CS 760: Machine Learning
Graphical Models - II

Kirthi Kandasamy

University of Wisconsin-Madison

March 29, 2023

Outline

- **Bayesian Networks Review**
  - Definition, examples, inference, learning
- **Undirected Graphical Models**
  - Definitions, MRFs, exponential families
- **Structure learning**
  - Chow-Liu Algorithm
- **D-separation**

Outline

- **Bayesian Networks Review**
  - Definition, examples, inference, learning
- **Undirected Graphical Models**
  - Definitions, MRFs, exponential families
- **Structure learning**
  - Chow-Liu Algorithm
- **D-separation**

Bayesian Networks Example

- Consider the following 5 binary random variables:

  B = a burglary occurs at the house

  E = an earthquake occurs at the house

  A = the alarm goes off

  J = John calls to report the alarm

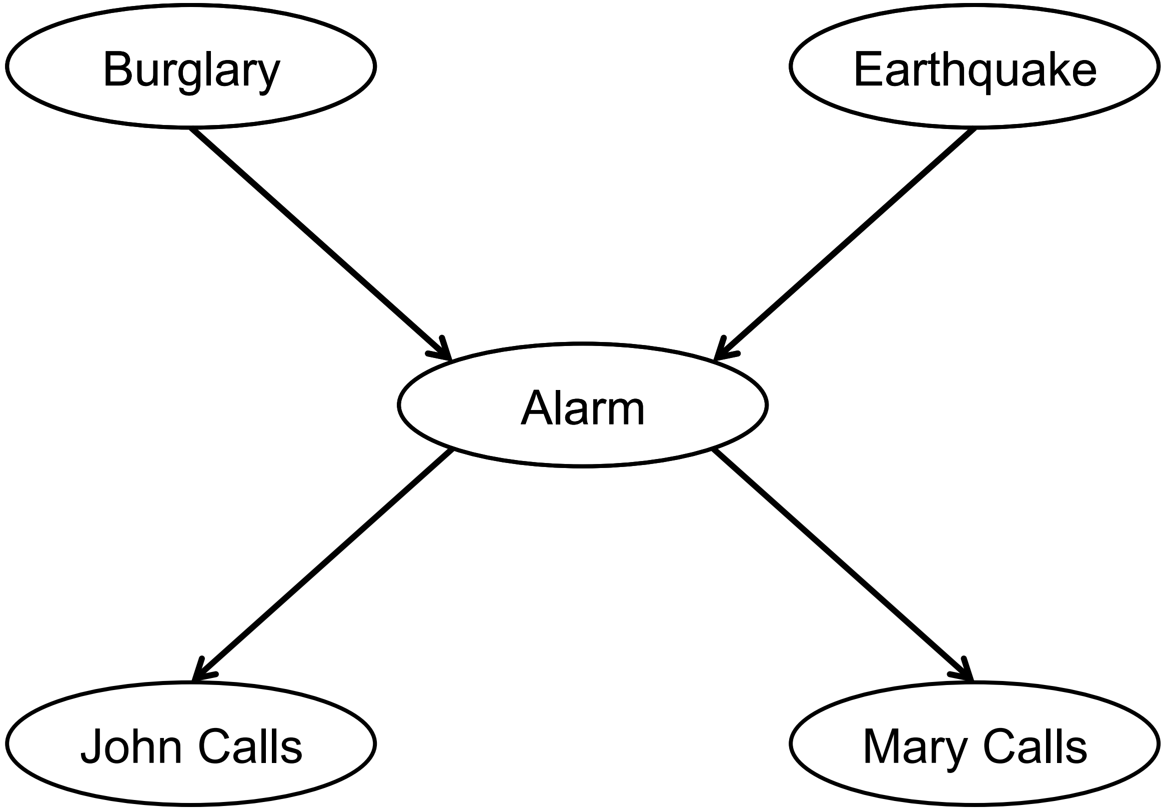
  M = Mary calls to report the alarm

- Suppose the Burglary or Earthquake can trigger Alarm, and Alarm can trigger John's call or Mary's call

- Now we want to answer queries like what is $P(B \mid M, J)$?

Bayesian Networks Example

- Set up a network that shows how random variables influence others:

