Artificial Intelligence

CS 165A

Jan 23 2020

Instructor: Prof. Yu-Xiang Wang

Today

→ Bayesian Network (Ch 14)

Announcement

- Extension of HW1 submission deadline to Friday 23:59.
  - No extensions will be given to future homeworks.

- HW2 will be released tonight.

- Homework is the most important component of this course.
  - Research shows that active learning is more effective.
  - It is expected to be long and challenging.
  - You have two full weeks of time. Start early!

Student feedback

- Feedback from the last lecture
  - Lecture is going at a "great pace"
  - "Ease your assumption on prior knowledge"
  - "Review conditional independence … how it reduces # of parameters in discussion sections"
  - "Color-blindness"
- On Piazza:
  - "Can we go over HW1 Q6 in the lecture please?"

Color blindness accessible palettes

Accessible palettes

So what colors *should* you use? The colorpicker tool above is intended to give the freedom to choose your own colors while making sure that your color palette is accessible. But to get you started, here are some ideas. Here are 8 pairs of contrasting colors which maintain their contrast for people who are colorblind. Click on any of them to load it into the color palette selection tool above.

HEX #FFC20A **R** 255 **G** 194 **B** 10 **HEX** #0C7BDC **R** 12 **G** 123 **B** 220	**HEX** #994F00 **R** 153 **G** 79 **B** 0 **HEX** #006CD1 **R** 0 **G** 108 **B** 209	**HEX** #E1BE6A **R** 225 **G** 190 **B** 106 **HEX** #40B0A6 **R** 64 **G** 176 **B** 166	**HEX** #E66100 **R** 230 **G** 97 **B** 0 **HEX** #5D3A9B **R** 93 **G** 58 **B** 155
HEX #1AFF1A **R** 26 **G** 255 **B** 26 **HEX** #4B0092 **R** 75 **G** 0 **B** 146	**HEX** #FEFE62 **R** 254 **G** 254 **B** 98 **HEX** #D35FB7 **R** 211 **G** 95 **B** 183	**HEX** #005AB5 **R** 0 **G** 90 **B** 181 **HEX** #DC3220 **R** 220 **G** 50 **B** 32	**HEX** #1A85FF **R** 26 **G** 133 **B** 255 **HEX** #D41159 **R** 212 **G** 17 **B** 89

https://davidmathlogic.com/colorblind/

Homework 1 Q6

- "SGD does not converge, results look random"
  - There are so many steps
  - Creating the vocabulary
  - Feature extraction
  - Gradient implementation
  - Optimization
  - Hyperparameter choices

- "What did I do wrongly?"

- These are exactly what you will run into in practice!

Gradient Descent and Stochastic Gradient Descent

- Optimization problem to solve in Q6

$$\min_{\Theta} \sum_{i=1}^{n} L(\Theta, (\mathbf{x}_i, y_i)) + \lambda \|\Theta\|_F^2.$$

- From Slide 36 of Lecture 4

Gradient descent

$$\theta_{t+1} = \theta_t - \eta_t \nabla f(\theta_t)$$

Stochastic gradient descent

$$\theta_{t+1} = \theta_t - \eta_t \hat{\nabla} f(\theta_t)$$

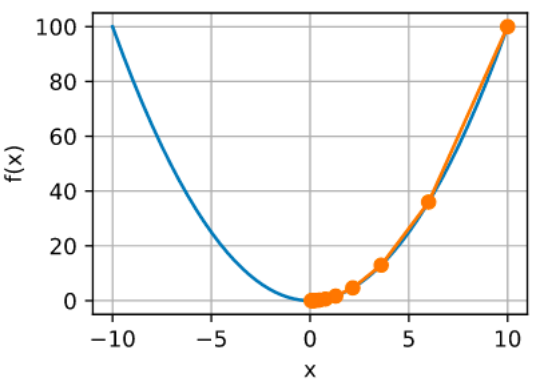
(Slide 41 for Lecture 4) How to choose the step sizes / learning rates?

- In theory:
  - Gradient decent: 1/L where L is the **Lipschitz constant** of the function we minimize.
  - SGD: $\sum_t \eta_t = \infty, \sum_t \eta_t^2 < \infty$
    - e.g. $\eta_t \in [1/t, 1/\sqrt{t})$

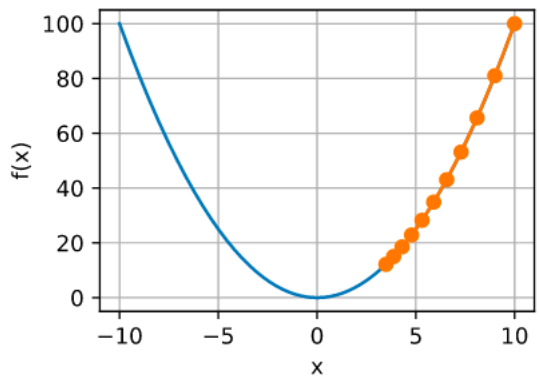
- In practice:
  - Use cross-validation on a subsample of the data.
  - Fixed learning rate for SGD is usually fine.
  - If it diverges, decrease the learning rate.
  - If for extremely small learning rate, it still diverges, check if your gradient implementation is correct.

Learning rate η choices in GD and SGD

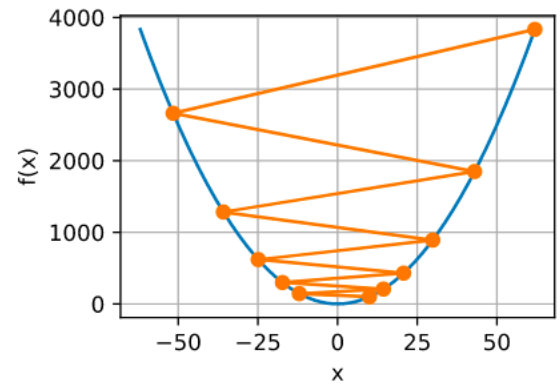
OKish Learning rate



Small Learning rate



Learning rate too large



1D-example from D2L book, Chapter 11.

http://d2l.ai/chapter_optimization/gd.html#gradient-descent-in-one-dimension

Recap of the last lecture

- Probability notations
  - Distinguish between events and random variables, apply rules of probabilities
- Representing a joint-distribution
  - number of parameters exponential in the number of variables
  - Calculating marginals and conditionals from the joint-distribution.
- Conditional independences and factorization of joint-distributions
  - Saves parameters, often exponential improvements

Tradeoffs in our model choices

Fully Independent

$P(X_1, X_2, \ldots, X_N)$

$= P(X_1)\, P(X_2) \ldots P(X_N)$

$O(N)$

?

Fully general

$P(X_1, X_2, \ldots, X_N)$

$O(e^N)$

Space / computation efficiency

Expressiveness

Idea:

1.Independent groups of variables?

2.Conditional independences?

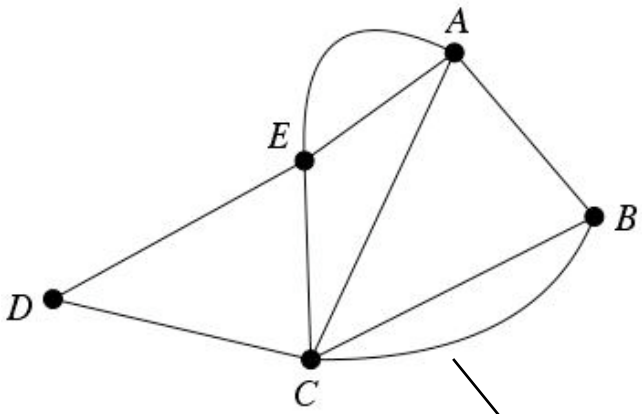
Benefit of conditional independence

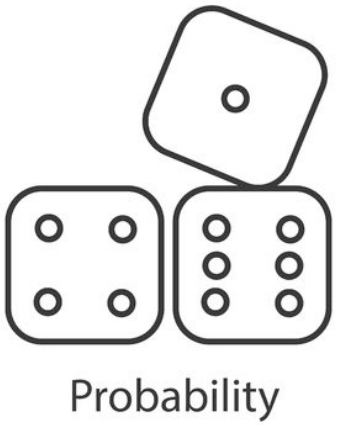
- If some variables are <u>conditionally independent</u>, the joint probability can be specified with many fewer than 2^N-1 numbers (or 3^N-1, or 10^N-1, or…)
- For example: (for binary variables W, X, Y, Z)
  - $P(W,X,Y,Z) = P(W)\, P(X|W)\, P(Y|\cancel{W},X)\, P(Z|\cancel{W},\cancel{X},Y)$
    - 1 + 2 + 4 + 8 = 15 numbers to specify
  - But if Y and W are independent given X, and Z is independent of W and X given Y, then
    - $P(W,X,Y,Z) = P(W)\, P(X|W)\, P(Y|X)\, P(Z|Y)$
      - 1 + 2 + 2 + 2 = 7 numbers
- This is often the case in real problems.

