Translation as Weighted Deduction

Adam Lopez
University of Edinburgh

Moses	Hiero
Koehn et al., ACL 2007	Chiang, CL 2007

Moses

Koehn et al., ACL 2007

Hiero

Chiang, CL 2007

Moses

Koehn et al., ACL 2007

Hiero

Chiang, CL 2007

Moses

Koehn et al., ACL 2007

30.7

Hiero

Chiang, CL 2007

32.6

Lopez, Coling 2008

Moses	Hiero
Koehn et al., ACL 2007	Chiang, CL 2007
phrase-based	hierarchical phrase-based
15 features	5 features
stack decoding	cube pruning
30.7	32.6

Lopez, Coling 2008

	Moses Koehn et al., ACL 2007	Hiero Chiang, CL 2007
rules	phrase-based	hierarchical phrase-based
parameters	15 features	5 features
search	stack decoding	cube pruning
	30.7	**32.6**

Lopez, Coling 2008

rules	phrase-based	hierarchical phrase-based
parameters	15 features	5 features
search	stack decoding	cube pruning

rules	phrase-based	hierarchical phrase-based
parameters	5 features	5 features
search	cube pruning	cube pruning

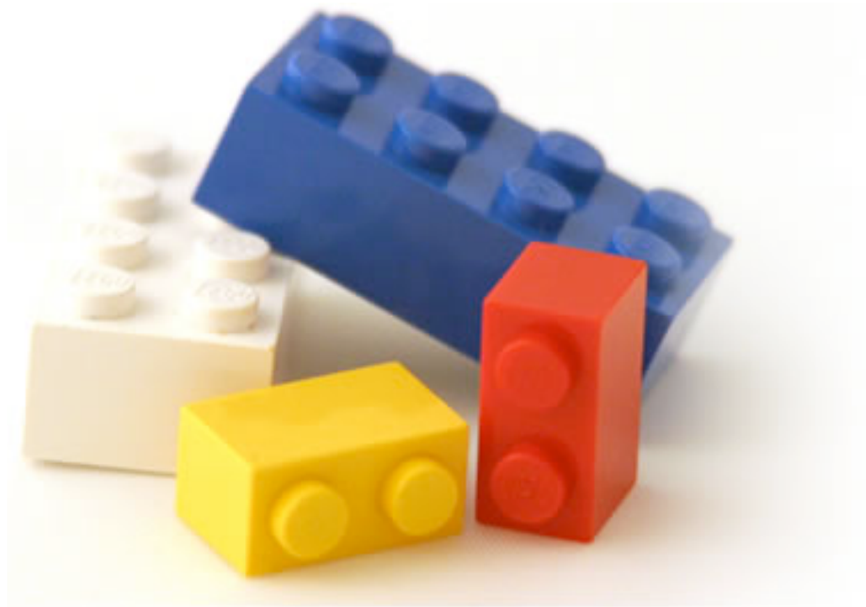
rules	phrase-based	synchronous TAG
parameters	5 features	5 features
search	cube pruning	cube pruning

This talk *is not* about

How to improve your BLEU score by 1.9.

This talk *is* about

<u>Building</u> and **<u>analyzing</u>** translation models and algorithms in a modular way.



rules	phrase-based
parameters	15 features
search	stack decoding

rules	phrase-based	deductive logic
parameters	15 features	
search	stack decoding	

rules	phrase-based	deductive logic
parameters	15 features	semiring
search	stack decoding	

rules	phrase-based	deductive logic
parameters	15 features	semiring
search	stack decoding	(hyper)graph algorithms

北 风 呼啸

北 风 呼啸

北/north	风/wind	呼啸/whistles
北/northerly	风/winds	呼啸/strong

- word-to-word translation
- no reordering

北 风 呼啸

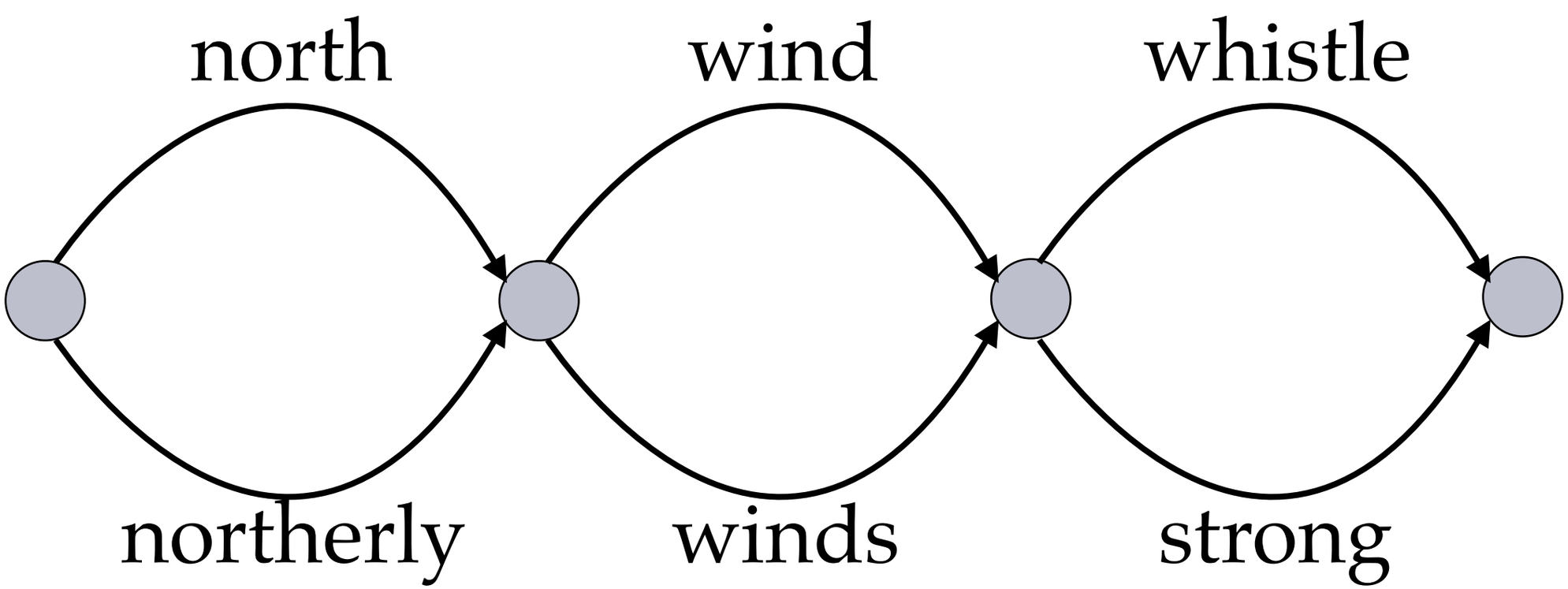
北/north	风/wind	呼啸/whistles
北/northerly	风/winds	呼啸/strong

north wind whistles	northerly wind whistles
north wind strong	northerly winds whistles
north winds whistles	northerly winds strong
north winds strong	northerly wind strong

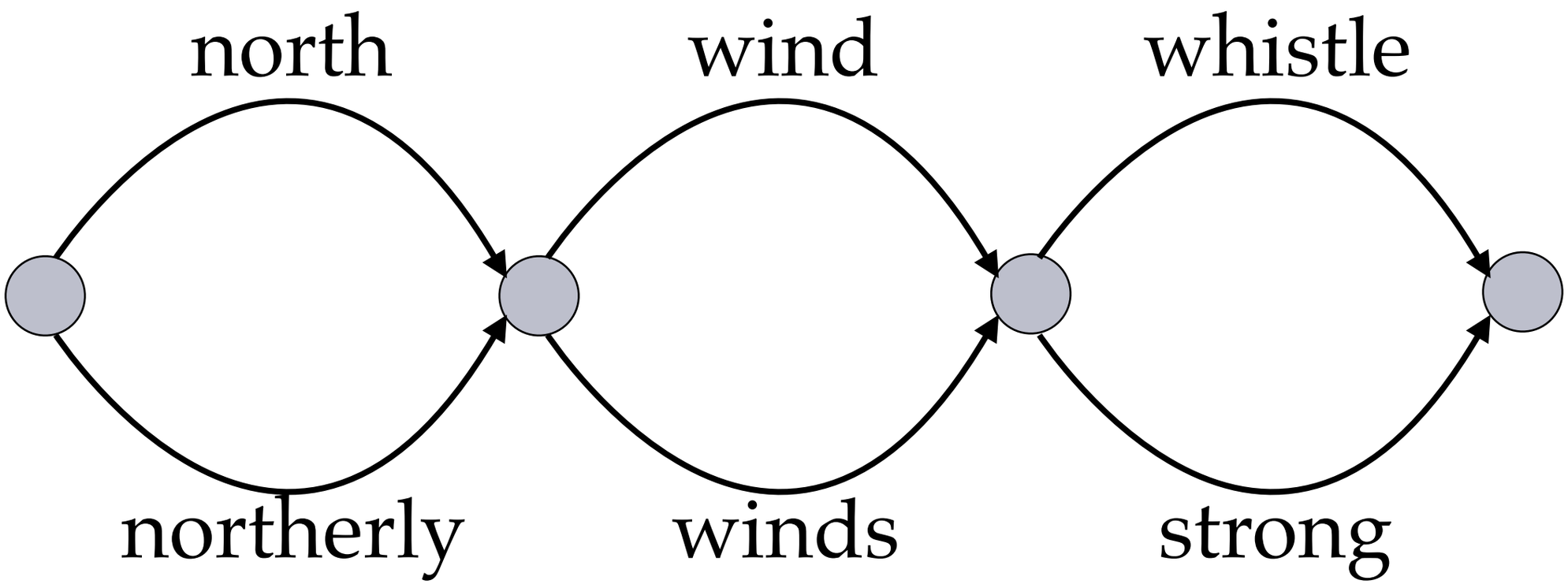
notice: complexity is $O(2^L)$ for sentence length L

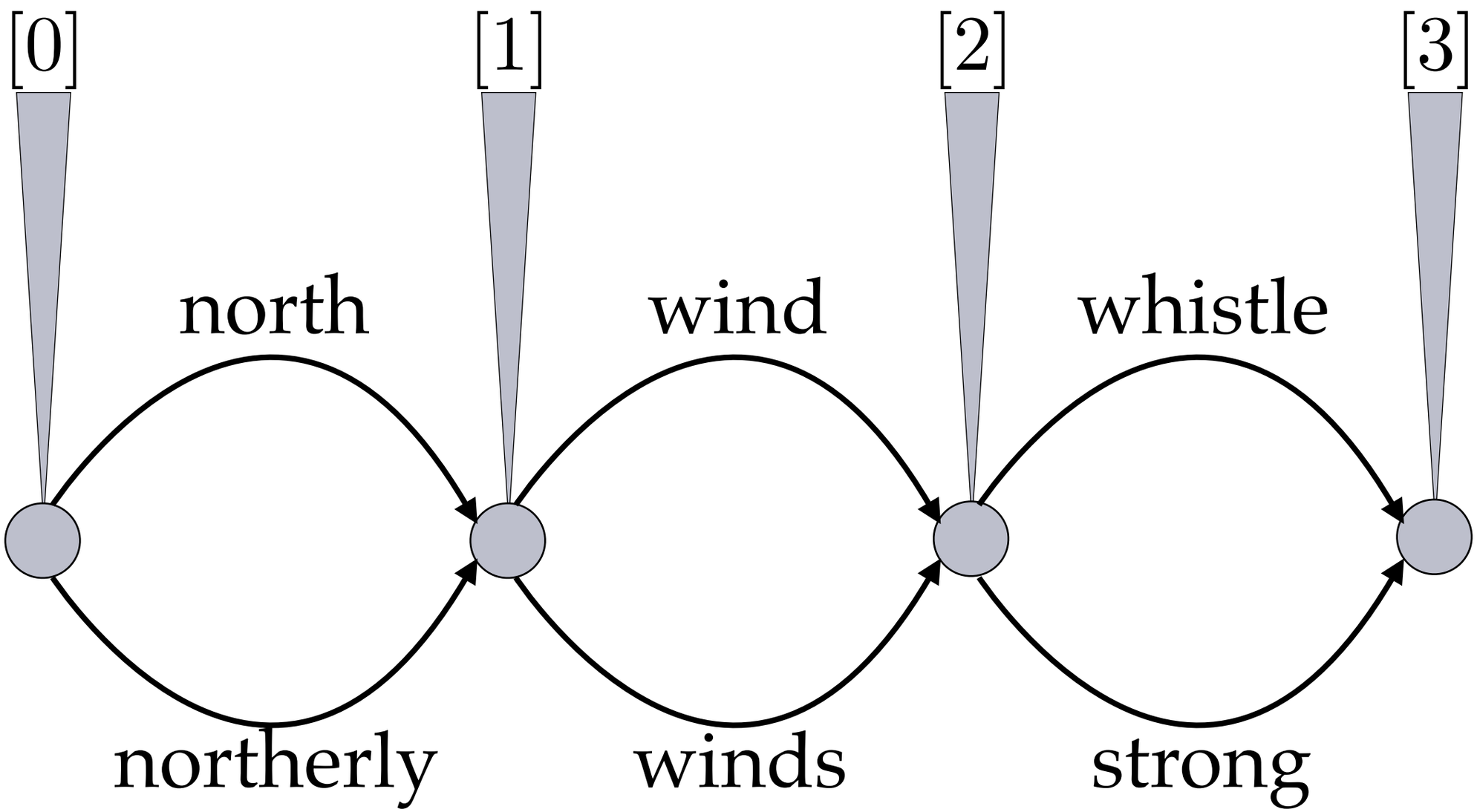
北 风 呼啸

北/north (.8)	风/wind (.6)	呼啸/whistles (.7)
北/northerly (.2)	风/winds (.4)	呼啸/strong (.3)