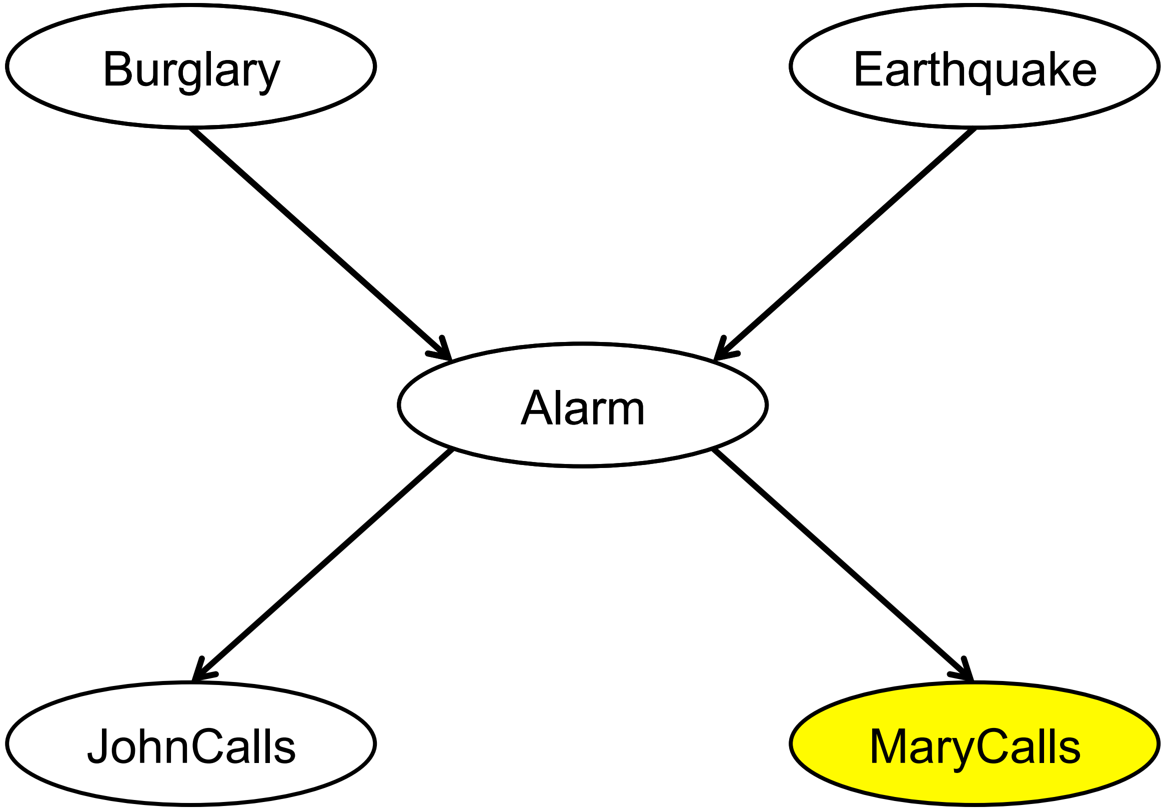


Bayesian Networks Example

- Set up a network that shows how random variables influence others:

P (B)

t	f
0.001	0.999



P (E)

t	f
0.001	0.999

P (A | B, E)

B	*E*	t	f
t	t	0.95	0.05
t	f	0.94	0.06
f	t	0.29	0.71
f	f	0.001	0.999

P (J | A)

A	t	f
t	0.9	0.1
f	0.05	0.95

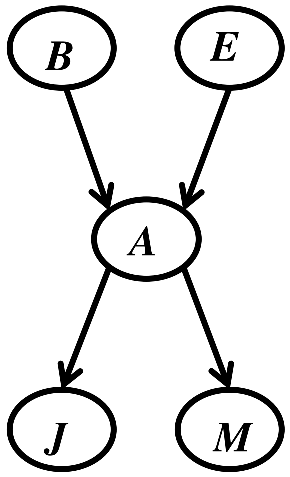
P (M | A)

A	t	f
t	0.7	0.3
f	0.01	0.99

Inference by Enumeration

- Let a denote $\boldsymbol{A}$=true, and $\neg a$ denote $\boldsymbol{A}$=false
- Suppose we're given the query: $P(b \mid j, m)$

  "probability the house is being burglarized given that John and Mary both called"
- From the graph structure we can first compute:



$$P(b, j, m) = \sum_{e,\neg e} \sum_{a,\neg a} P(b)P(E)P(A \mid b, E)P(j \mid A)P(m \mid A)$$

sum over possible values for E and A variables (e, $\neg e$, a, $\neg a$)

Parameter Learning: MLE

- **Goal**: infer the parameters of the CPDs
- As usual, can use MLE

$$L(\theta : D, G) = P(D \mid G, \theta) = \prod_{d \in D} P(x_1^{(d)}, x_2^{(d)}, \ldots, x_n^{(d)})$$

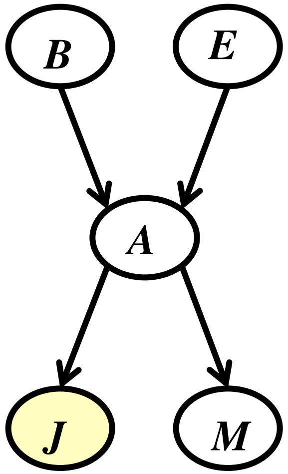
$$= \prod_{d \in D} \prod_i P(x_i^{(d)} \mid Parents(x_i^{(d)}))$$

$$= \prod_i \underbrace{\left(\prod_{d \in D} P(x_i^{(d)} \mid Parents(x_i^{(d)})) \right)}_{\text{independent parameter learning problem for each CPD}}$$

independent parameter learning problem for each CPD

Parameter Learning: MLE Example

- **Goal**: infer the parameters of the CPDs
- Consider estimating the CPD parameters for B and J in the alarm network given the following data set



B	E	A	J	M
f	f	f	t	f
f	t	f	f	f
f	f	f	t	t
t	f	f	f	t
f	f	t	t	f
f	f	t	f	t
f	f	t	t	t
f	f	t	t	t

$$P(b) = \frac{1}{8} = 0.125$$

$$P(\neg b) = \frac{7}{8} = 0.875$$

$$P(j \mid a) = \frac{3}{4} = 0.75$$

$$P(\neg j \mid a) = \frac{1}{4} = 0.25$$

$$P(j \mid \neg a) = \frac{2}{4} = 0.5$$

$$P(\neg j \mid \neg a) = \frac{2}{4} = 0.5$$

Break & Quiz

Quiz

Can the Naïve Bayes' model be represented as a Bayesian network?

If no, explain why. If yes, draw the network.

Ans: Yes

Outline

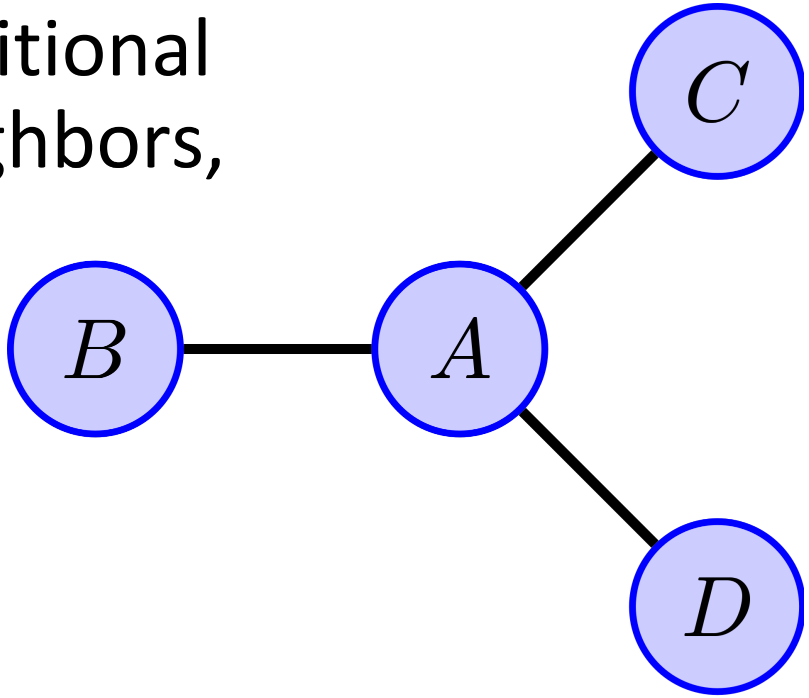
- **Bayesian Networks Review**
  - Definition, examples, inference, learning
- **Undirected Graphical Models**
  - Definitions, MRFs, exponential families
- **Structure learning**
  - Chow-Liu Algorithm
- **D-separation**