Today

- Bayesian Networks
- Examples
- d-separation, Bayes Ball algorithm
- Examples

Graphical models come out of the marriage of graph theory and probability theory

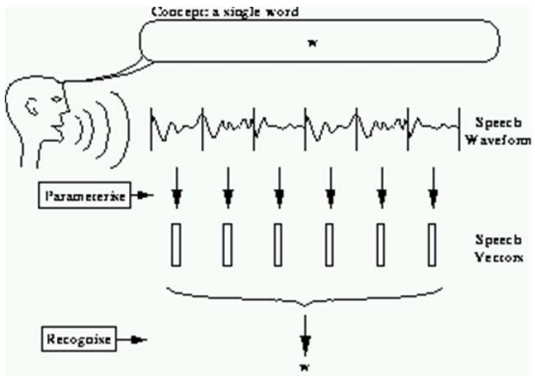




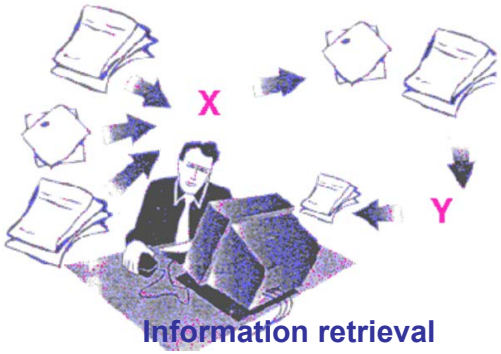
Directed Graph => Bayesian Networks / Belief Networks

Undirected Graph => Markov Random Fields

Used as a modeling tool. Many applications!

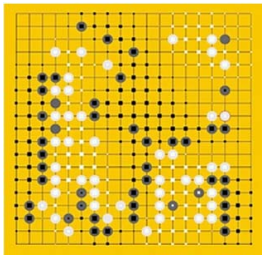


Speech recognition

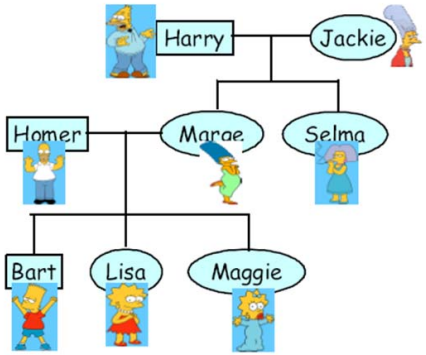


Information retrieval

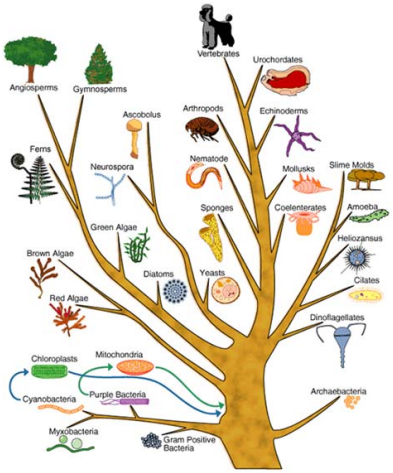
Computer vision



Games



Pedigree



Evolution

Robotic control

Planning

(Slides from Prof. Eric Xing)

Two ways to think about Graphical Models

- A particular factorization of a joint distribution
  - $P(X,Y,Z) = P(X)\ P(Y|X)\ P(Z|Y)$
- A collection of conditional independences
  - $\{ X \perp Z \mid Y, \ldots \}$

Represented using a graph!

Belief Networks

a.k.a. Probabilistic networks, Belief nets, Bayes nets, etc.

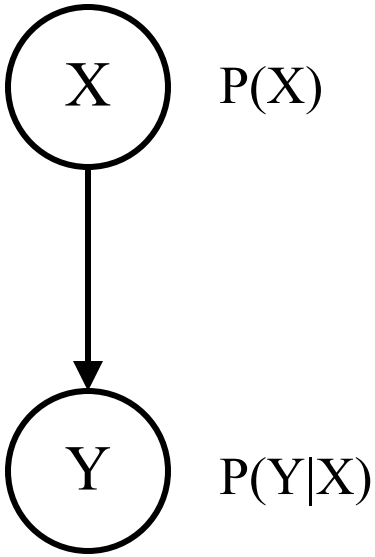
- Belief network
  - A data structure (depicted as a graph) that represents the dependence among variables and allows us to concisely specify the joint probability distribution
  - The graph itself is known as an "influence diagram"

- A belief network is a **directed acyclic graph** where:
  - The nodes represent the set of random variables (one node per random variable)
  - Arcs between nodes represent *influence*, or *causality*
    - A link from node X to node Y means that X "directly influences" Y
  - Each node has a ***conditional probability table*** (CPT) that defines **P(node | parents)**

Example

- Random variables X and Y
  - X – It is raining
  - Y – The grass is wet
- X has an *effect* on Y

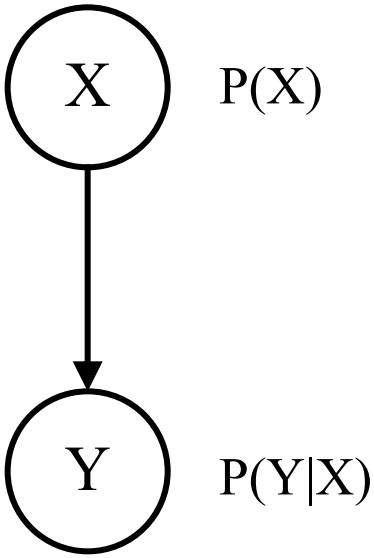
  Or, Y is a *symptom* of X
- Draw two nodes and link them



- Define the CPT for each node
  - P(X) and P(Y | X)
- Typical use: we observe Y and we want to query P(X | Y)
  - Y is an *evidence variable*
  - X is a *query variable*

We can write everything we want as a function of the CPTs. Try it!

- What is P(X | Y)?
  - Given that we know the CPTs of each node in the graph



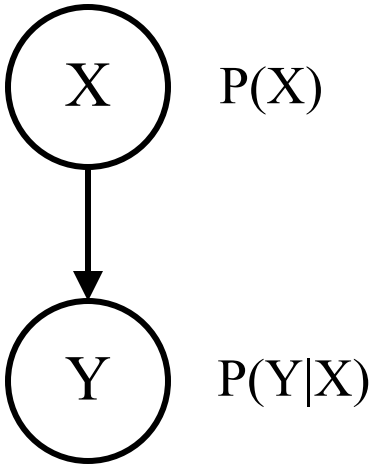
$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

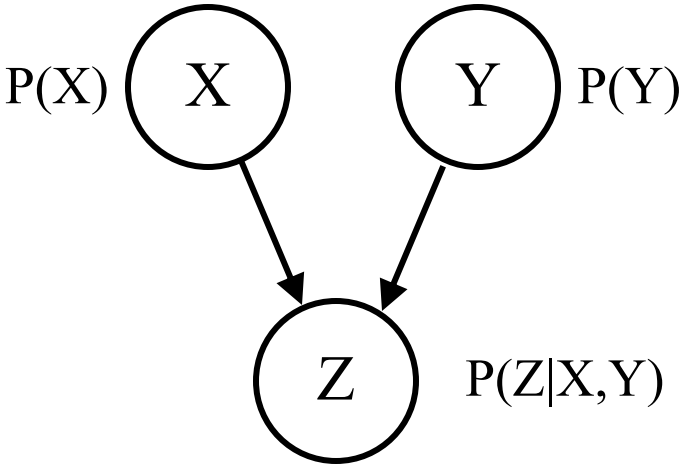
$$= \frac{P(Y \mid X)P(X)}{\sum_X P(X,Y)}$$

$$= \frac{P(Y \mid X)P(X)}{\sum_X P(Y \mid X)P(X)}$$

Belief nets represent the joint probability

- The joint probability function can be calculated directly from the network
  - It's the product of the CPTs of all the nodes
  - $P(var_1, \ldots, var_N) = \prod_i P(var_i | Parents(var_i))$





