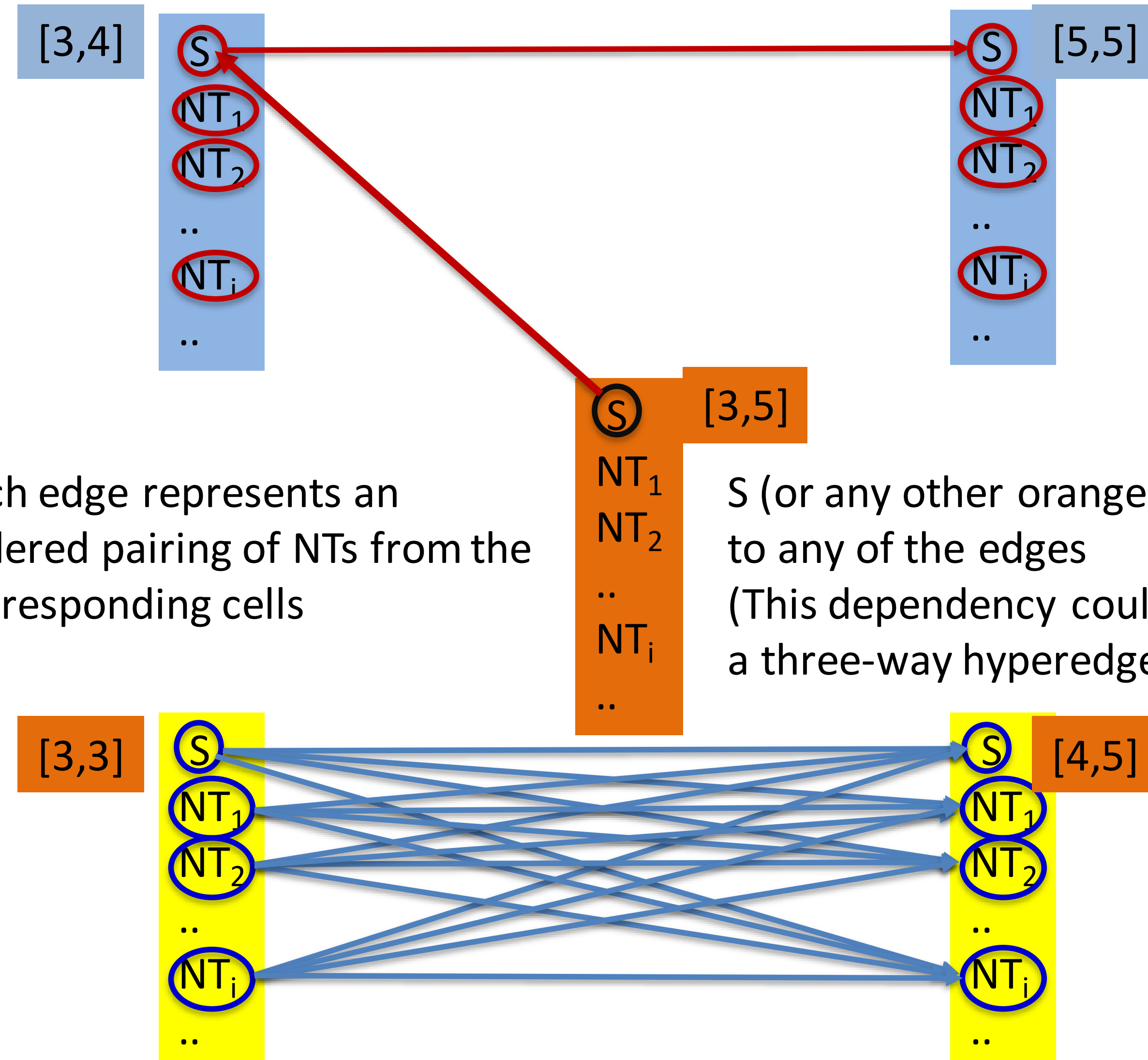
Each edge represents an ordered pairing of NTs from the corresponding cells



S (or any other orange NT) may expand out to any of the edges
(This dependency could be represented by a three-way hyperedge)



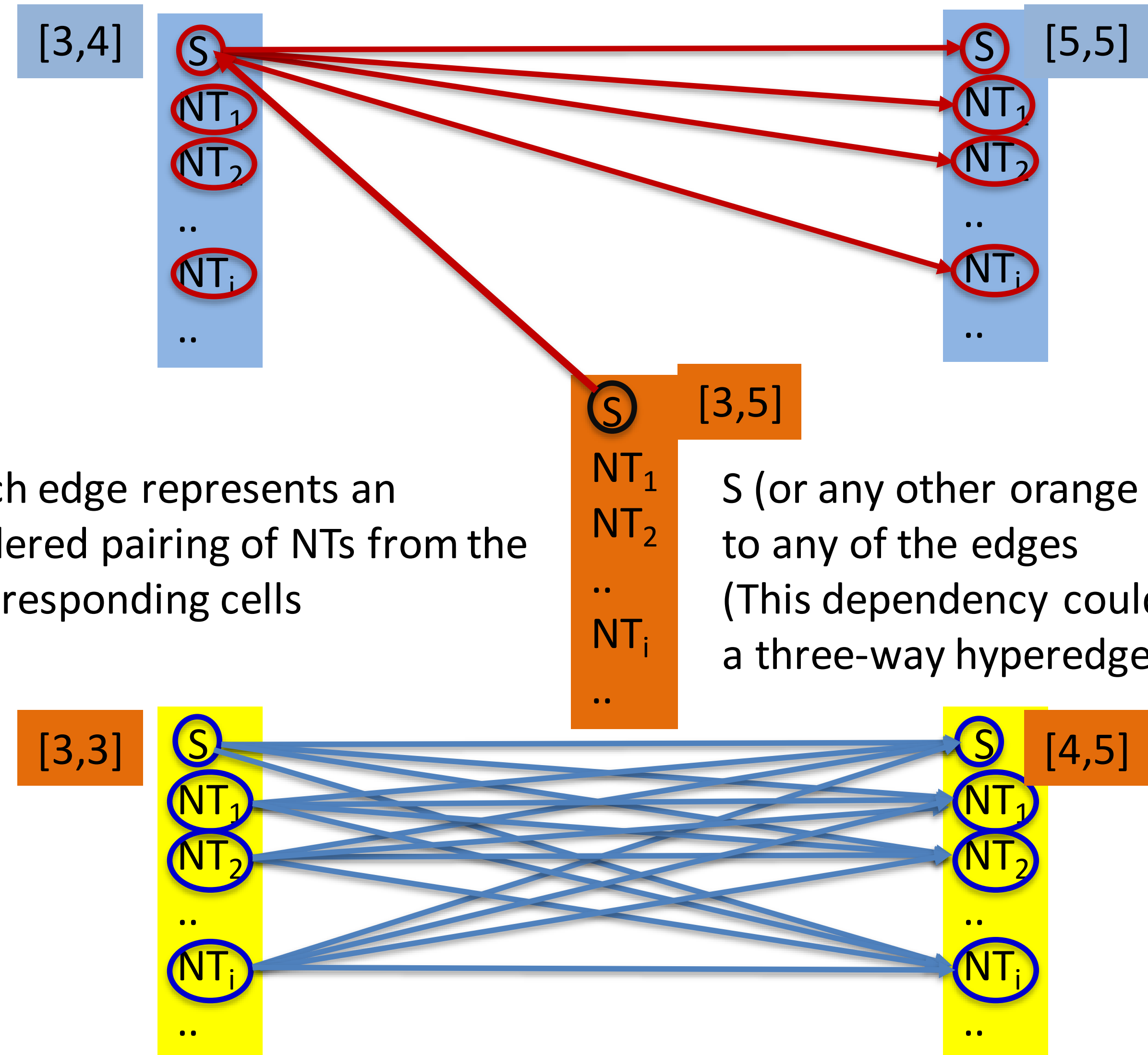
- Ways in which NT could occur at (i, j)



Each edge represents an ordered pairing of NTs from the corresponding cells

S (or any other orange NT) may expand out to any of the edges
(This dependency could be represented by a three-way hyperedge)

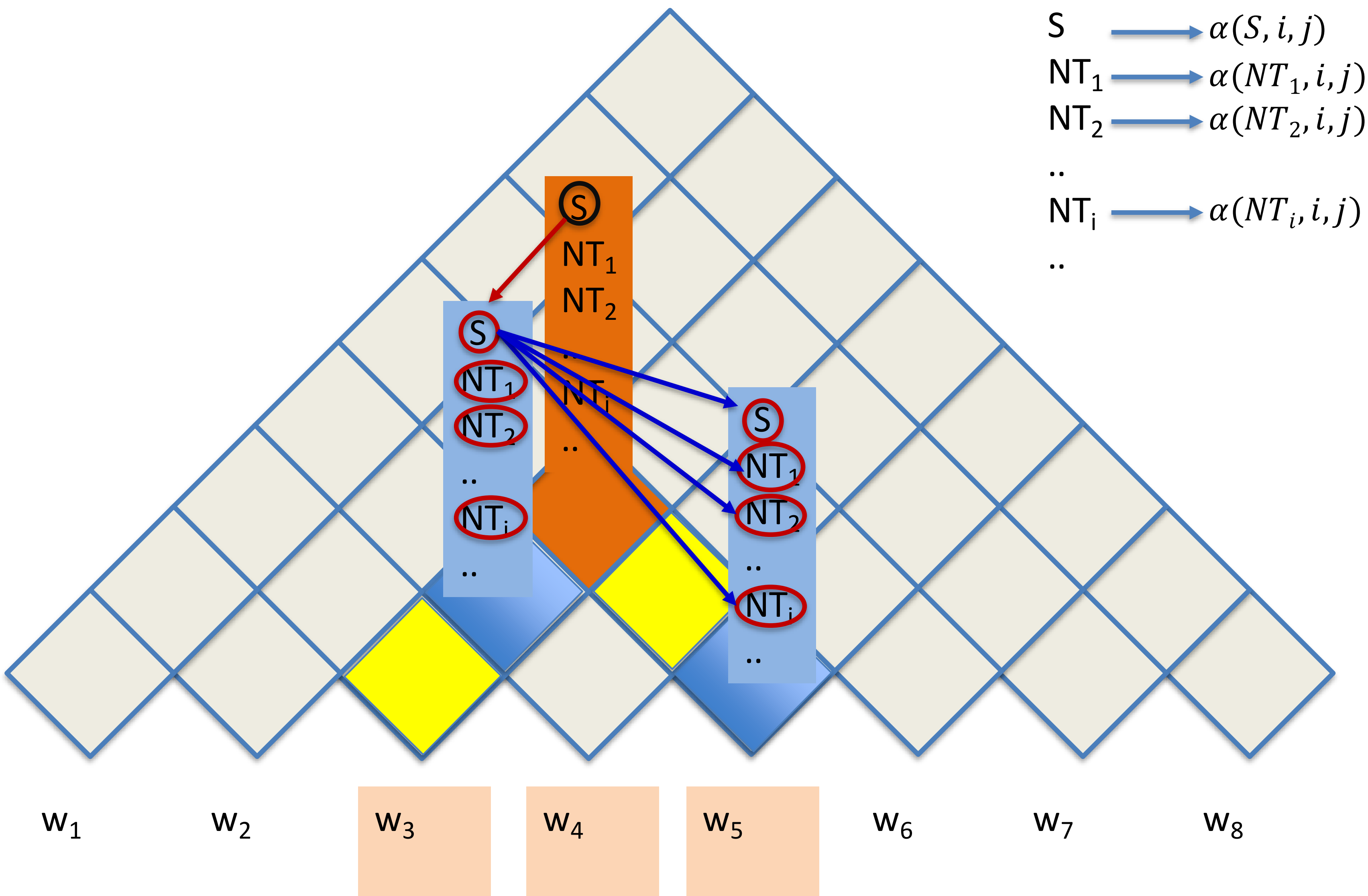
$$P(S \rightarrow w_3 \ldots w_5) = P(S \rightarrow S\ S) P(S \rightarrow w_3 \ldots w_4) P(S \rightarrow w_5) + \cdots$$



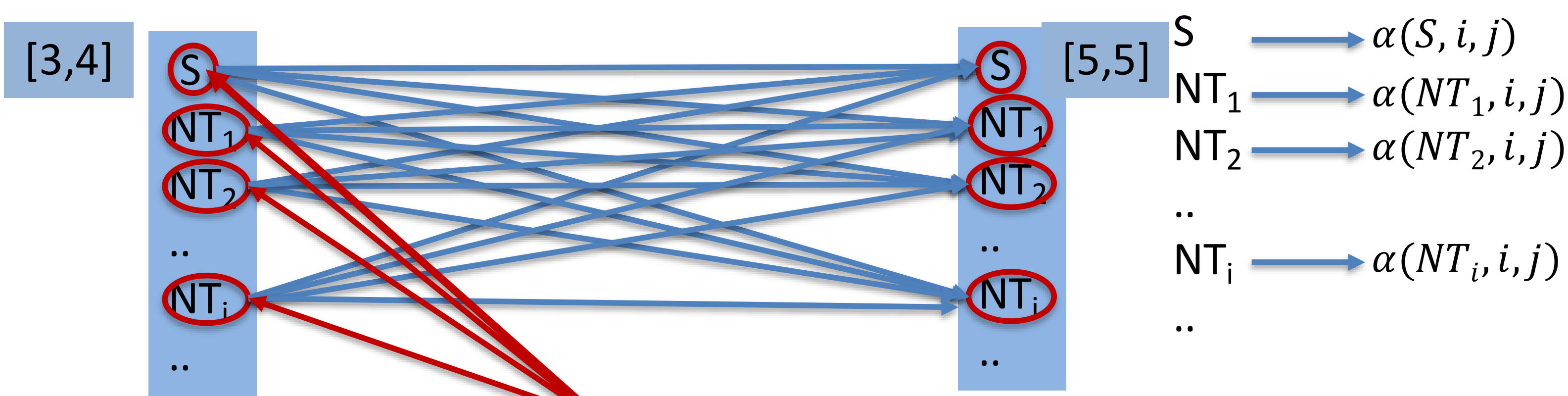
Each edge represents an ordered pairing of NTs from the corresponding cells

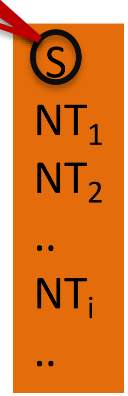
S (or any other orange NT) may expand out to any of the edges
(This dependency could be represented by a three-way hyperedge)

$$P(S \to w_3 \ldots w_5) = \sum_{NT} P(S \to S\ NT) P(S \to w_3 \ldots w_4) P(NT \to w_5) + \cdots$$



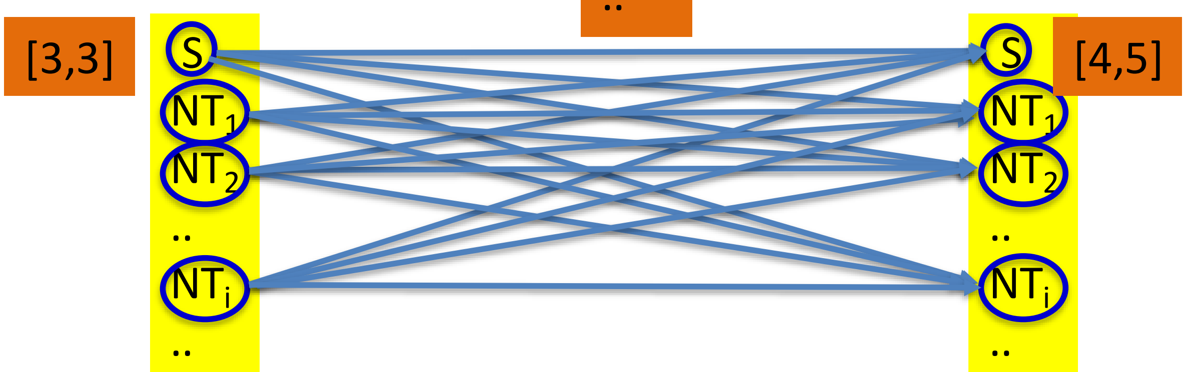
- Ways in which NT could occur at (i, j)



