CS 583: Probabilistic Graphical Models

Topic: Variable Elimination
Chapter: 9

Mustafa Bilgic

🔗 http://www.cs.iit.edu/~mbilgic

🐦 https://twitter.com/bilgicm

TASK

- Given a graphical model over $\mathcal{X}$ (structure and parameters)

- Compute $P(\mathbf{Y} \mid \mathbf{e})$, where $\mathbf{Y} \subseteq \mathcal{X}$ and $\mathbf{E} \subseteq \mathcal{X}$

- There are a number of approaches

  - Exact inference

    - Variable elimination

    - Belief propagation

  - Approximate inference

    - Sampling

- Today, we'll cover variable elimination

VARIABLE ELIMINATION

- $P(\boldsymbol{Y} \mid \boldsymbol{e}) = P(\boldsymbol{Y}, \boldsymbol{e}) / P(\boldsymbol{e})$

- $\boldsymbol{W} = \mathcal{X} - \boldsymbol{Y} - \boldsymbol{E}$

- $P(\boldsymbol{y}, \boldsymbol{e}) = \Sigma_{\boldsymbol{w}} P(\boldsymbol{y}, \boldsymbol{e}, \boldsymbol{w})$

- $P(\boldsymbol{e}) = \Sigma_{\boldsymbol{y},\boldsymbol{w}} P(\boldsymbol{y}, \boldsymbol{e}, \boldsymbol{w})$

- Or, better yet: $P(\boldsymbol{e}) = \Sigma_{\boldsymbol{y}} P(\boldsymbol{y}, \boldsymbol{e})$

$P(Y, E) = \Sigma_W P(Y, E, W)$

- P(**Y, E, W**) can be represented as
  - $\prod P(X_i \mid Pa(X_i))$
  - $1/Z \prod \phi(\boldsymbol{D}_i)$

- The problem with $P(\boldsymbol{y}, \boldsymbol{e}) = \Sigma_w P(\boldsymbol{y}, \boldsymbol{e}, \boldsymbol{w})$ is that the joint representation is exponential
  - The very first problem we were trying to avoid

4

COMPLEXITY

- Unfortunately, exact inference is $\mathcal{NP}$-hard in worst case

  - Proof: pages 288 and 289. Reduction from 3-SAT

- Approximate inference is also $\mathcal{NP}$-hard in worst case

  - Proof: pages 291 and 292.

- Good news:

  - In general, we care about the cases we encounter in practice, not the worst case scenario

KEY IDEA

- Summation can be moved inside

- $\Sigma_x \Sigma_y \, x*y = \Sigma_x x*(\Sigma_y \, y)$

- If *x* has *n* and *y* has *m* possible values, how many operations are needed, if we use

  - $\Sigma_x \Sigma_y \, x*y$?

  - $\Sigma_x x*(\Sigma_y \, y)$?

OUTLINE

- First, focus on Bayesian networks
  - Simple linear chains
  - More complex structures

- Two cases
  - Marginal queries: $\boldsymbol{E} = \varnothing$
  - Conditional queries: $\boldsymbol{E} \neq \varnothing$

VARIABLE ELIMINATION

- $\mathcal{X}$: all variables, $\boldsymbol{Y}$: query variables, $\boldsymbol{E}$: evidence variables, $\boldsymbol{W} = \mathcal{X} - \boldsymbol{Y} - \boldsymbol{E}$: remaining variables

1. Write down the joint for $P(\mathcal{X})$

2. Set $X_i \in \boldsymbol{E}$ to their values

3. Pick an order for $X_j \in \boldsymbol{W}$

4. Sum out each X_j from the joint

   a) Multiply the factors $\phi(X_j, Z_1), \ldots, \phi(X_j, Z_k)$ to create $\psi(X_j, Z_1, \ldots, Z_k)$

   b) Sum out X_j from $\psi(X_j, Z_1, \ldots, Z_k)$ to create $\tau(Z_1, \ldots, Z_k)$

5. What remains is $\tau(\boldsymbol{Y}, \boldsymbol{e})$. Normalize it to get $P(\boldsymbol{Y} \mid \boldsymbol{e})$.

8

EXAMPLES – LINEAR CHAIN BNS

- $A \to B$

  - P(B) =?