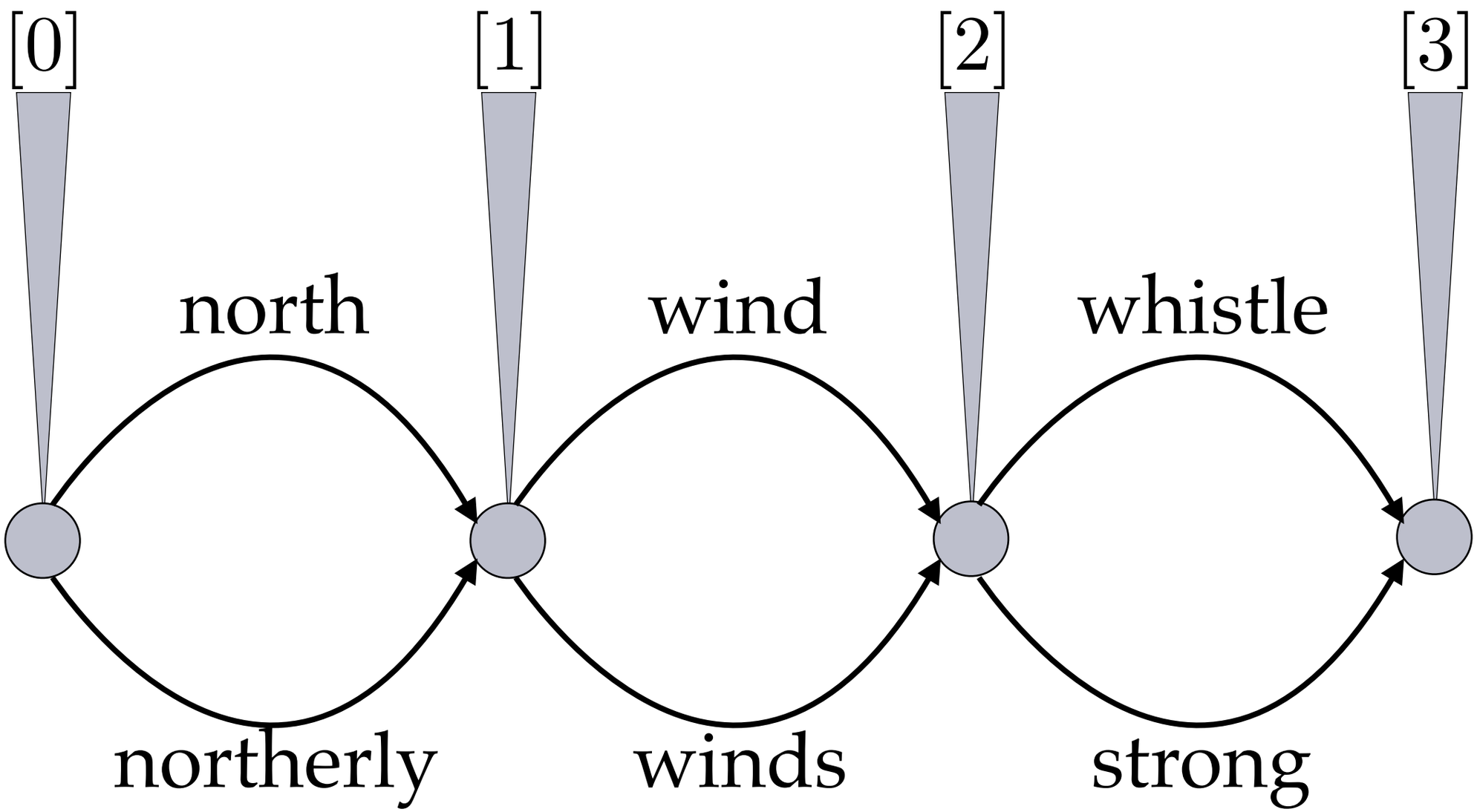


complexity is $O(L)$ for sentence length L

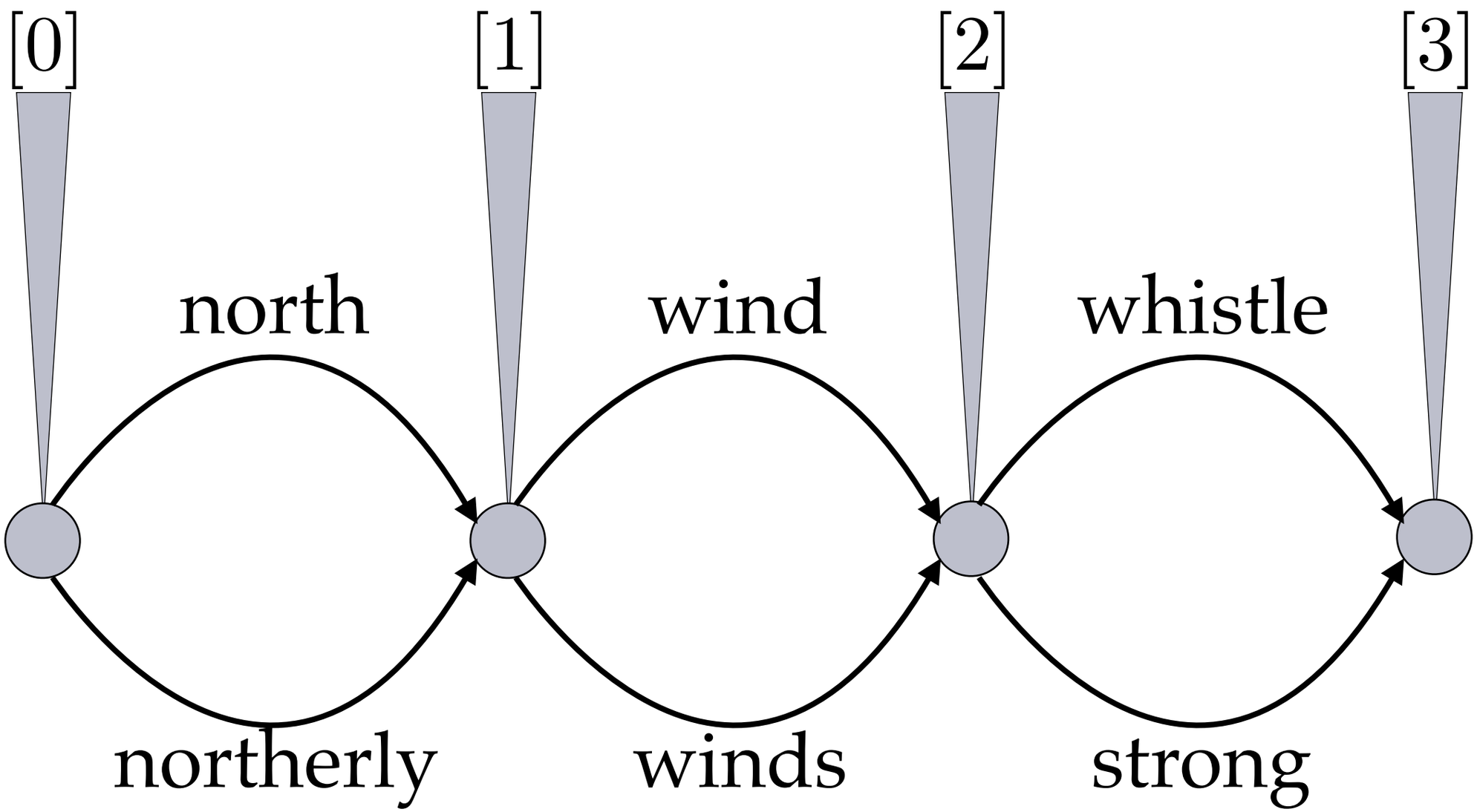

north
wind
whistle
northerly
winds
strong


[0]
[1]
[2]
[3]
north
wind
whistle
northerly
winds
strong

$$\frac{[1] \quad R(\text{风}/wind)}{[2]}$$

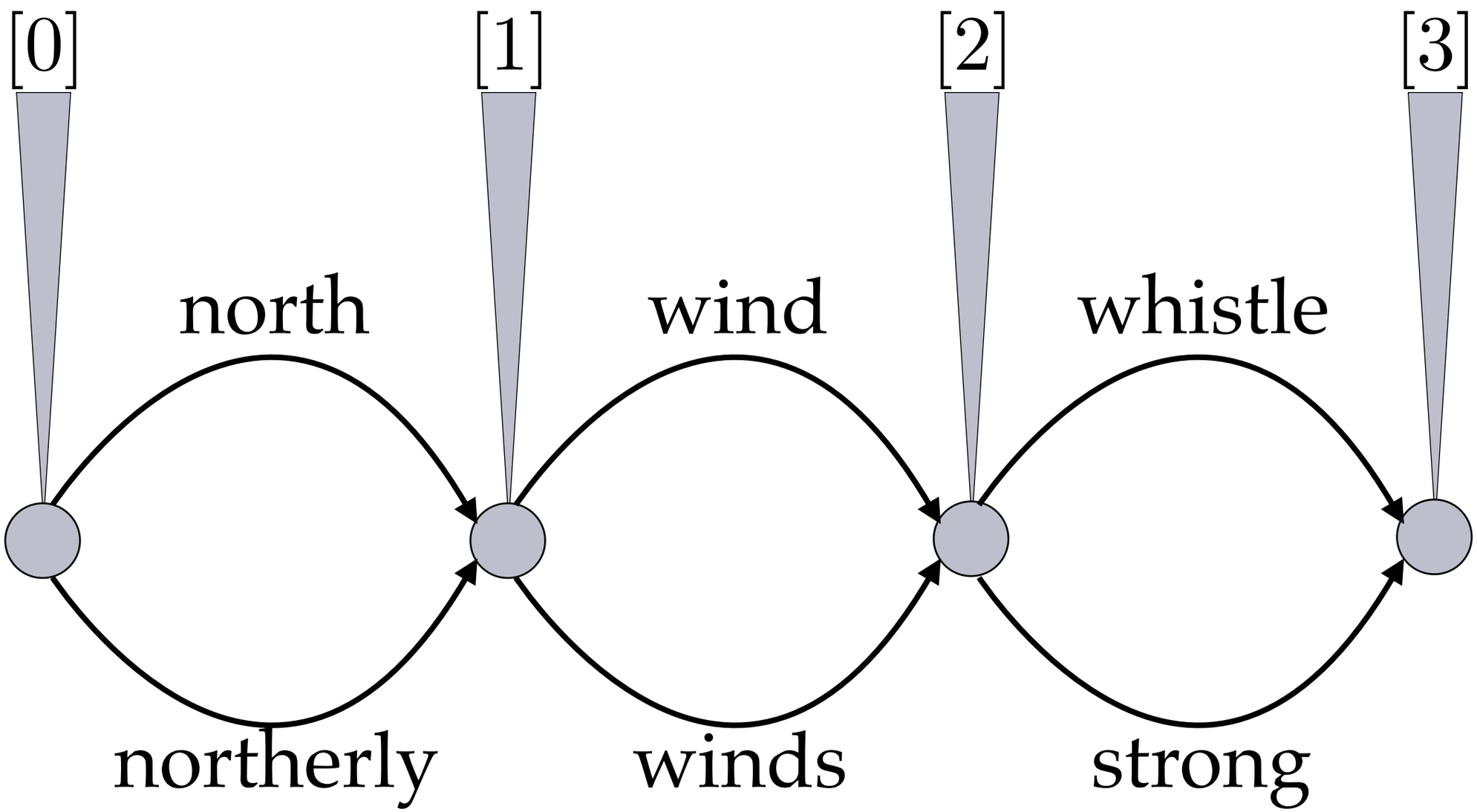


$$\frac{[1] \quad R(\text{风}/wind)}{[2]} \qquad [i]$$

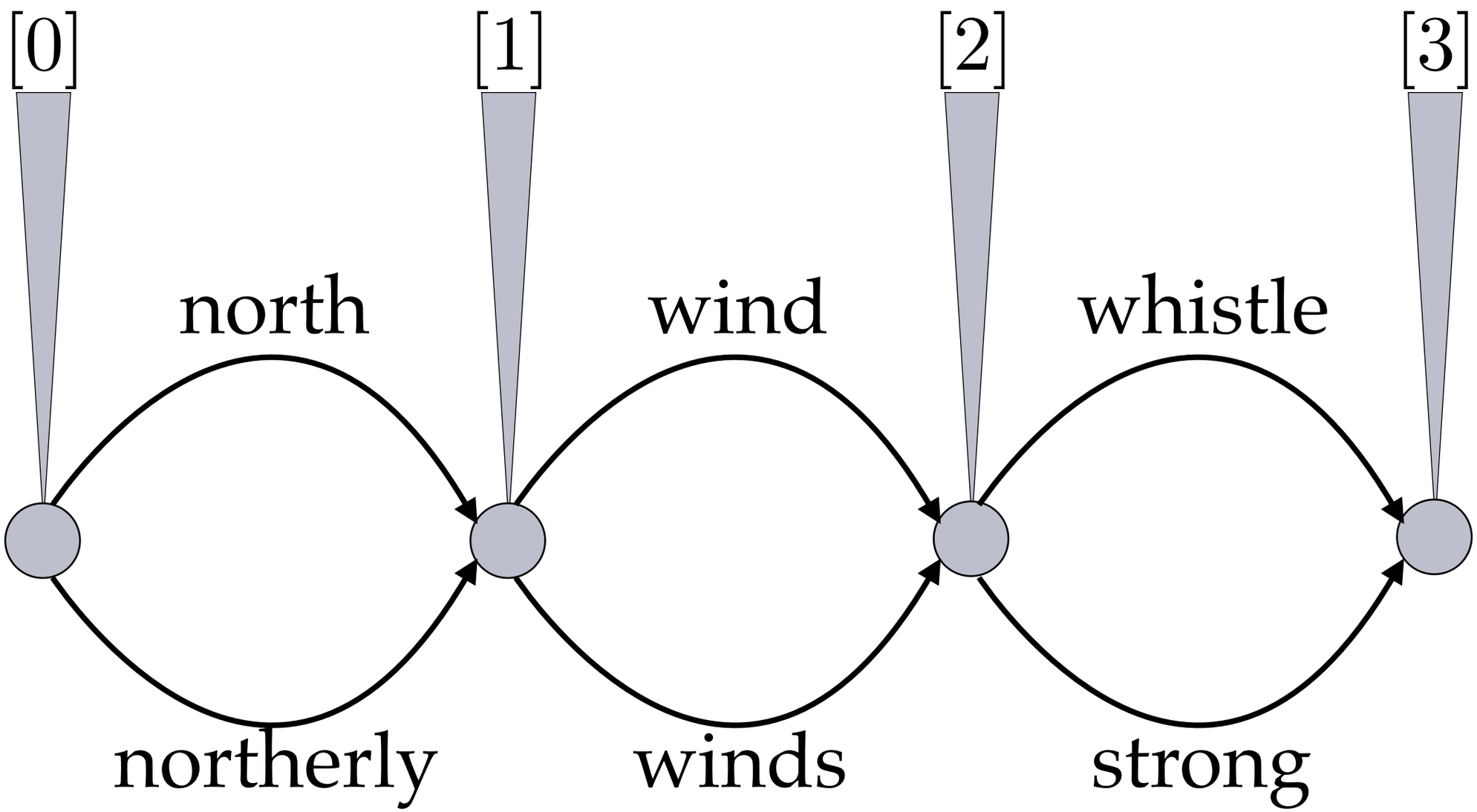


$$\frac{[1] \quad R(\text{风}/wind)}{[2]}$$

$$[i] \quad R(f_{i+1}/e_j)$$

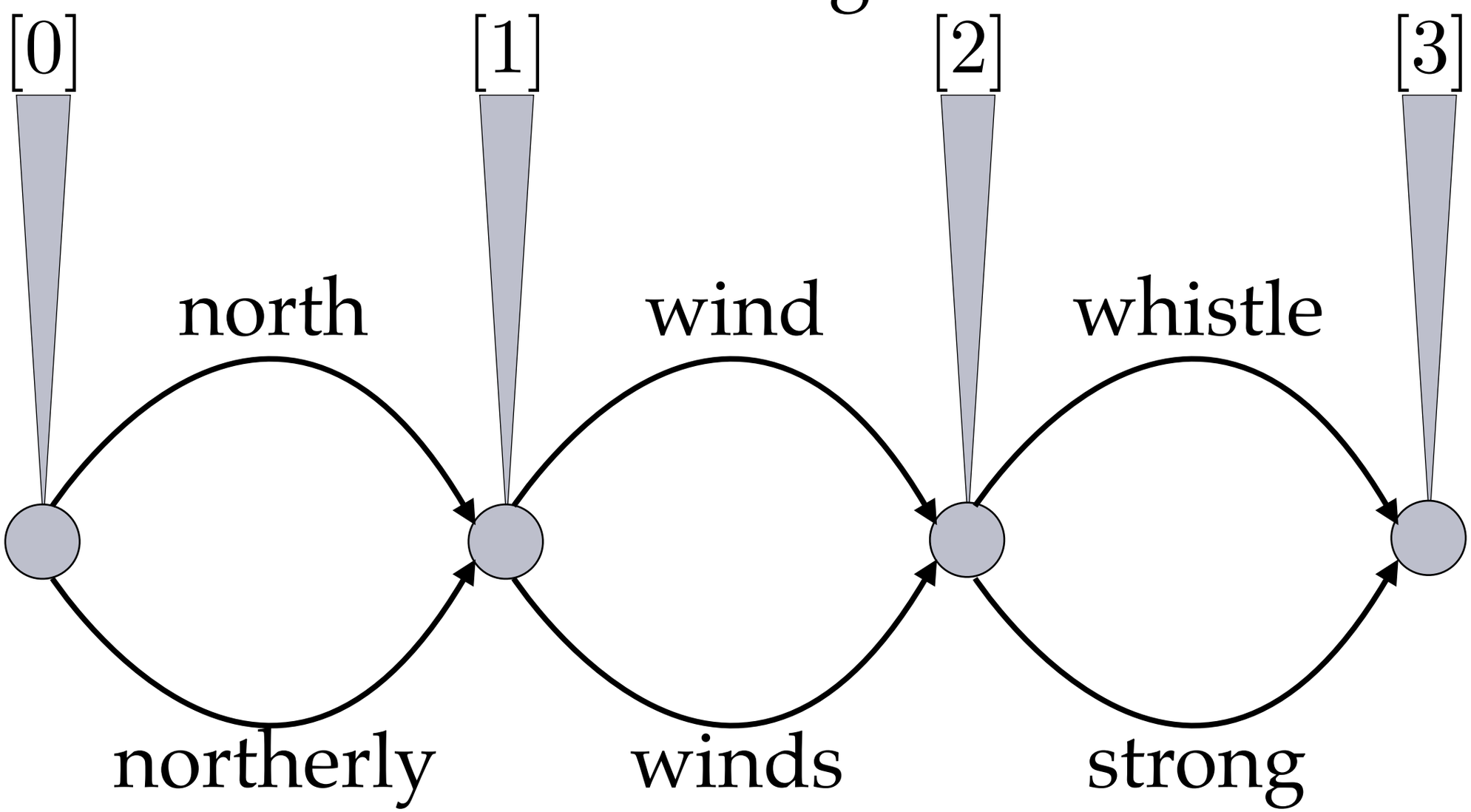


$$\frac{[1] \quad R(\text{风}/wind)}{[2]} \qquad \frac{[i] \quad R(f_{i+1}/e_j)}{[i+1]}$$



$$\frac{[1] \quad R(\text{风}/wind)}{[2]} \qquad \frac{[i] \quad R(f_{i+1}/e_j)}{[i+1]}$$