Undirected Graphical Models

- Still want to encode conditional independence, but not in a causal way (ie, no parents, direction)
  - **Why**? Allows for modeling other distributions that Bayes nets can't, allows for other algorithms

- Graph directly encodes a type of conditional independence. If nodes i,j are not neighbors,

$$X_i \perp X_j | X_{V \setminus \{i,j\}}$$



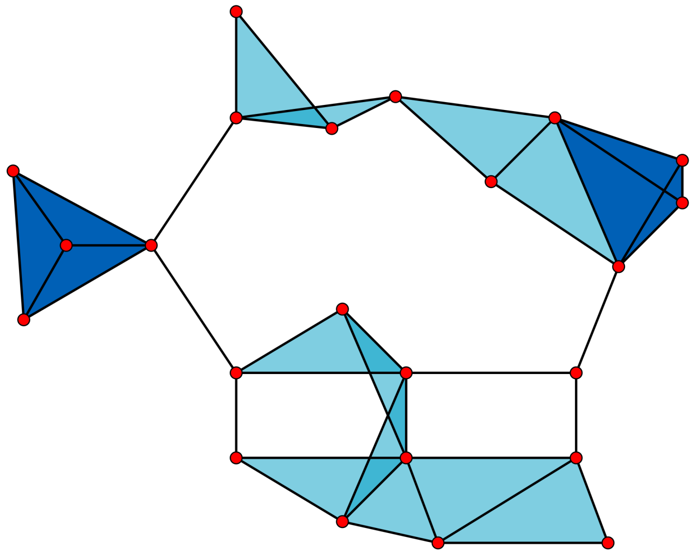
Markov Random Fields

- A particularly popular kind of undirected model. As above, can describe in terms of:
  - 1. Conditional independence: $X_i \perp X_j | X_{V \setminus \{i,j\}}$
  - 2. Factorization. (Clique: maximal fully-connected subgraphs)
    - Bayes nets: factorize over CPTs with **parents**; MRFs: factorize over **cliques**

$$P(X) = \frac{1}{Z} \prod_{C \in \text{cliques}(G)} \phi_C(X_C)$$

Partition function

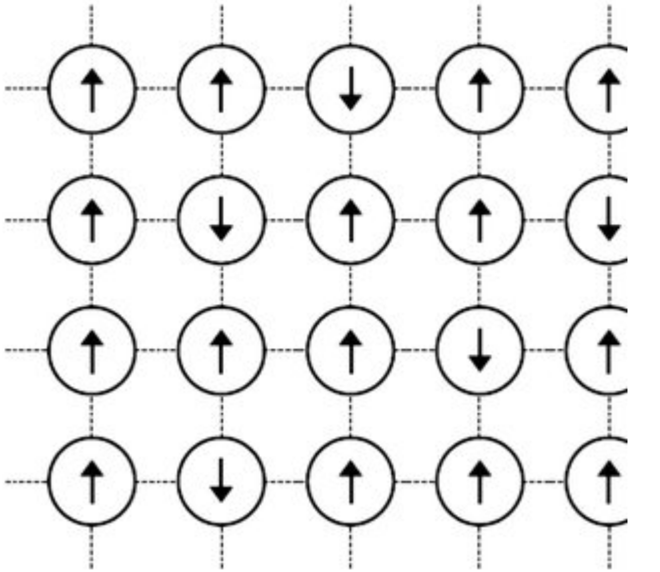
Potential functions



Ising Models

- Ising models: a particular kind of MRF usually written in exponential form
  - Popular in statistical physics
  - **Idea**: pairwise interactions (biggest cliques of size 2)

$$P(x_1, \ldots, x_d) = \frac{1}{Z} \exp\Big(\sum_{(i,j) \in E} \theta_{ij} x_i x_j \Big)$$



Khudier and Fawaz

- Challenges:
  - Compute partition function
  - Perform inference/marginalization

Outline

- **Bayesian Networks Review**
  - Definition, examples, inference, learning
- **Undirected Graphical Models**
  - Definitions, MRFs, exponential families
- **Structure learning**
  - Chow-Liu Algorithm
- **D-separation**

Structure Learning

- Generally a hard problem, many approaches.
  - Exponentially (or worse) many structures in # variables
  - Can either use heuristics or restrict to some tractable subset of networks. Ex: **trees**

- Chow-Liu Algorithm
  - Learns a BN with a tree structure that **maximizes the likelihood of the training data**

1. Compute weight $I(X_i, X_j)$ of each possible edge (X_i, X_j)
2. Find maximum weight spanning tree (MST)

Chow-Liu: Computing weights

- Use mutual information to calculate edge weights

$$I(X,Y) = \sum_{x \in \text{values}(X)} \sum_{y \in \text{values}(Y)} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- The probabilities are calculated empirically using data

Chow-Liu: Finding MST

- Many algorithms for calculating MST (e.g Kruskal's, Prim's)

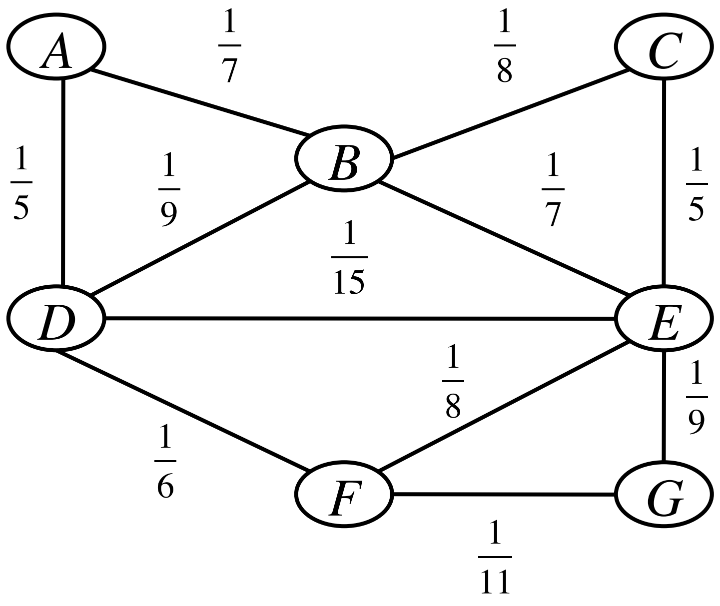
- Kruskal's algorithm

given: graph with vertices V and edges E

$E_{new} \leftarrow \{ \}$
for each (u, v) in E ordered by weight (from high to low)
{
  remove (u, v) from E
  if adding (u, v) to E_{new} does not create a cycle
    add (u, v) to E_{new}
}
return V and E_{new} which represent an MST