  - P(A) =?

- $X_1 \to X_2 \to \dots \to X_n$

  - P(X_i) where $1 \leq i \leq n$

- How many operations are needed if we compute the full joint distribution vs. if we use variable elimination?
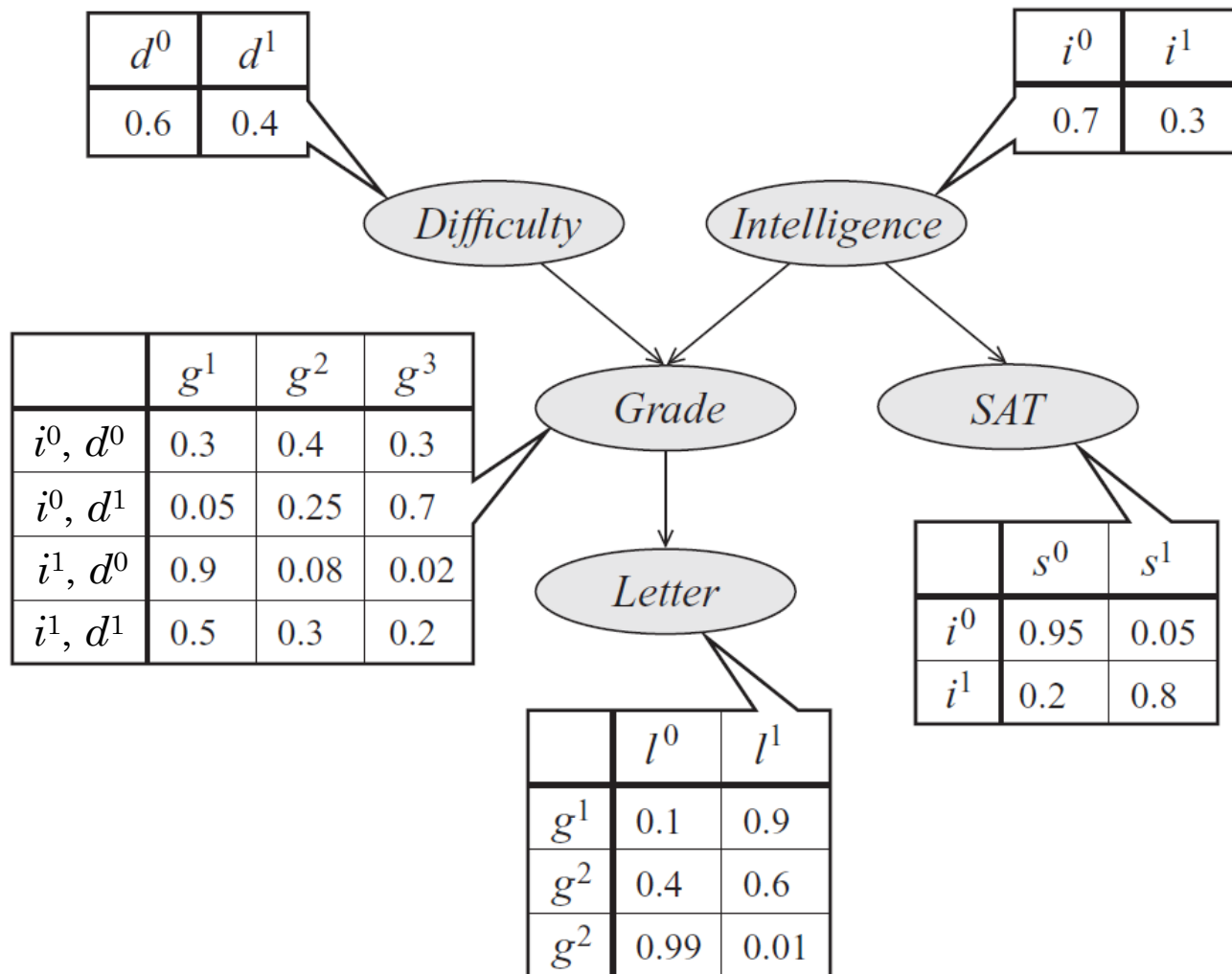
STUDENT NETWORK EXAMPLE

d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

Difficulty   **Intelligence**

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

Grade   **SAT**

Letter

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

P(D)
P(I)
P(S)
P(G)
P(L)