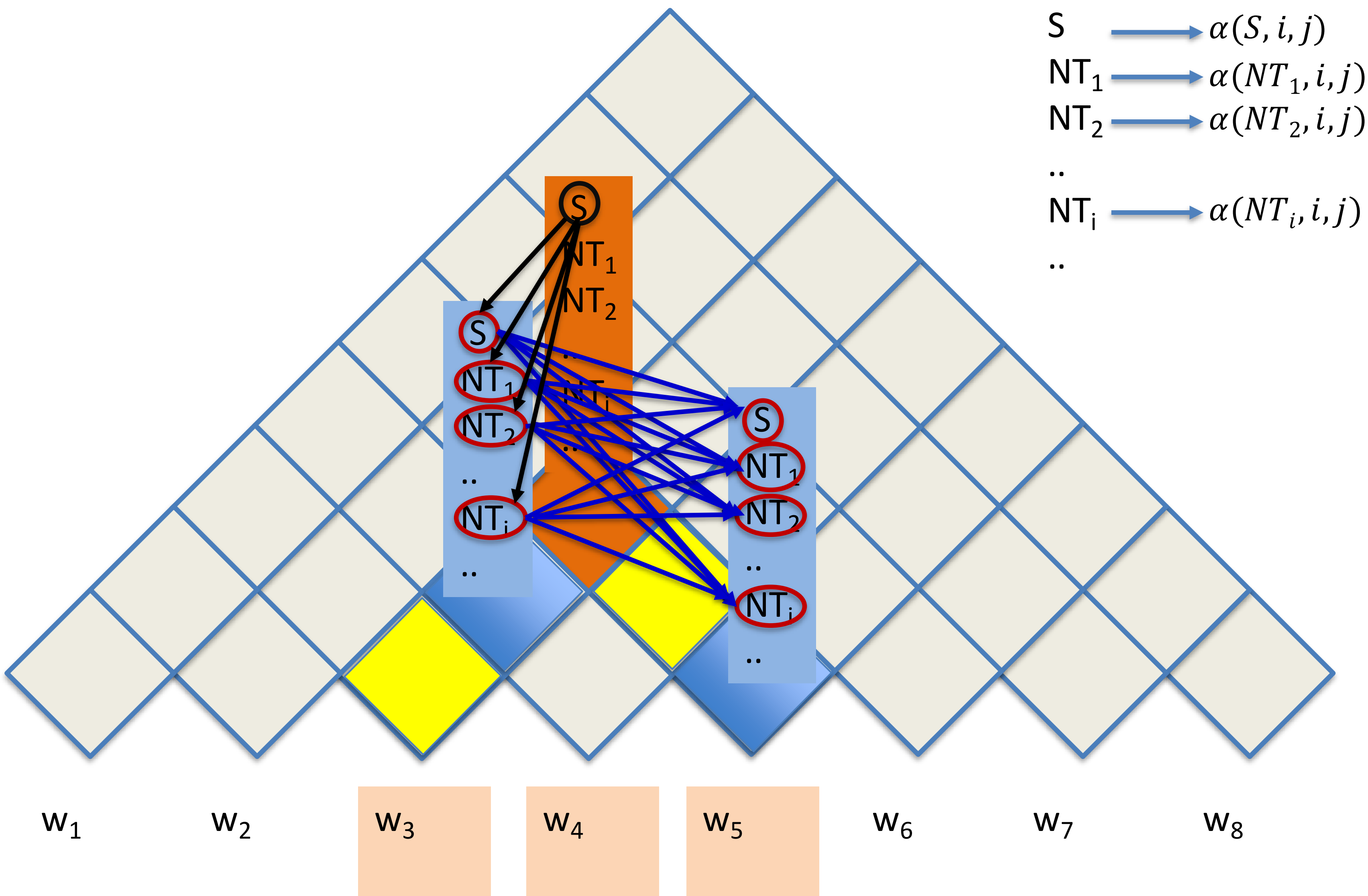


Each edge represents an ordered pairing of NTs from the corresponding cells

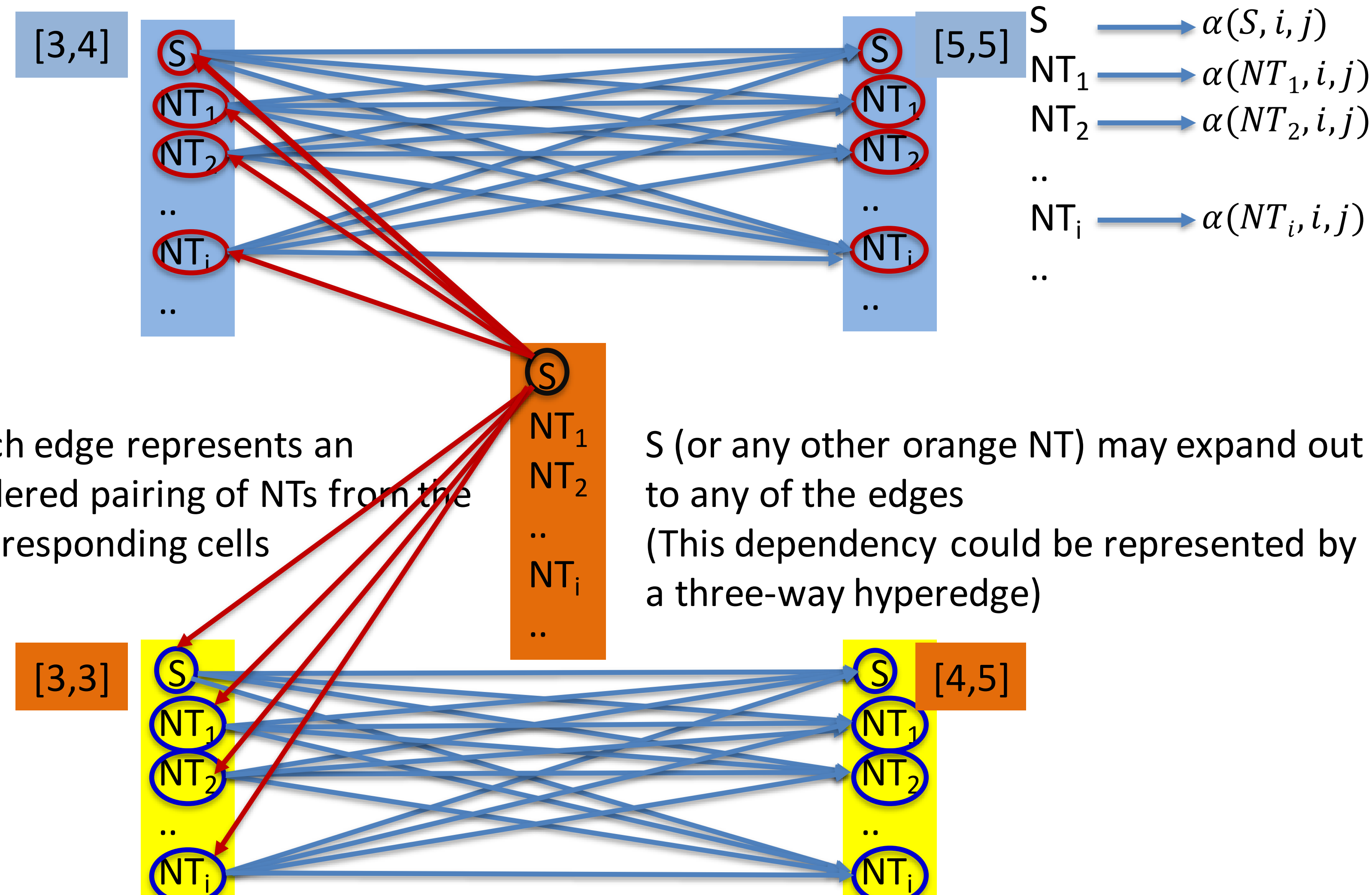
S (or any other orange NT) may expand out to any of the edges
(This dependency could be represented by a three-way hyperedge)



$$P(S \rightarrow w_3 \ldots w_5) = \sum_{NT_a} \sum_{NT_b} P(S \rightarrow NT_a NT_b) P(NT_a \rightarrow w_3 \ldots w_4) P(NT_b \rightarrow w_5) + \cdots$$



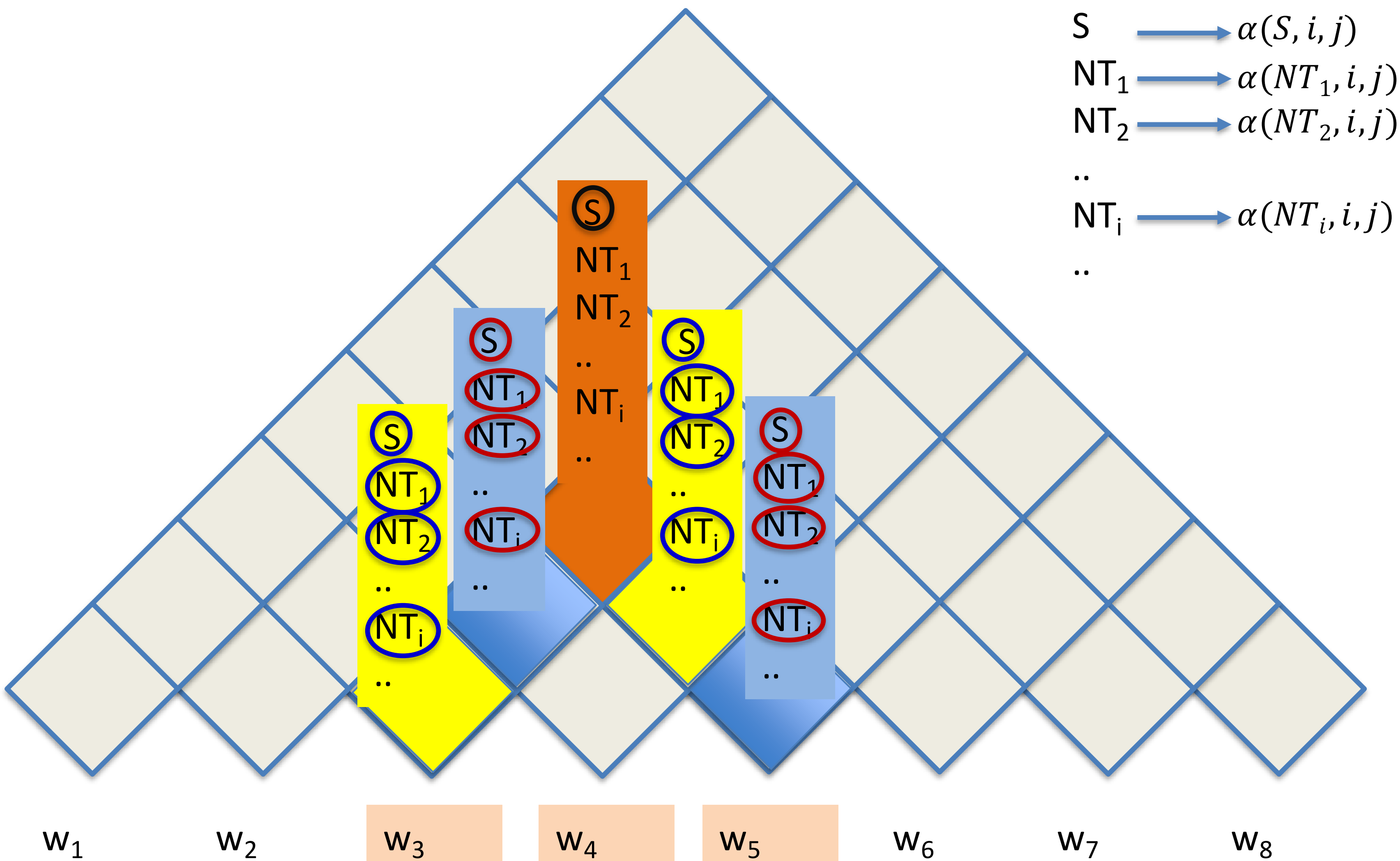
$$P(S \to w_3 \ldots w_5) = \sum_{NT_a,} \sum_{NT_b} P(S \to NT_a NT_b) P(NT_a \to w_3 \ldots w_4) P(NT_b \to w_5) + \cdots$$



Each edge represents an ordered pairing of NTs from the corresponding cells

S (or any other orange NT) may expand out to any of the edges
(This dependency could be represented by a three-way hyperedge)

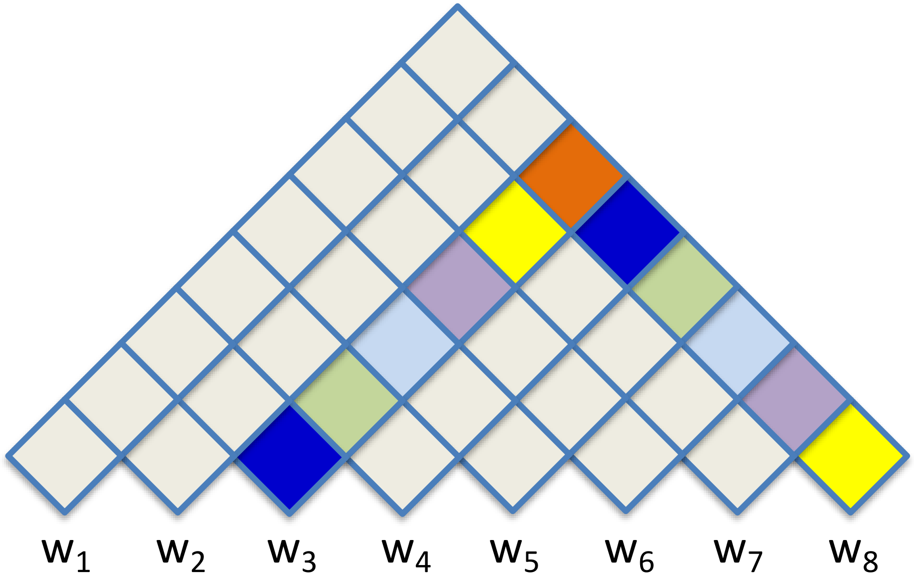
$$p(S \to w_3 \ldots w_5) = \sum_{NT_a,} \sum_{NT_b} P(S \to NT_a NT_b) P(NT_a \to w_3 \ldots w_4) P(NT_b \to w_5) +$$
$$\sum_{NT_a,} \sum_{NT_b} P(S \to NT_a NT_b) P(NT_a \to w_3) P(NT_b \to w_4 \ldots w_5)$$



$$p(S \rightarrow w_3 \dots w_5) = \sum_{NT_a,} \sum_{NT_b} P(S \rightarrow NT_a NT_b) P(NT_a \rightarrow w_3 \dots w_4) P(NT_b \rightarrow w_5) +$$

$$\sum_{NT_a,} \sum_{NT_b} P(S \rightarrow NT_a NT_b) P(NT_a \rightarrow w_3) P(NT_b \rightarrow w_4 \dots w_5)$$

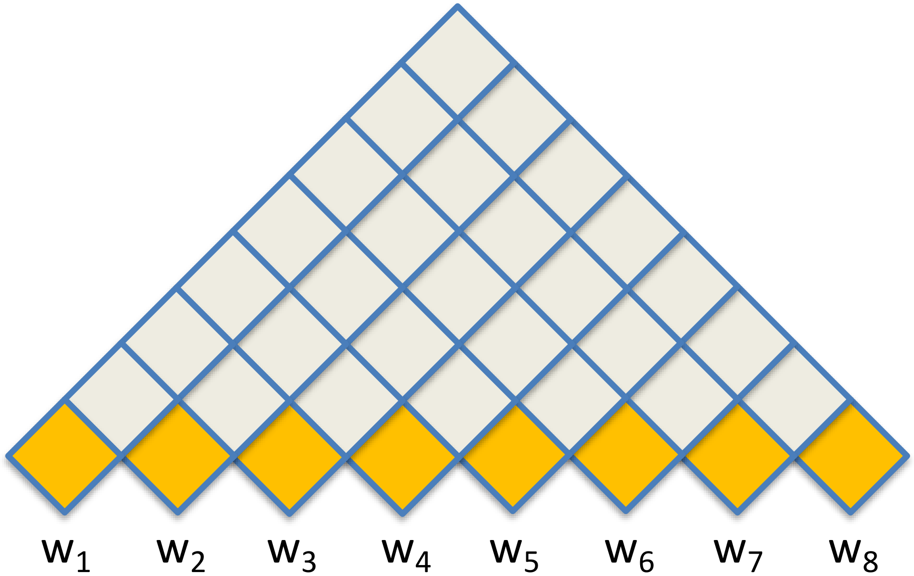
More generally



$$p(NT \rightarrow w_i \dots w_j) = \sum_{k} \sum_{NT_a,} \sum_{NT_b} P(NT \rightarrow NT_a NT_b) P(NT_a \rightarrow w_i \dots w_k) P(NT_b \rightarrow w_{k+1} \dots w_j)$$