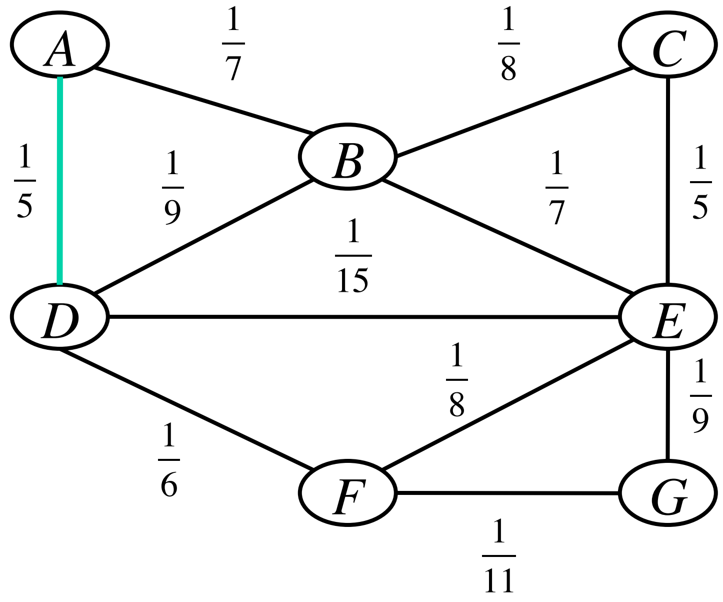
Chow-Liu: Example

- First, calculate empirical mutual information for each pair and calculate edge weights.
  - Graph is usually fully connected (using a non-complete graph for clarity)

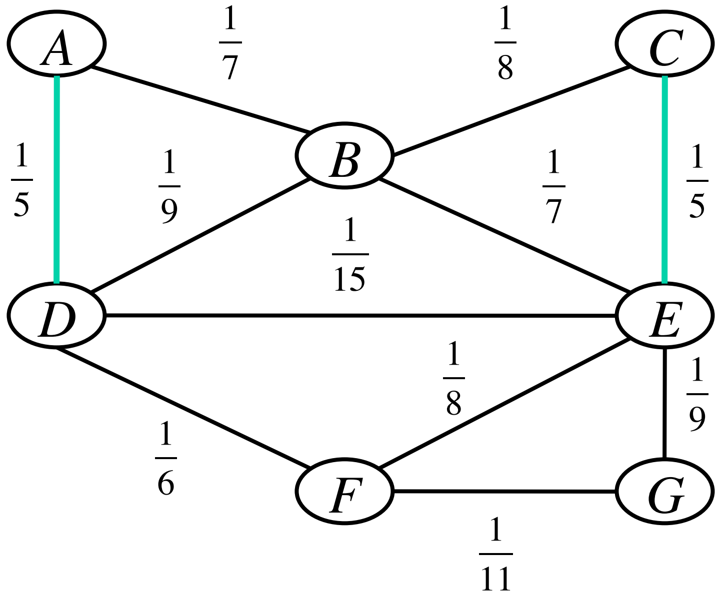


Chow-Liu: Example (cont'd)

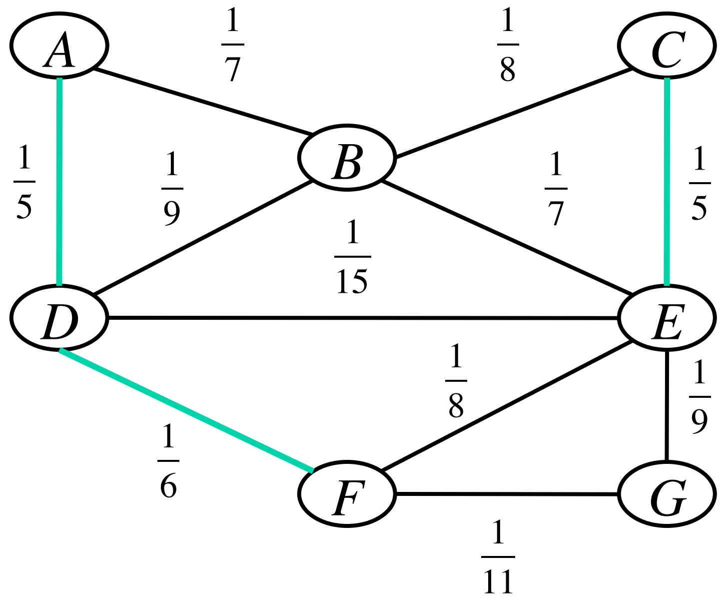
i.



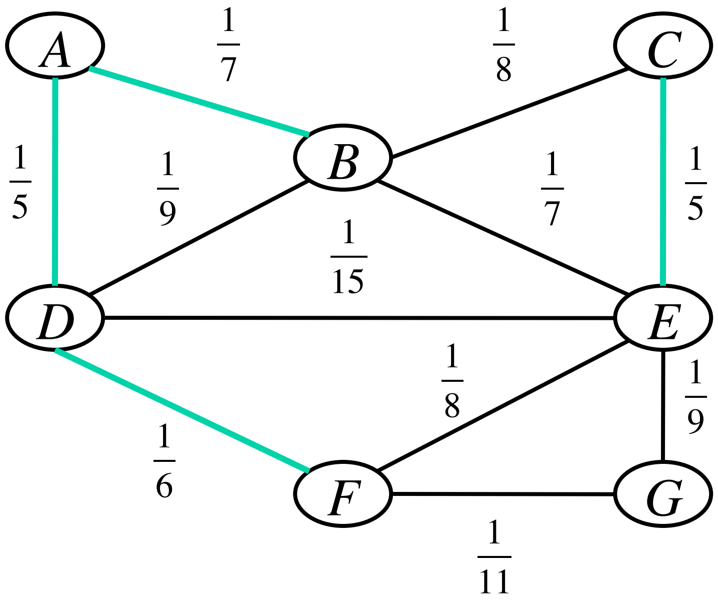
ii.



iii.