i ranges over sentence length



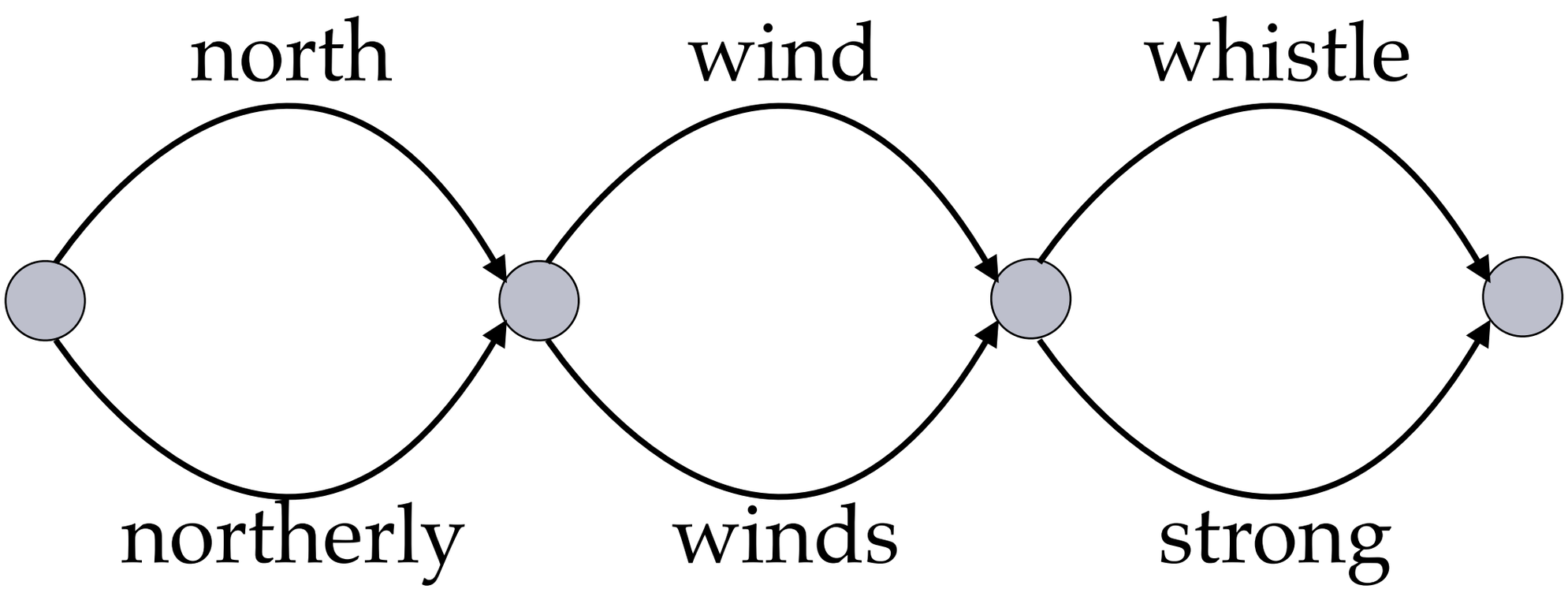
Determine complexity from inspection

McAllester, Proc. Static Analysis 1999

$$\frac{[i] \quad R(f_{i+1}/e_j)}{[i+1]}$$

Viterbi: $\langle [0,1], \max, \times \rangle$ Boolean: $\langle \{\top, \bot\}, \cup, \cap \rangle$

sum: $\langle [0,1], +, \times \rangle$ *Reverse* (outside) values



Compute many quantities on same graph

Goodman, CL 1999

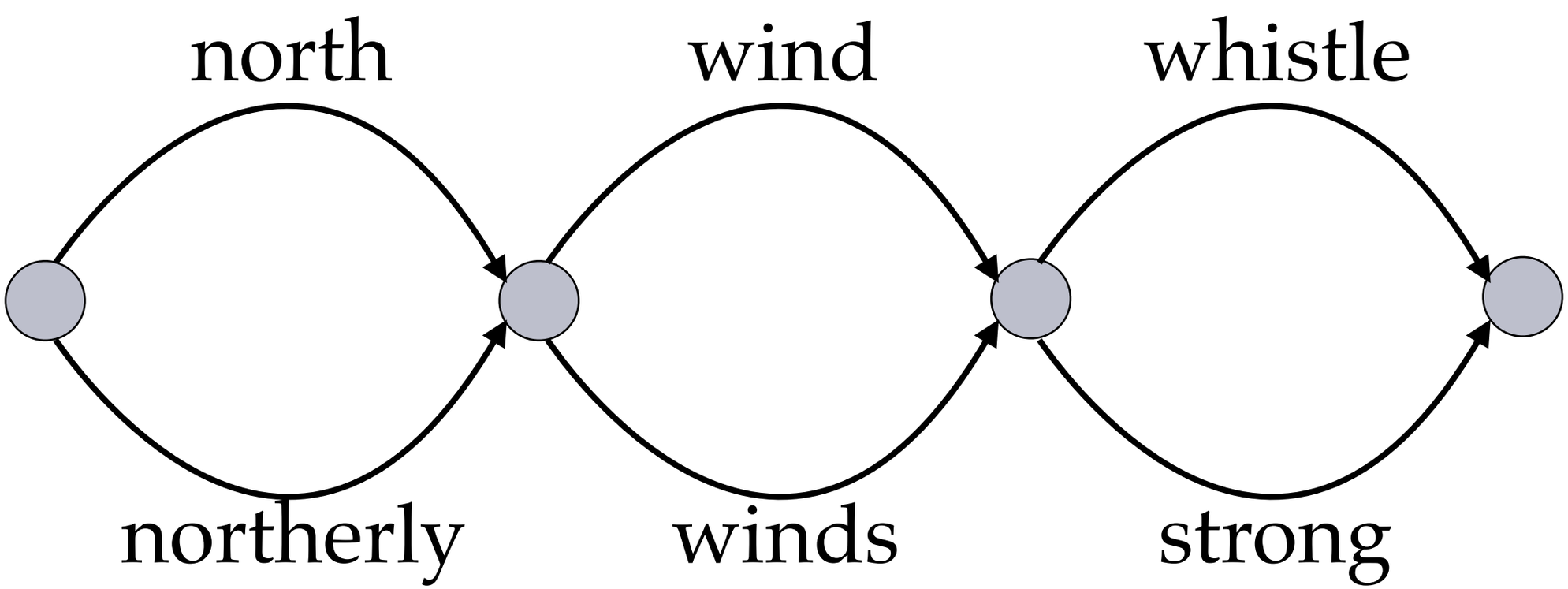
$$\frac{[i] \quad R(f_{i+1}/e_j)}{[i+1]}$$

Expectation semiring

Eisner 2002

Approximation semiring

Gimpel & Smith 2009



Compute many quantities on same graph

Goodman, CL 1999

Basic Idea

- Supply a logic and a semiring, get a complete algorithm.
- Does it work for most translation models?

$$\frac{[i'', V] \quad R(f_{i+1} \ldots f_{i'} / e_j \ldots e_{j'})}{[i', V \vee 0^i 1^{i'-i} 0^{I-i'}]} V \wedge 0^i 1^{i'-i} 0^{I-i'} = 0^I, |i - i''| \leq d$$

previous last position translated

previous coverage vector

phrase pair

distortion limit

$$\frac{[i'', V] \quad R(f_{i+1} \ldots f_{i'} / e_j \ldots e_{j'})}{[i', V \vee 0^i 1^{i'-i} 0^{I-i'}]} \quad V \wedge 0^i 1^{i'-i} 0^{I-i'} = 0^I, \; |i - i''| \leq d$$

last position translated

coverage vector

only translate previously untranslated words

Phrase-based Models

Phrase-based Models

Max distortion d

see, e.g. Moore & Quirk 2007

Window length d

Moses (Hoang & Koehn, pc)

First d uncovered

see, e.g. Tillman & Ney 2003,
Zens & Ney 2004

Phrase-based Models

Max distortion d

see, e.g. Moore & Quirk 2007

$$\frac{[i'', V] \quad R(f_{i+1} \ldots f_{i'} / e_j \ldots e_{j'})}{[i', V \vee 0^i 1^{i'-i} 0^{I-i'}]} V \wedge 0^i 1^{i'-i} 0^{I-i'} = 0^I, |i - i''| \leq d$$

Window length d

Moses (Hoang & Koehn, pc)

$$\frac{[i, C] \quad R(f_{i+1} \ldots f_{i'} / e_j \ldots e_{j'})}{[i', C \ll i' - i]} C \wedge 1^{i'-i} 0^{d-i'+i} = 0^d, i' - i \leq d$$

$$\frac{[i, C] \quad R(f_{i'} \ldots f_{i''} / e_j \ldots e_{j'})}{[i, C \vee 0^{i'-i} 1^{i''-i'} 0^{d-i''+i}]} C \wedge 0^{i'-i} 1^{i''-i'} 0^{d-i''+i} = 0^d, i'' - i \leq d$$

First d uncovered

see, e.g. Tillman & Ney 2003, Zens & Ney 2004

$$\frac{[i, U] \quad R(f_{i'} \ldots f_{i''} / e_j \ldots e_{j'})}{[i'', U - [i', i''] \vee [i'', i'' + d - |U - [i', i'']|]]} i' > i, f_{i+1} \in U$$

$$\frac{[i, U] \quad R(f_{i'} \ldots f_{i''} / e_j \ldots e_{j'})}{[i, U - [i', i''] \vee [\max(U \vee i) + 1, \max(U \vee i) + 1 + d - |U - [i', i'']|]]} i' < i, [f_{i'}, f_{i''}] \subset U$$

Phrase-based Models

Max distortion d

see, e.g. Moore & Quirk 2007

Window length d

Moses (Hoang & Koehn, pc)

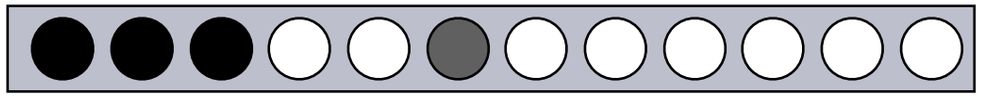
First d uncovered

see, e.g. Tillman & Ney 2003,
Zens & Ney 2004

Phrase-based Models

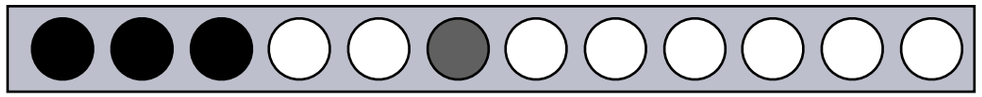
Max distortion d

see, e.g. Moore & Quirk 2007



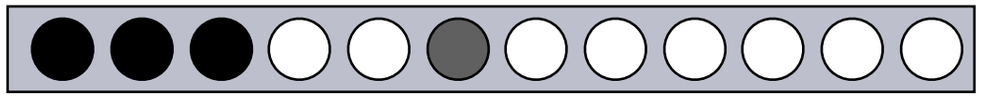
Window length d

Moses (Hoang & Koehn, pc)



First d uncovered

see, e.g. Tillman & Ney 2003,
Zens & Ney 2004

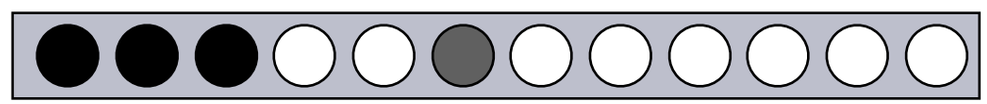


Phrase-based Models

$d = 3$

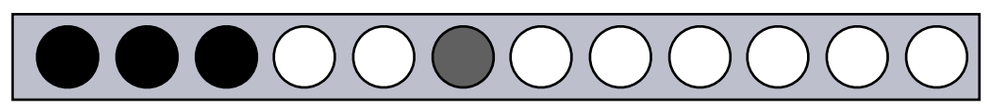
Max distortion d

see, e.g. Moore & Quirk 2007



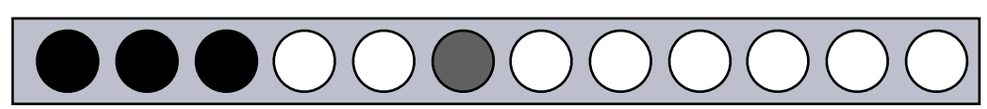
Window length d

Moses (Hoang & Koehn, pc)



First d uncovered

see, e.g. Tillman & Ney 2003,
Zens & Ney 2004

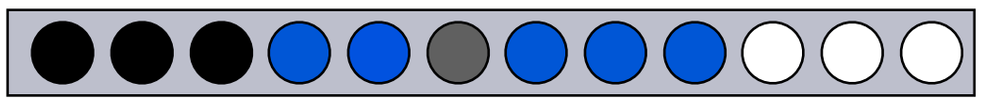


Phrase-based Models

$d = 3$

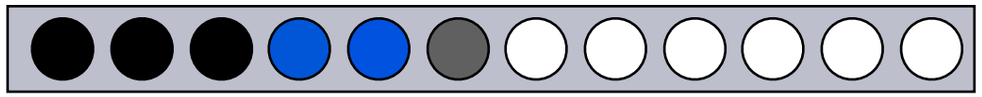
Max distortion d

see, e.g. Moore & Quirk 2007



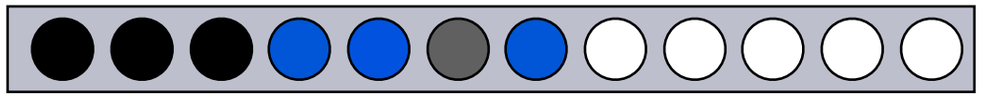
Window length d

Moses (Hoang & Koehn, pc)



First d uncovered

see, e.g. Tillman & Ney 2003,
Zens & Ney 2004

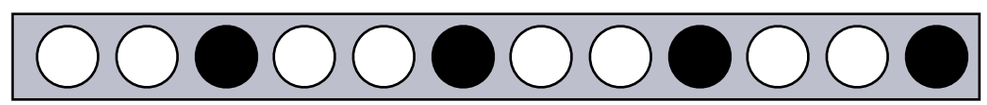


Phrase-based Models

$d = 3$

Max distortion d

see, e.g. Moore & Quirk 2007

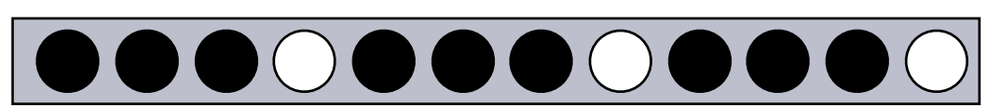


Window length d

Moses (Hoang & Koehn, pc)

First d uncovered

see, e.g. Tillman & Ney 2003,
Zens & Ney 2004

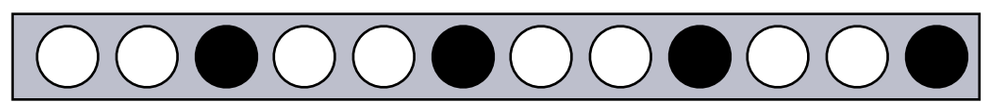


Phrase-based Models

$d = 3$

Max distortion d

see, e.g. Moore & Quirk 2007



$O(n^3 d^2)$

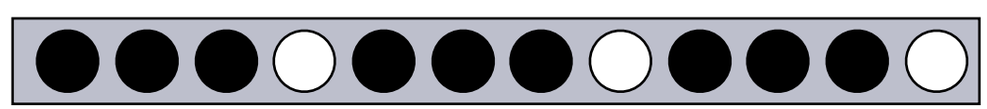
Window length d

Moses (Hoang & Koehn, pc)

$O(nd^2 2^d)$

First d uncovered

see, e.g. Tillman & Ney 2003, Zens & Ney 2004



$O(nd\binom{n}{d+1})$

Phrase-based Models

<u>These models are not the same.</u>

- Each can generate translations that the other cannot (regardless of d).
- Different complexities.
- Reported results will be impossible to replicate with your (different) strategy.

Good News

- Most translation models are a few lines of deductive logic.
- Computation of any semiring for free.

Good News

- Most translation models are a few lines of deductive logic.
- Computation of any semiring for free.
- You might conclude: give a logic and a semiring, get a complete algorithm.