$P(X,Y) = P(X)\ P(Y|X)$

$P(X,Y,Z) = P(X)\ P(Y)\ P(Z|X,Y)$

Three steps in modelling with Belief Networks

1. Choose variables in the environments, represent them as nodes.

2. Connect the variables by inspecting the “direct influence”: cause-effect

3. Fill in the probabilities in the CPTs.

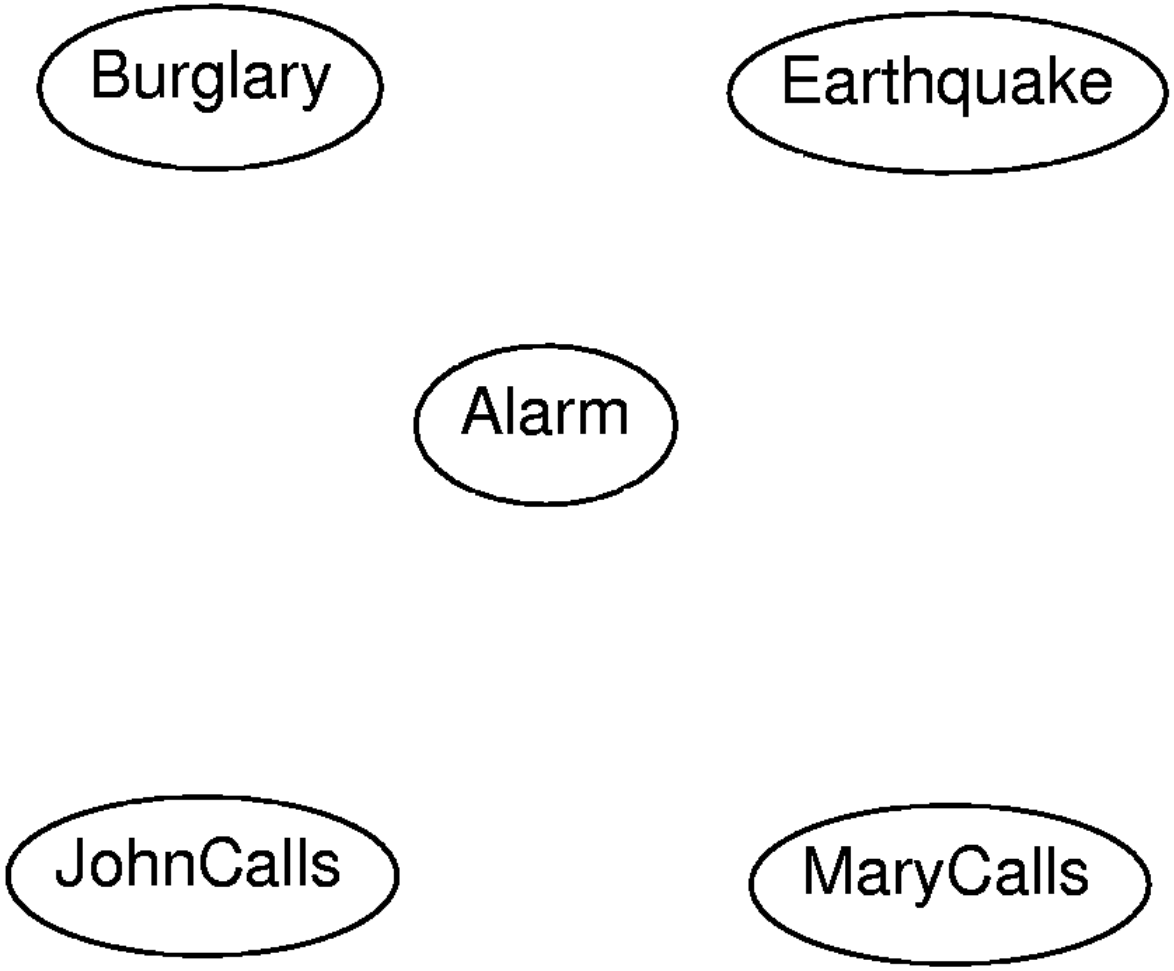
Example: Modelling with Belief Net

I'm at work and my neighbor John called to say my home alarm is ringing, but my neighbor Mary didn't call. The alarm is sometimes triggered by minor earthquakes. Was there a burglar at my house?

- Random (boolean) variables:
  - JohnCalls, MaryCalls, Earthquake, Burglar, Alarm
- The belief net shows the causal links
- This defines the joint probability
  - P(JohnCalls, MaryCalls, Earthquake, Burglar, Alarm)
- What do we want to know?

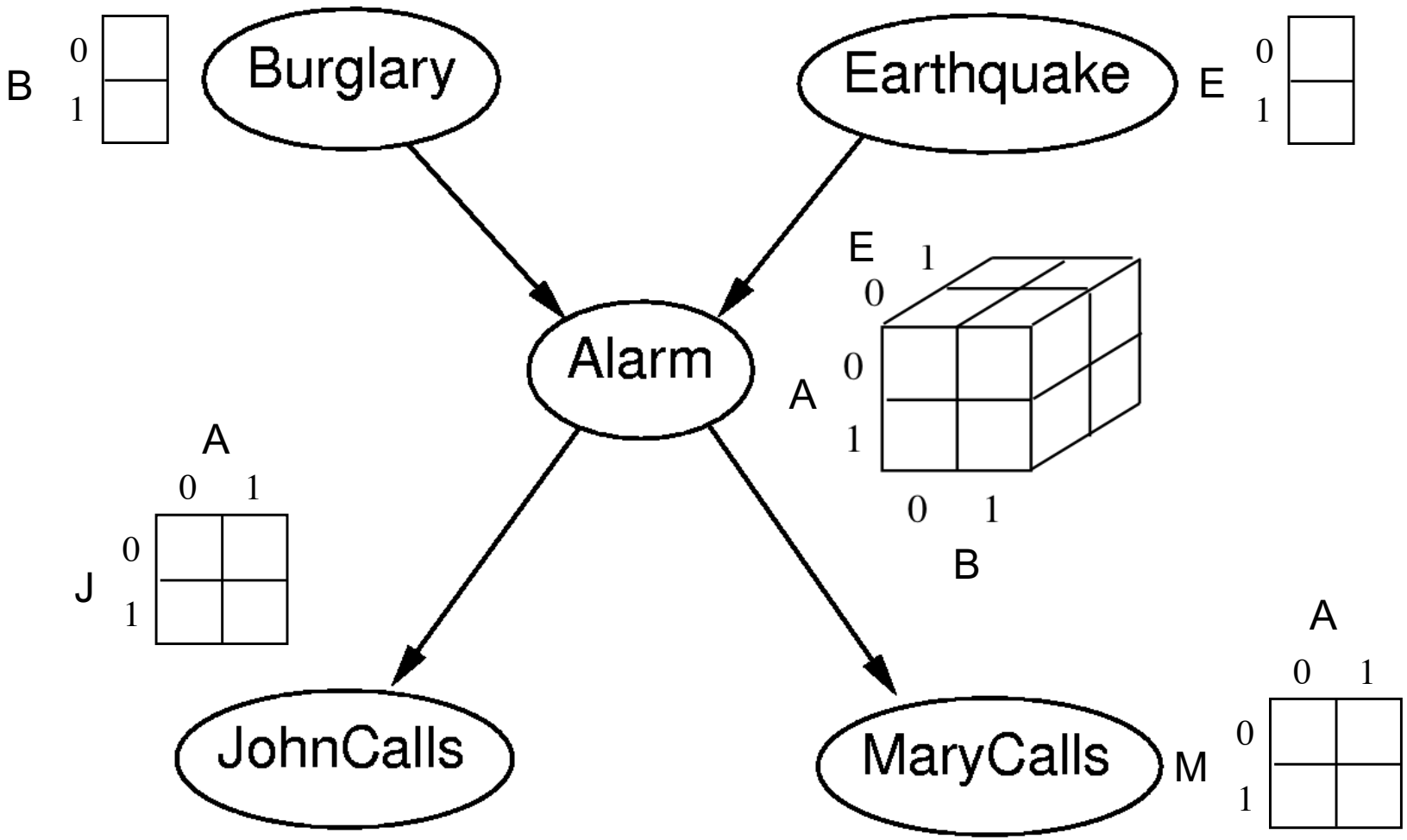
$$\mathbf{P(B \mid J, \neg M)}$$

How should we connect the nodes? (3 min discussion)



Links and CPTs?