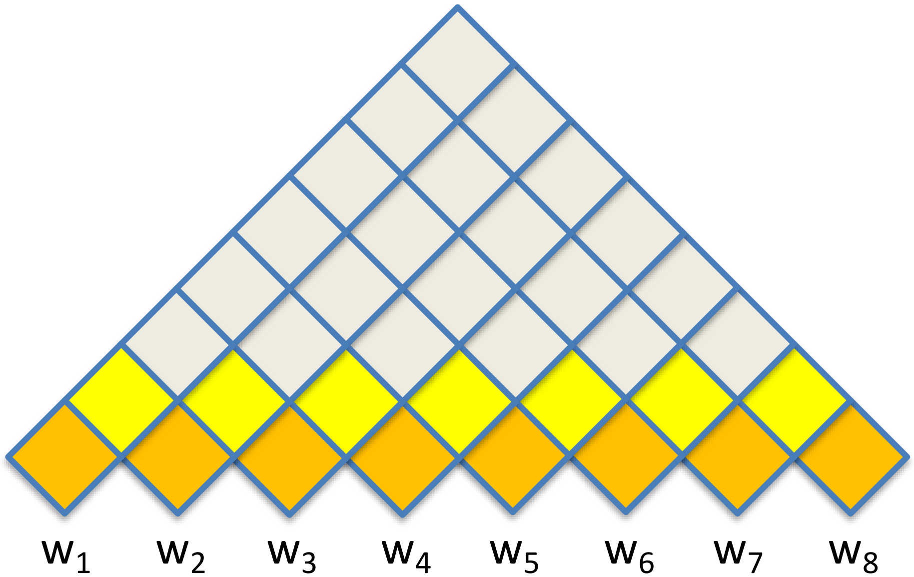
$$\alpha(NT, i, j) = \sum_{i \le k \le j} \sum_{NT_a,} \sum_{NT_b} P(NT \rightarrow NT_a NT_b) \alpha(NT_a, i, k) \alpha(NT_b, k+1, j)$$

The Inside Algorithm



For i = 1..N
    For all NT:
        $\alpha(NT, i, i) = P(NT \rightarrow w_i)$

The Inside Algorithm

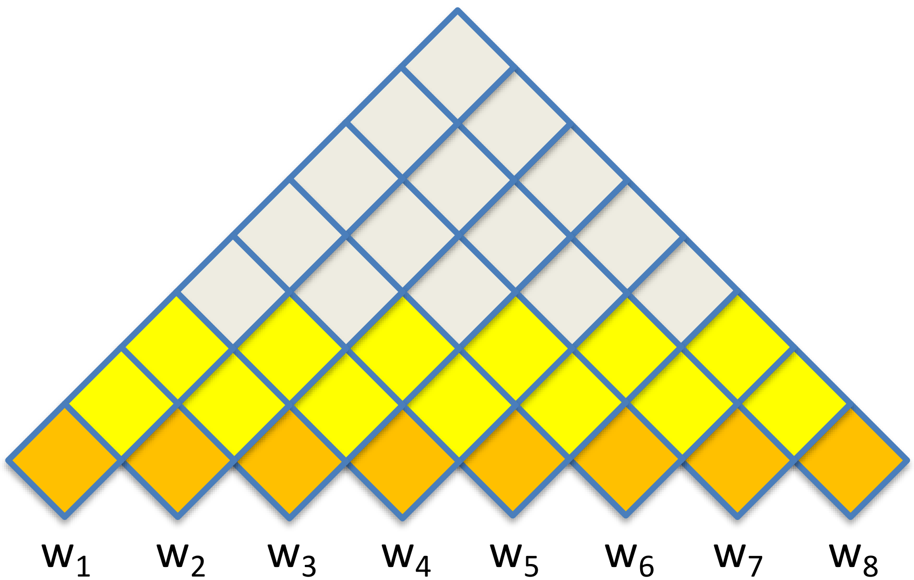


For $i = 1..N$
    For all NT:
        $\alpha(NT, i, i) = P(NT \rightarrow w_i)$

For $l = 1..N-1$
    For $i = 1..N-l$
        $j = i + l$
        For all NT:

$$\alpha(NT, i, j) = \sum_{i \le k \le j} \sum_{NT_a,} \sum_{NT_b} P(NT \rightarrow NT_a NT_b)\, \alpha(NT_a, i, k)\, \alpha(NT_b, k+1, j)$$

The Inside Algorithm

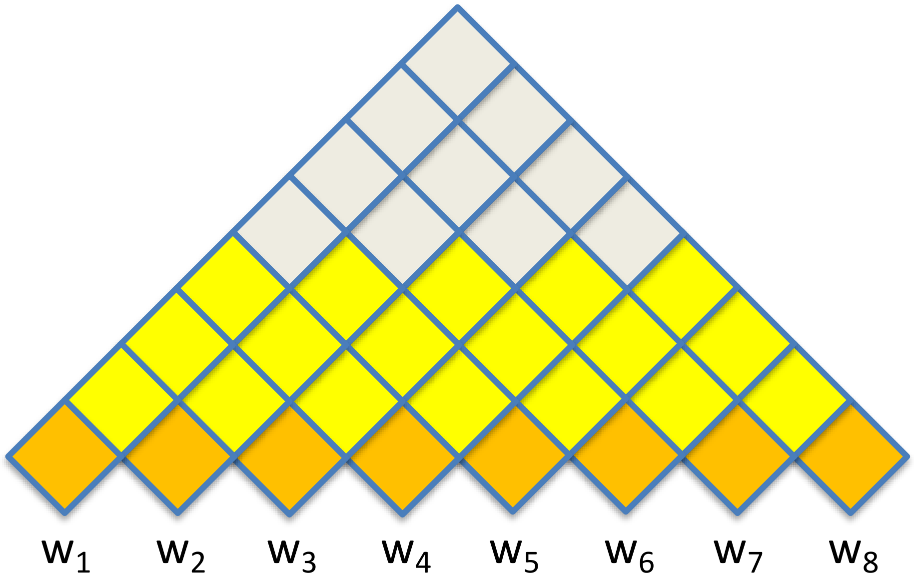


For $i = 1..N$
For all NT:
$\alpha(NT, i, i) = P(NT \to w_i)$

For $l = 1..N-1$
For $i = 1..N-l$
$j = i + l$
For all NT:

$$\alpha(NT, i, j) = \sum_{i \le k \le j} \sum_{NT_a,} \sum_{NT_b} P(NT \to NT_a NT_b)\alpha(NT_a, i, k)\alpha(NT_b, k+1, j)$$

The Inside Algorithm

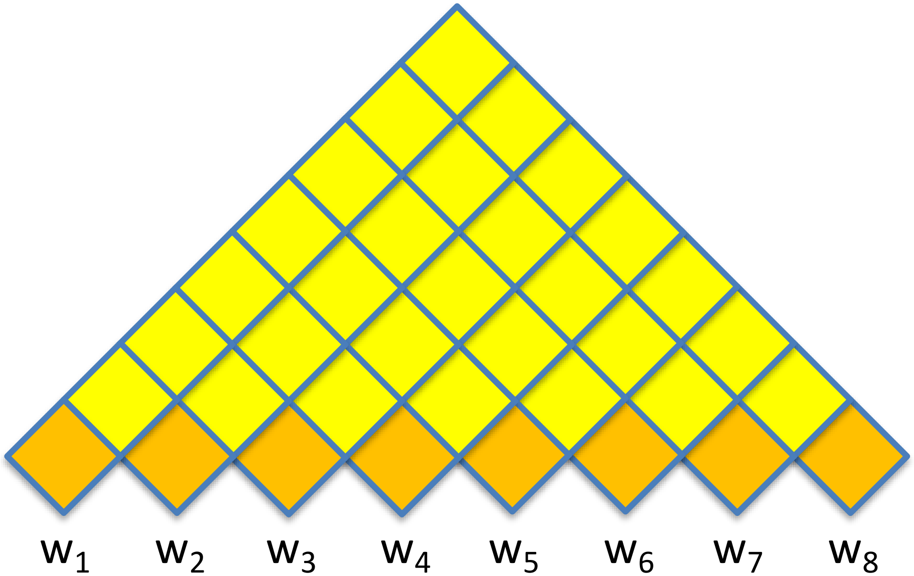


For $i = 1..N$
  For all NT:
    $\alpha(NT, i, i) = P(NT \rightarrow w_i)$

For $l = 1..N-1$
  For $i = 1..N-l$
    $j = i + l$
    For all NT:

$$\alpha(NT, i, j) = \sum_{i \le k \le j} \sum_{NT_a,} \sum_{NT_b} P(NT \rightarrow NT_a NT_b)\alpha(NT_a, i, k)\alpha(NT_b, k+1, j)$$

The Inside Algorithm

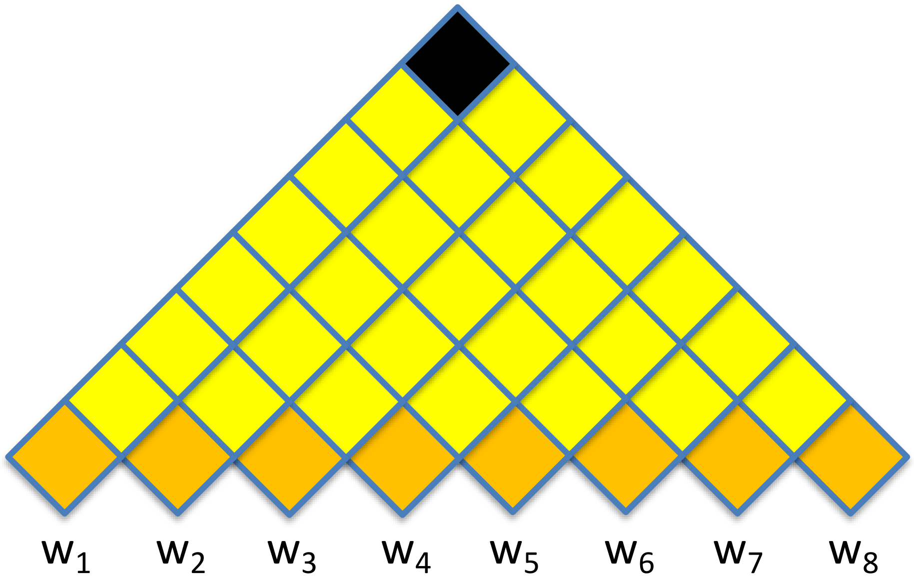


For $i = 1..N$
For all NT:
$\alpha(NT, i, i) = P(NT \rightarrow w_i)$

For $l = 1..N-1$
For $i = 1..N-l$
$j = i + l$
For all NT:

$$\alpha(NT, i, j) = \sum_{i \leq k \leq j} \sum_{NT_a,} \sum_{NT_b} P(NT \rightarrow NT_a NT_b)\alpha(NT_a, i, k)\alpha(NT_b, k+1, j)$$

The Inside Algorithm



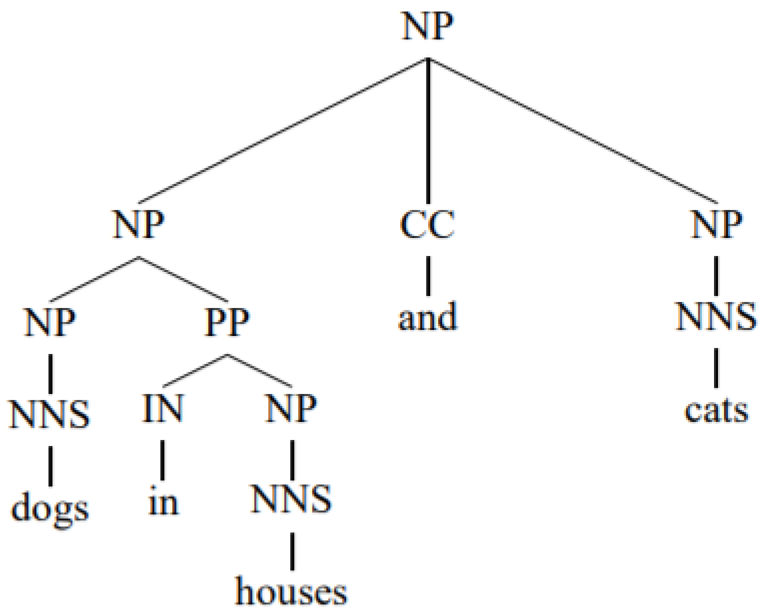
For $i = 1..N$
For all NT:
$\alpha(NT, i, i) = P(NT \rightarrow w_i)$

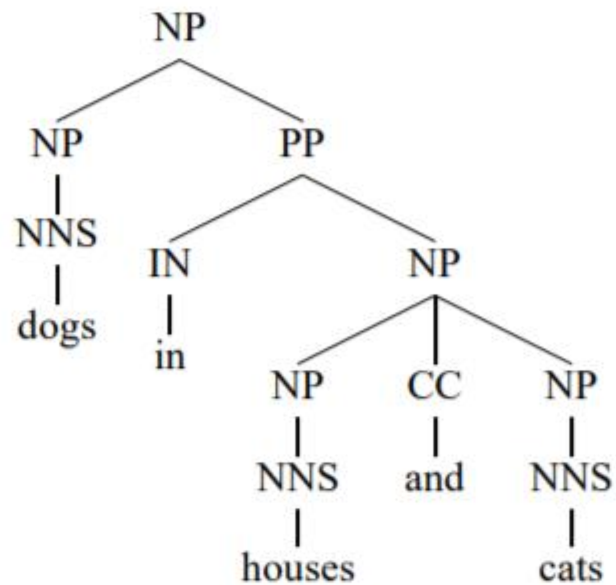
$$P(w_1 \ldots w_N) = \alpha(S, 1, N)$$

For $l = 1..N-1$
For $i = 1..N-l$
$j = i + l$
For all NT:

$$\alpha(NT, i, j) = \sum_{i \le k \le j} \sum_{NT_a,} \sum_{NT_b} P(NT \rightarrow NT_a NT_b) \alpha(NT_a, i, k) \alpha(NT_b, k+1, j)$$

Inferences we would like to make..

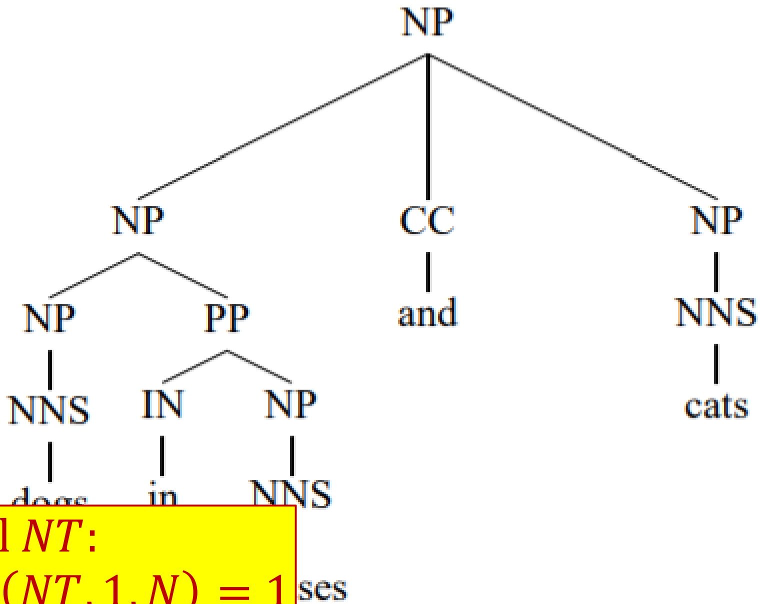


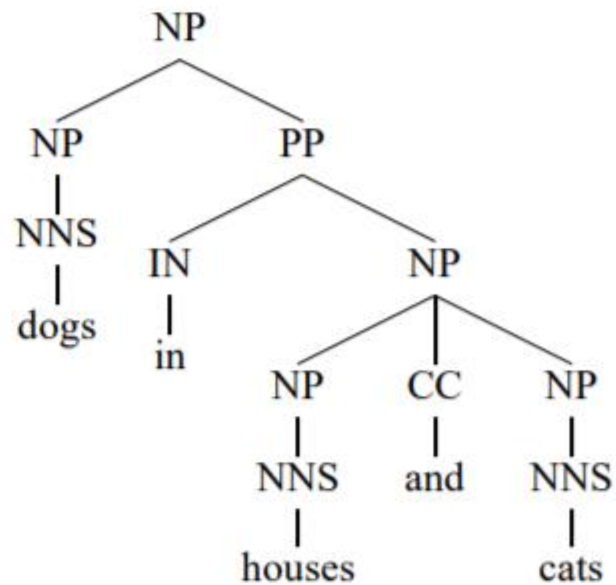


Done

- Which of the probability of "dogs in houses and cats"
  - P("dogs in houses and cats")
- What is the probability that "houses and cats" is a clause by itself?
  - P("houses and cats" = clause | "dogs in houses and cats")
- What is the probability that its an *NP?*
  - P("houses and cats" = NP | "dogs in houses and cats")
- Is there a *PP* in the sentence?
  - P(PP | "dogs in houses and cats")

Inferences we would like to make..