10
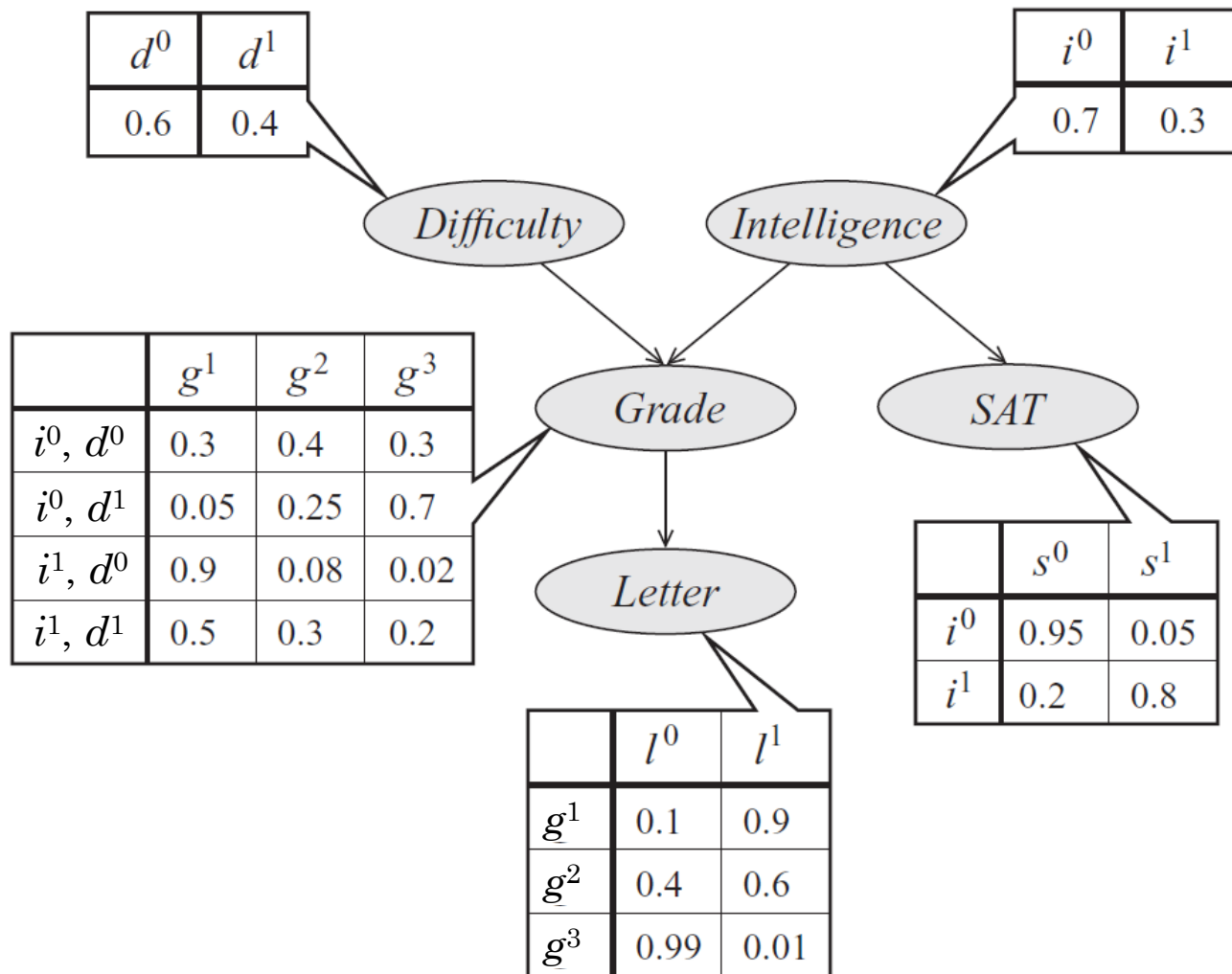
P($Y | E$) EXAMPLES

- $A \rightarrow B$
  - P($B | A$=t)
  - P(A=t $| B$=t)
- $X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_n$
  - P($X_i | X_j=x_j$)
  - P($X_i | \boldsymbol{X_j}=\boldsymbol{x_j}$)