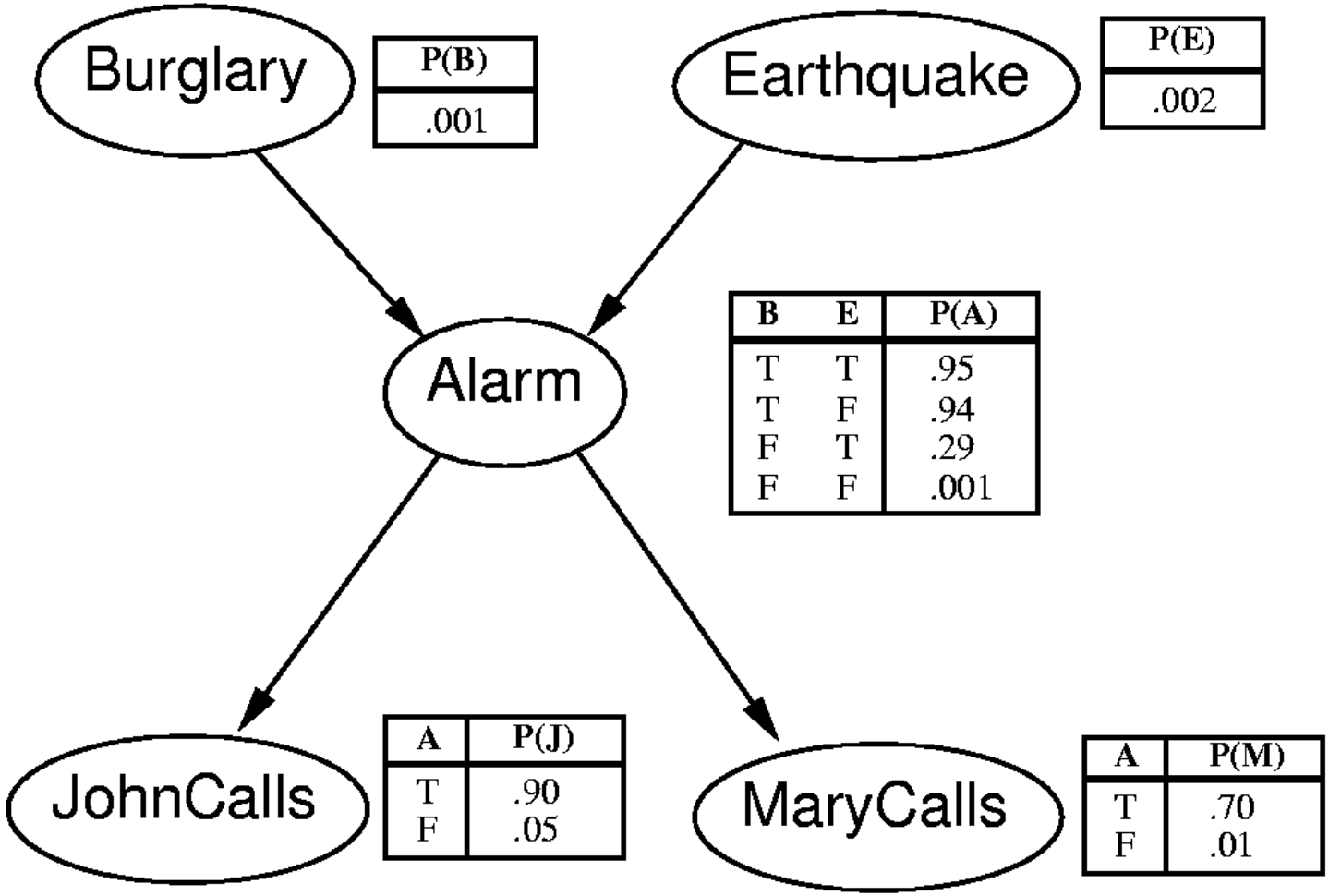
What are the CPTs? What are their dimensions?



Question: How to fill values into these CPTs?

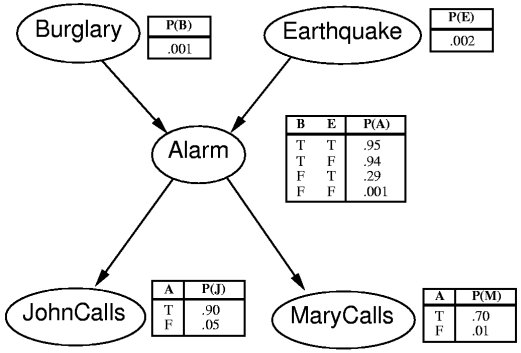
Ans: Specify by hands. Learn from data (e.g., MLE).

Example



Joint probability? $P(J, \neg M, A, B, \neg E)$?

Calculate P(J, ¬M, A, B, ¬E)



Read the joint pf from the graph:

$P(J, M, A, B, E) = P(B)\ P(E)\ P(A|B,E)\ P(J|A)\ P(M|A)$

Plug in the desired values:

$P(J, \neg M, A, B, \neg E) = P(B)\ P(\neg E)\ P(A|B,\neg E)\ P(J|A)\ P(\neg M|A)$

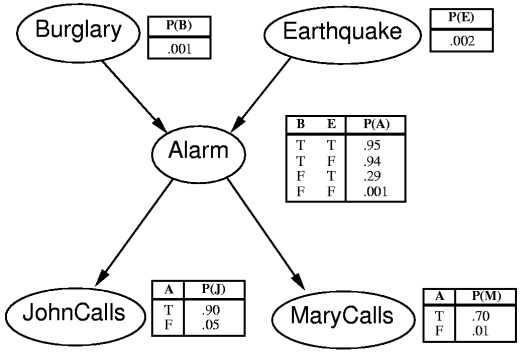
$= 0.001 * 0.998 * 0.94 * 0.9 * 0.3$

$= 0.0002532924$

How about P(B | J, ¬M) ?

Remember, this means P(B=true | J=true, M=false)

Calculate P(B | J, ¬M)



$$P(B \mid J, \neg M) = \frac{P(B, J, \neg M)}{P(J, \neg M)}$$

By marginalization:

$$= \frac{\sum_i \sum_j P(J, \neg M, A_i, B, E_j)}{\sum_i \sum_j \sum_k P(J, \neg M, A_i, B_j, E_k)}$$

$$= \frac{\sum_i \sum_j P(B)P(E_j)P(A_i \mid B, E_j)P(J \mid A_i)P(\neg M \mid A_i)}{\sum_i \sum_j \sum_k P(B_j)P(E_k)P(A_i \mid B_j, E_k)P(J \mid A_i)P(\neg M \mid A_i)}$$

Quick checkpoint

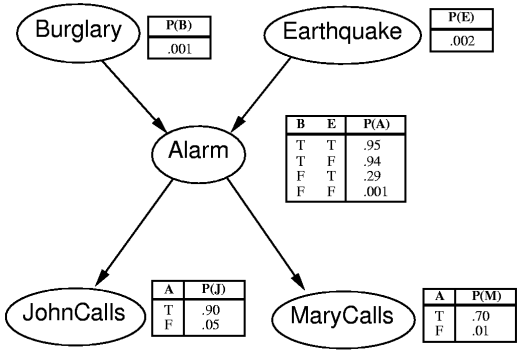
- Belief Net as a modelling tool

- By inspecting the cause-effect relationships, we can draw directed edges based on our domain knowledge

- The product of the CPTs give the joint distribution
  - We can calculate P(A | B) for any A and B
  - The factorization makes it computationally more tractable

What else can we get?

Example: Conditional Independence



- Conditional independence is seen here
  - P(JohnCalls | MaryCalls, Alarm, Earthquake, Burglary) = P(JohnCalls | Alarm)
  - So JohnCalls is independent of MaryCalls, Earthquake, and Burglary, given Alarm
- Does this mean that an earthquake or a burglary do not influence whether or not John calls?
  - No, but the influence is already accounted for in the Alarm variable
  - JohnCalls is <u>conditionally</u> independent of Earthquake, but not <u>absolutely</u> independent of it

***This conclusion is independent to values of CPTs!**

Question

If X and Y are independent, are they therefore independent given any variable(s)?

I.e., if $P(X, Y) = P(X)\,P(Y)$ [i.e., if $P(X|Y) = P(X)$], can we conclude that

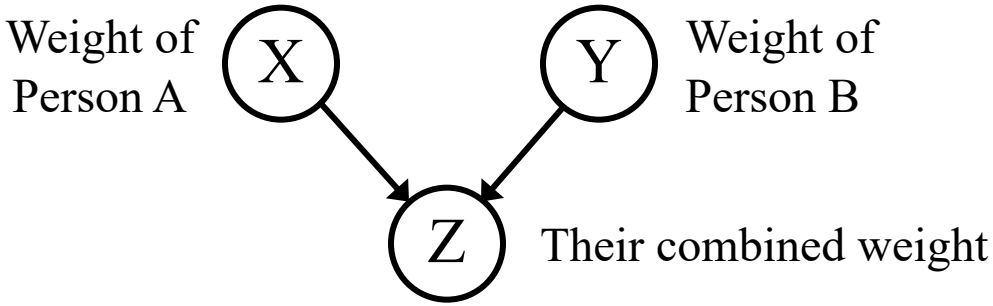
$P(X \mid Y, Z) = P(X \mid Z)$?

Question

If X and Y are independent, are they therefore independent given any variable(s)?

I.e., if $P(X, Y) = P(X)\,P(Y)$ [i.e., if $P(X|Y) = P(X)$], can we conclude that
$P(X \mid Y, Z) = P(X \mid Z)$?