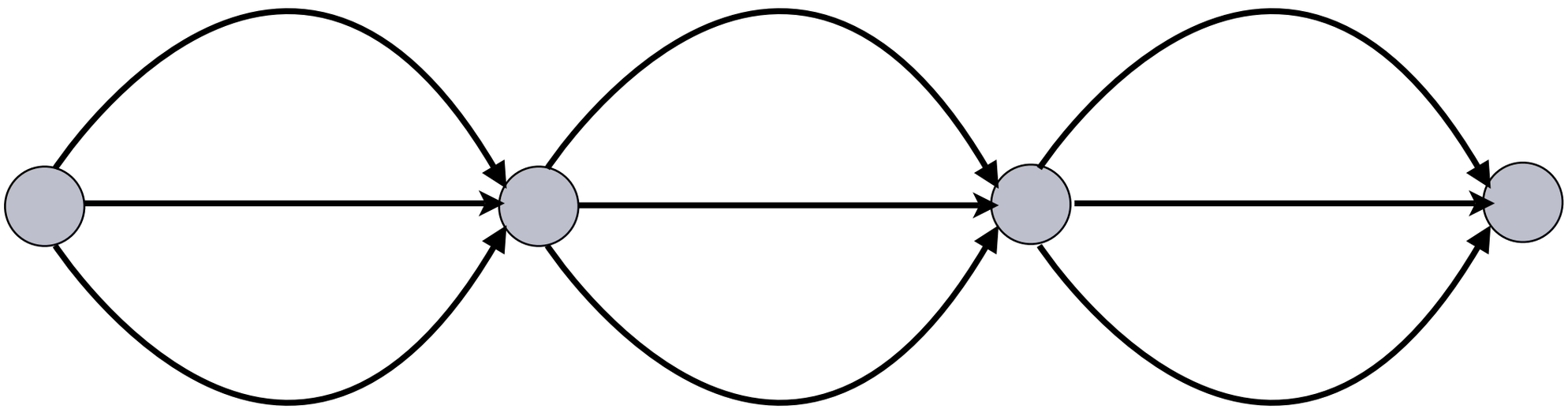
Result

- Given:
  - A logic
  - A semiring
- Get: a complete algorithm

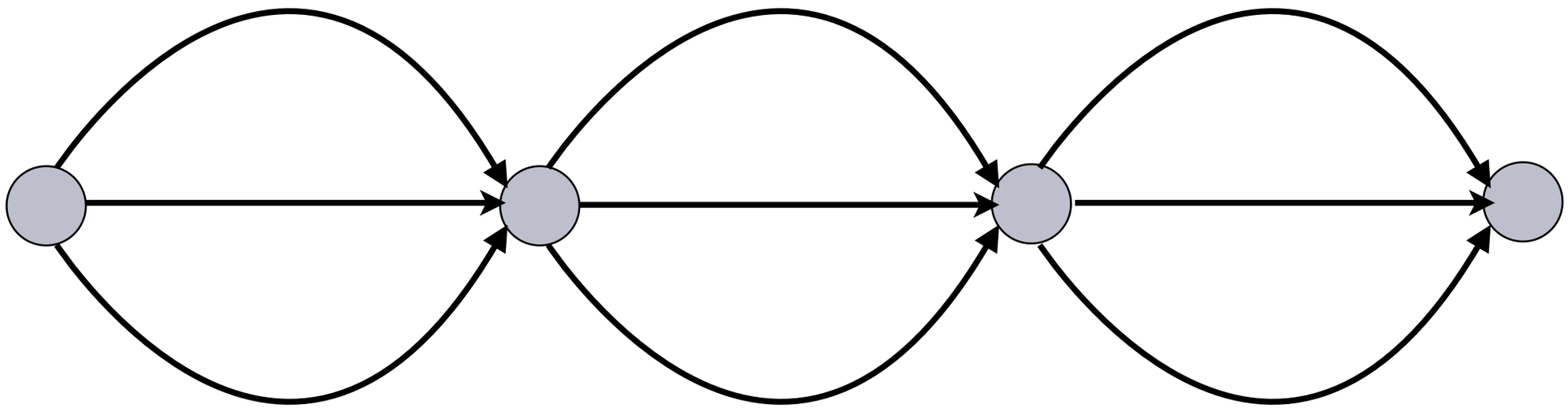
Bad News

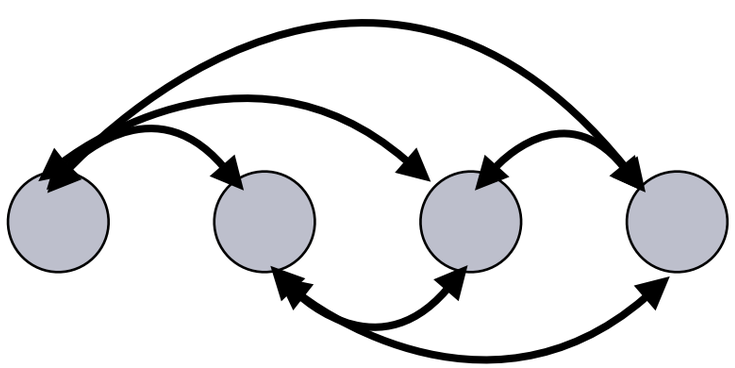
- Our models use non-local features.
- We need approximate search algorithms (and we need to be able to tweak them).

Non-local features

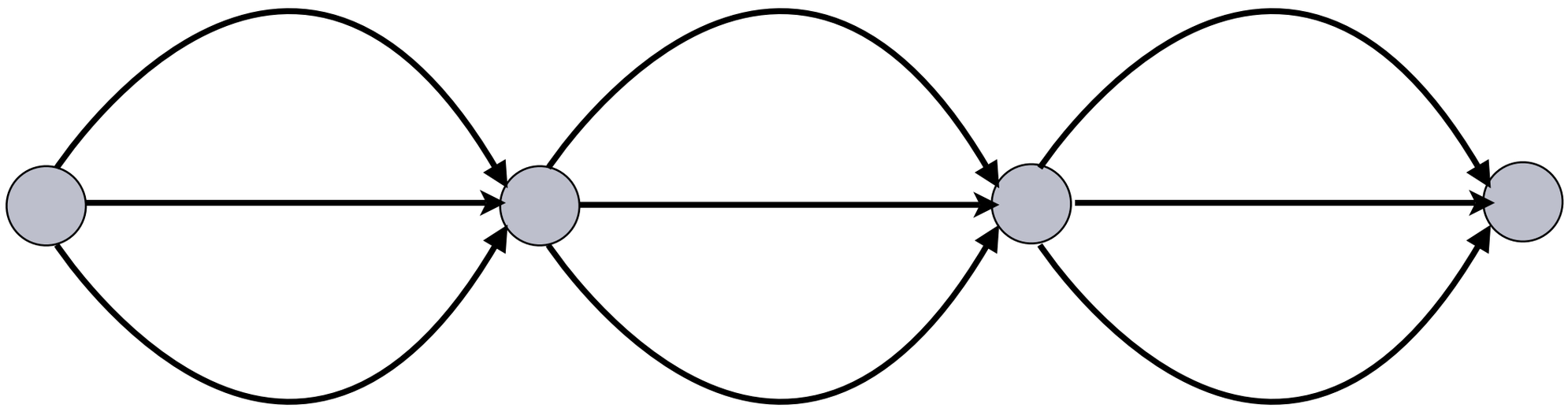


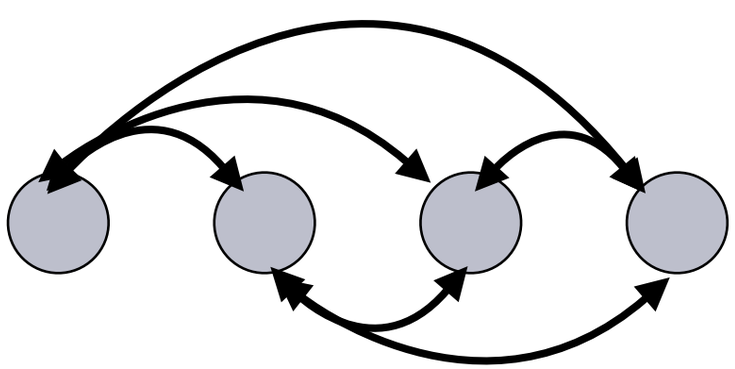
Non-local features





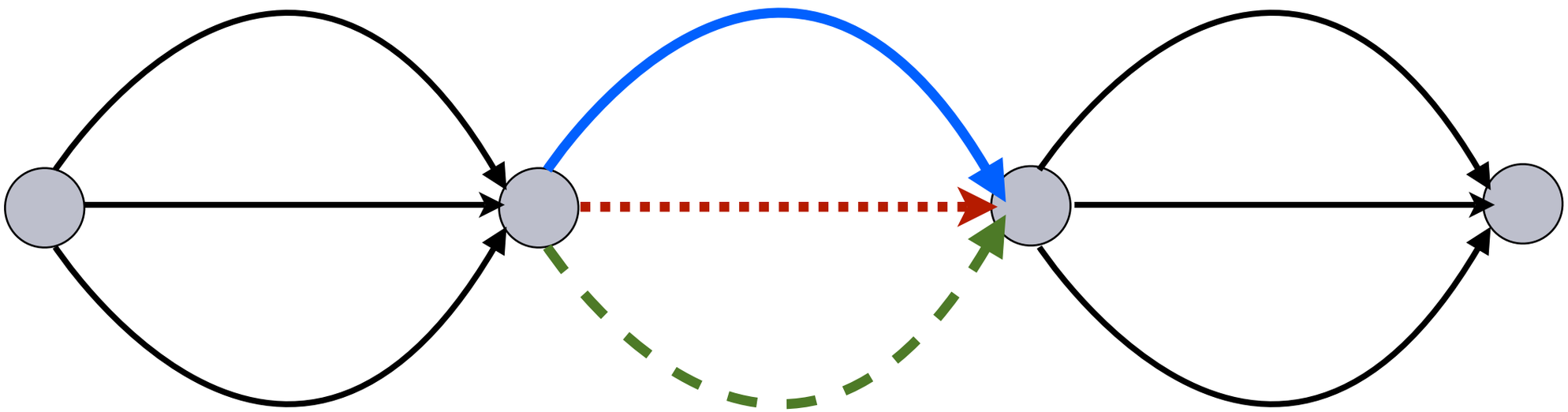
Non-local features



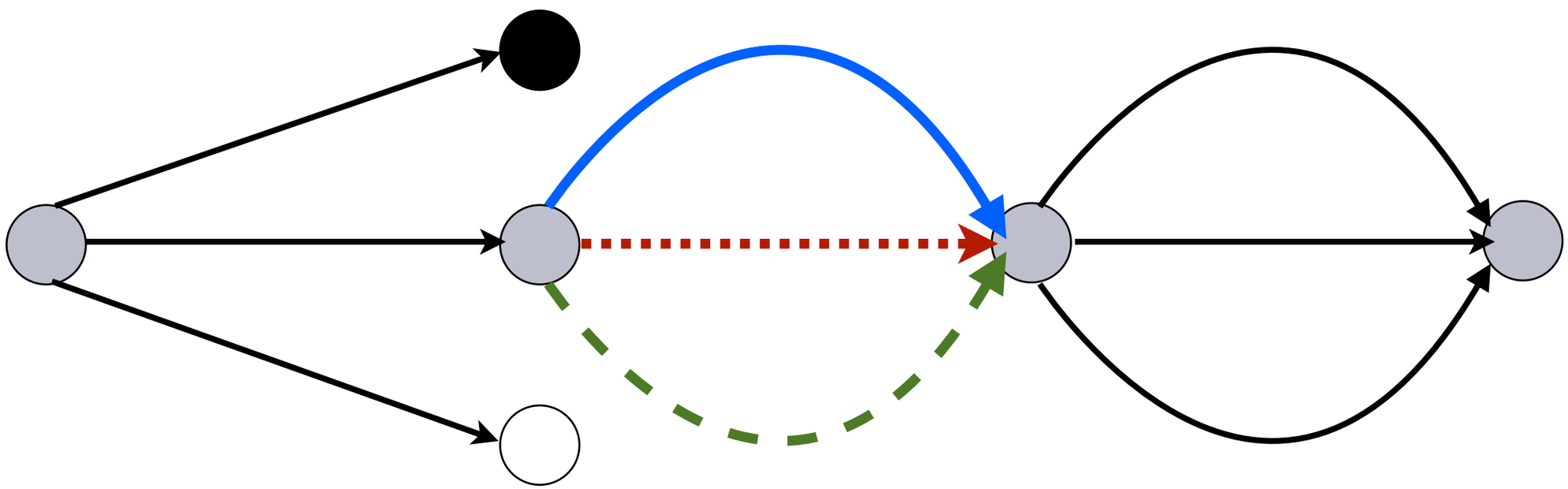


$$\frac{[e_q, \ldots, e_{q+n-2}]R(e_q, \ldots, e_{q+n-1})}{[e_{q+1}, \ldots, e_{q+n-1}]}$$

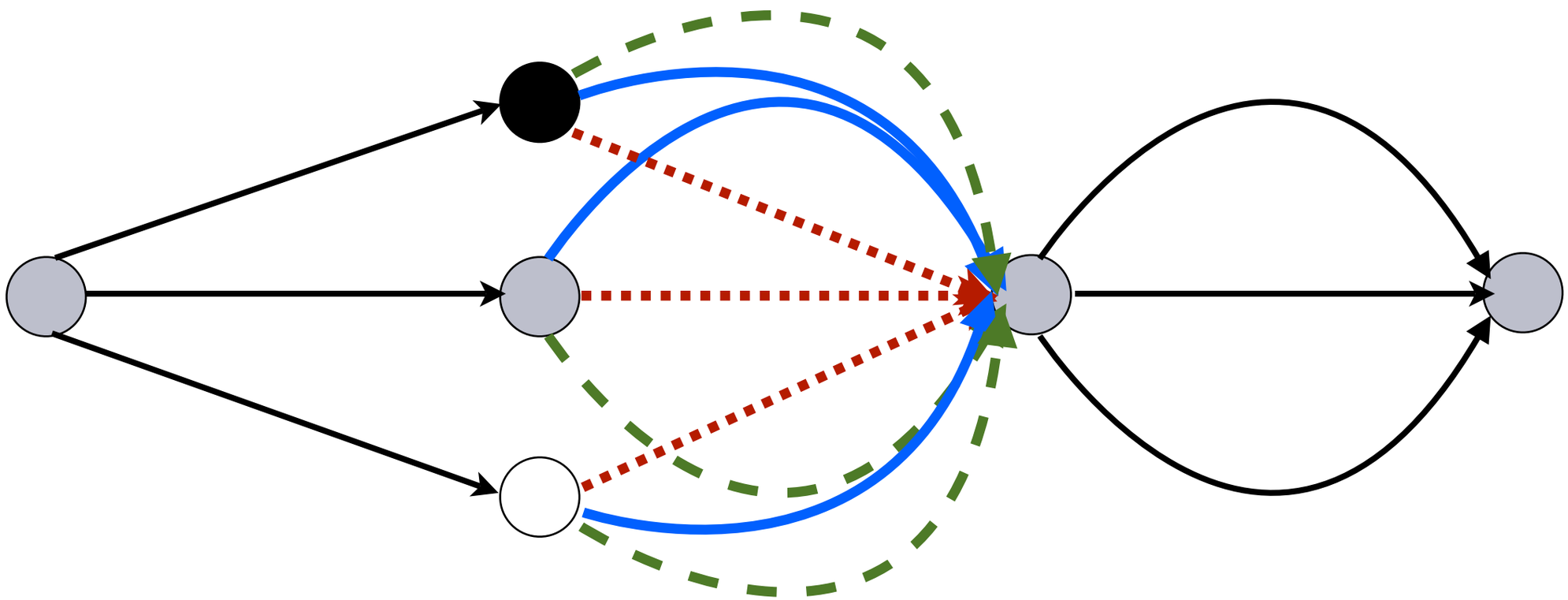
Non-local features



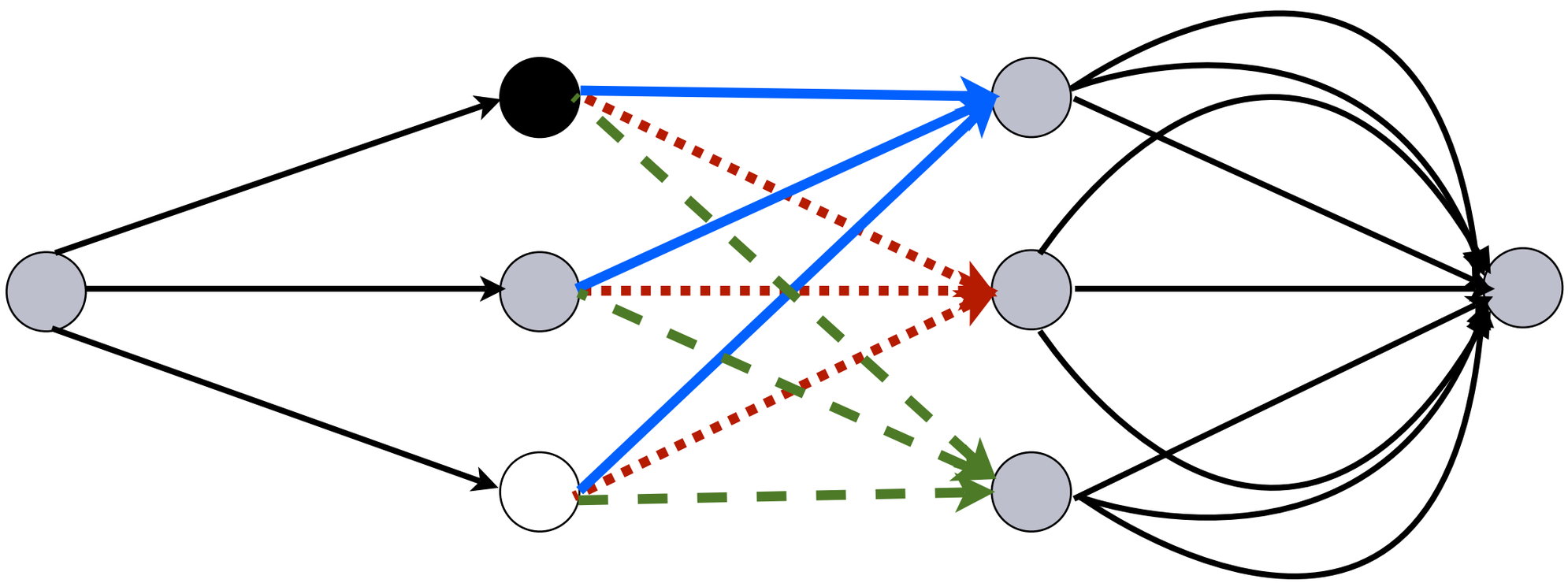
Non-local features



Non-local features



Non-local features



Non-local features

minimal logic

$$\frac{[i] \quad R(f_{i+1} \ldots f_{i'} / e_j \ldots e_{j'})}{[i']}$$

Non-local features

minimal logic

$$\frac{[i] \quad R(f_{i+1}...f_{i'}/e_j...e_{j'})}{[i']}$$

complete logic

$$\frac{[i, e_{j-n+1}, ..., e_{j-1}] \quad R(f_{i+1}...f_{i'}/e_j...e_{j'})R(e_{j-n+1}, ..., e_j)...R(e_{j'-n+1}...e_{j'})}{[i', e_{j'-n+2}...e_{j'}]}$$