STUDENT NETWORK EXAMPLE

d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

Difficulty *Intelligence*

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

Grade *SAT*

Letter

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

$P(S \mid I=i^0)$
$P(D \mid I=i^0)$
$P(I \mid G=g^1)$
$P(D \mid G=g^1)$
$P(D \mid I=i^0, G=g^1)$

12

P(L) – ORDER: I, S, D, G

Variable	All Factors	Participates	New Factor After *	# *s	New Factor After +	# +s	# Ops
I	P(I), P(D), P(S|I), P(G|D,I), P(L|G)	P(I), P(S|I), P(G|D, I)	$\psi_1(G,D,S,I)$	2*3*2*2*2= 48	$\tau_1(G,D,S)$	1*3*2*2=12	60
S	P(D), P(L|G), $\tau_1(G,D,S)$	$\tau_1(G,D,S)$	$\psi_2(G,D,S)$	0	$\tau_2(G,D)$	1*3*2=6	6
D	P(D), P(L|G), $\tau_2(G,D)$	P(D), $\tau_2(G,D)$	$\psi_3(G,D)$	1*3*2=6	$\tau_3(G)$	1*3	9
G	P(L|G), $\tau_3(G)$	P(L|G), $\tau_3(G)$	$\psi_4(L,G)$	1*2*3=6	$\tau_4(L)$	2*2=4	10
Normalize	$\tau_4(L)$					1	3 (2 divs)
Total							88

13

P(L) – ORDER: S, I, D, G

Variable	All Factors	Participates	New Factor After *	# *s	New Factor After +	# +s	# Ops
S	P(I), P(D), P(S|I), P(G|D,I), P(L|G)	P(S|I)	$\psi_1(I,S)$	0	$\tau_1(I)$	1*2=2	2
I	P(I), P(D), P(G|D,I), P(L|G) $\tau_1(I)$	P(I), P(G|D,I), $\tau_1(I)$	$\psi_2(G,D,I)$	2*3*2*2=24	$\tau_2(G,D)$	1*3*2=6	30
D	P(D), P(L|G), $\tau_2(G,D)$	P(D), $\tau_2(G,D)$	$\psi_3(G,D)$	1*3*2=6	$\tau_3(G)$	1*3	9
G	P(L|G), $\tau_3(G)$	P(L|G), $\tau_3(G)$	$\psi_4(L,G)$	1*2*3=6	$\tau_4(L)$	2*2=4	10
Normalize	$\tau_4(L)$					1	3 (2 divs)
Total							54

Markov network example

A	B	$\phi(A,B)$
T	T	5
T	F	1
F	T	1
F	F	3

B	C	$\phi(B,C)$
T	T	1
T	F	2
F	T	6
F	F	1

A	B	C	$\phi(A,B)*\phi(B,C)$	P(A,B,C)
T	T	T	5	0.11
T	T	F	10	0.22
T	F	T	6	0.13
T	F	F	1	0.02
F	T	T	1	0.02
F	T	F	2	0.04
F	F	T	18	0.39
F	F	F	3	0.07
Z			46	1.00