The answer is **no**, and here's a counter example:



$$P(X \mid Y) = P(X)$$

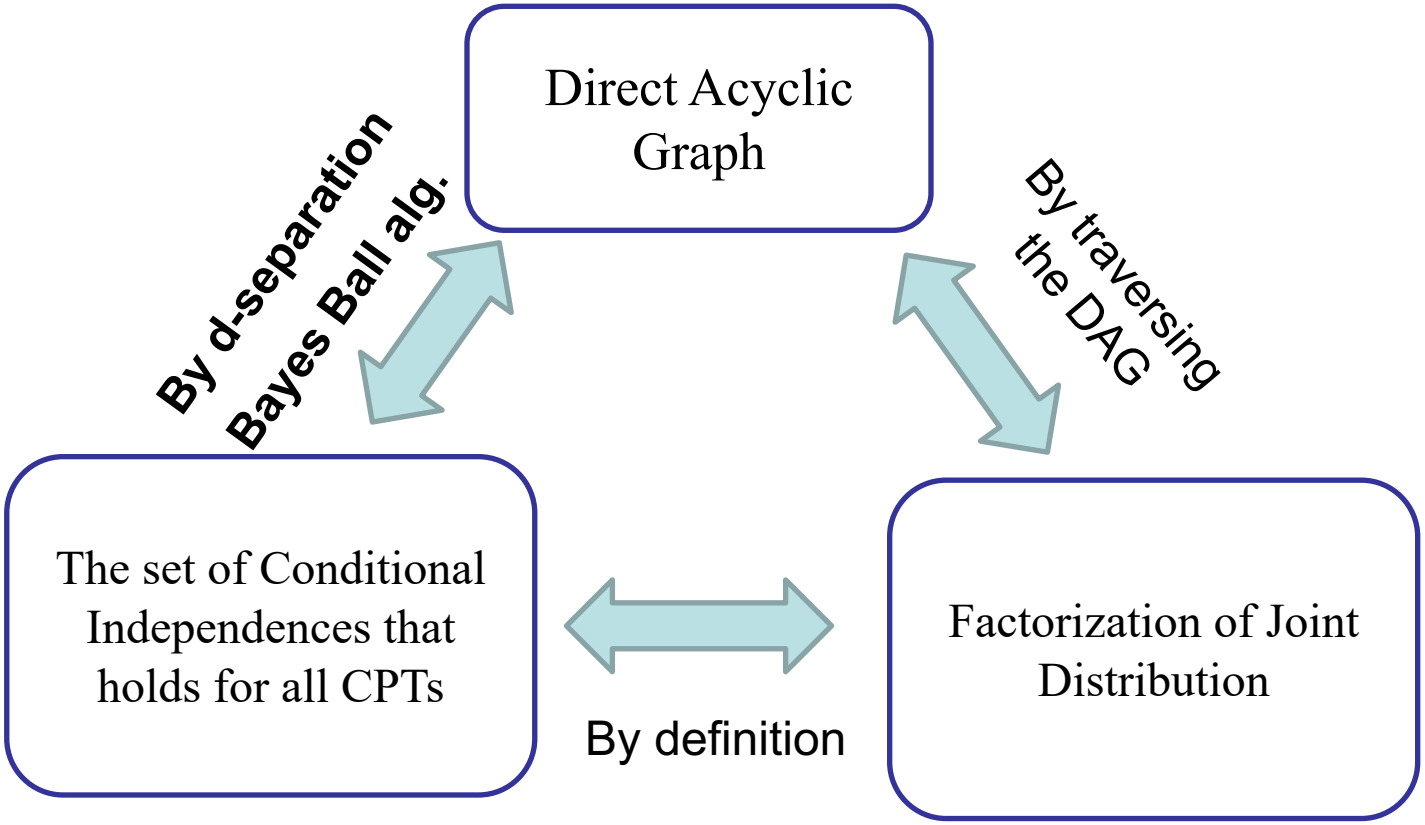
$$P(X \mid Y, Z) \neq P(X \mid Z)$$

Note: Even though Z is a deterministic function of X and Y, it is still a random variable with a probability distribution

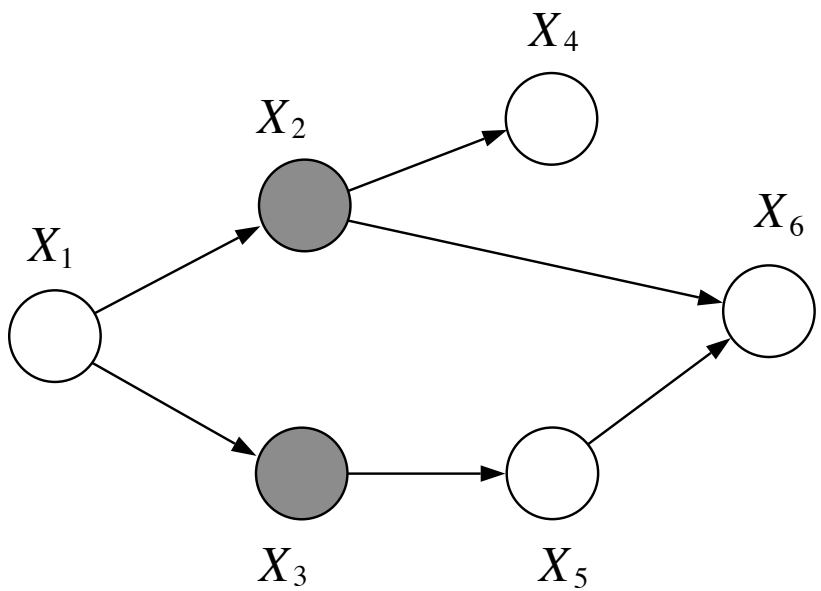
***Again: This conclusion is independent to values of CPTs!**

Big question: Is there a general way that we can answer questions about conditional independences by just inspecting the graphs?

- Turns out the answer is "Yes!"



Intuition: the graph and the edges controls the information flow, if there is no path that the information can flow from one-node to another, we say these two nodes are independent..

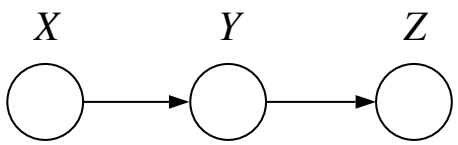


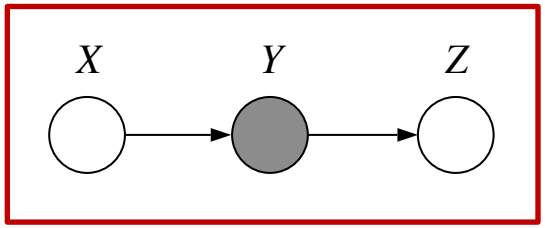
"Shading" denotes "observing" or "conditioning on" that variables.

Figure 2.3: The nodes X_2 and X_3 separate X_1 from X_6.

From Chapter 2 of Jordan book "Introduction to Probabilistic Graphical Models"

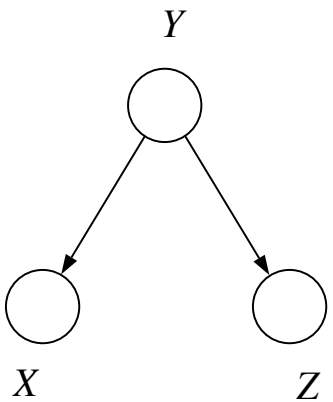
d-separation in three canonical graphs

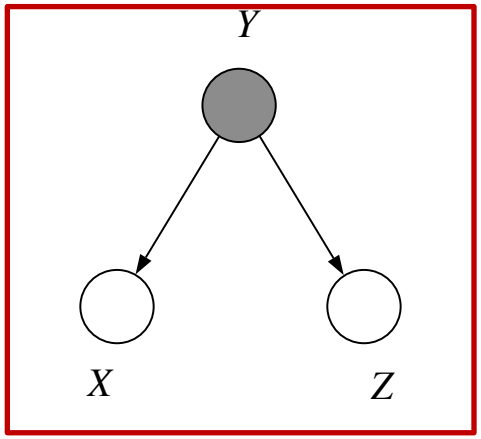




X ⊥ Z | Y

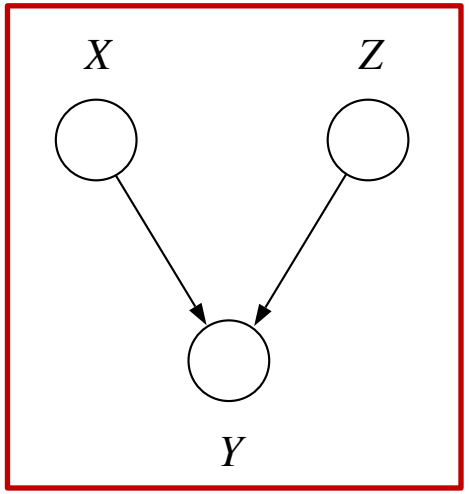
"X and Z are d-separated by the observation of Y."