Non-local features

minimal logic

$$\frac{[i] \quad R(f_{i+1} \ldots f_{i'} / e_j \ldots e_{j'})}{[i']}$$

complete logic

$$\frac{[i, e_{j-n+1}, \ldots, e_{j-1}] \quad R(f_{i+1} \ldots f_{i'} / e_j \ldots e_{j'}) R(e_{j-n+1}, \ldots, e_j) \ldots R(e_{j'-n+1} \ldots e_{j'})}{[i', e_{j'-n+2} \ldots e_{j'}]}$$

Deductive logics provide useful tools to manipulate search algorithms

PRODUCT (Cohen et al. ICLP 2009)

Fold-Unfold (Eisner & Blatz 2006; Johnson 2007)

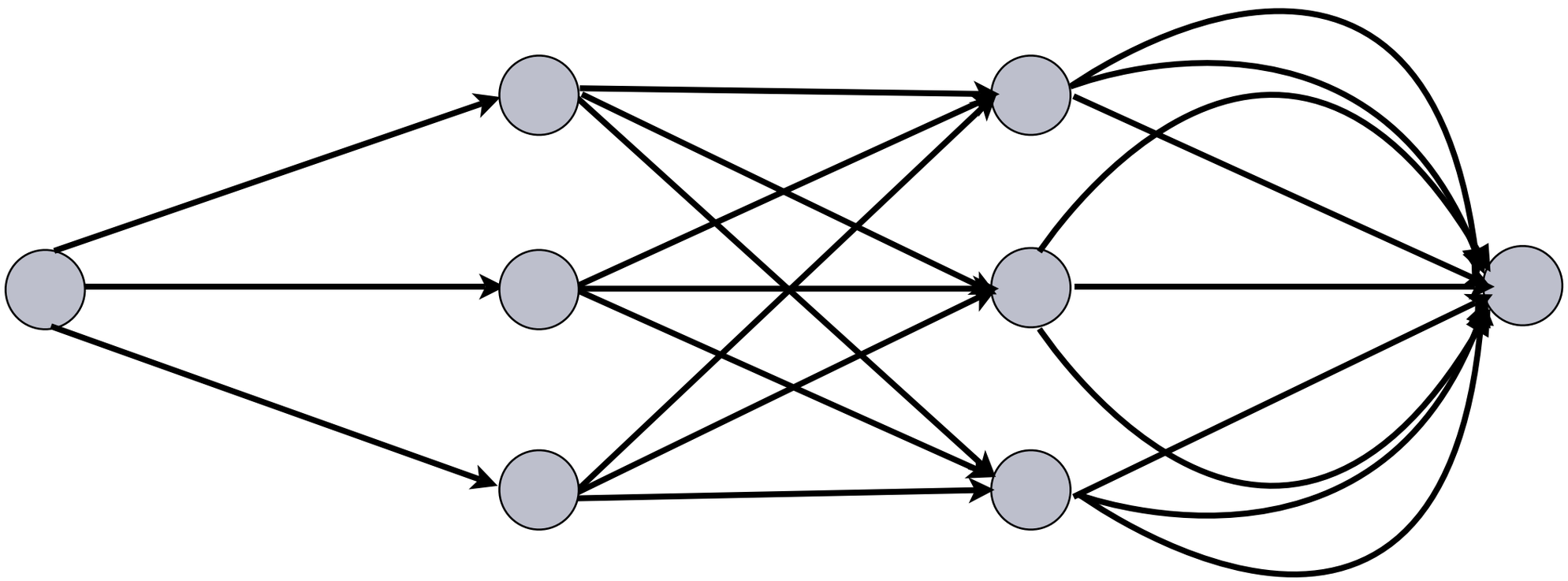
Result

- Given:
  - A complete logic
  - A semiring
- Get: a complete algorithm

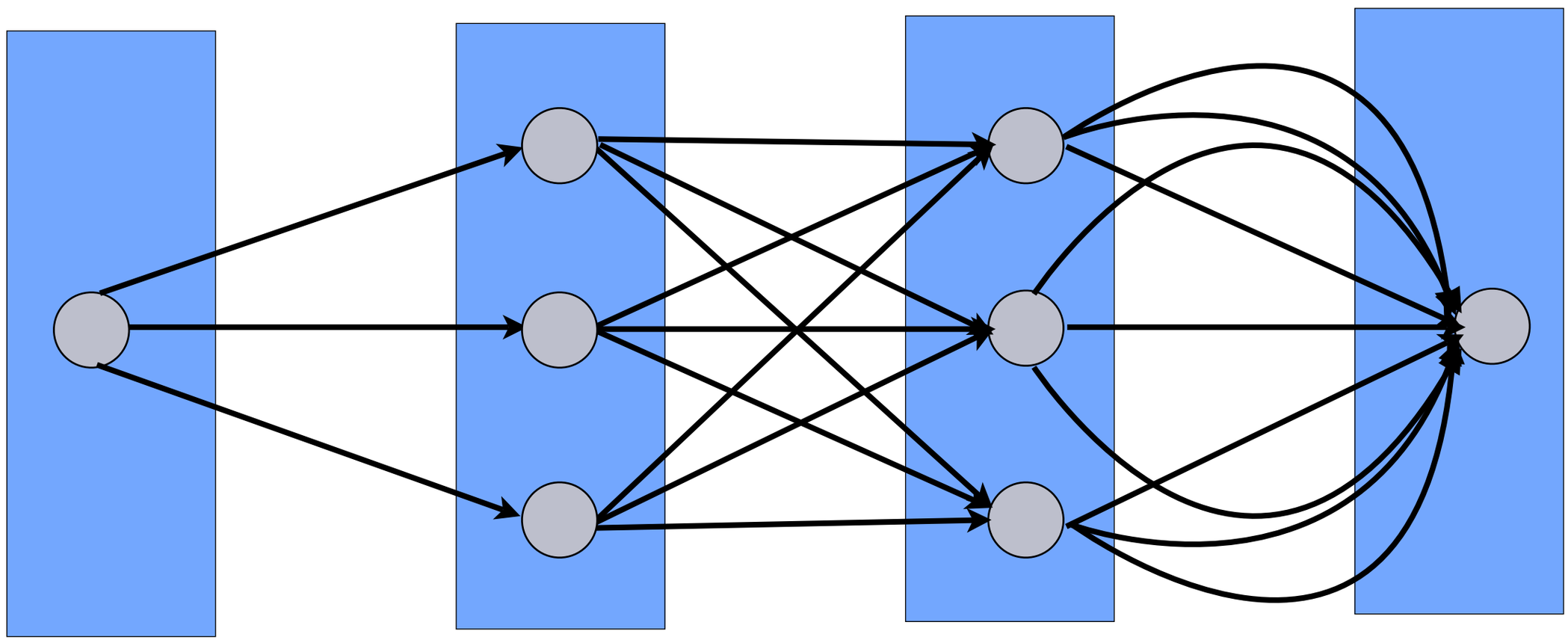
Result

- Given:
  - A complete logic
  - A semiring
- Get: a complete algorithm
- Problem: how to deal with approximate search?