A	B	P(A,B)
T	T	0.33
T	F	0.15
F	T	0.07
F	F	0.46

B	C	P(B,C)
T	T	0.13
T	F	0.26
F	T	0.52
F	F	0.09

A	P(A)
T	0.48
F	0.52

B	P(B)
T	0.39
F	0.61

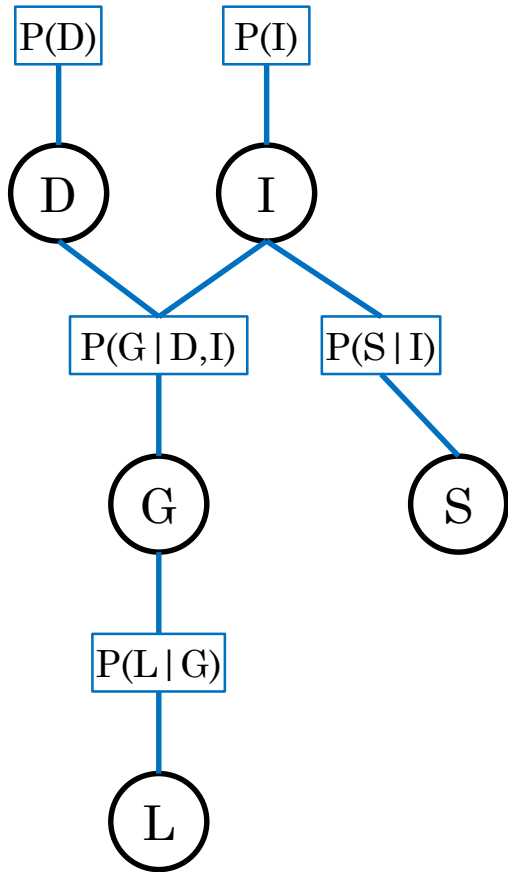| C	P(C)
T	0.65
F	0.35

15

ELIMINATION AS GRAPH TRANSFORMATION

- Eliminating X
  - Multiply all the factors X participates in
  - Sum out X

- Graph transformation (need to be moralized first)
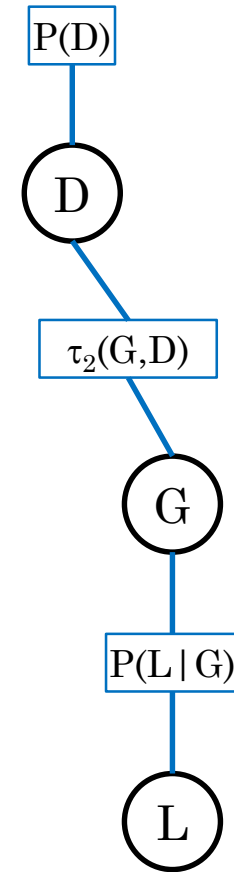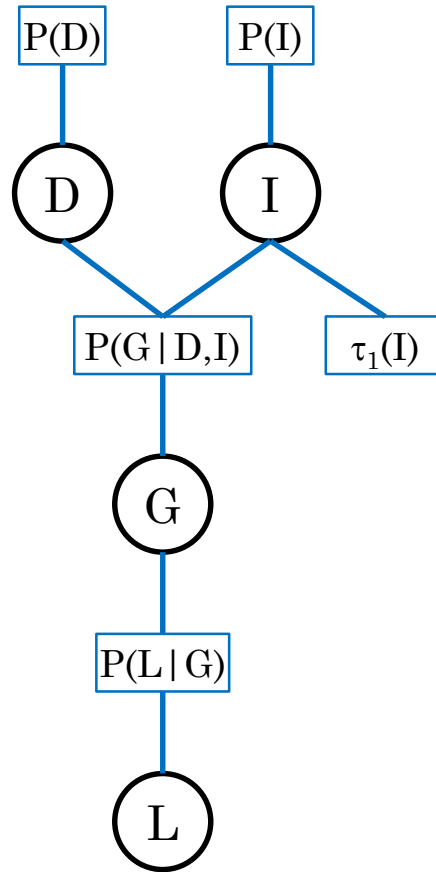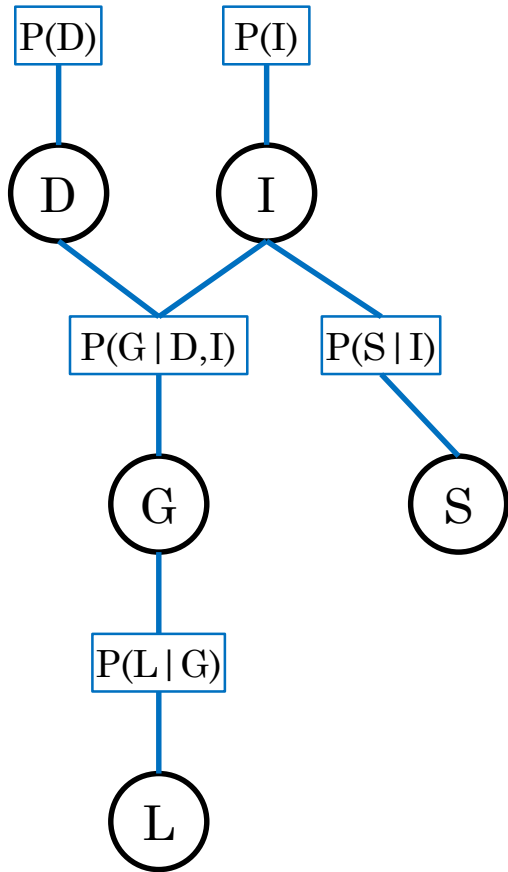 - Connect all of X's neighbors
  - Remove X