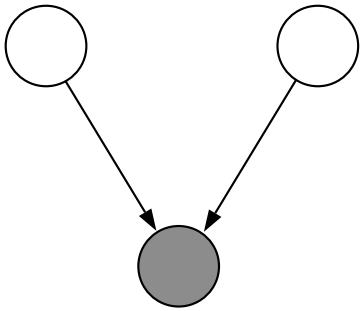




X ⊥ Z | Y

"X and Z are d-separated by the observation of Y."

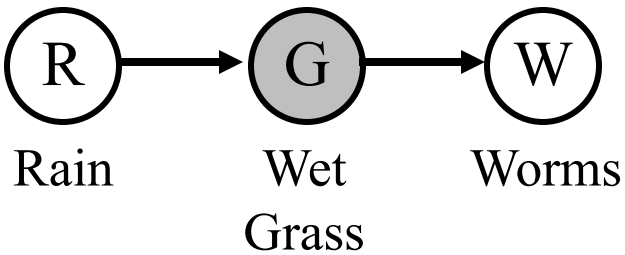




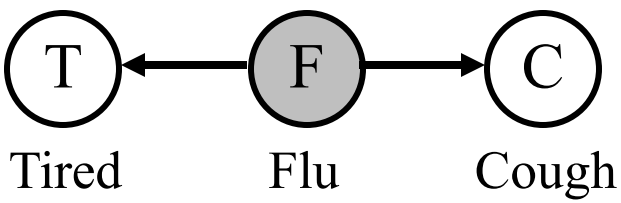
X ⊥ Z | Y

"X and Z are d-separated by NOT observing Y nor any descendants of Y."

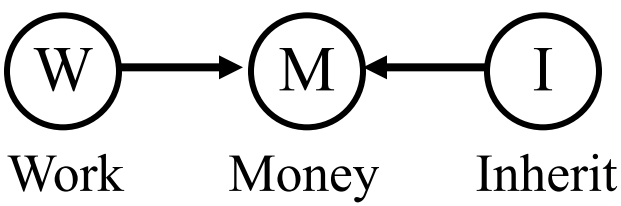
Examples



$$P(W \mid R, G) = P(W \mid G)$$



$$P(T \mid C, F) = P(T \mid F)$$



$$P(W \mid I, M) \neq P(W \mid M)$$

$$P(W \mid I) = P(W)$$

The Bayes Ball algorithm

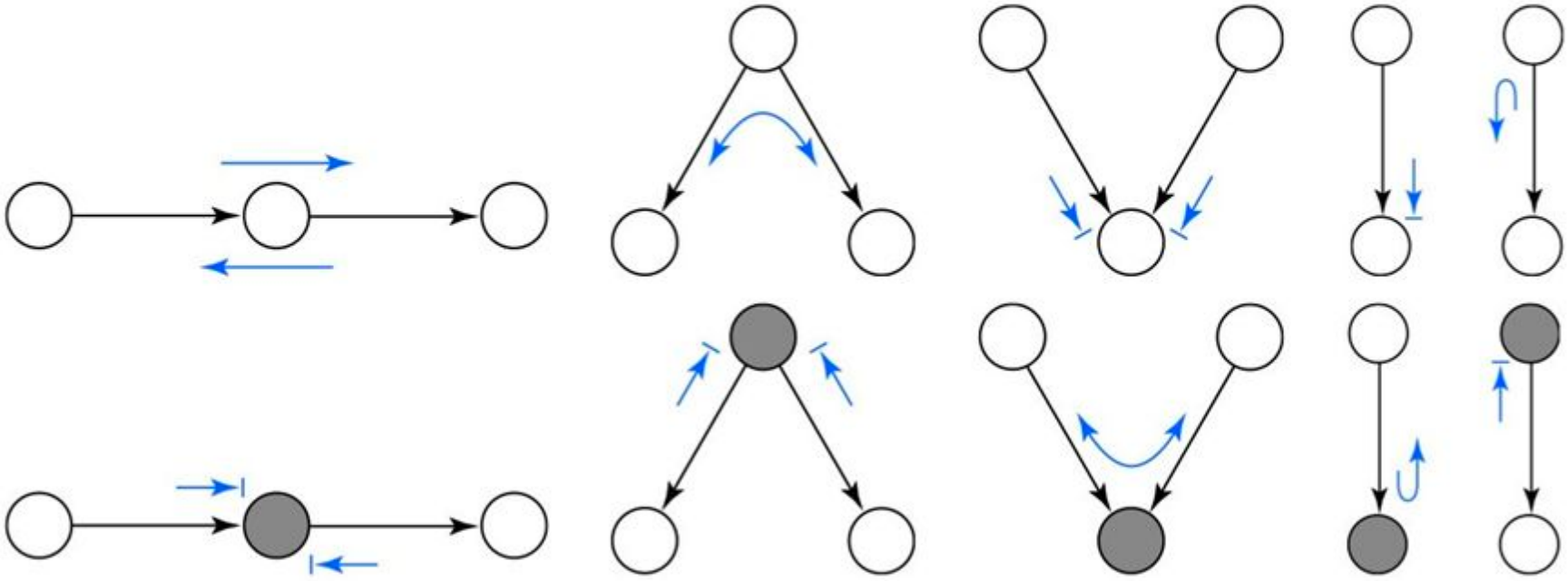
- Let **X**, **Y**, **Z** be "***groups***" of nodes / set / subgraphs.

- Shade nodes in **Y**
- Place a "ball" at each node in **X**
- Bounce balls around the graph according to **rules**

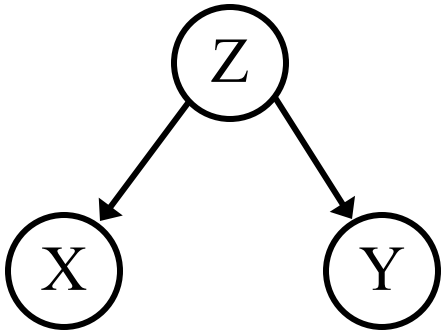
- If no ball reaches any node in **Z**, then declare

$$\mathbf{X} \perp \mathbf{Z} \mid \mathbf{Y}$$

The Ten Rules of Bayes Ball Algorithm



Examples



X – wet grass

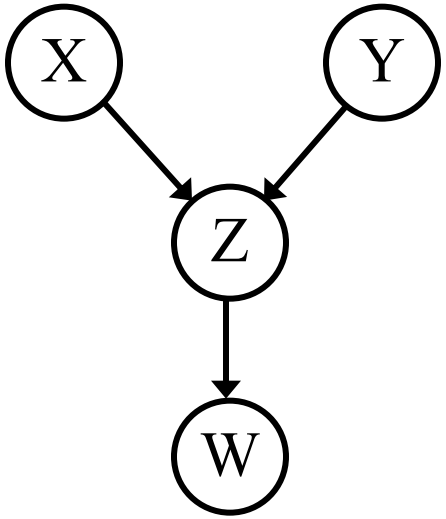
Y – rainbow

Z – rain

$P(X, Y) \neq P(X)\,P(Y)$

$P(X \mid Y, Z) = P(X \mid Z)$

Are X and Y ind.? Are X and Y cond. ind. given…?



X – rain

Y – sprinkler

Z – wet grass

W – worms

$P(X, Y) = P(X)\,P(Y)$

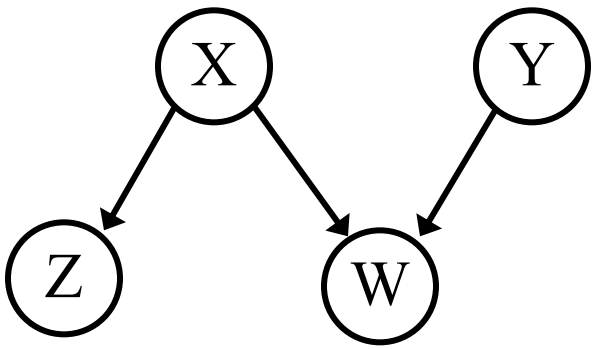
$P(X \mid Y, Z) \neq P(X \mid Z)$

$P(X \mid Y, W) \neq P(X \mid W)$

Examples

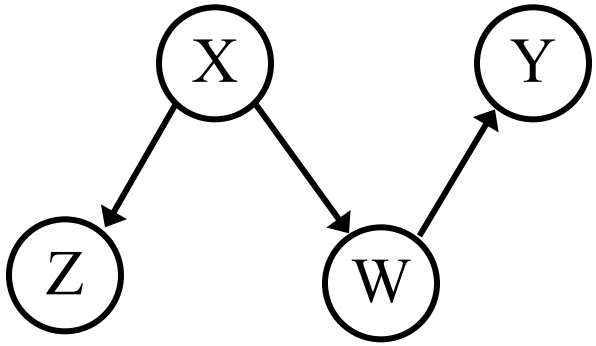
Are X and Y independent?