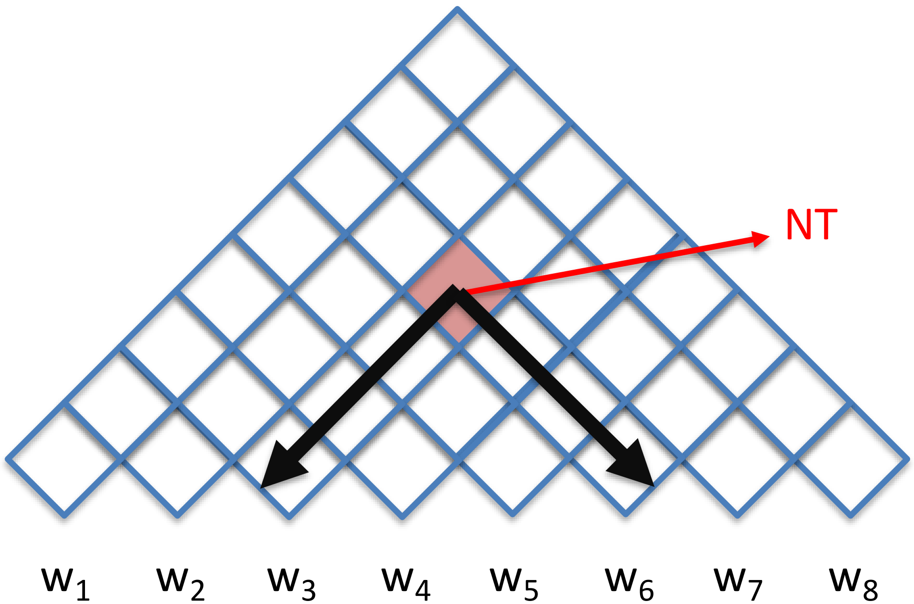


For all NT:
$$\beta(NT, 1, N) = 1$$

Done

- Which of the probability of "dogs in houses and cats"
  - P("dogs in houses and cats")
- What is the probability that "houses and cats" is a clause by itself?
  - P("houses and cats" = clause | "dogs in houses and cats")
- What is the probability that its an *NP?*
  - P("houses and cats" = NP | "dogs in houses and cats")
- Is there a *PP* in the sentence?
  - P(PP | "dogs in houses and cats")

The Conditional Probability



- What we desire to compute:
  - $P(NT \in c(i,j)|W)$: Probability that the cell spanning words $i \ldots j$ contains the specific nonterminal NT, given the observed word sequence W
    - The probability that $w_i \ldots w_j$ were produced by NT given the entire word sequence W

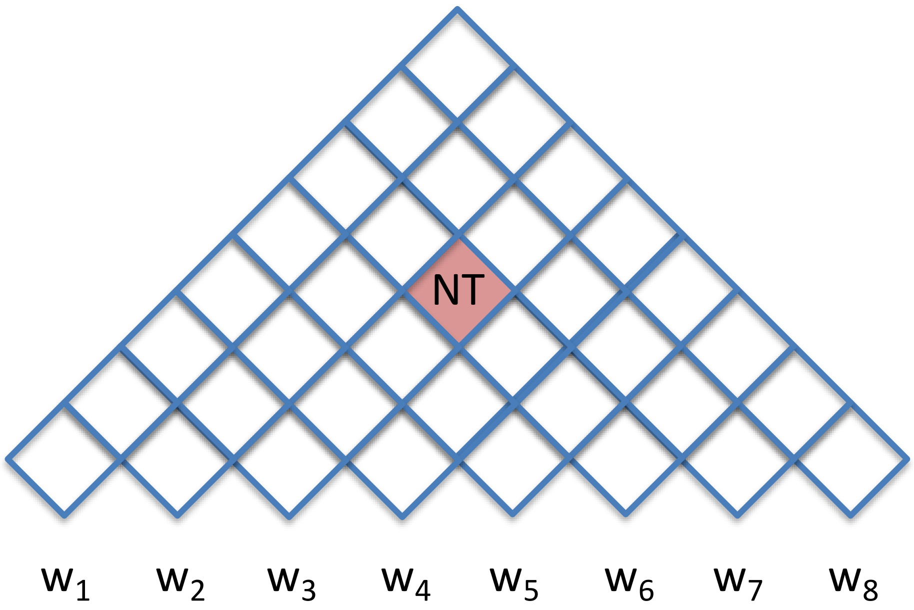
Conditional vs Joint

- $P(NT \in c(i,j)|W) = \frac{P(NT \in c(i,j),W)}{P(W)}$

- We know how to compute the denominator
- So we must compute:

$$P(NT \in c(i,j), W)$$

The Joint Probability

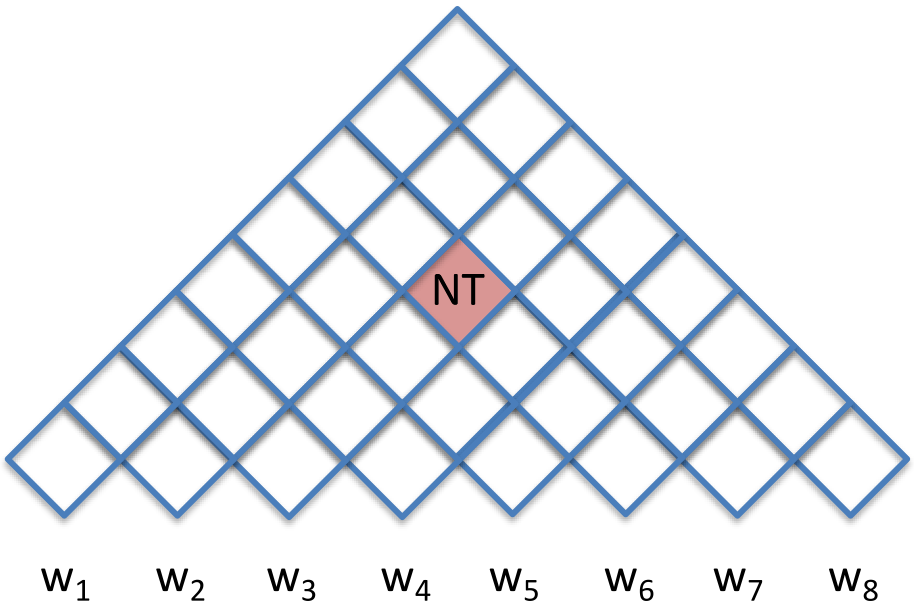


- $P(NT \in c(i,j), w_1 \dots w_N)$ is the total probability of the entire word sequence *AND* that cell $c(i,j)$ contains NT

$$P(NT \in c(i,j), w_1 \dots w_N) = P\left(NT \rightarrow w_i \dots w_j, w_1 \dots w_N\right)$$

$$= P\left(NT \rightarrow w_i \dots w_j, w_1 \dots w_{i-1}, w_{j+1} \dots w_N\right)$$

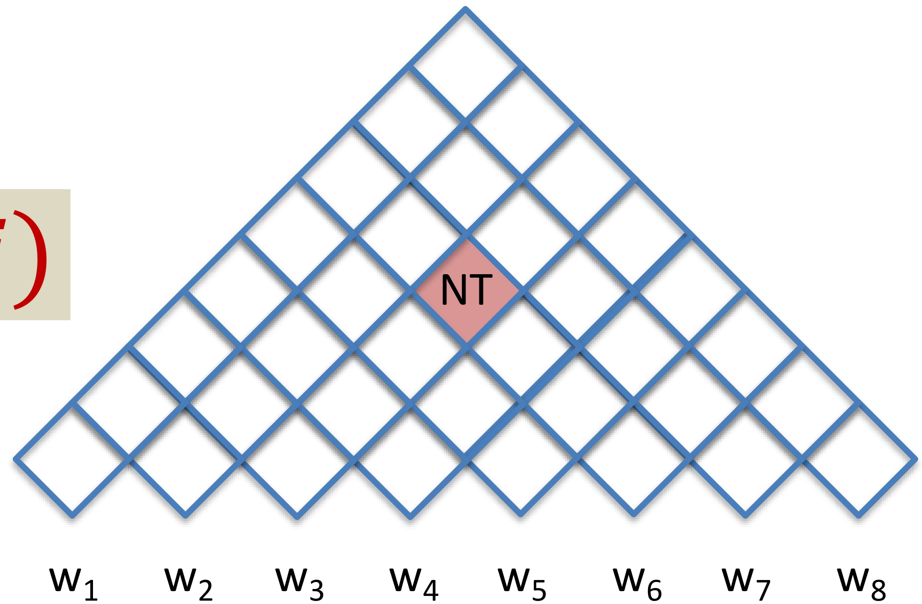
The Joint Probability



- $P\left(NT \to w_i \dots w_j, w_1 \dots w_{i-1}, w_{j+1} \dots w_N\right) =$
$$P\left(NT \to w_i \dots w_j\right) P\left(w_1 \dots w_{i-1}, w_{j+1} \dots w_N, c(i,j) = NT\right)$$
- Note: The second term on the RHS explicitly takes advantage of the fact that for a CFG the NT isolates the rest of the sentence from the words produced by the NT

The Joint Probability

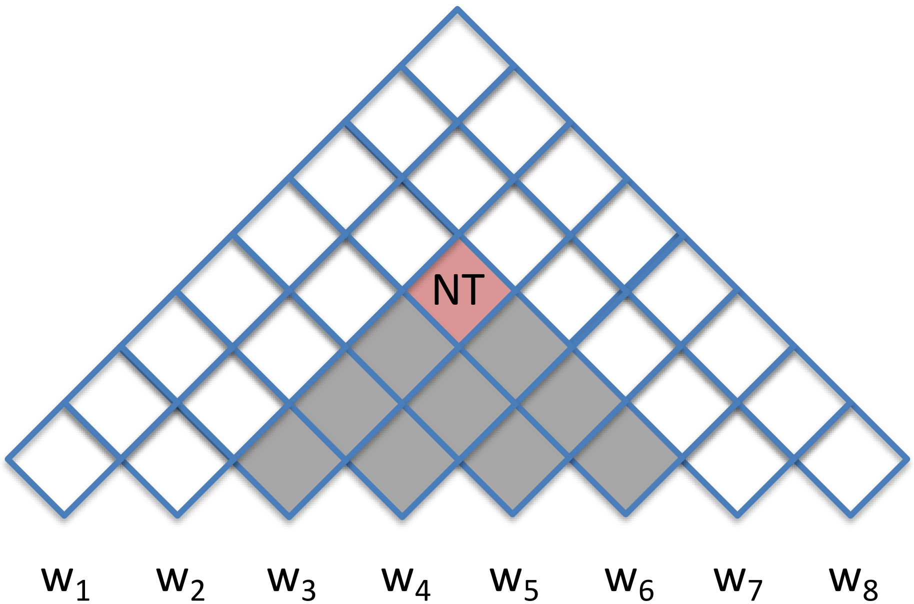
$\alpha\,(NT, i, j)$



We need to compute this

- $P\big(NT \rightarrow w_i \ldots w_j, w_1 \ldots w_{i-1}, w_{j+1} \ldots w_N\big) =$
  $$P\big(NT \rightarrow w_i \ldots w_j\big) P\big(w_1 \ldots w_{i-1}, w_{j+1} \ldots w_N, c(i,j) = NT\big)$$
- Note: The second term on the RHS explicitly takes advantage of the fact that for a CFG the NT isolates the rest of the sentence from the words produced by the NT

The *Outside* Probability

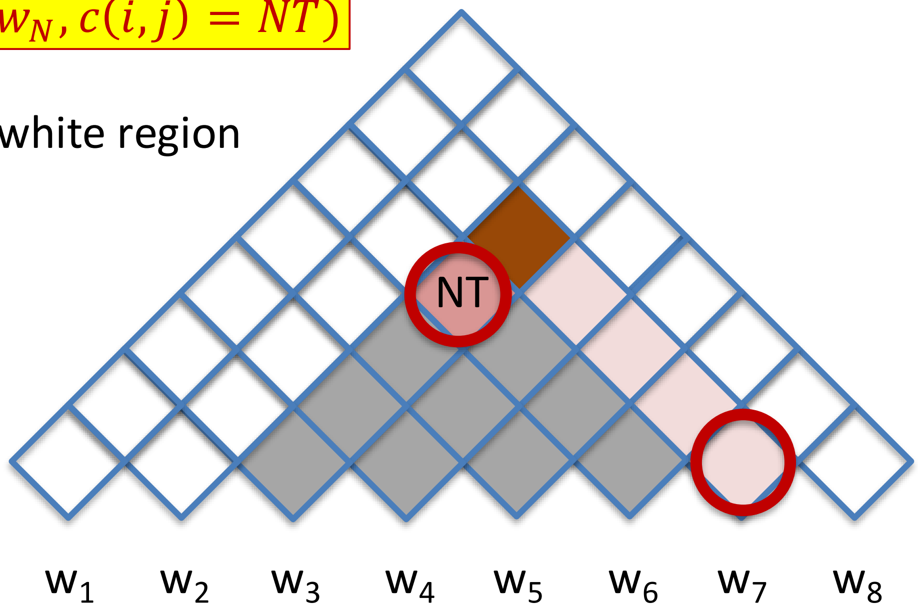


- $P\left(w_1 \ldots w_{i-1}, w_{j+1} \ldots w_N, c(i,j) = NT\right)$
  – The probability of the words under the white region of the grid, conditioned on the pink node taking value *NT*

The *Outside* Probability

$P(w_1 \ldots w_{i-1}, w_{j+1} \ldots w_N, c(i,j) = NT)$

Need probability of white region given the pink NT

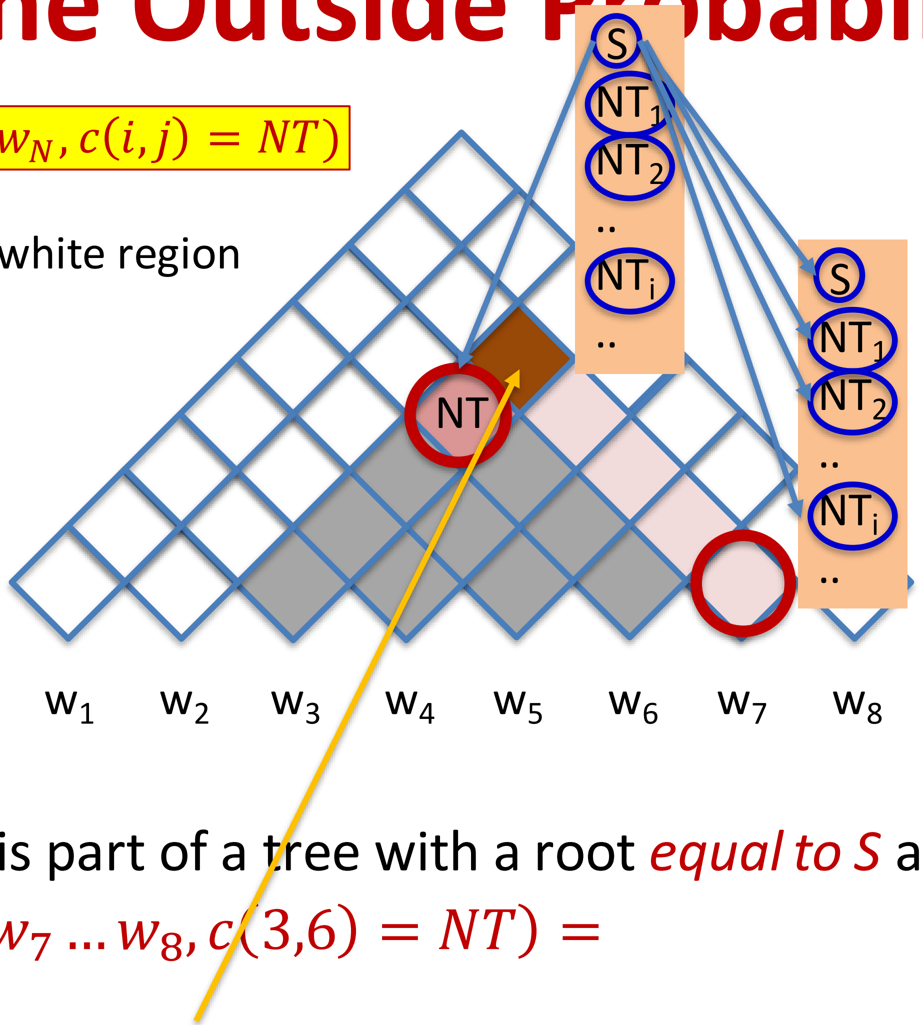


- Option 1: NT is part of a tree with a root at the Brown cell (w_2 .. w_7)
  - w_8 is not part of the tree
  - Must generate w_1..w_2, w_8 *outside* the tree