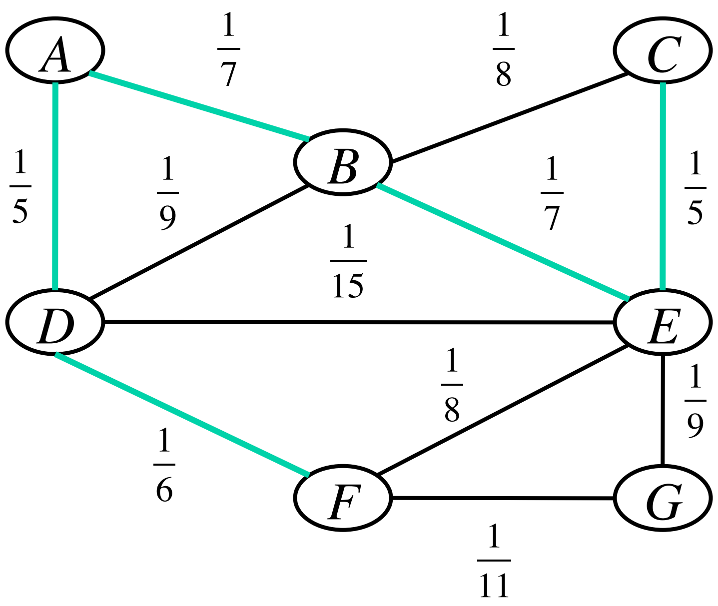


iv.

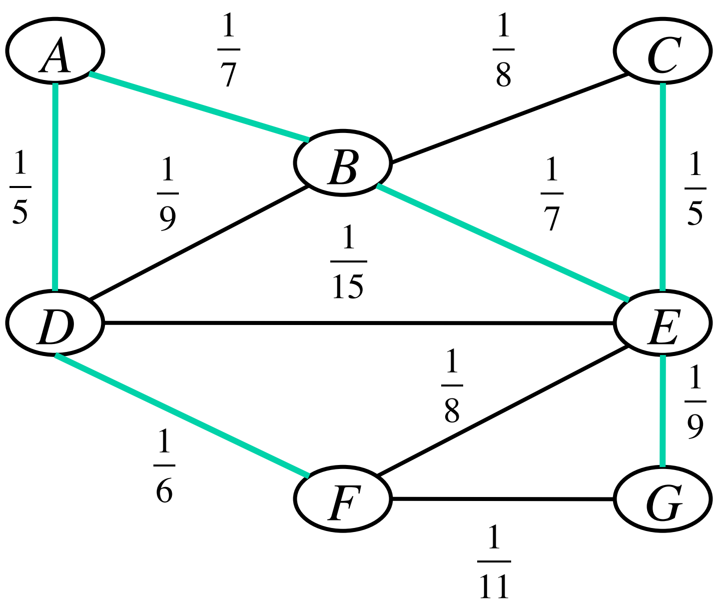


Chow-Liu: Example (cont'd)

v.



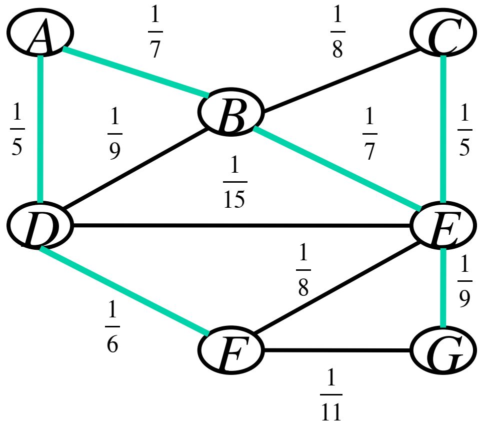
vi.

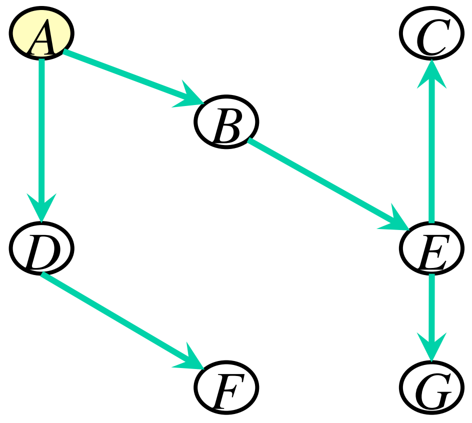


Chow-Liu Algorithm

1. Finding tree structures is a ‘second order’ approximation
   - First order: product of marginals $$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i)$$
   - Second order: allow conditioning on one variable
$$P(X_1, \ldots, X_n) = P(X_1) \prod_{i=2}^{n} P(X_i | X_{i-1})$$

2. To assign directions in a Bayes’ network, pick a root and making everything directed from root (may require domain expertise)

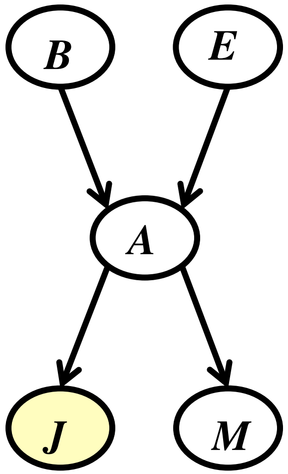




Outline

- **Bayesian Networks Review**
  - Definition, examples, inference, learning
- **Undirected Graphical Models**
  - Definitions, MRFs, exponential families
- **Structure learning**
  - Chow-Liu Algorithm
- **D-separation**

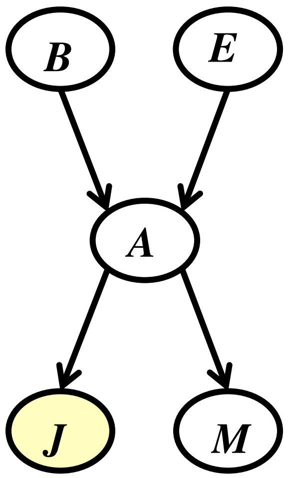
D-separation in Bayesian Networks



- Which of the following are true?
  1. $J \perp\!\!\!\perp M$
  2. $J \perp\!\!\!\perp M \mid A$
  3. $B \perp\!\!\!\perp J$
  4. $B \perp\!\!\!\perp J \mid A$
  5. $B \perp\!\!\!\perp E$
  6. $B \perp\!\!\!\perp E \mid A$