Are X and Y conditionally independent given Z?



X – rain
Y – sprinkler
Z – rainbow
W – wet grass

$P(X,Y) = P(X)\,P(Y)$ Yes
$P(X \mid Y, Z) = P(X \mid Z)$ Yes



X – rain
Y – sprinkler
Z – rainbow
W – wet grass

$P(X,Y) \neq P(X)\,P(Y)$ No
$P(X \mid Y, Z) \neq P(X \mid Z)$ No

Conditional Independence

- Where are conditional independences here?

Radio and Ignition, given Battery?

Yes

Radio and Starts, given Ignition?

Yes

Gas and Radio, given Battery?

Yes

Gas and Radio, given Starts?

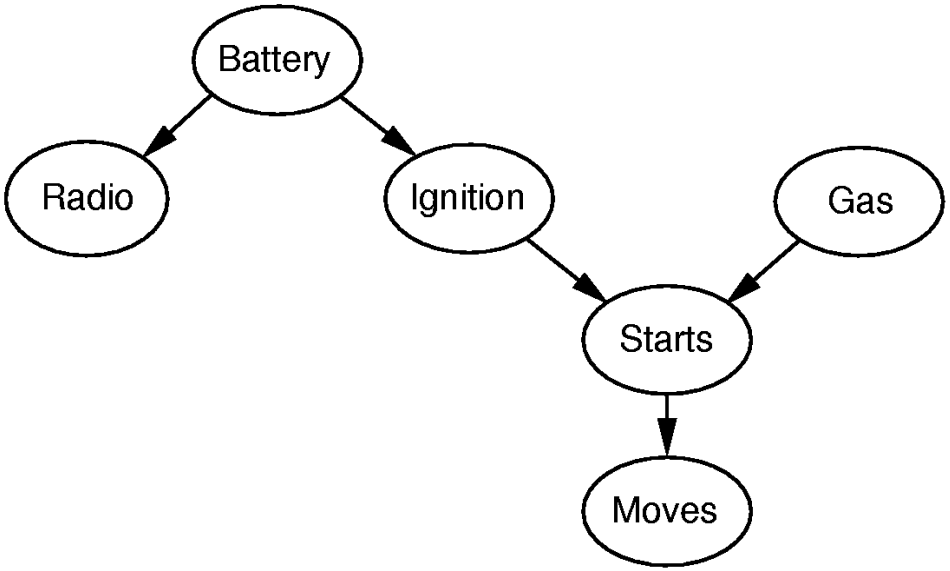
No

Gas and Radio, given nil?

Yes

Gas and Battery, given Moves?

No (why?)



Summary of the today

- Encode knowledge / structures using a DAG
- How to check conditional independence algebraically by the factorizations?
- How to read off conditional independences from a DAG
  - d-separation, Bayes Ball algorithm

(**More examples in supplementary slides, and in the discussion: Hidden Markov Models, AIMA 15.3**)

Additional resources about PGM

- Recommended: Ch.2 Jordan book. AIMA Ch. 13-14.

- More readings:
  - Koller's PGM book: https://www.amazon.com/Probabilistic-Graphical-Models-Daphne-Koller/dp/B007YXTT12
  - Probabilistic programming: http://probabilistic-programming.org/wiki/Home

- Software for PGMs and modeling and inference:
  - Stan: https://mc-stan.org/
  - JAGS: http://mcmc-jags.sourceforge.net/

Upcoming lectures

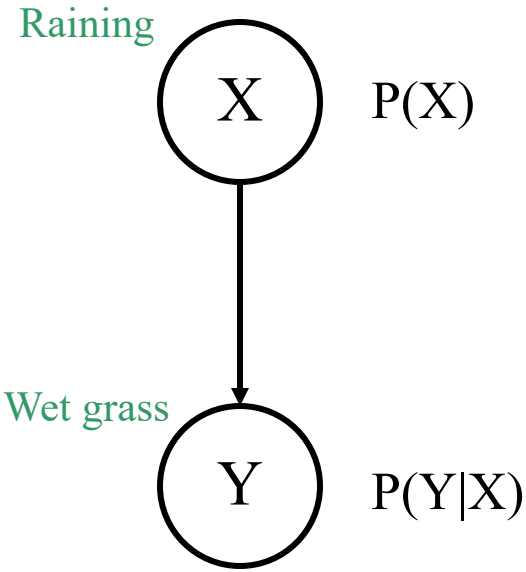
- Jan 28: Finish Graphical models. Start "Search"
- Jan 30: Search algorithms
- Feb 4: Minimax search and game playing
- Feb 6: Finish "search" + Midterm review. HW2 Due.

- Recommended readings on search:
  - AIMA Ch 3, Ch 5.1-5.3.

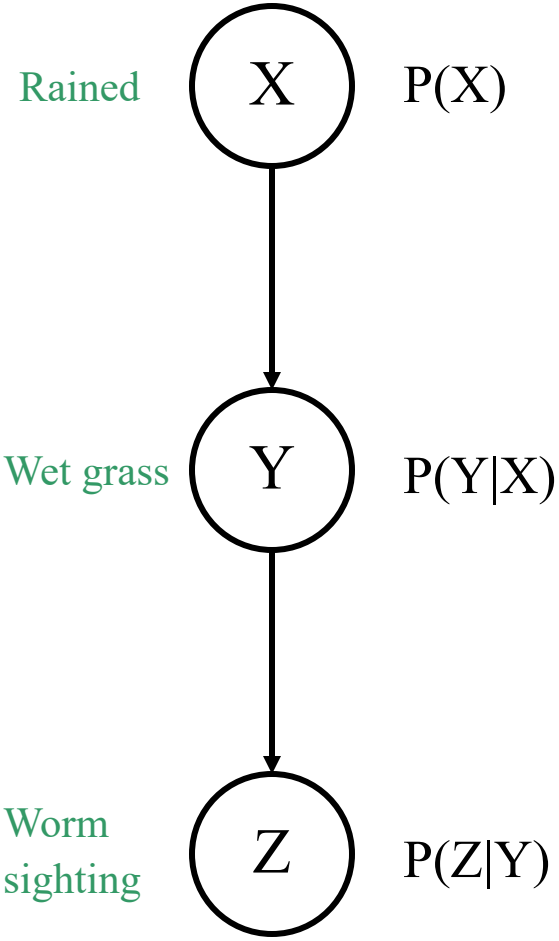
---------- Supplementary Slides ----------

- More examples
- Practice questions

Representing any inference as CPTs



ASK P(X|Y)



ASK P(X|Z)

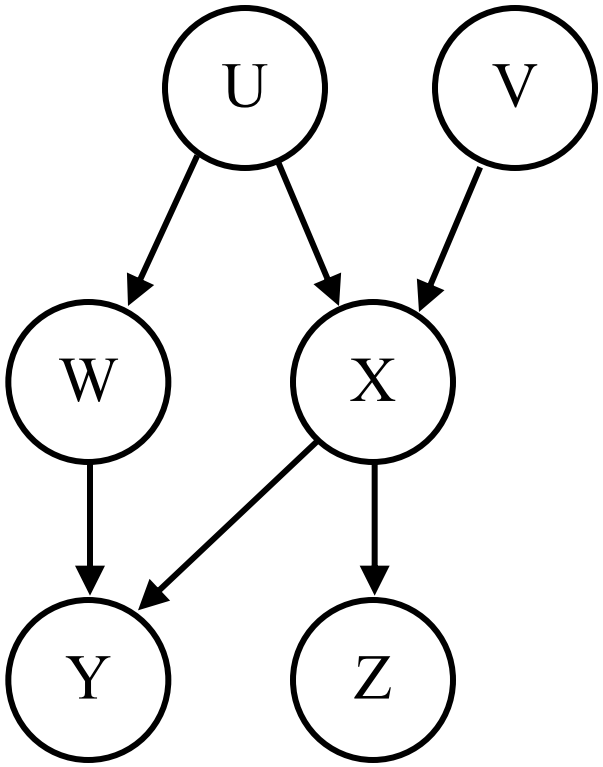
Quiz

1. What is the joint probability distribution of the random variables described by this belief net?
   - I.e., what is P(U, V, W, X, Y, Z)?