The Outside Probability

$P(w_1 \ldots w_{i-1}, w_{j+1} \ldots w_N, c(i,j) = NT)$

Need probability of white region given the pink NT



- Option 1: *NT* is part of a tree with a root *equal to S* at the Brown cell

$$P(w_1 \ldots w_2, w_7 \ldots w_8, c(3,6) = NT) =$$

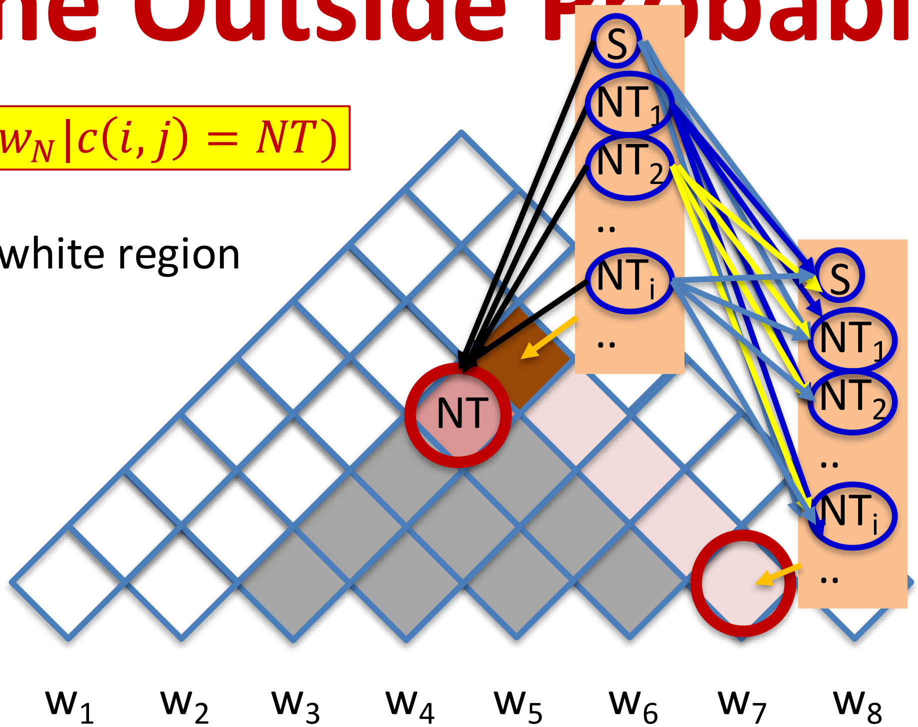
$$P(w_1 \ldots w_2, w_8, c(3,7) = S) \sum_{NT_b} P(S \to NT\ NT_b) P(NT_b \to w_7) + \cdots$$

Outside probability of (3,7)

The Outside Probability

$P(w_1 \dots w_{i-1}, w_{j+1} \dots w_N | c(i,j) = NT)$

Need probability of white region given the pink NT



- Option 1: *NT* is part of a tree *with a root at the Brown cell*

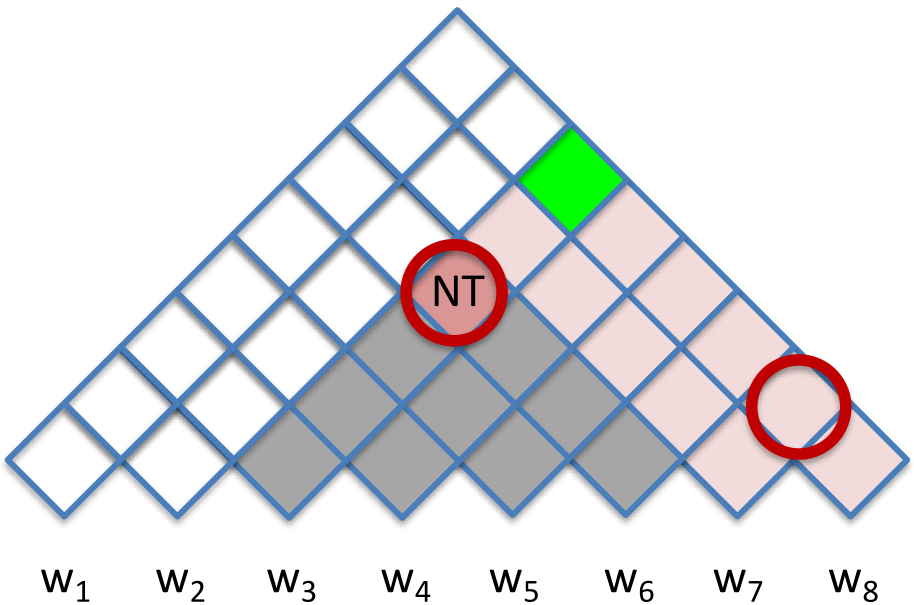
$$P(w_1 \dots w_2, w_7 \dots w_8, c(3,6) = NT) =$$

$$\sum_{NT_a} P(w_1 \dots w_2, w_8, c(3,7) = NT_a) \sum_{NT_b} P(NT_a \to NT\ NT_b) P(NT_b \to w_7) + \cdots$$

The Outside Probability

$$P(w_1 \dots w_2, w_7 \dots w_8, c(3,6) = NT) =$$

$$\sum_{NT_a} P(w_1 \dots w_2, w_8, c(3,7) = NT_a) \sum_{NT_b} P(NT_a \to NT\ NT_b) P(NT_b \to w_7) + \cdots$$



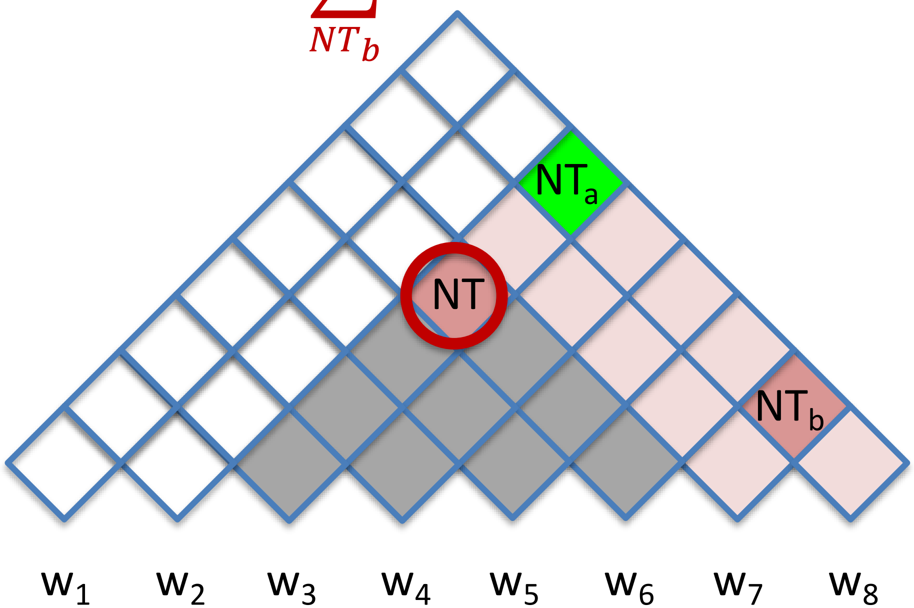
- Option 2: *NT* is part of a tree *with a root at the Green cell*
  - *Note the counterpart cell of NT under the green root*

The Outside Probability

$$P(w_1 \dots w_2, w_7 \dots w_8, c(3,6) = NT) =$$

$$\sum_{NT_a} P(w_1 \dots w_2, w_8, c(3,7) = NT_a) \sum_{NT_b} P(NT_a \to NT\ NT_b) P(NT_b \to w_7) +$$

$$\sum_{NT_a} P(w_1 \dots w_2, c(3,8) = NT_a) \sum_{NT_b} P(NT_a \to NT\ NT_b) P(NT_b \to w_7 \dots w_8) + \cdots$$



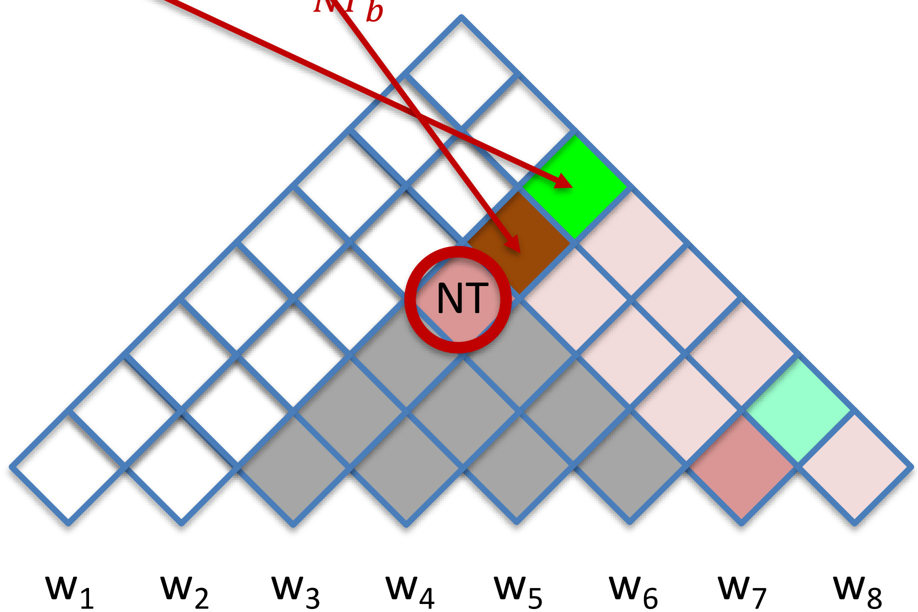
- Option 2: *NT* is part of a tree *with a root at the Green cell*
  - *Note the counterpart cell of NT under the green root*

The Outside Probability

$$P(w_1 \ldots w_2, w_7 \ldots w_8, c(3,6) = NT) =$$

$$\sum_{NT_a} P(w_1 \ldots w_2, w_8, c(3,7) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b) P(NT_b \rightarrow w_7) +$$

$$\sum_{NT_a} P(w_1 \ldots w_2, c(3,8) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b) P(NT_b \rightarrow w_7 \ldots w_8) + \cdots$$



- Option 2: *NT* is part of a tree *with a root at the Green cell*
  - *Note the counterpart cell of NT under the green root*

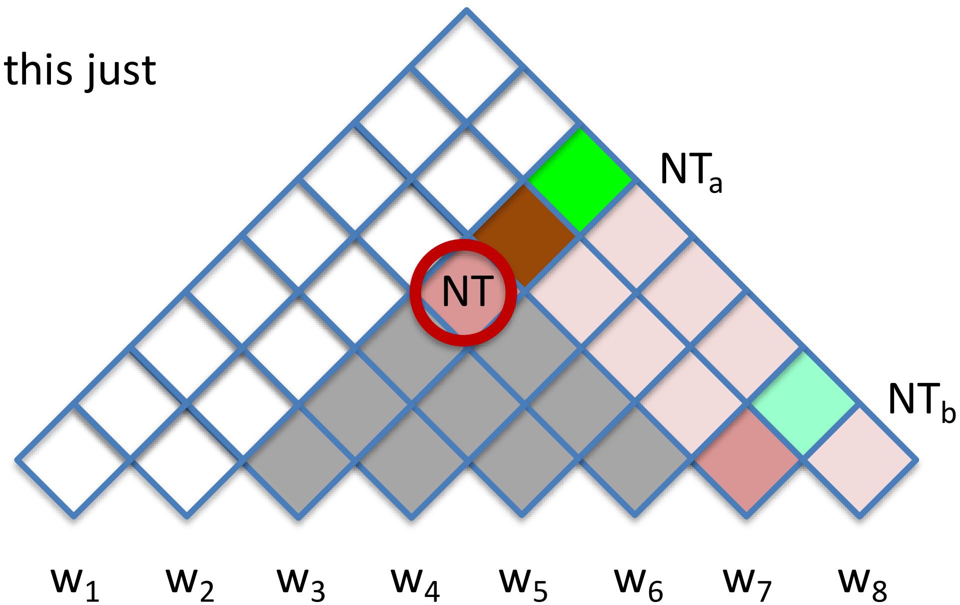
The Outside Probability

$$P(w_1 \dots w_2, w_7 \dots w_8, c(3,6) = NT) =$$

$$\sum_{k=7}^{8} \sum_{NT_a} P(w_1 \dots w_2, w_{k+1} \dots w_8, c(3,k) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b) P(NT_b \rightarrow w_7 \dots w_k)$$

$$+ \cdots$$

Note, if $k + 1 > 8$, this just becomes $w_1 \dots w_2$



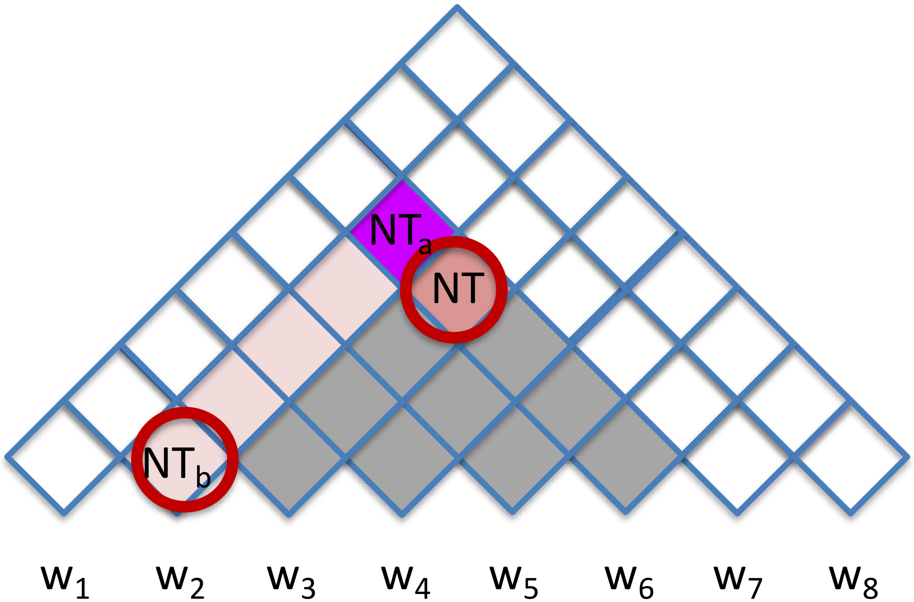
- Option 2: *NT* is part of a tree *with a root at the Green cell*
  - *Note the counterpart cell of NT under the green root*

The Outside Probability

$$P(w_1 \dots w_2, w_7 \dots w_8, c(3,6) = NT) =$$

$$\sum_{k=7}^{8} \sum_{NT_a} P(w_1 \dots w_2, w_{k+1} \dots w_8, c(3,k) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b) P(NT_b \rightarrow w_7 \dots w_k)$$

$$+ \sum_{NT_a} P(w_1, w_7 \dots w_8, c(2,6) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ ) P(NT_b \rightarrow w_2)\ + \cdots$$



- Option 3: *NT* is part of a tree *with a root at the purple cell*
  - *Note the counterpart cell of NT under the purple root*
  - *Now the outside part is w_1, $w_7 \dots w_8$*