Search



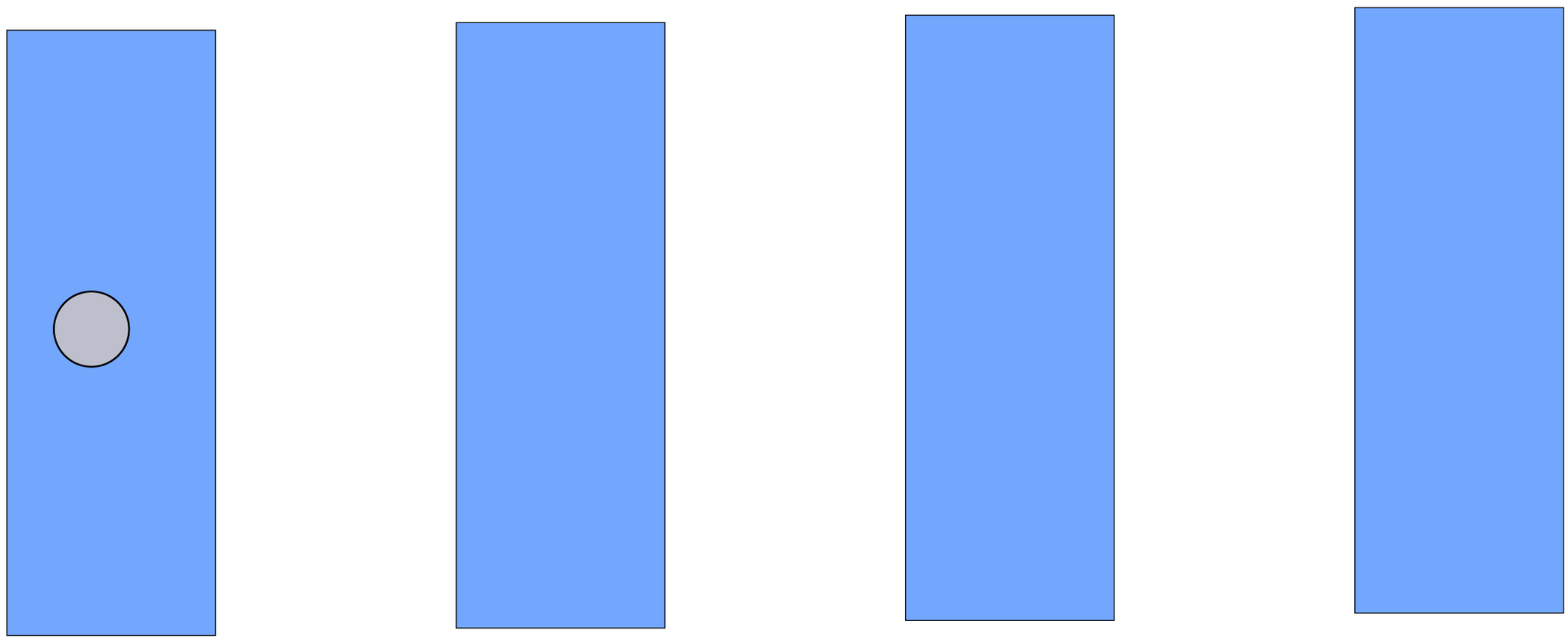
Search



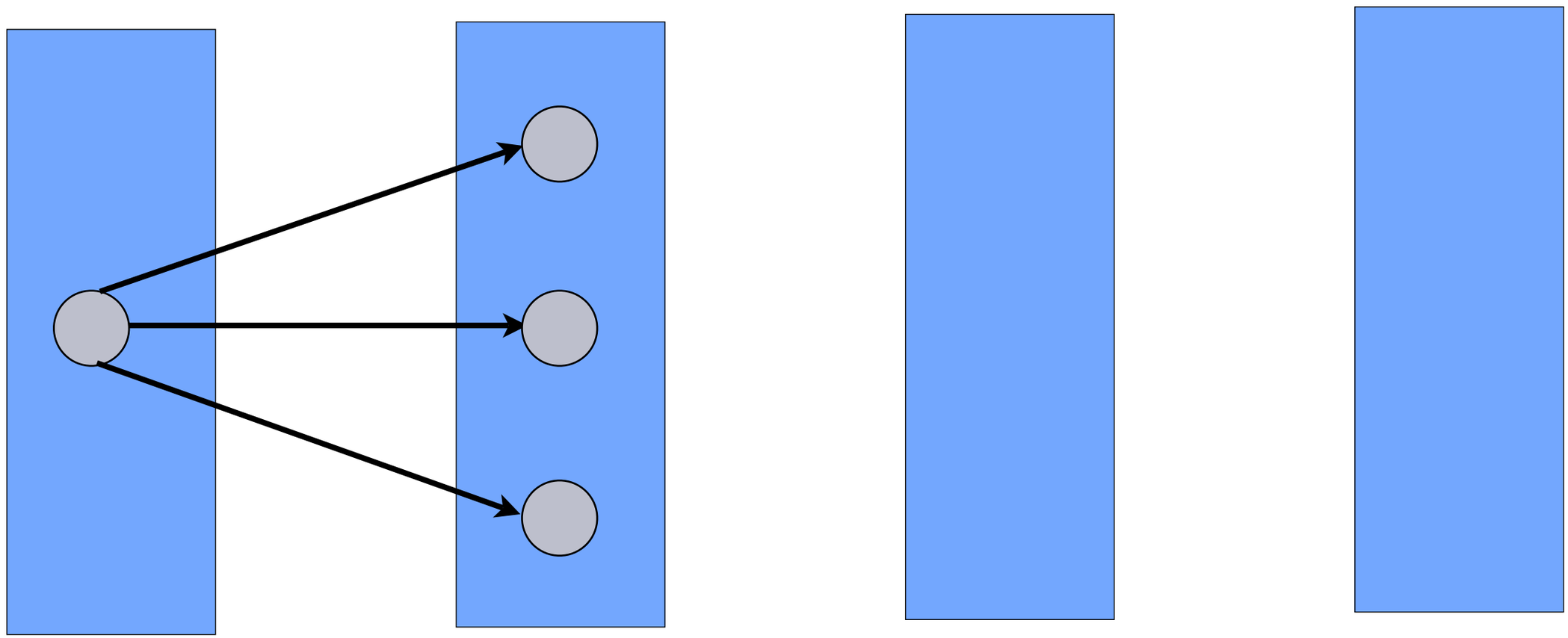
stack decoding

Koehn 2004

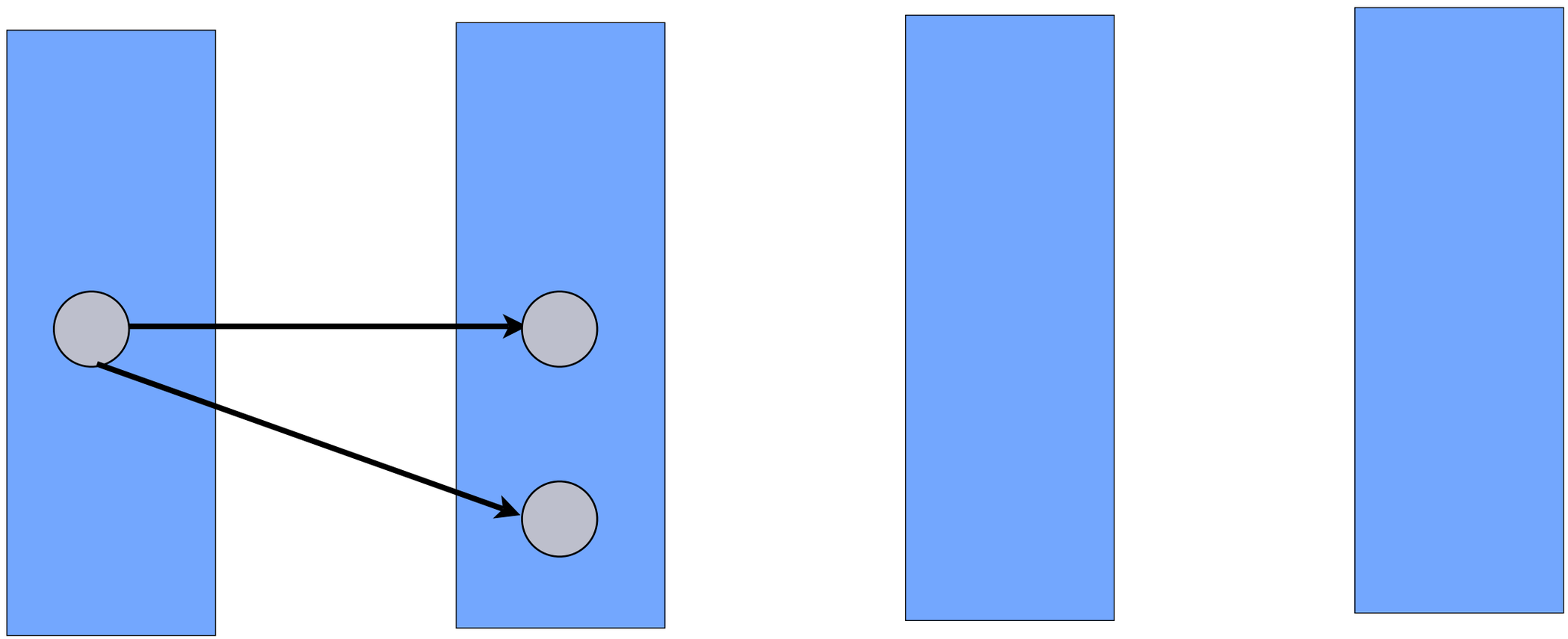
Search



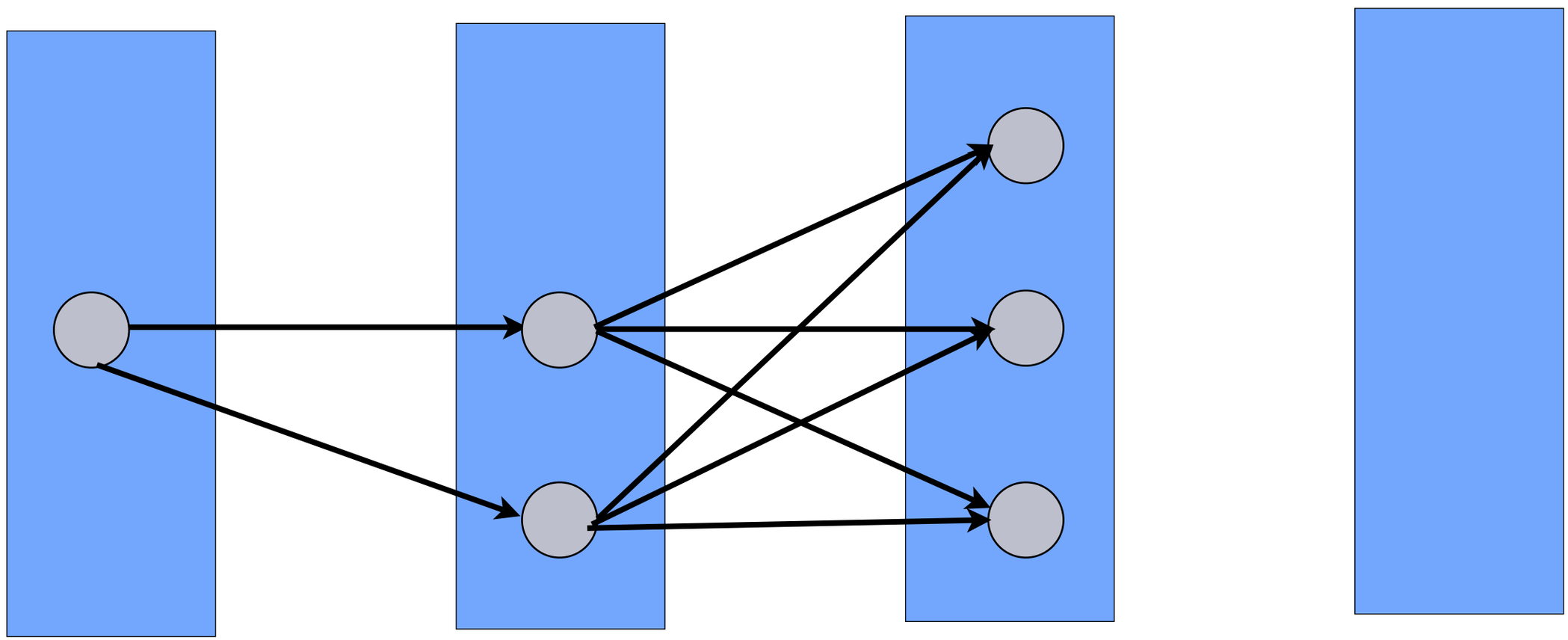
Search



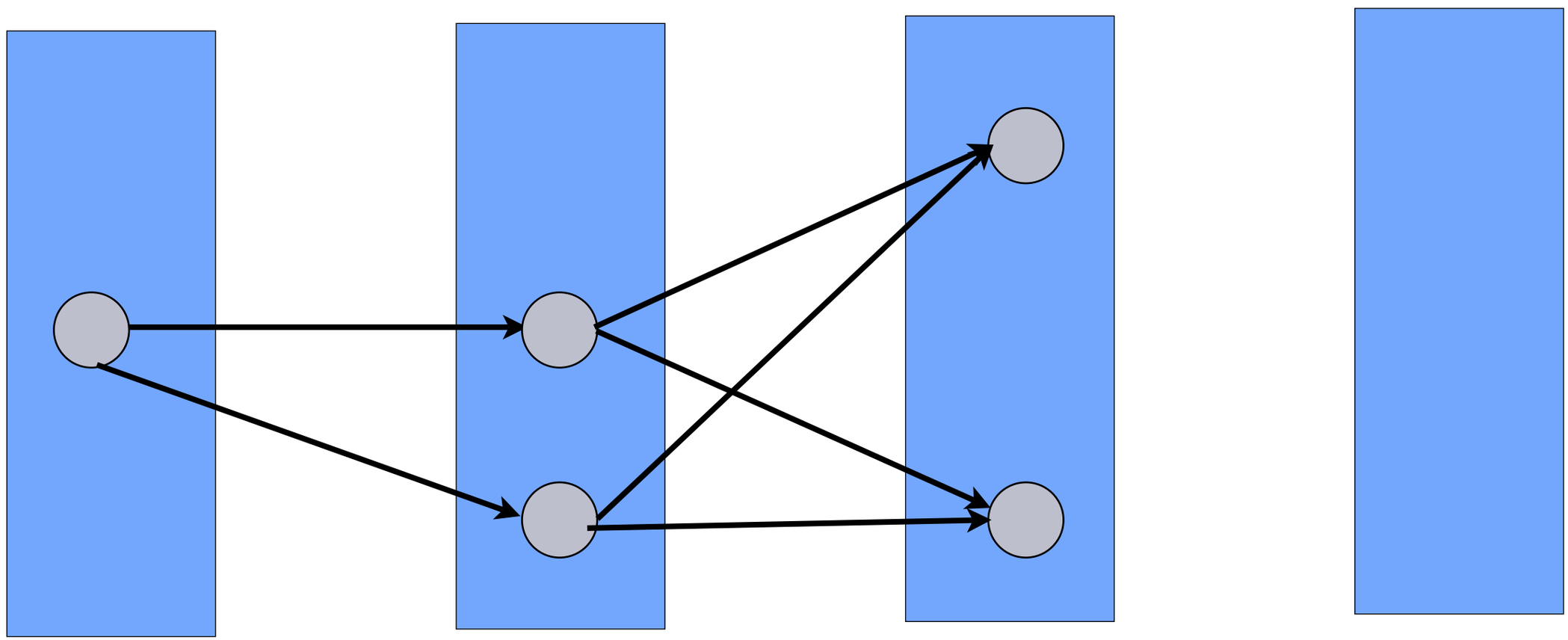
Search



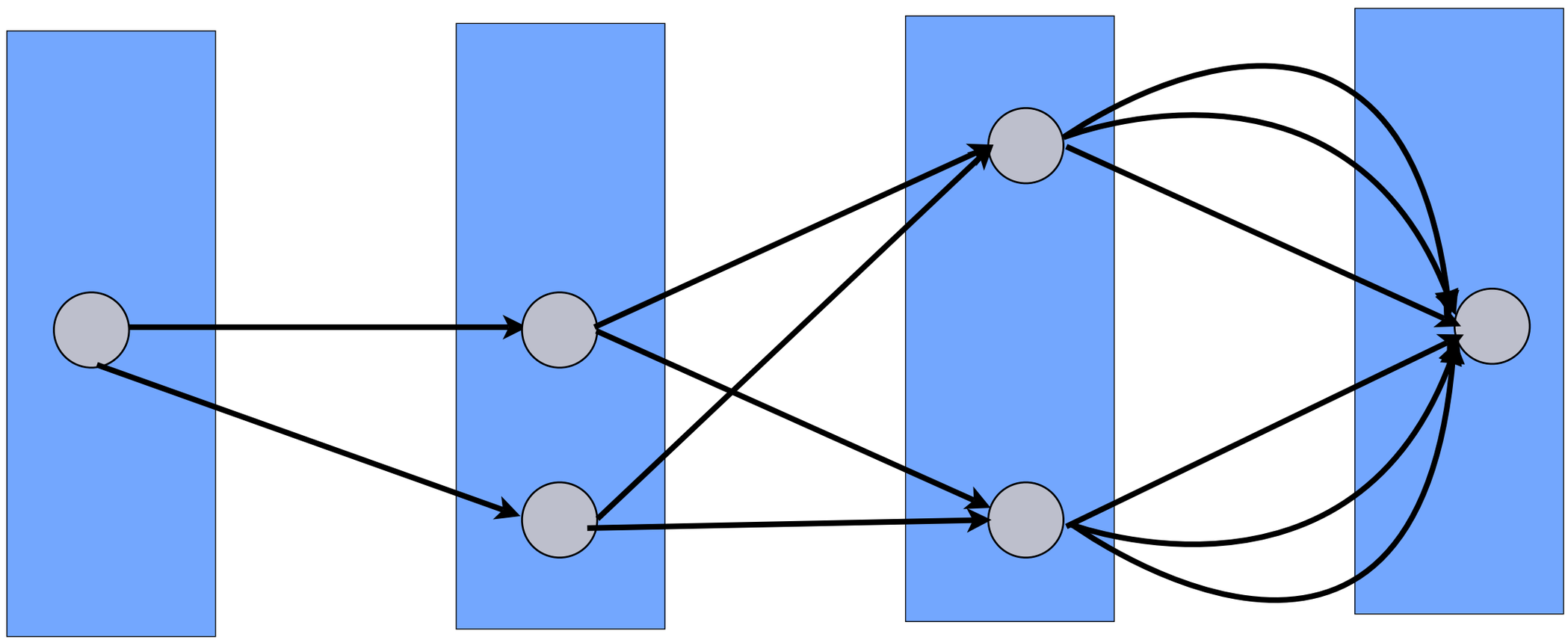
Search



Search



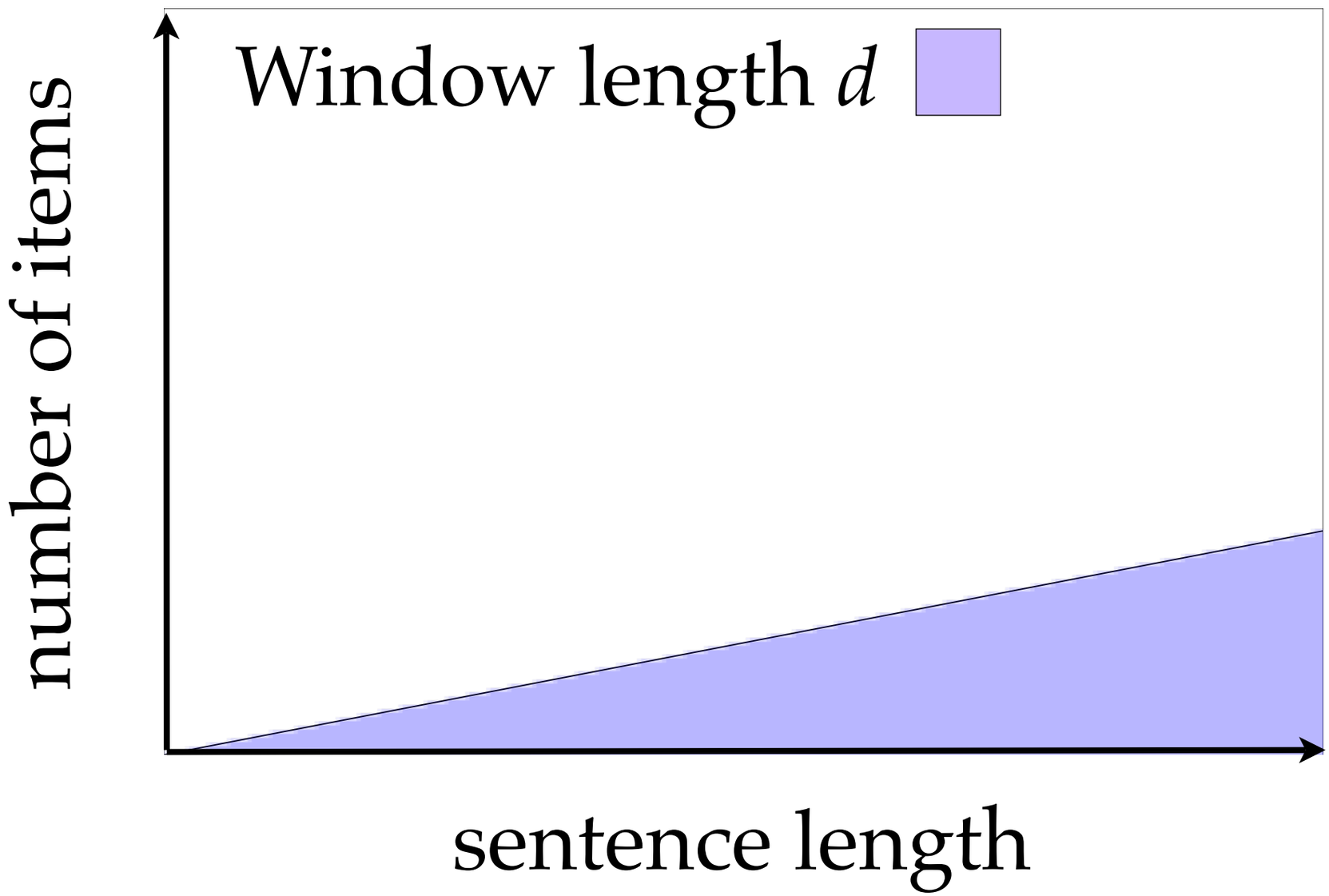
Search



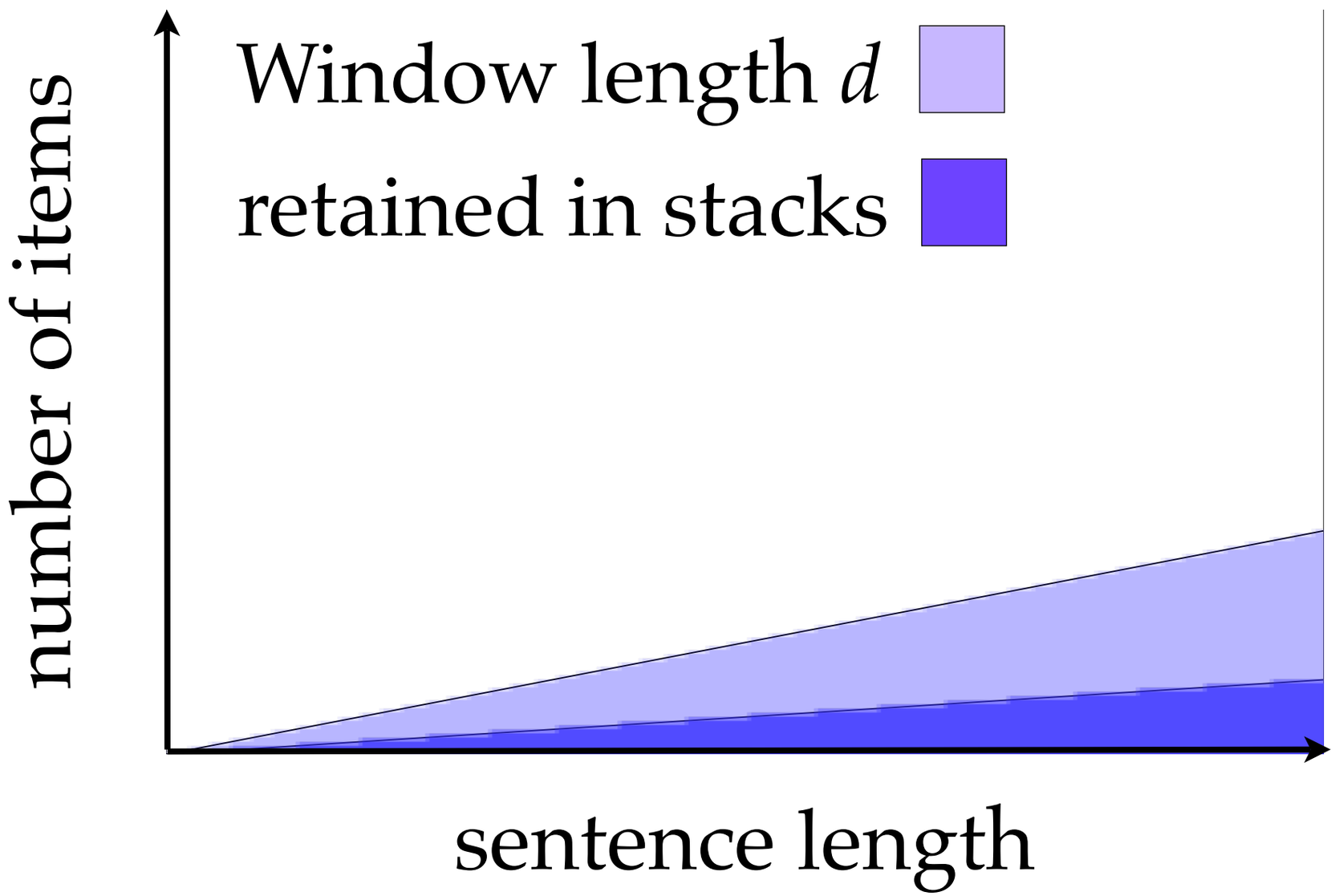
Result

- Given:
  - A complete logic
  - A semiring
  - A stack predicate
  - Pruning parameters
- Get: a complete algorithm