16

REPRESENTATION
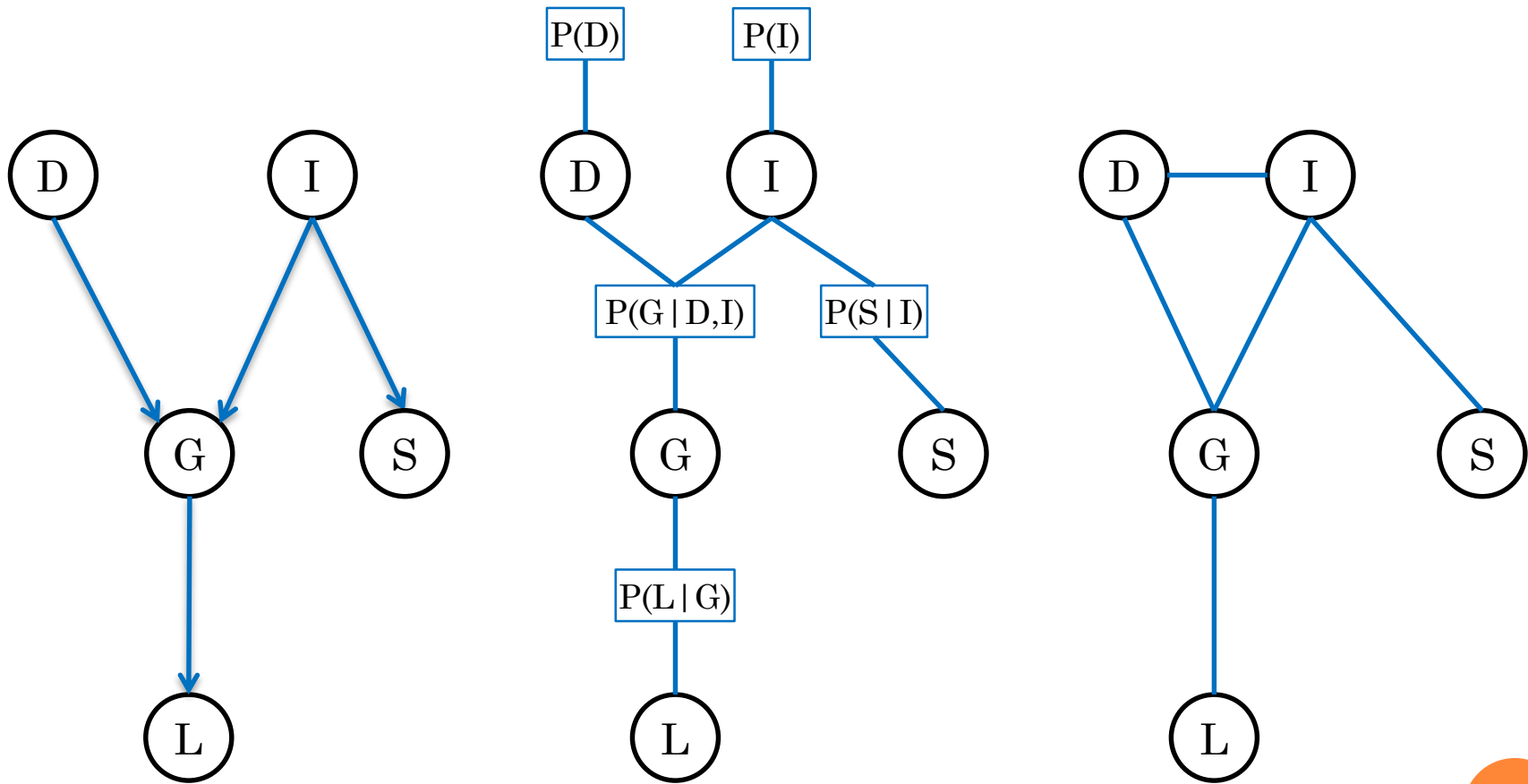
IF WE FIRST ELIMINATE *I* THEN *S*

CS 583 – Probabilistic Graphical Models – Illinois Institute of Technology
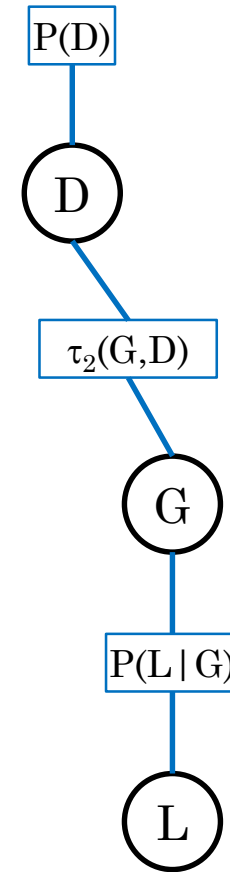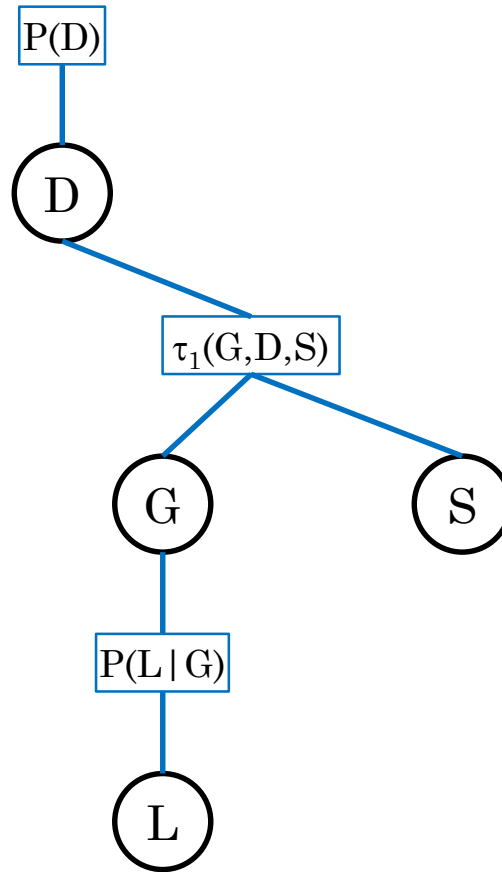
IF WE FIRST ELIMINATE *S* THEN *I*

FINDING GOOD ELIMINATION ORDERINGS

- Finding the best order is NP-hard
  - Best = optimal time and space complexity

- Heuristics
  - Min-neighbors
  - Min-fill
  - Weighted versions of min-neighbors and min-fill

IRRELEVANT NODES IN BNS

- $\mathcal{X}$: all variables, Y: query variables, E: evidence variables, $W = \mathcal{X} - Y - E$: remaining variables

- A node $X_i \in W$ is irrelevant for the query $\mathrm{P}(Y|e)$ if it can be removed from the network without effecting the value of $\mathrm{P}(Y|e)$

- Obvious:

  - If $Z \subseteq W$ is d-separated from Y given E, then Z is irrelevant

- Perhaps less obvious:

  - Let Z be ancestors of $Y \cup E$. Then $W \setminus Z$ is irrelevant