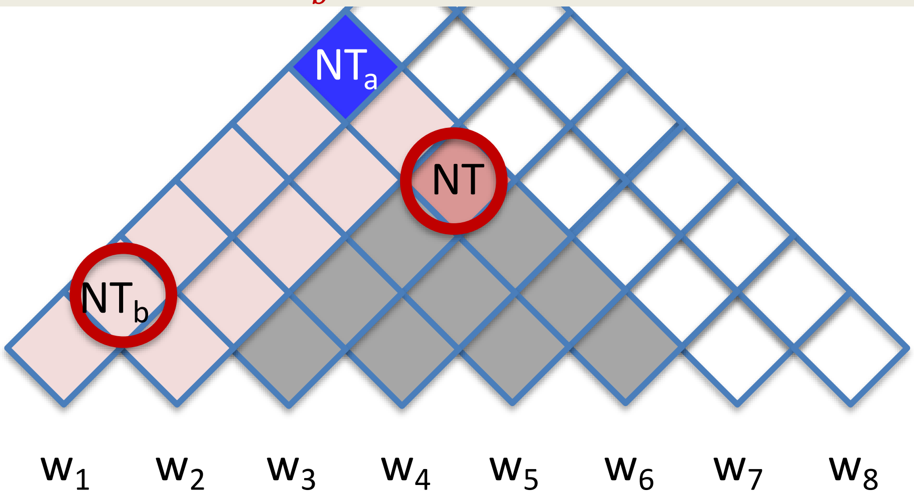
The Outside Probability

$$P(w_1 \dots w_2, w_7 \dots w_8, c(3,6) = NT) =$$

$$\sum_{k=7}^{8} \sum_{NT_a} P(w_1 \dots w_2, w_{k+1} \dots w_8, c(3,k) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b) P(NT_b \rightarrow w_7 \dots w_k)$$

$$+ \sum_{NT_a} P(w_1, w_7 \dots w_8, c(2,6) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ ) P(NT_b \rightarrow w_2)$$

$$+ \sum_{NT_a} P(w_7 \dots w_8, c(1,6) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ ) P(NT_b \rightarrow w_1 \dots w_2)$$



- Option 4: *NT* is part of a tree *with a root at the blue cell*
  - *Note the counterpart cell of NT under the blue root*
  - *Now the outside part is $w_7 \dots w_8$*

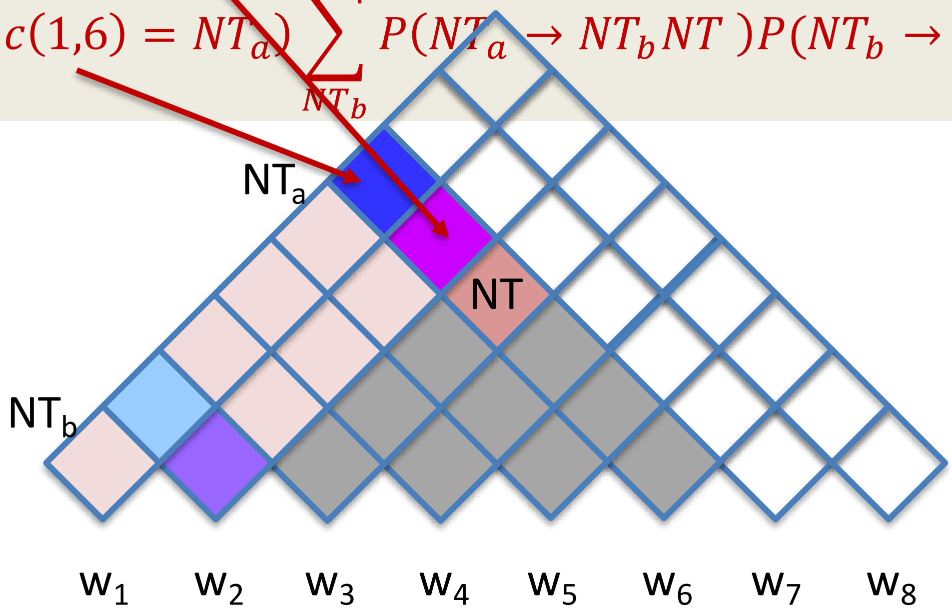
The Outside Probability

$$P(w_1 \ldots w_2, w_7 \ldots w_8, c(3,6) = NT) =$$

$$\sum_{k=7}^{8} \sum_{NT_a} P(w_1 \ldots w_2, w_{k+1} \ldots w_8, c(3,k) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b) P(NT_b \rightarrow w_7 \ldots w_k)$$

$$+ \sum_{NT_a} P(w_1, w_7 \ldots w_8, c(2,6) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ ) P(NT_b \rightarrow w_2)$$

$$+ \sum_{NT_a} P(w_7 \ldots w_8, c(1,6) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ ) P(NT_b \rightarrow w_1 \ldots w_2)$$



- Option 4: *NT* is part of a tree *with a root at the blue cell*
  - *Note the counterpart cell of NT under the blue root*
  - *Now the outside part is $w_7 \ldots w_8$*

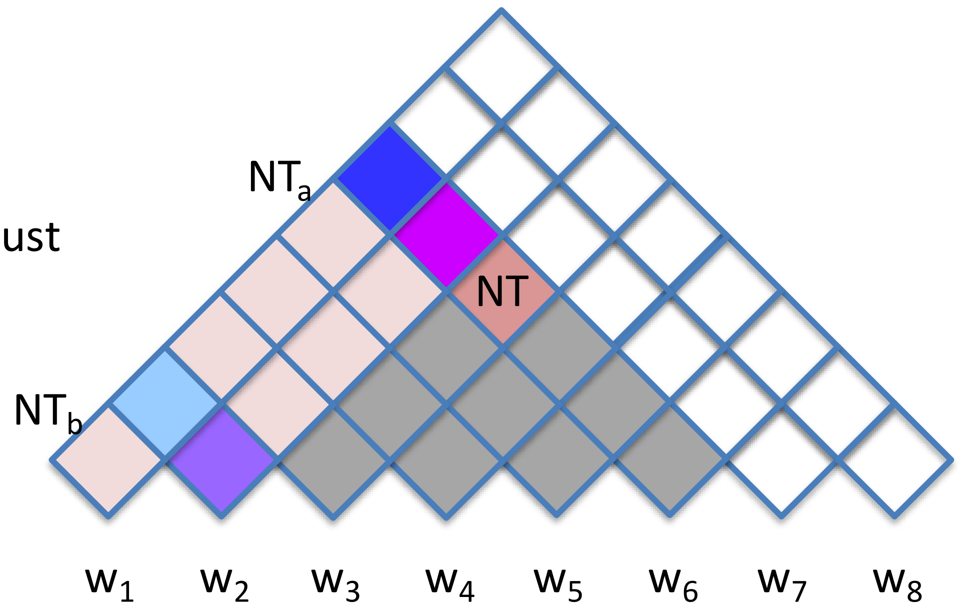
The Outside Probability

$$P(w_1 \dots w_2, w_7 \dots w_8, c(3{,}6) = NT) =$$

$$\sum_{k=7}^{8} \sum_{NT_a} P(w_1 \dots w_2, w_{k+1} \dots w_8, c(3,k) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b) P(NT_b \rightarrow w_7 \dots w_k)$$

$$+ \sum_{k=1}^{2} \sum_{NT_a} P(w_1 \dots w_{k-1}, w_7 \dots w_8, c(k,6) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ ) P(NT_b \rightarrow w_k \dots w_2)$$

Note, if $k < 1$, this just becomes $w_7 \dots w_8$



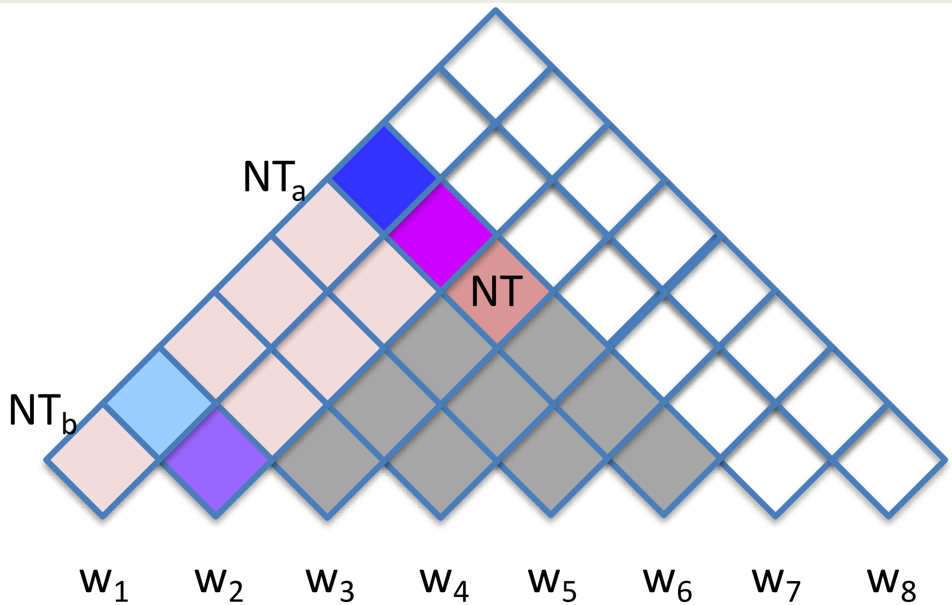
- Option 4: *NT* is part of a tree *with a root at the blue cell*
  - *Note the counterpart cell of NT under the blue root*
  - *Now the outside part is $w_7 \dots w_8$*

The Outside Probability

$$P\left(w_1 \dots w_{i-1}, w_{j+1} \dots w_N, c(i,j) = NT\right) =$$

$$\sum_{k=j+1}^{N} \sum_{NT_a} P(w_1 \dots w_{i-1}, w_{k+1} \dots w_N, c(i,k) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b) P(NT_b \rightarrow w_{j+1} \dots w_k)$$

$$+ \sum_{k=1}^{i} \sum_{NT_a} P\left(w_1 \dots w_{k-1}, w_{j+1} \dots w_N, c(k,j) = NT_a\right) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ ) P(NT_b \rightarrow w_k \dots w_{i-1})$$



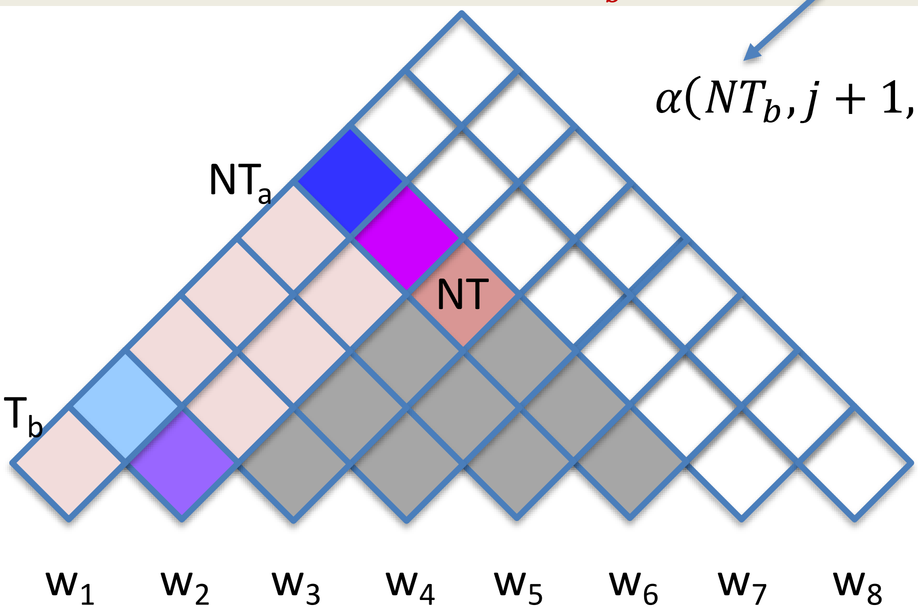
- Generic equation

The Outside Probability

$$P\left(w_1 \ldots w_{i-1}, w_{j+1} \ldots w_N, c(i,j) = NT\right) =$$

$$\sum_{k=j+1}^{N} \sum_{NT_a} P(w_1 \ldots w_{i-1}, w_{k+1} \ldots w_N, c(i,k) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b) \underbrace{P(NT_b \rightarrow w_{j+1} \ldots w_k)}_{\alpha(NT_b, j+1, k)}$$

$$+ \sum_{k=1}^{i} \sum_{NT_a} P\left(w_1 \ldots w_{k-1}, w_{j+1} \ldots w_N, c(k,j) = NT_a\right) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ ) \underbrace{P(NT_b \rightarrow w_k \ldots w_{i-1})}_{\alpha(NT_b, k, i-1)}$$



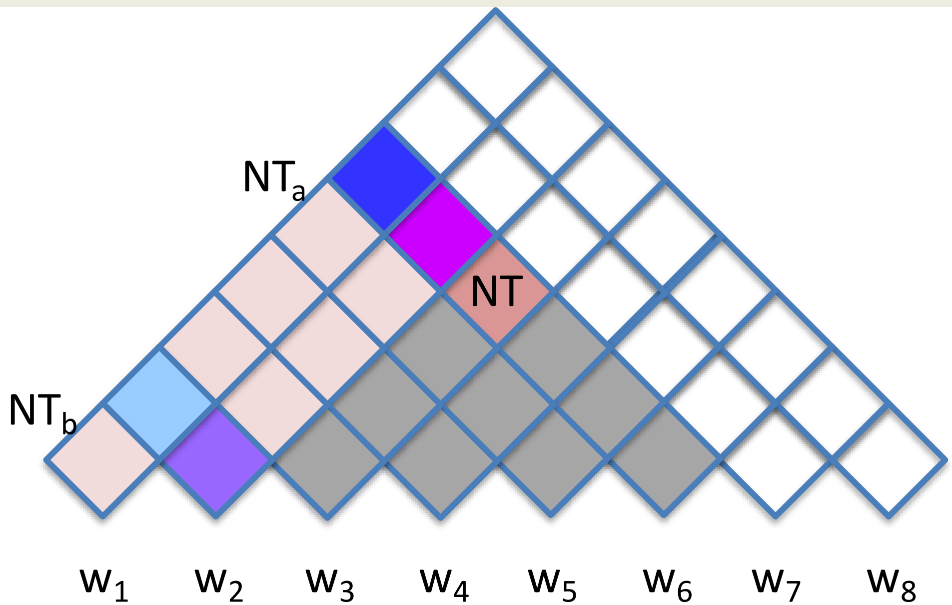
- Generic equation

The Outside Probability

$$P\left(w_1 \ldots w_{i-1}, w_{j+1} \ldots w_N, c(i,j) = NT\right) =$$

$$\sum_{k=j+1}^{N} \sum_{NT_a} P(w_1 \ldots w_{i-1}, w_{k+1} \ldots w_N, c(i,k) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b) \alpha(NT_b, j+1, k)$$

$$+ \sum_{k=1}^{i} \sum_{NT_a} P\left(w_1 \ldots w_{k-1}, w_{j+1} \ldots w_N, c(k,j) = NT_a\right) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ ) \alpha(NT_b, k, i-1)$$



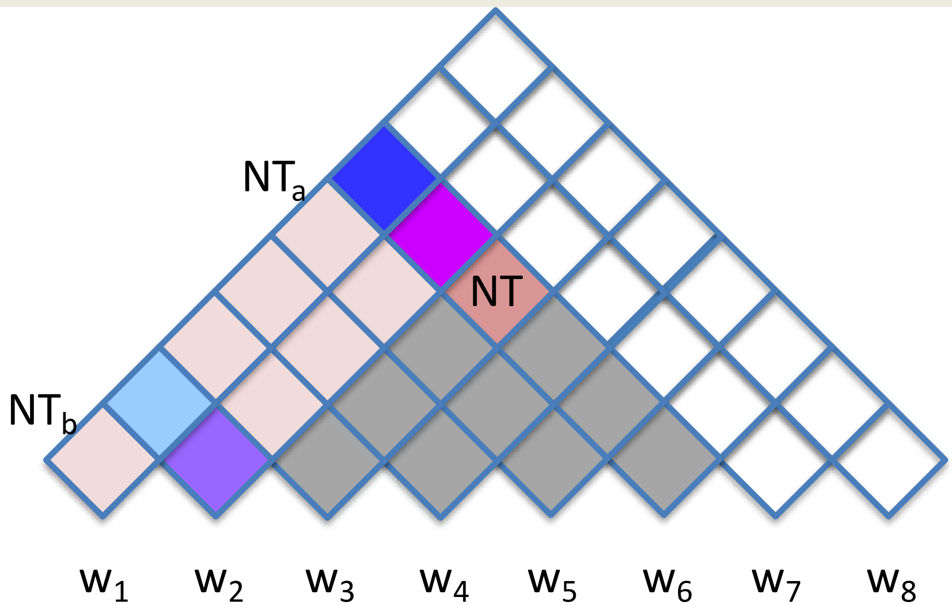
- Generic equation

The Outside Probability

$$P\big(w_1 \ldots w_{i-1}, w_{j+1} \ldots w_N, c(i,j) = NT\big) =$$

$$\sum_{k=j+1}^{N} \sum_{NT_a} P(w_1 \ldots w_{i-1}, w_{k+1} \ldots w_N, c(i,k) = NT_a) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b)\alpha(NT_b, j+1, k)$$

$$+ \sum_{k=1}^{i} \sum_{NT_a} P\big(w_1 \ldots w_{k-1}, w_{j+1} \ldots w_N, c(k,j) = NT_a\big) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ )\alpha(NT_b, k, i-1)$$