D-separation in Bayesian Networks

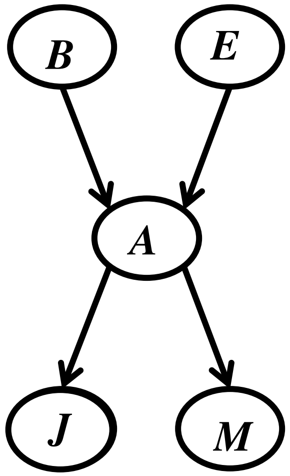
- Still want to encode conditional independence, but not in a,



- Which of the following are true?
  1. $J \perp\!\!\!\perp M$ (False)
  2. $J \perp\!\!\!\perp M \mid A$ (True)
  3. $B \perp\!\!\!\perp J$ (False)
  4. $B \perp\!\!\!\perp J \mid A$ (True)
  5. $B \perp\!\!\!\perp E$ (True)
  6. $B \perp\!\!\!\perp E \mid A$ (False)

D-separation in Bayesian Networks

- D-separation: A formal way to answer questions of conditional independence:
  - E.g. $J \perp\!\!\!\perp M \mid A$, $J \perp\!\!\!\perp E \mid B, M$ etc.



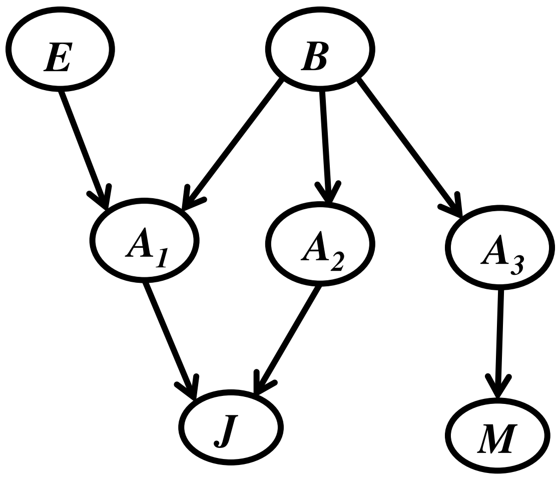
- Triples: Any 3 connected vertices
- We say that a triple is active if
  - (Causal chain): $X \rightarrow Y \rightarrow Z$ (Y is unobserved)
  - (Common cause): $X \leftarrow Y \rightarrow Z$ (Y is unobserved)
  - (Common effect): $X \rightarrow Y \leftarrow Z$ (Y or any descendent of Y is observed)

- An (undirected) path is active if all of it's triples are active.

D-separation in Bayesian Networks

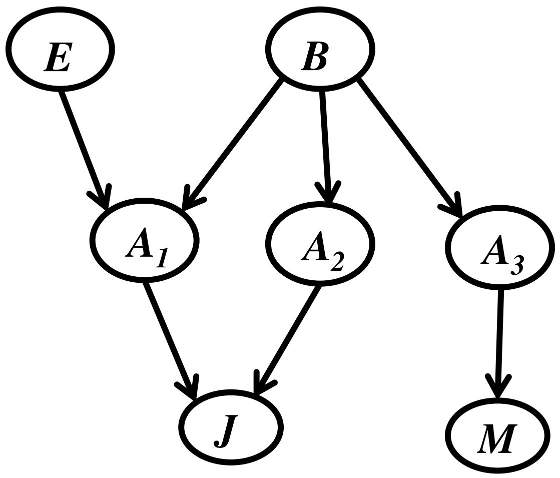
- Goal: Answer queries of the form: $A \perp\!\!\!\perp B \mid \{C, D, \ldots\}$
- D-separation Algorithm:
  - For all (undirected) paths from A to B
    - Check if path is active (i.e all triples are active)
      - Return “$A \perp\!\!\!\perp B \mid \{C, D, \ldots\}$ is **not** guaranteed”
  - If all paths are inactive:
    - Return “$A \perp\!\!\!\perp B \mid \{C, D, \ldots\}$ is true”

D-separation Examples



- Are the following conditional independences guaranteed?
  1. $B \perp\!\!\!\perp M$
  2. $B \perp\!\!\!\perp M \mid A_3$
  3. $E \perp\!\!\!\perp B$
  4. $E \perp\!\!\!\perp B \mid A_1$
  5. $E \perp\!\!\!\perp B \mid A_2$
  6. $E \perp\!\!\!\perp B \mid J$
  7. $A_1 \perp\!\!\!\perp A_2$
  8. $A_1 \perp\!\!\!\perp A_2 \mid E$
  9. $A_2 \perp\!\!\!\perp A_3 \mid B$
  10. $J \perp\!\!\!\perp M$
  11. $J \perp\!\!\!\perp M \mid A_3$