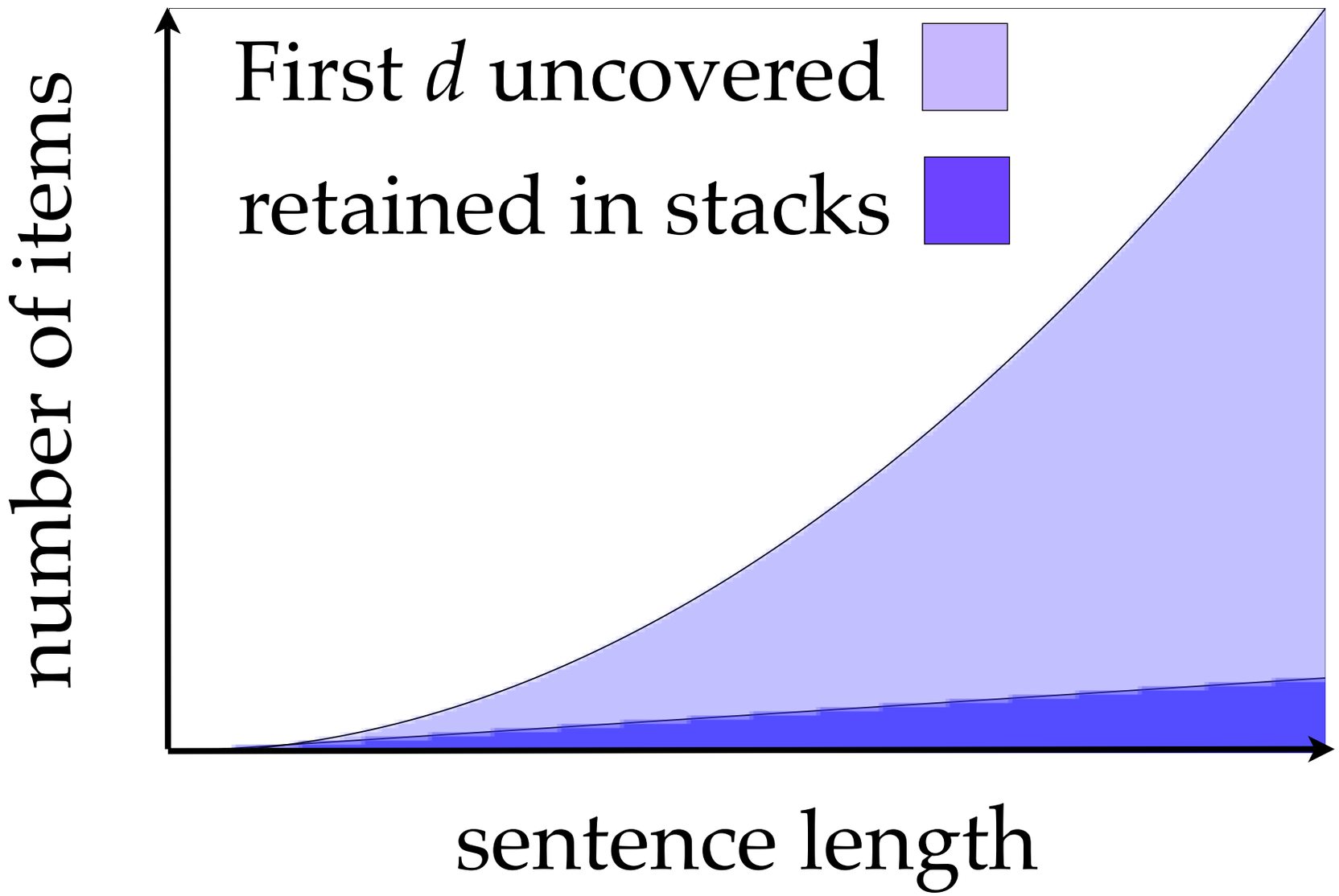
Stack Pruning Effects



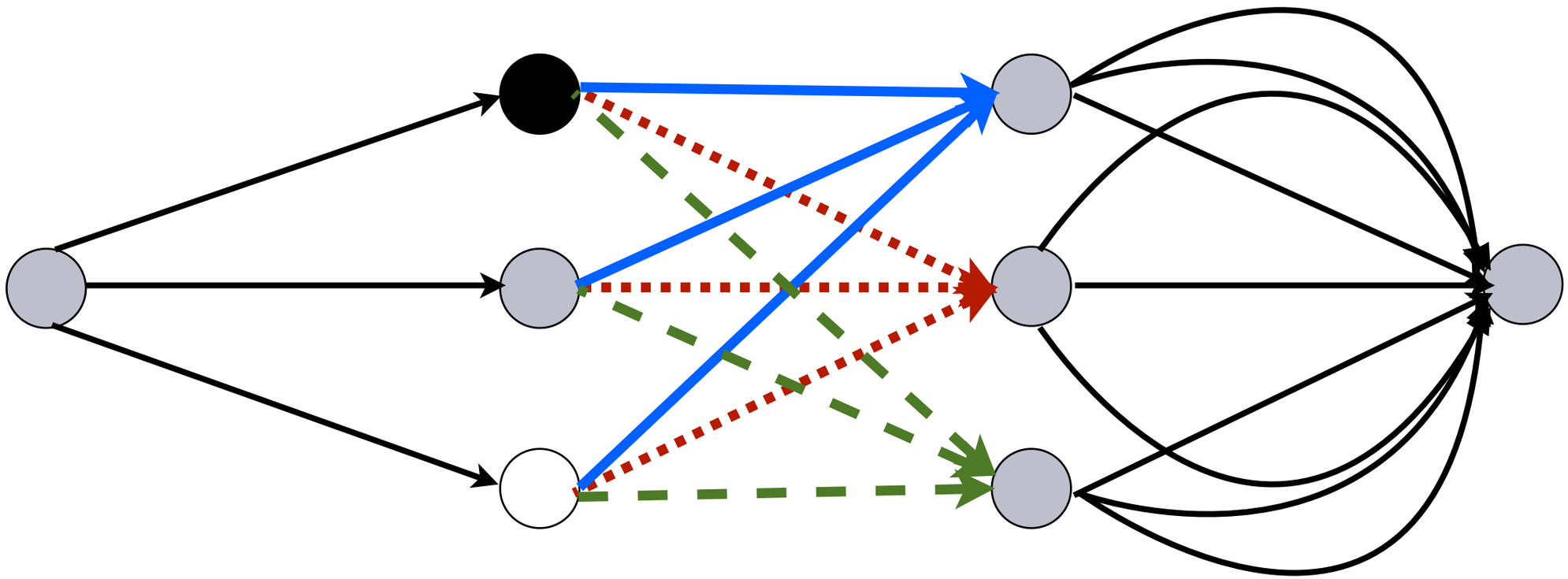
Stack Pruning Effects

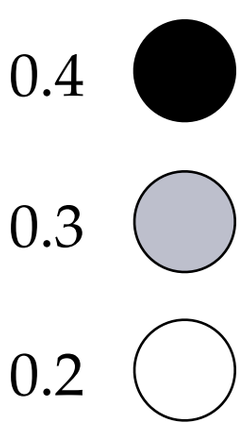


Stack Pruning Effects

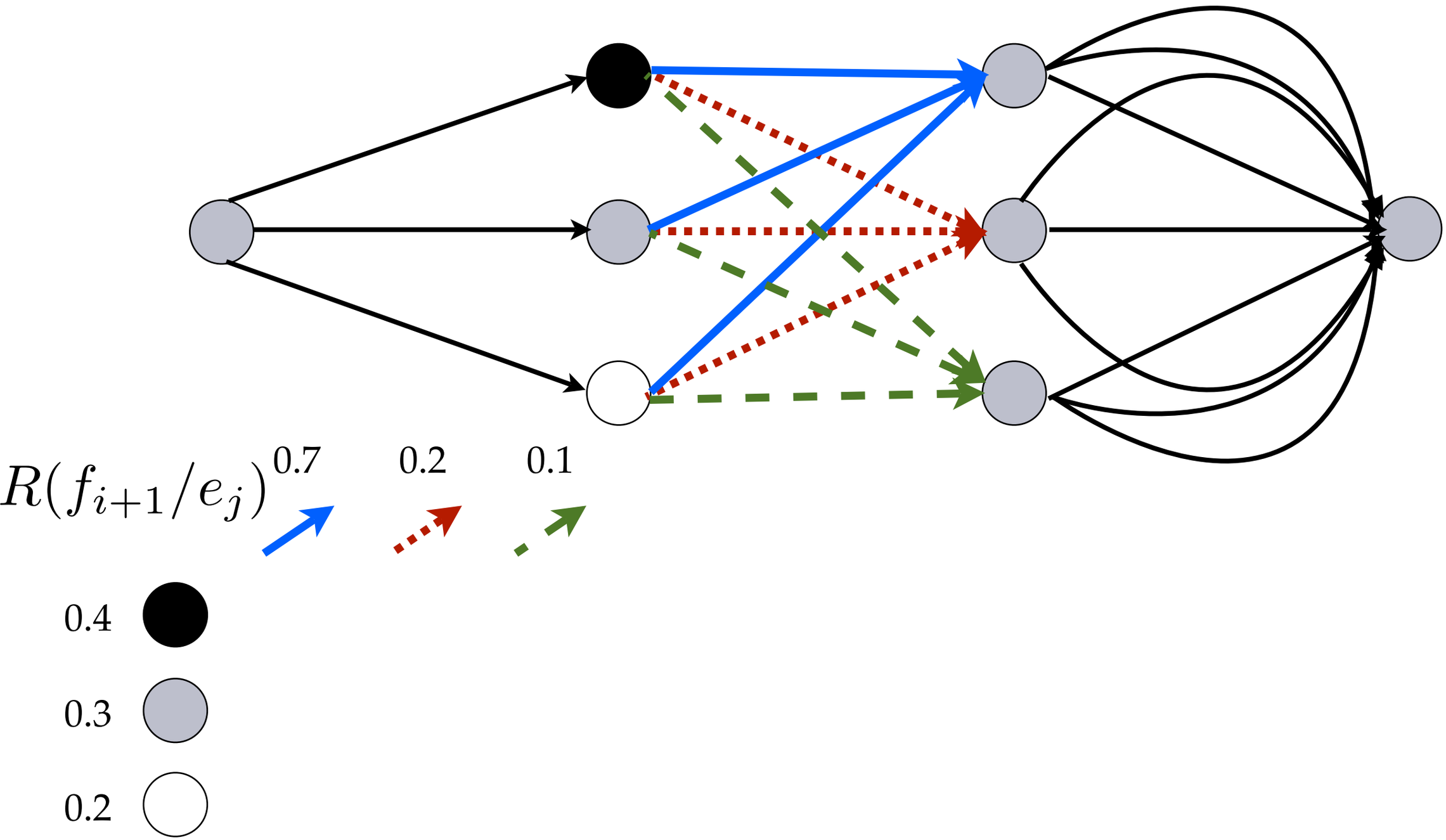


Search

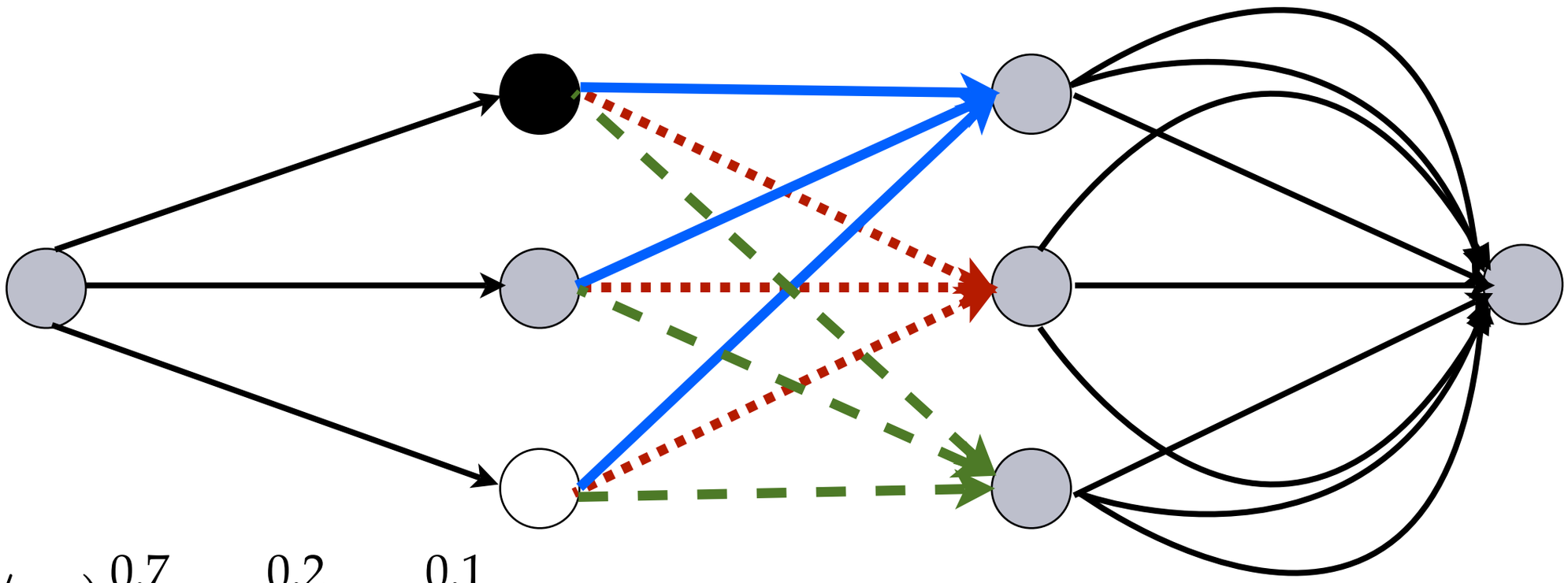




Search



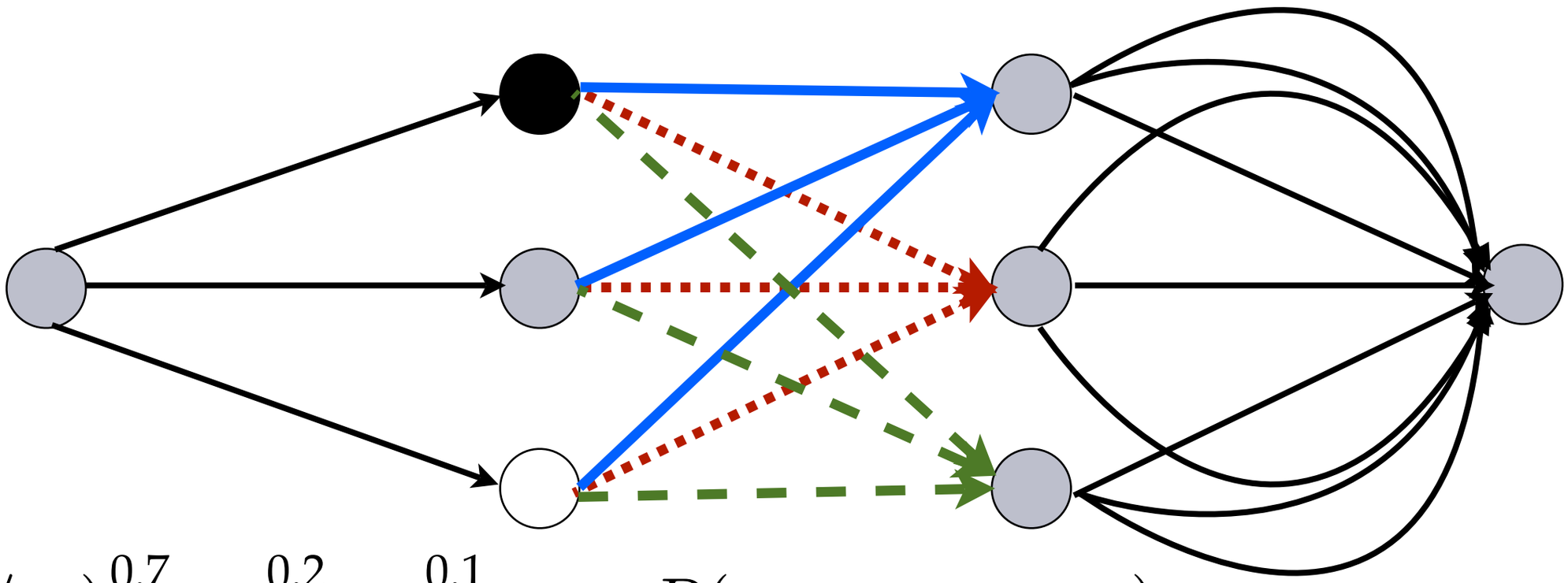
Search



$R(f_{i+1}/e_j)$

	0.7	0.2	0.1
0.4	0.28	0.08	0.04
0.3	0.21	0.06	0.03
0.2	0.14	0.04	0.02

Search



$R(f_{i+1}/e_j)$

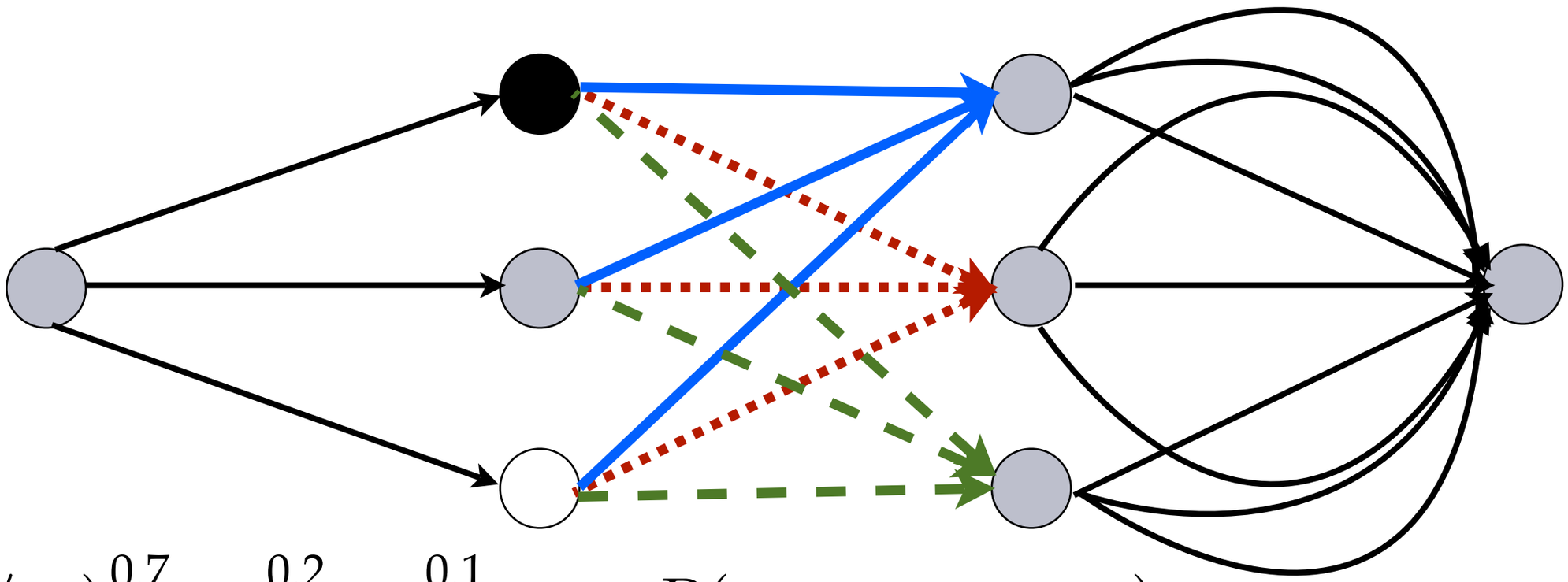
	0.7	0.2	0.1
0.4	0.28	0.08	0.04
0.3	0.21	0.06	0.03
0.2	0.14	0.04	0.02