- Let

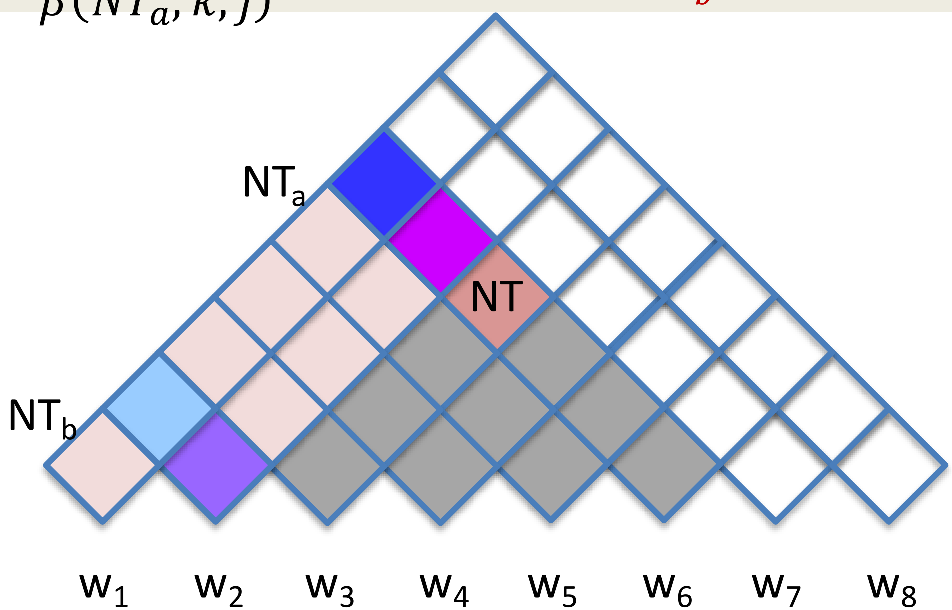
$$P\big(w_1 \ldots w_{i-1}, w_{j+1} \ldots w_N, c(i,j) = NT\big) = \beta(NT, i, j)$$

The Outside Probability

$$\underbrace{P\big(w_1 \dots w_{i-1}, w_{j+1} \dots w_N, c(i,j) = NT\big)}_{\beta(NT,i,j)} =$$

$$\sum_{k=j+1}^{N} \sum_{NT_a} \underbrace{P(w_1 \dots w_{i-1}, w_{k+1} \dots w_N, c(i,k) = NT_a)}_{\beta(NT_a,i,k)} \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b)\alpha(NT_b, j+1, k)$$

$$+ \sum_{k=1}^{i} \sum_{NT_a} \underbrace{P\big(w_1 \dots w_{k-1}, w_{j+1} \dots w_N, c(k,j) = NT_a\big)}_{\beta(NT_a,k,j)} \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ )\alpha(NT_b, k, i-1)$$

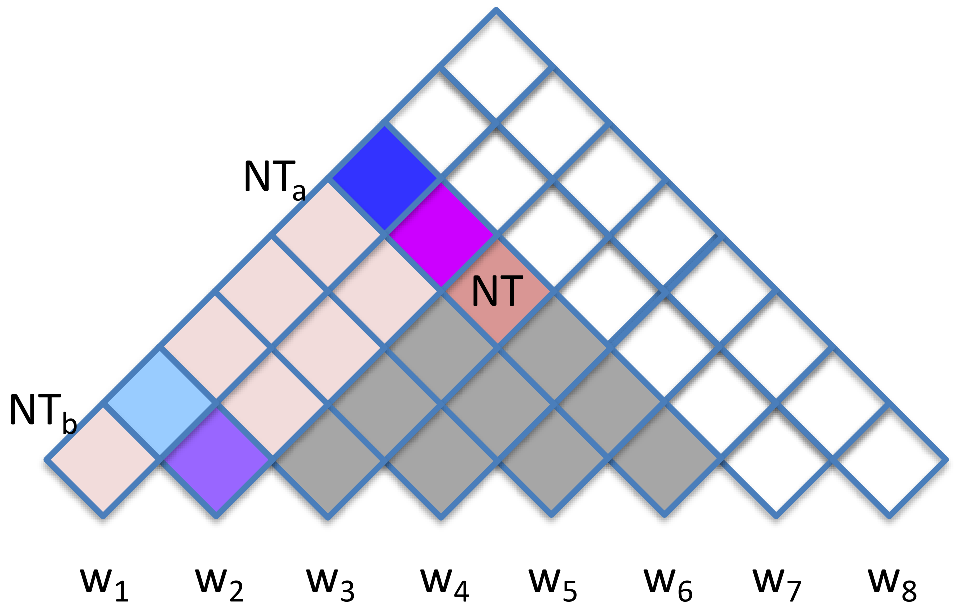


- Let

$$P\big(w_1 \dots w_{i-1}, w_{j+1} \dots w_N, c(i,j) = NT\big) = \beta(NT, i, j)$$

The Outside Probability

$$\beta(NT,i,j) = \sum_{k=j+1}^{N} \sum_{NT_a} \beta(NT_a,i,k) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b)\alpha(NT_b,j+1,k)$$
$$+ \sum_{k=1}^{i} \sum_{NT_a} \beta(NT_a,k,j) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ )\alpha(NT_b,k,i-1)$$

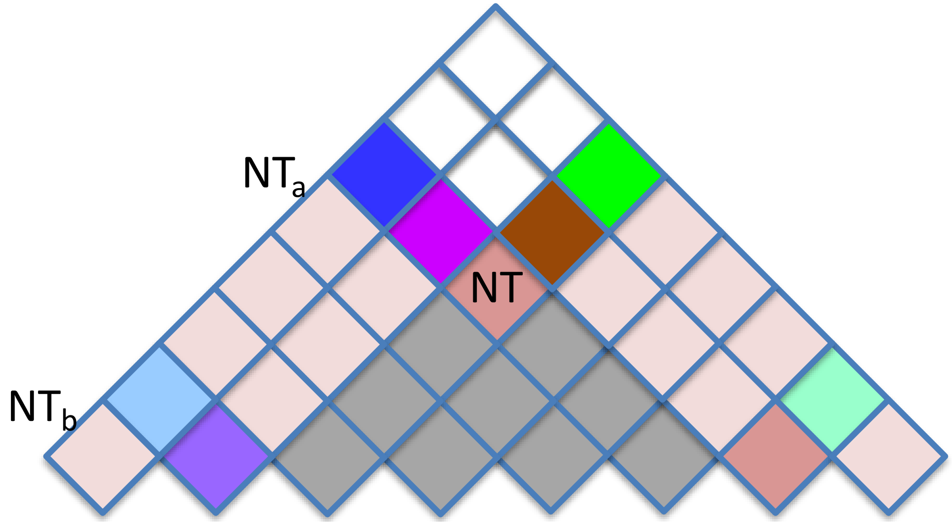


- Let

$$P\left(w_1 \dots w_{i-1}, w_{j+1} \dots w_N, c(i,j) = NT\right) = \beta(NT,i,j)$$

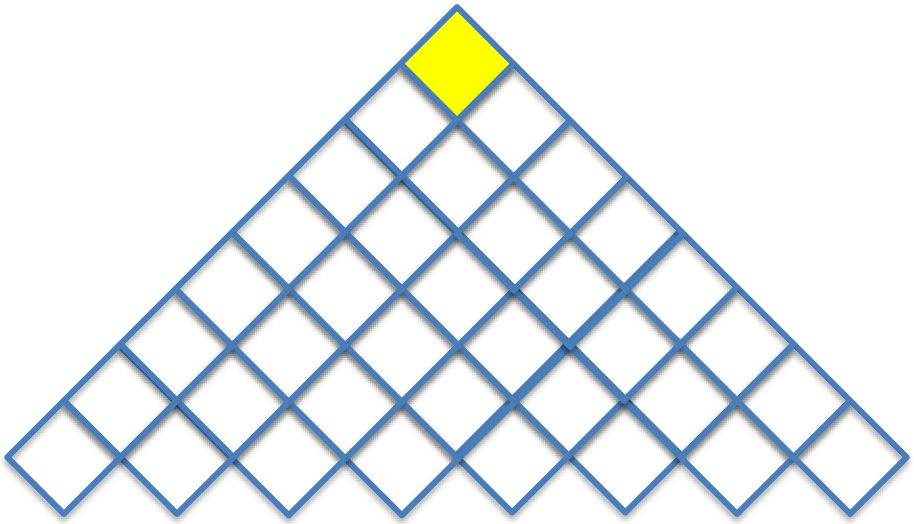
The Outside Probability

$$\beta(NT,i,j) = \sum_{k=j+1}^{N} \sum_{NT_a} \beta(NT_a,i,k) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b)\alpha(NT_b,j+1,k)$$
$$+ \sum_{k=1}^{i} \sum_{NT_a} \beta(NT_a,k,j) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ )\alpha(NT_b,k,i-1)$$



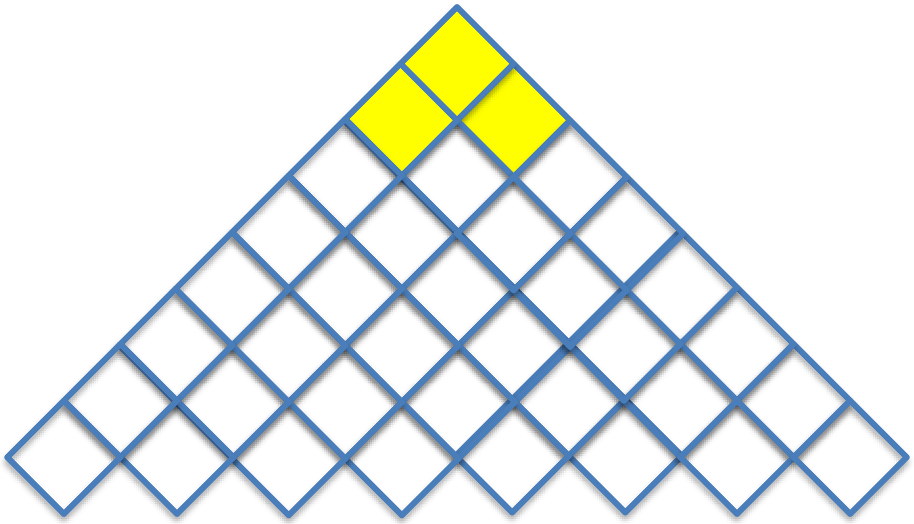
- **Note:** The computation of any outside probability β depends only on other betas *above* it and alphas *below* it
  - Beta computation requires preliminary computation of inside probabilities (alphas)
  - Given alpha, betas can now be computed recursively

The Outside Recursion



For all NT:
$\beta(NT, 1, N) = 1$

The Outside Recursion



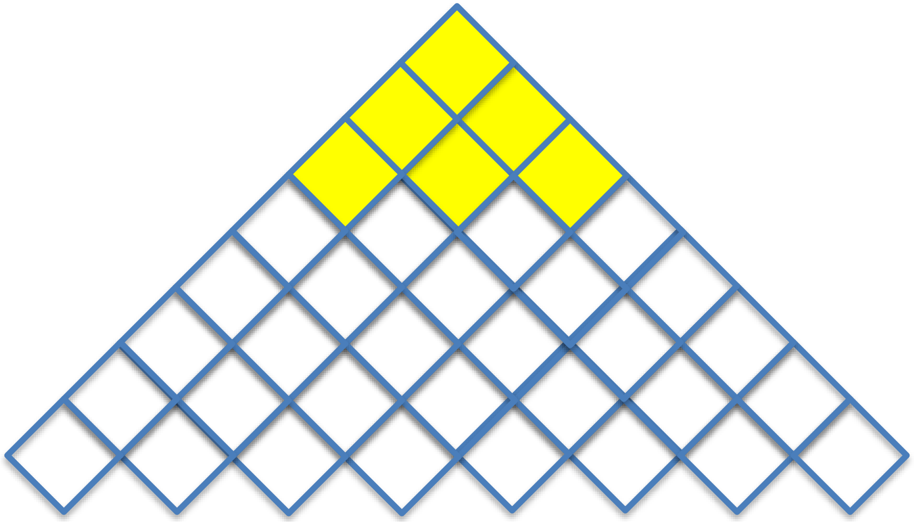
For all NT:
$\beta(NT, 1, N) = 1$

For $l = N - 2\ downto\ 1$
For $i = 1..N - l$
$j = i + l$
For all NT:

$$\beta(NT, i, j) = \sum_{k=j+1}^{N} \sum_{NT_a} \beta(NT_a, i, k) \sum_{NT_b} P(NT_a \to NT\ NT_b)\alpha(NT_b, j+1, k)$$

$$+ \sum_{k=1}^{i} \sum_{NT_a} \beta(NT_a, k, j) \sum_{NT_b} P(NT_a \to NT_b NT\ )\alpha(NT_b, k, i-1)$$

The Outside Recursion

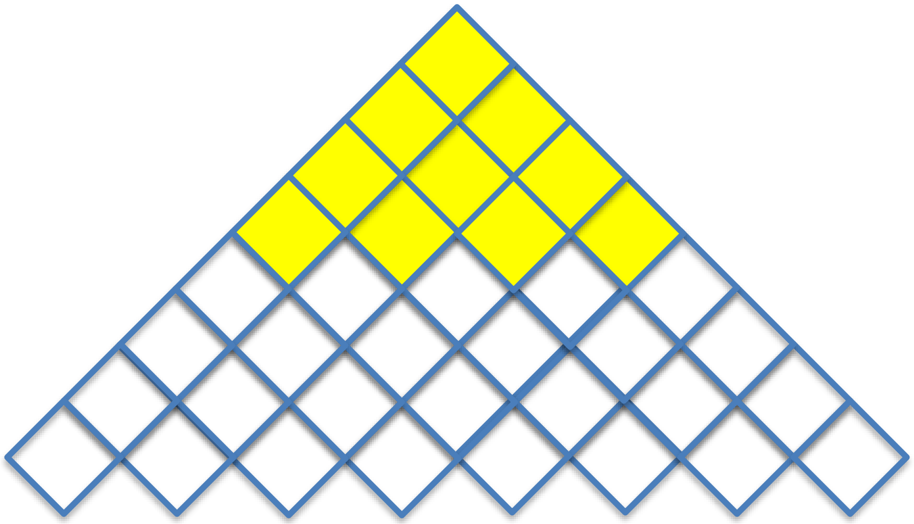


For all NT:
$\beta(NT, 1, N) = 1$

For $l = N - 2 \; downto \; 1$
For $i = 1..N - l$
$j = i + l$
For all NT:

$$\beta(NT, i, j) = \sum_{k=j+1}^{N} \sum_{NT_a} \beta(NT_a, i, k) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b) \alpha(NT_b, j+1, k)$$
$$+ \sum_{k=1}^{i} \sum_{NT_a} \beta(NT_a, k, j) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ ) \alpha(NT_b, k, i-1)$$