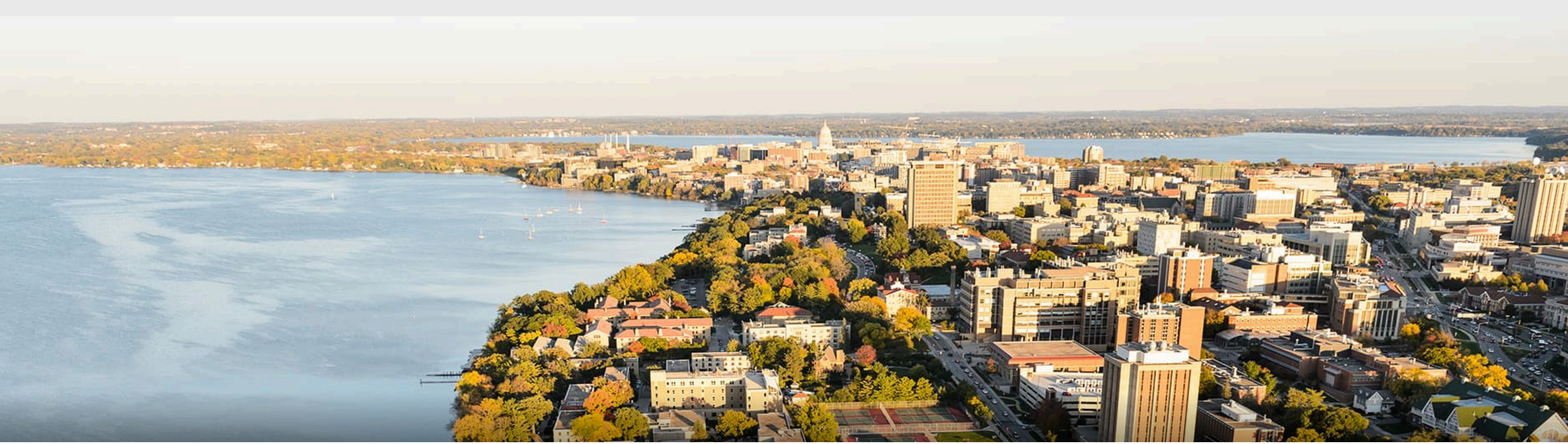
D-separation Examples



- Are the following conditional independences guaranteed?
  1. $B \perp\!\!\!\perp M$ (False)
  2. $B \perp\!\!\!\perp M \mid A_3$ (True)
  3. $E \perp\!\!\!\perp B$ (True)
  4. $E \perp\!\!\!\perp B \mid A_1$ (False)
  5. $E \perp\!\!\!\perp B \mid A_2$ (True)
  6. $E \perp\!\!\!\perp B \mid J$ (False)
  7. $A_1 \perp\!\!\!\perp A_2$ (False)
  8. $A_1 \perp\!\!\!\perp A_2 \mid E$ (False)
  9. $A_2 \perp\!\!\!\perp A_3 \mid B$ (True)
  10. $J \perp\!\!\!\perp M$ (False)
  11. $J \perp\!\!\!\perp M \mid A_3$ (True)

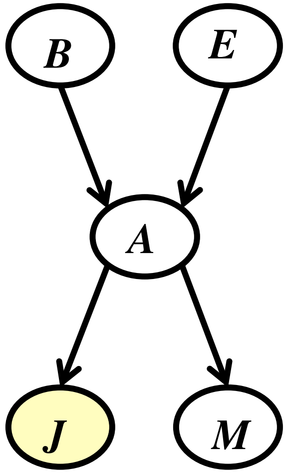
D-separation in Bayesian Networks

- Goal: Answer queries of the form: $A \perp\!\!\!\perp B \mid \{C, D, \ldots\}$
- D-separation Algorithm:
  - For all (undirected) paths from A to B
    - Check if path is active (i.e all triples are active)
      - Return "$A \perp\!\!\!\perp B \mid \{C, D, \ldots\}$ is **not** guaranteed"
  - If all paths are inactive:
    - Return "$A \perp\!\!\!\perp B \mid \{C, D, \ldots\}$ is true"



Break & Quiz

Quiz



True or False:

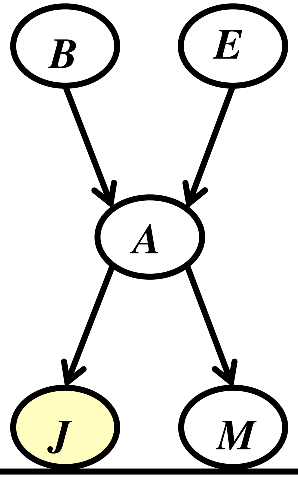
Bayesian networks can be used for unsupervised learning only. They cannot be used for supervised learning.

Ans: False

Quiz

You are given data of the form $\{B_i, E_i, J_i, M_i\}_i$. That is, you observe all variables except A.

1. Can you still learn the parameters via MLE?
2. If yes, what algorithm will you use?



B	E	A	J	M
f	f	?	f	f
f	f	?	t	f
t	f	?	t	t
f	f	?	f	t
f	t	?	t	f
f	f	?	f	t
t	t	?	t	t
f	f	?	f	f
f	f	?	t	f
f	f	?	f	t

Ans:

1. Yes

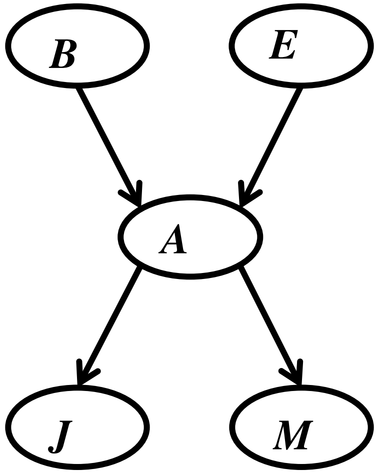
2. EM

Example: EM for parameter learning

suppose we're given the following <u>initial</u> BN and training set

P(B)
0.1

P(E)
0.2



B	*E*	*P(A)*
t	**t**	**0.9**
t	**f**	**0.6**
f	**t**	**0.3**
f	**f**	**0.2**

A	*P(J)*
t	**0.9**
f	**0.2**

A	*P(M)*
t	**0.8**
f	**0.1**

B	*E*	*A*	*J*	*M*
f	f	?	f	f
f	f	?	t	f
t	f	?	t	t
f	f	?	f	t
f	t	?	t	f
f	f	?	f	t
t	t	?	t	t
f	f	?	f	f
f	f	?	t	f
f	f	?	f	t