x

$R(e_{j-n+1}, \ldots, e_j)$

0.5	0.4	0.5
0.9	0.3	0.6
0.5	0.3	0.4

Search



$R(f_{i+1}/e_j)$ 0.7 0.2 0.1

$R(e_{j-n+1}, \ldots, e_j)$

0.4	0.28	0.08	0.04
0.3	0.21	0.06	0.03
0.2	0.14	0.04	0.02

x

0.5	0.4	0.5
0.9	0.3	0.6
0.5	0.3	0.4

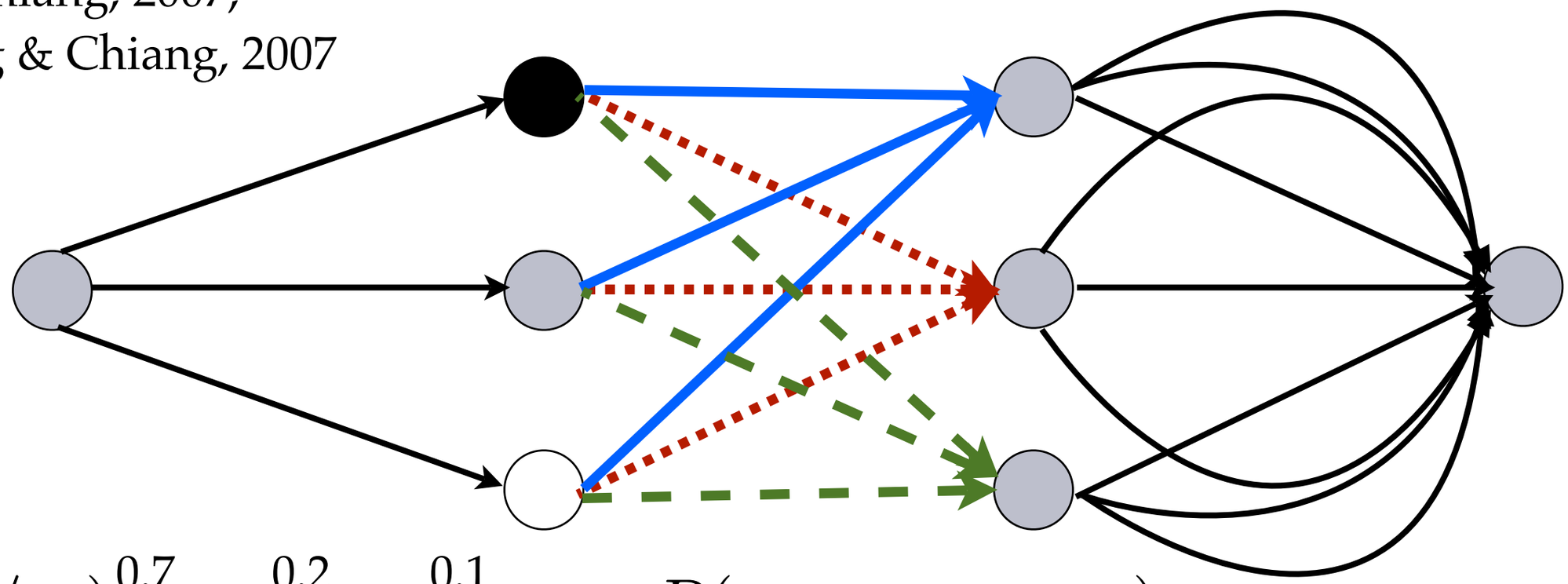
=

0.14	0.03	0.02
0.18	0.02	0.02
0.7	0.01	0.01

Search

Cube Pruning

Chiang, 2007;
Huang & Chiang, 2007



$R(f_{i+1}/e_j)$ 0.7 0.2 0.1

$R(e_{j-n+1}, \ldots, e_j)$

0.4	0.28	0.08	0.04
0.3	0.21	0.06	0.03
0.2	0.14	0.04	0.02

x

0.5	0.4	0.5
0.9	0.3	0.6
0.5	0.3	0.4

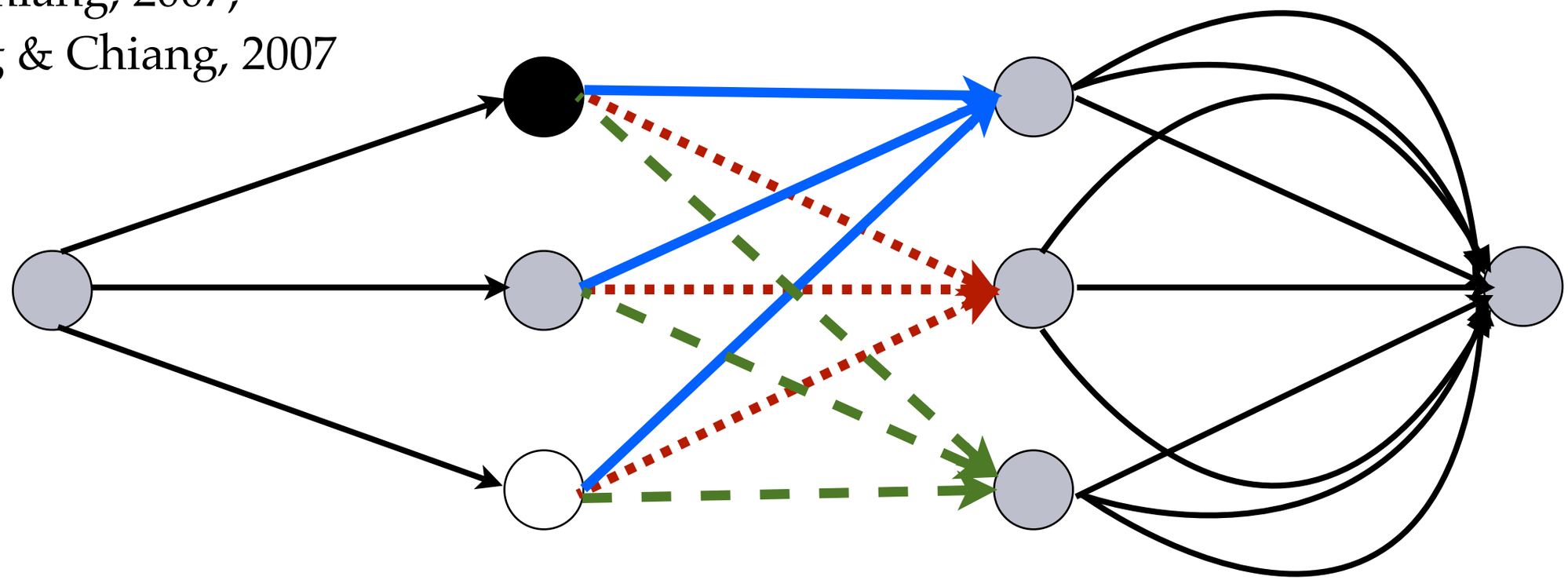
=

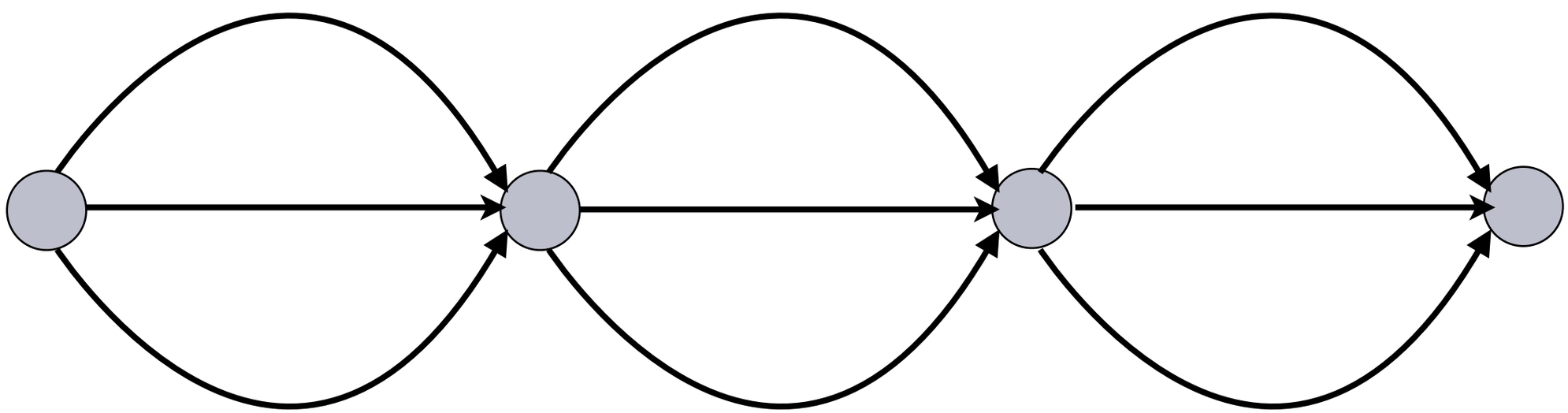
0.14	0.03	0.02
0.18	0.02	0.02
0.7	0.01	0.01

Search

Cube Pruning

Chiang, 2007;
Huang & Chiang, 2007

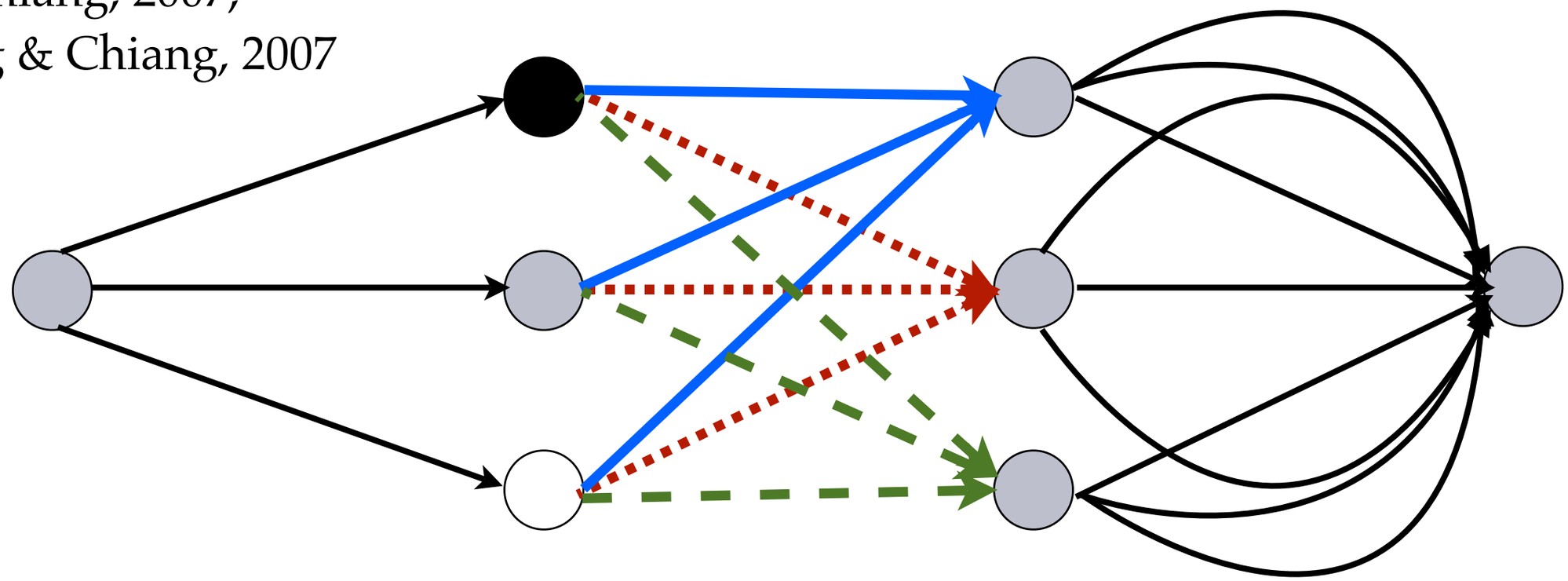


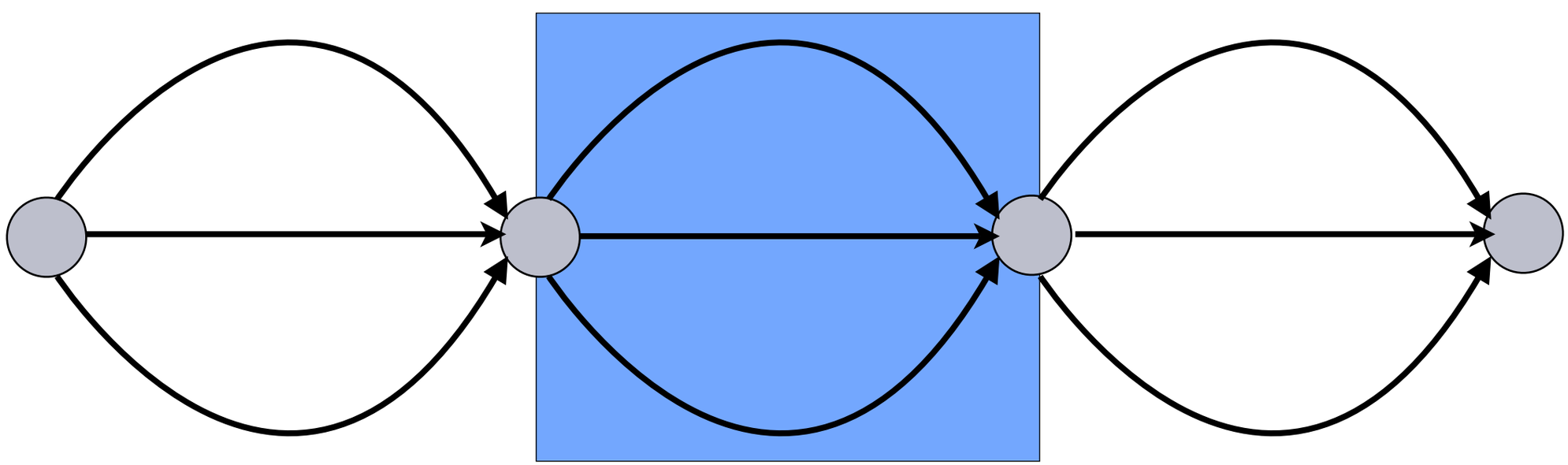


Search

Cube Pruning

Chiang, 2007;
Huang & Chiang, 2007





Result

- Given:
  - A minimal logic
  - A complete logic
  - A semiring
  - Pruning parameters
- Get: a complete algorithm

Conclusion

- Translation can easily be cast in the deductive framework.
  - Analysis reveals inconsistencies.
- Modify models with logic transforms.
  - Easy to describe non-local features.
- Search strategies can be incorporated into deductive systems.

Future Work

- Other approximate search strategies.
- Modular implementation.
- Exploration of novel models.

Thanks

Phil Blunsom
Chris Callison-Burch
Chris Dyer
Hieu Hoang
Martin Kay
Philipp Koehn
Josh Schroeder
Lane Schwartz