The Outside Recursion

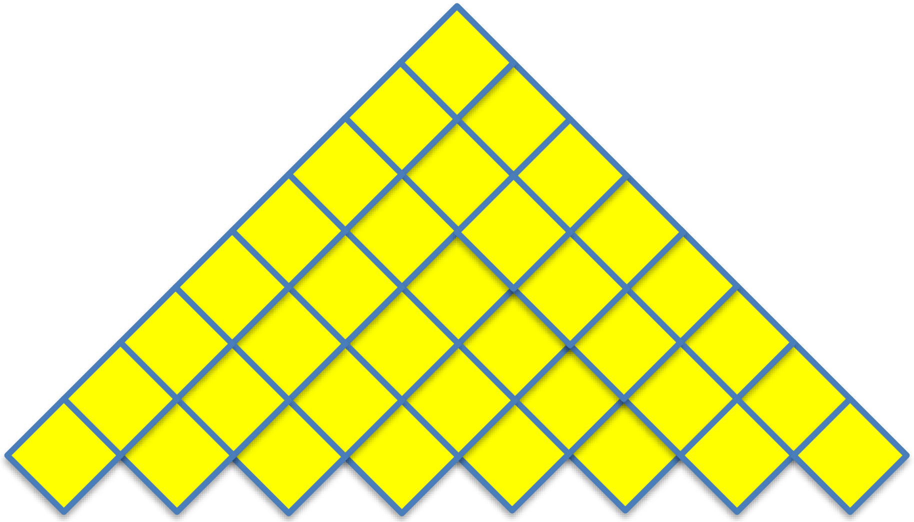


For all NT:
$\beta(NT, 1, N) = 1$

For $l = N - 2$ *downto* 1
For $i = 1..N - l$
$j = i + l$
For all NT:

$$\beta(NT, i, j) = \sum_{k=j+1}^{N} \sum_{NT_a} \beta(NT_a, i, k) \sum_{NT_b} P(NT_a \to NT\ NT_b)\alpha(NT_b, j+1, k)$$
$$+ \sum_{k=1}^{i} \sum_{NT_a} \beta(NT_a, k, j) \sum_{NT_b} P(NT_a \to NT_b NT\ )\alpha(NT_b, k, i-1)$$

The Outside Recursion



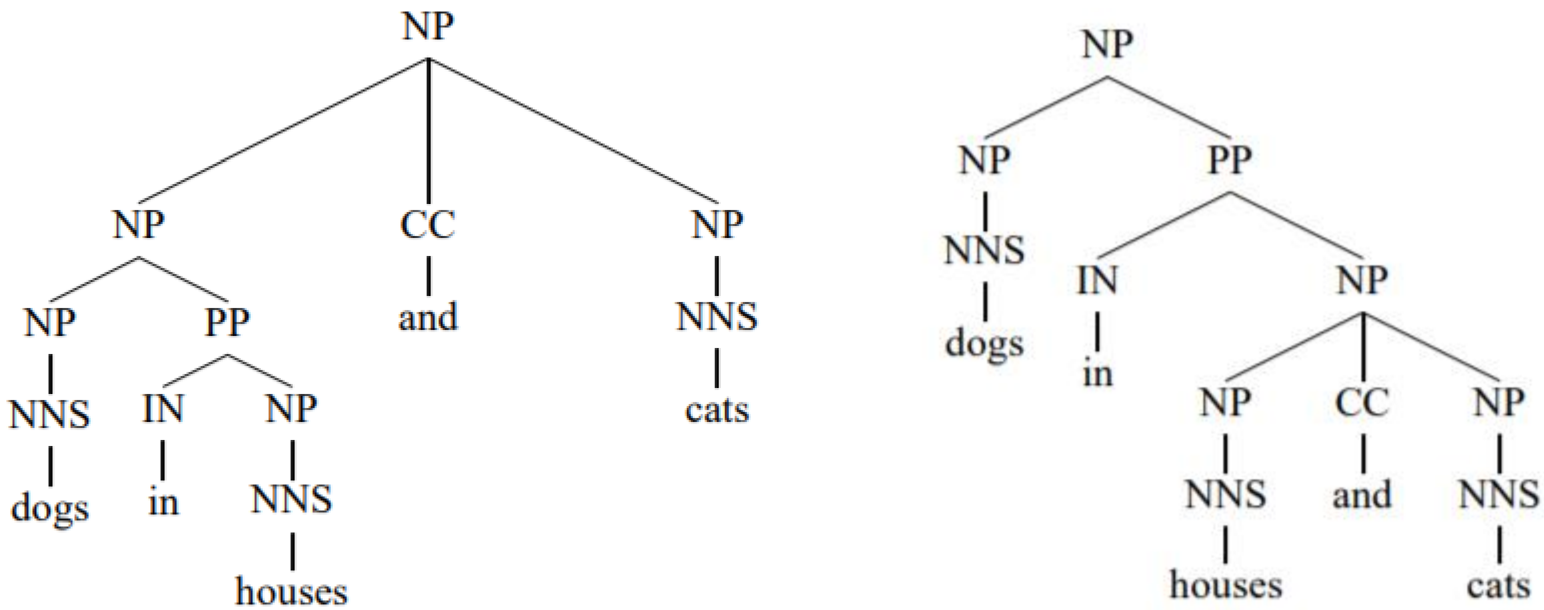
For all NT:
$\beta(NT, 1, N) = 1$

For $l = N - 2$ $downto$ 1
For $i = 1..N - l$
$j = i + l$
For all NT:

$$\beta(NT, i, j) = \sum_{k=j+1}^{N} \sum_{NT_a} \beta(NT_a, i, k) \sum_{NT_b} P(NT_a \rightarrow NT\ NT_b)\alpha(NT_b, j+1, k)$$

$$+ \sum_{k=1}^{i} \sum_{NT_a} \beta(NT_a, k, j) \sum_{NT_b} P(NT_a \rightarrow NT_b NT\ )\alpha(NT_b, k, i-1)$$

Inferences we would like to make..



Done

- Which of the probability of “dogs in houses and cats”
  - P(“dogs in houses and cats”)
- What is the probability that “houses and cats” is a clause by itself?
  - P(“houses and cats” = clause | “dogs in houses and cats”)
- What is the probability that its an *NP?*
  - P(“houses and cats” = NP | “dogs in houses and cats”)
- Is there a *PP* in the sentence?
  - P(PP | “dogs in houses and cats”)

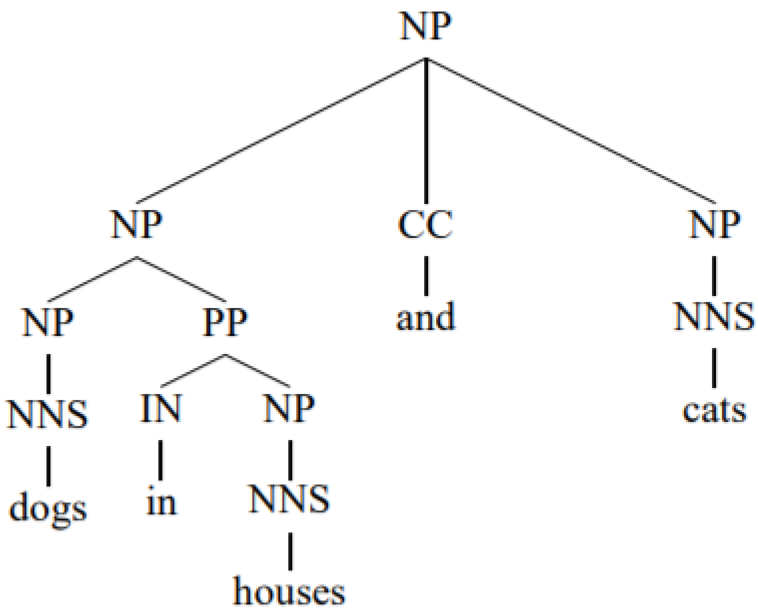
Posterior Marginal

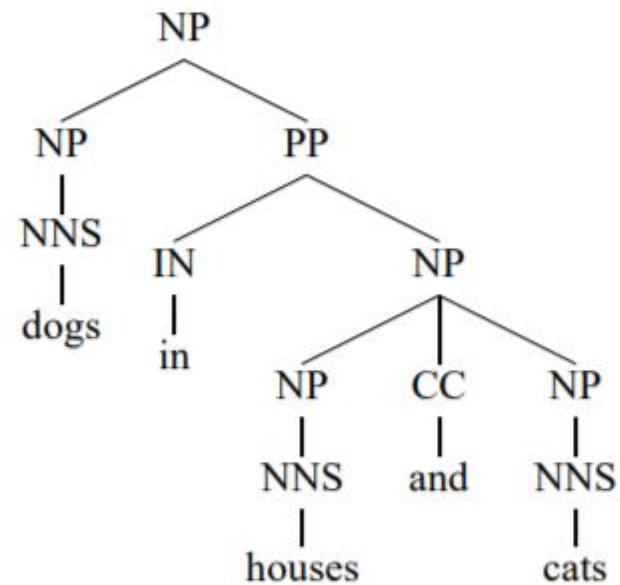
$$P(c(i,j) = NT, w_1 \dots w_N) = \alpha(NT,i,j)\beta(NT,i,j)$$

- The posterior marginal is:

$$P(c(i,j) = NT|W) = \frac{\alpha(NT,i,j)\beta(NT,i,j)}{\alpha(S,1,N)}$$

Inferences we would like to make..





Done

- Which of the probability of “dogs in houses and cats”
  - P(“dogs in houses and cats”)
- What is the probability that “houses and cats” is a clause by itself?
  - P(“houses and cats” = clause | “dogs in houses and cats”)
- What is the probability that its an *NP?*
  - P(“houses and cats” = NP | “dogs in houses and cats”)

Done

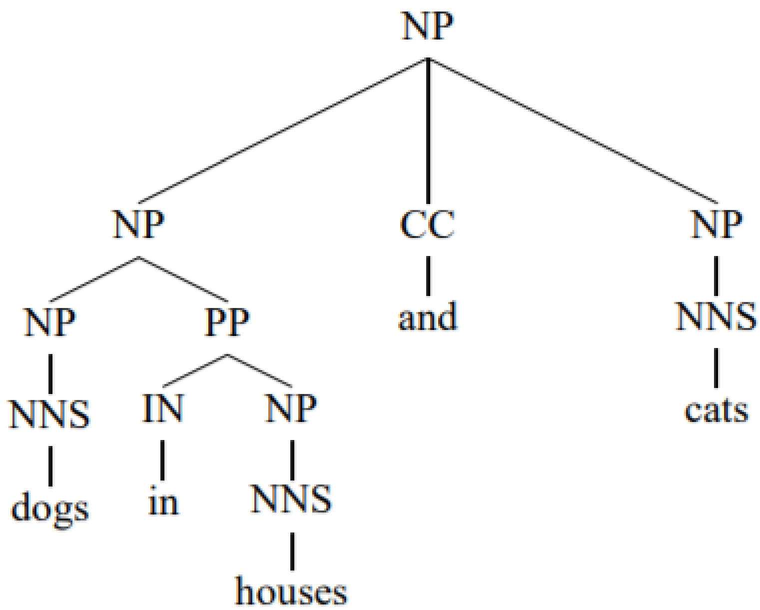
- Is there a *PP* in the sentence?
  - P(PP | “dogs in houses and cats”)

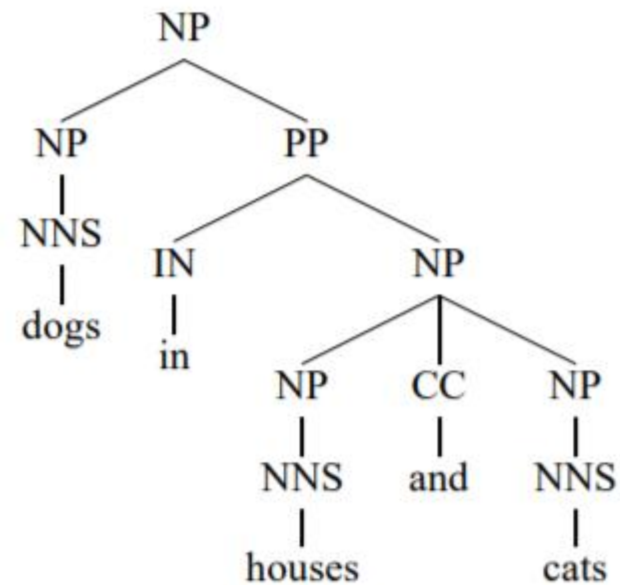
Posterior Marginal

- The posterior marginal that $w_i \ldots w_j$ is a constituent:

$$P(c(i,j)|W) = \sum_{NT} \frac{\alpha(NT,i,j)\beta(NT,i,j)}{\alpha(S,1,N)}$$

Inferences we would like to make..





Done

- Which of the probability of "dogs in houses and cats"
  – P("dogs in houses and cats")

Done

- What is the probability that "houses and cats" is a clause by itself?
  – P("houses and cats" = clause | "dogs in houses and cats")

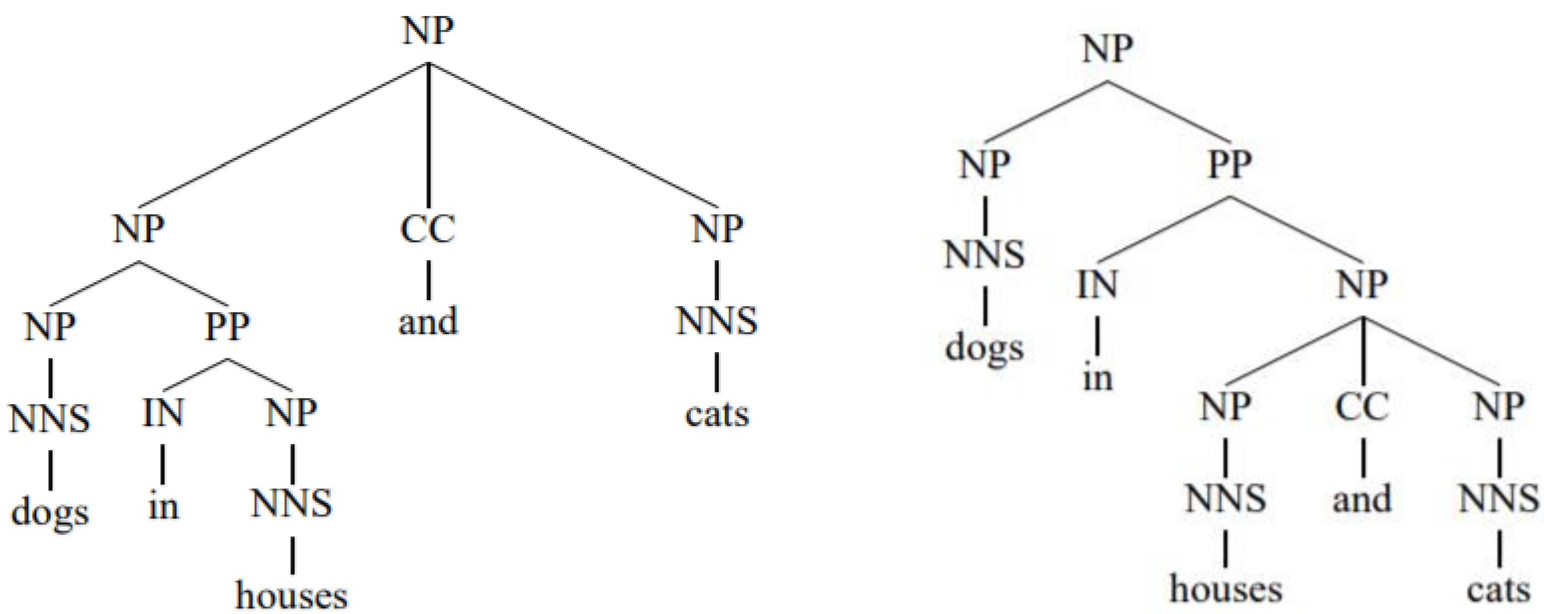
Done

- What is the probability that its an *NP?*
  – P("houses and cats" = NP | "dogs in houses and cats")
- Is there a *PP* in the sentence?
  – P(PP | "dogs in houses and cats")

Rule marginals

- $P(PP|W;\ \mathcal{G}) = \dfrac{\alpha(S,1,N;\mathcal{G}\backslash \text{PP}\ rules)}{\alpha(S,1,N;\mathcal{G})}$

Inferences we would like to make..



- Does the sentence have both a VP and a PP?
  - Exercise for you..

Posterior Marginals

- Marginal inference question for PCFGs
  - Given **w**, what is the probability of having a constituent of type Z from *i* to *j*?
  - Given **w**, what is the probability of having a constituent of ***any*** type from *i* to *j*?
  - Given **w**, what is the probability of using rule Z -> XY to derive the span from *i* to *j*?

In Conclusion

- Have looked at a few ways of arriving at posterior marginal inferences for fininte-state and context-free grammars
- Similar approach extends to dependency grammars
  - If you can use DP and you can write probabilistic rules, you can derive probabilistic inferences
- Possibly one of the biggest uses for these methods is *learning*
  - Applicable in EM methods to *learn* grammars
  - Not a topic for today..