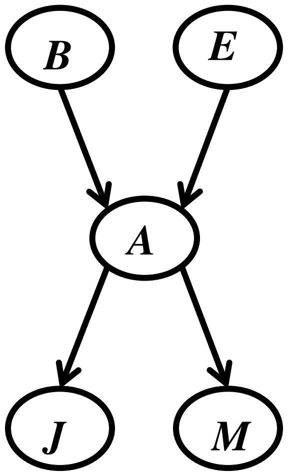
E-step

$$P(a \mid \neg b, \neg e, \neg j, \neg m)$$

$$P(\neg a \mid \neg b, \neg e, \neg j, \neg m)$$

P(B)
0.1

P(E)
0.2



B	*E*	*P(A)*
t	t	0.9
t	f	0.6
f	t	0.3
f	f	0.2

A	*P(J)*
t	0.9
f	0.2

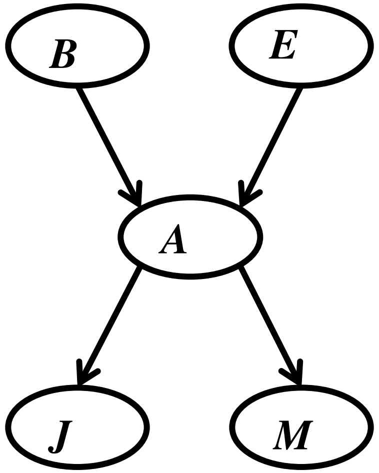
A	*P(M)*
t	0.8
f	0.1

B	*E*	*A*	*J*	*M*
f	f	t: 0.0069 f: 0.9931	f	f
f	f	t:0.2 f:0.8	t	f
t	f	t:0.98 f: 0.02	t	t
f	f	t: 0.2 f: 0.8	f	t
f	t	t: 0.3 f: 0.7	t	f
f	f	t:0.2 f: 0.8	f	t
t	t	t: 0.997 f: 0.003	t	t
f	f	t: 0.0069 f: 0.9931	f	f
f	f	t:0.2 f: 0.8	t	f
f	f	t: 0.2 f: 0.8	f	t

Example: E-step

P(B)
0.1

P(E)
0.2



B	*E*	*P(A)*
t	t	**0.9**
t	f	**0.6**
f	t	**0.3**
f	f	**0.2**

A	*P(J)*
t	**0.9**
f	**0.2**

A	*P(M)*
t	**0.8**
f	**0.1**

$$P(a \mid \neg b, \neg e, \neg j, \neg m)$$
$$= \frac{P(\neg b, \neg e, a, \neg j, \neg m)}{P(\neg b, \neg e, a, \neg j, \neg m) + P(\neg b, \neg e, \neg a, \neg j, \neg m)}$$
$$= \frac{0.9 \times 0.8 \times 0.2 \times 0.1 \times 0.2}{0.9 \times 0.8 \times 0.2 \times 0.1 \times 0.2 + 0.9 \times 0.8 \times 0.8 \times 0.8 \times 0.9}$$
$$= \frac{0.00288}{0.4176} = 0.0069$$

$$P(a \mid \neg b, \neg e, j, \neg m)$$
$$= \frac{P(\neg b, \neg e, a, j, \neg m)}{P(\neg b, \neg e, a, j, \neg m) + P(\neg b, \neg e, \neg a, j, \neg m)}$$
$$= \frac{0.9 \times 0.8 \times 0.2 \times 0.9 \times 0.2}{0.9 \times 0.8 \times 0.2 \times 0.9 \times 0.2 + 0.9 \times 0.8 \times 0.8 \times 0.2 \times 0.9}$$
$$= \frac{0.02592}{0.1296} = 0.2$$

M-step

re-estimate probabilities
using expected counts

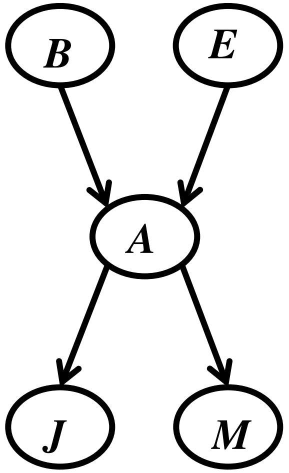
$$P(a \mid b, e) = \frac{E\#(a \wedge b \wedge e)}{E\#(b \wedge e)}$$

$$P(a \mid b, e) = \frac{0.997}{1}$$

$$P(a \mid b, \neg e) = \frac{0.98}{1}$$

$$P(a \mid \neg b, e) = \frac{0.3}{1}$$

$$P(a \mid \neg b, \neg e) = \frac{0.0069 + 0.2 + 0.2 + 0.2 + 0.0069 + 0.2 + 0.2}{7}$$



B	*E*	*P(A)*
t	t	0.997
t	f	0.98
f	t	0.3
f	f	0.145

re-estimate probabilities for
$P(J \mid A)$ and $P(M \mid A)$ in same way

B	*E*	*A*	*J*	*M*
f	f	t: 0.0069 f: 0.9931	f	f
f	f	t:0.2 f:0.8	t	f
t	f	t:0.98 f: 0.02	t	t
f	f	t: 0.2 f: 0.8	f	t
f	t	t: 0.3 f: 0.7	t	f
f	f	t:0.2 f: 0.8	f	t
t	t	t: 0.997 f: 0.003	t	t
f	f	t: 0.0069 f: 0.9931	f	f
f	f	t:0.2 f: 0.8	t	f
f	f	t: 0.2 f: 0.8	f	t

M-step

re-estimate probabilities using expected counts

$$P(j \mid a) = \frac{E\#(a \wedge j)}{E\#(a)}$$

$$P(j \mid a) = \frac{0.2 + 0.98 + 0.3 + 0.997 + 0.2}{0.0069 + 0.2 + 0.98 + 0.2 + 0.3 + 0.2 + 0.997 + 0.0069 + 0.2 + 0.2}$$

$$P(j \mid \neg a) = \frac{0.8 + 0.02 + 0.7 + 0.003 + 0.8}{0.9931 + 0.8 + 0.02 + 0.8 + 0.7 + 0.8 + 0.003 + 0.9931 + 0.8 + 0.8}$$

B	E	A	J	M
f	f	t: 0.0069 f: 0.9931	f	f
f	f	t:0.2 f:0.8	t	f
t	f	t:0.98 f: 0.02	t	t
f	f	t: 0.2 f: 0.8	f	t
f	t	t: 0.3 f: 0.7	t	f
f	f	t:0.2 f: 0.8	f	t
t	t	t: 0.997 f: 0.003	t	t
f	f	t: 0.0069 f: 0.9931	f	f
f	f	t:0.2 f: 0.8	t	f
f	f	t: 0.2 f: 0.8	f	t

Thanks Everyone!

Some of the slides in these lectures have been adapted/borrowed from materials developed by Mark Craven, David Page, Jude Shavlik, Tom Mitchell, Nina Balcan, Elad Hazan, Tom Dietterich, Pedro Domingos, Jerry Zhu, Yingyu Liang, Volodymyr Kuleshov, Fei-Fei Li, Justin Johnson, Serena Yeung, Pieter Abbeel, Peter Chen, Jonathan Ho, Aravind Srinivas, and Fred Sala