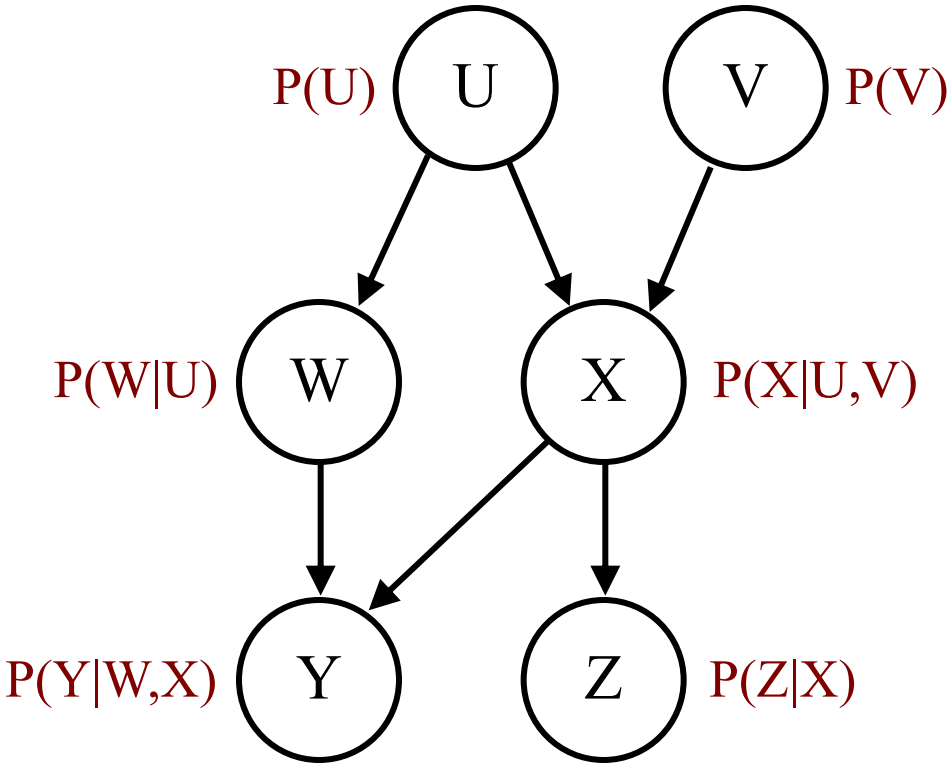
2. Variables W and X are
   a) Independent
   b) Independent given U
   c) Independent given Y

   (choose one)

3. If you know the CPTs, is it possible to compute P(Z | U)? P(U | Z)?



Review



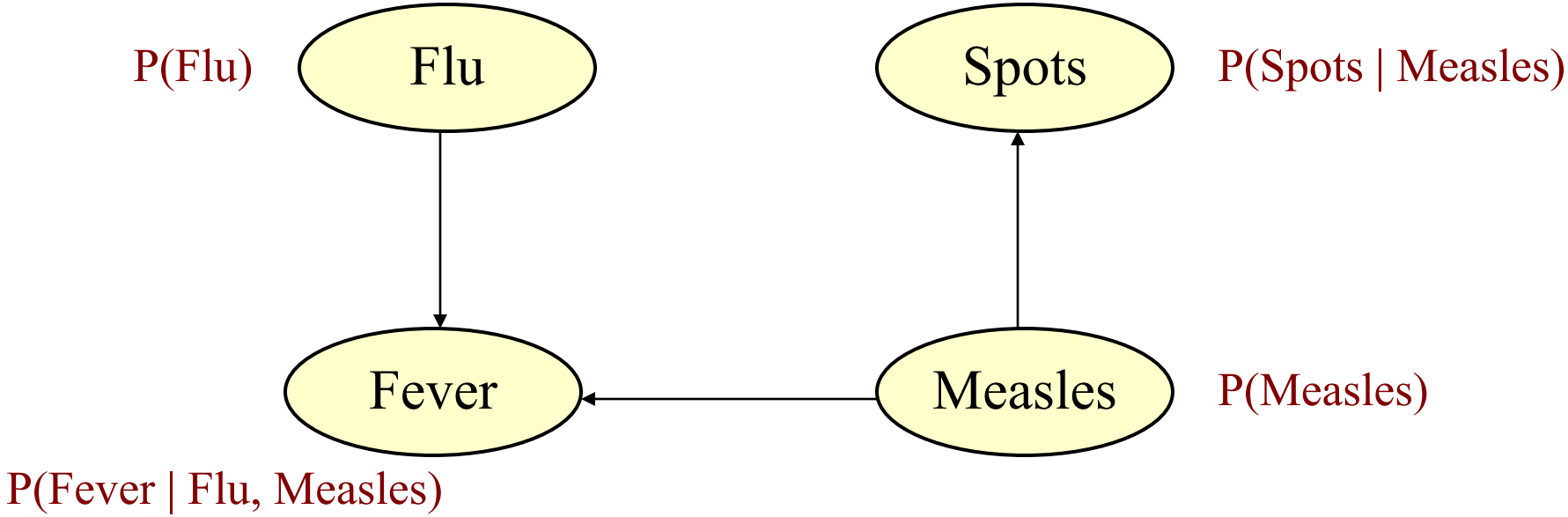
Given this Bayesian network:

1. What are the CPTs?
2. What is the joint probability distribution of all the variables?
3. How would we calculate P(X | W, Y, Z)?

P(U,V,W,X,Y,Z) = product of the CPTs

= P(U) P(V) P(W|U) P(X|U,V) P(Y|W,X) P(Z|X)

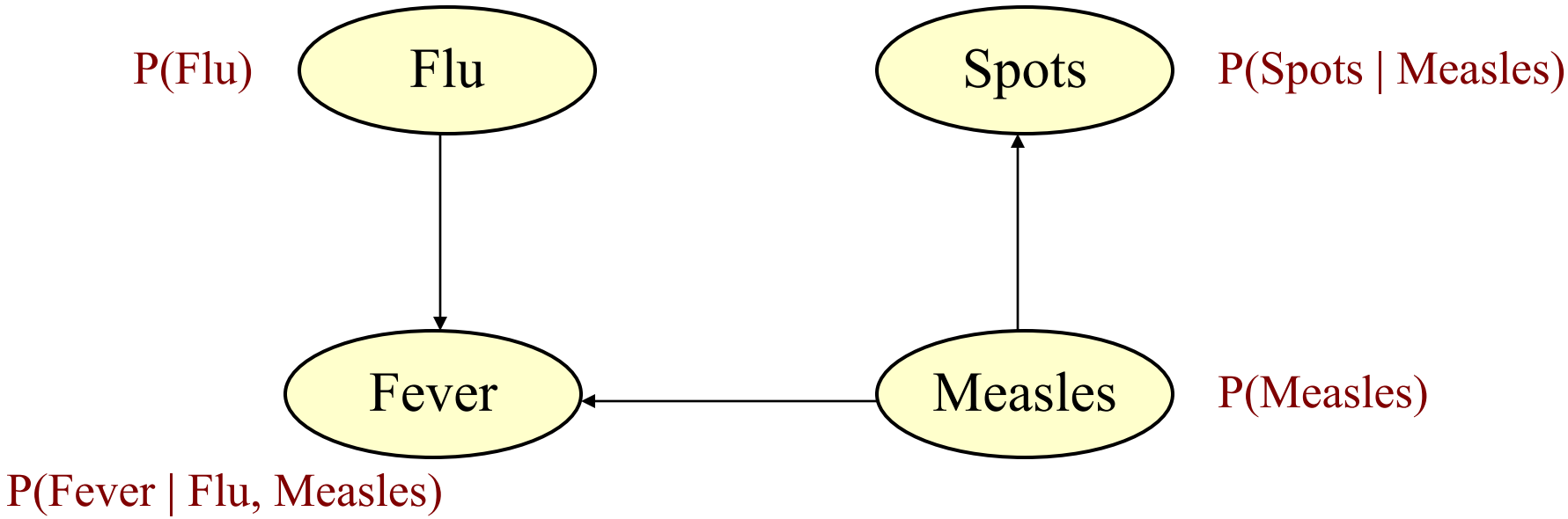
Example: Flu and measles



To create the belief net:

- Choose variables (evidence and query)
- Choose an ordering and create links (direct influences)
- Fill in probabilities (CPTs)

Example: Flu and measles



CPTs: P(Flu) = 0.01 P(Spots | Measles) = [0, 0.9]

P(Measles) = 0.001 P(Fever | Flu, Measles) = [0.01, 0.8, 0.9, 1.0]

Compute P(Flu | Fever) and P(Flu | Fever, Spots).
Are they equivalent?

Practical uses of Belief Nets

- Uses of belief nets:
  - Calculating probabilities of causes, given symptoms
  - Calculating probabilities of symptoms, given multiple causes
  - Calculating probabilities to be combined with an agent's utility function
  - Explaining the results of probabilistic inference to the user
  - Deciding which additional evidence variables should be observed in order to gain useful information
    - Is it worth the extra cost to observe additional evidence?
      - $P(X \mid Y)$ vs. $P(X \mid Y, Z)$
  - Performing sensitivity analysis to understand which aspects of the model affect the queries the most
    - How accurate must various evidence variables be? How will input errors affect the reasoning?