Bayesian Models

Dr. Collin F. Lynch

AI Academy: North Carolina State University

AI Academy

NC STATE
UNIVERSITY

Agenda

Complex Models

Conditioning

Chaining

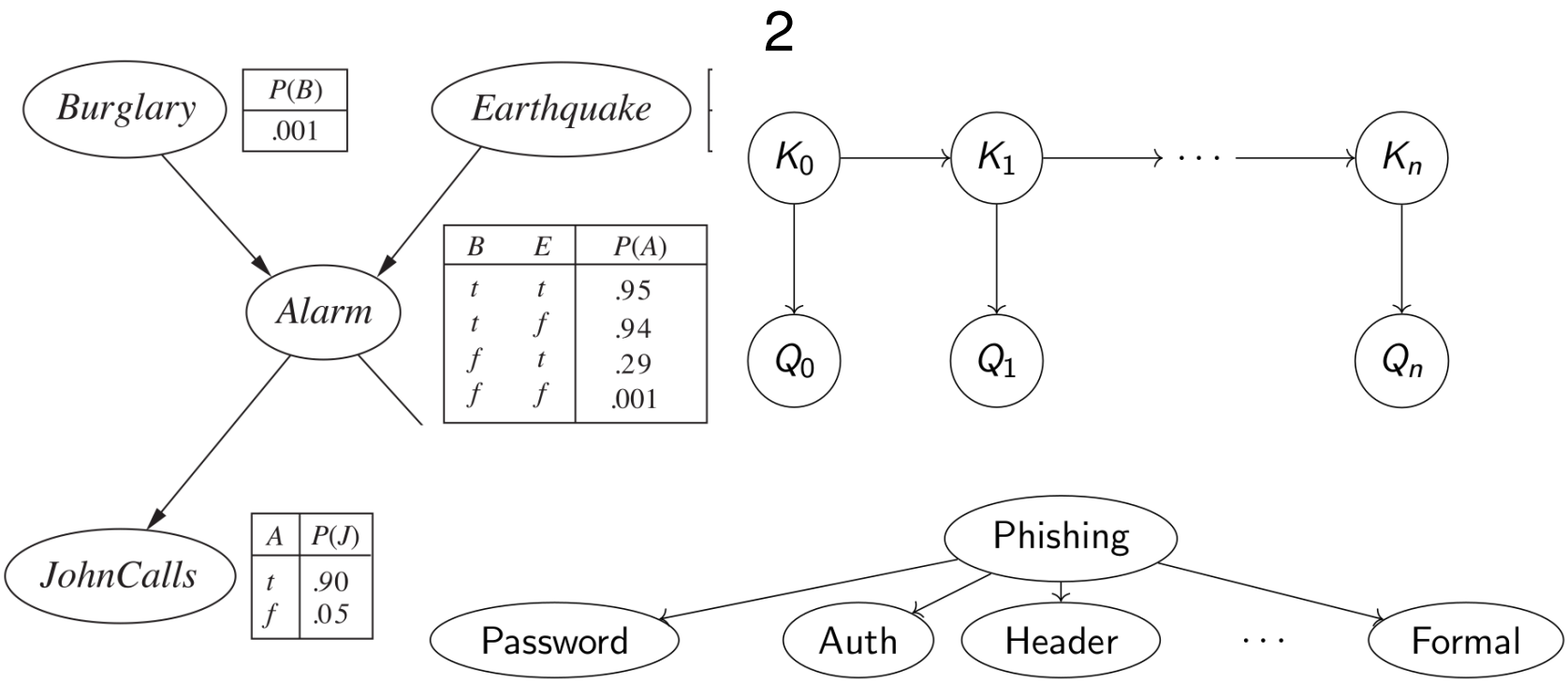
Queries

Independence

Complex Models

What would you do with probability?

Bayes' Nets


Burglary
P(B)
.001
Earthquake
Alarm
B E P(A)
t t .95
t f .94
f t .29
f f .001
JohnCalls
A P(J)
t .90
f .05
2
K_0
K_1
K_n
Q_0
Q_1
Q_n
Phishing
Password
Auth
Header
Formal

Distributions

EAT	$p(EAT)$
EAT=1	.001
EAT=2	.003
EAT=3	.014
EAT=4	.006
EAT=5	.008
EAT=6	.002
EAT=7	.093
EAT=8	.873
Total	1.000

Distributions

EAT	$p(EAT)$
EAT=1	.001
EAT=2	.003
EAT=3	.014
EAT=4	.006
EAT=5	.008
EAT=6	.002
EAT=7	.093
EAT=8	.873
Total	1.000

E	A	T	$p(E, A, T)$
t	t	t	.001
t	t	f	.003
t	f	t	.014
t	f	f	.006
f	t	t	.008
f	t	f	.002
f	f	t	.093
f	f	f	.873
Total			1.000

Marginalizing Joint Distributions

E	A	T	$p(E, A, T)$
t	t	t	.001
t	t	f	.003
t	f	t	.014
t	f	f	.006
f	t	t	.008
f	t	f	.002
f	f	t	.093
f	f	f	.873
			1.000

$$P(E) = P(E = t) = \sum_{A,T} P(E, A, T)$$
$$= .001 + .003 + .014 + .006 = .024$$

$$P(A) = .001 + .003 + .008 + .002 = .014$$
$$P(T) = .001 + .014 + .008 + .093 = .116$$

$$P(\neg T) = P(T = f) = 1 - P(T)$$
$$= .884 = .003 + .006 + .002 + .873$$

Marginalization (2)

$P(A, T) = \sum_E P(E, A, T)$

A	T	$p(A, T)$
t	t	.001 + .008 = .009
t	f	.003 + .002 = .005
f	t	.014 + .093 = .107
f	f	.006 + .873 = .879
		1.000

$P(A|T) = \frac{p(A,T)}{p(T)}$

A	T	$P(A\|T)$
t	t	.009/.116 = .078
t	f	.005/.884 = .006
f	t	.107/.116 = .922
f	f	.879/.884 = .994

So what?

- Given a full joint table we can calculate any query.
- Full joint tables are large $\Omega(2^n)$.
- Getting exact probabilities is *hard!*.

Conditioning

Bayes' Rule

$$p(A|B) = \frac{p(B|A) * p(A)}{p(B)}$$

Bayes' Rule

$$p(A|B) = \frac{p(B|A) * p(A)}{p(B)}$$

$$p(disease|symptom) = \frac{p(symptom|disease) * p(disease)}{p(symptom)}$$

Conditioning Rule

$$
\begin{aligned}
p(A) &= (p(A|B) * p(B)) + (p(A|\neg B) * p(\neg B)) && (1)\\
&= p(A \wedge B) + p(A \wedge \neg B) && (2)
\end{aligned}
$$

Independence

A and B are independent if any of the following hold:

- $p(A, B) = p(A) * p(B)$
- $p(A|B) = p(A)$
- $p(B|A) = p(B)$

Thus you gain nothing by knowing about B or A relative to the other.

Conditional Independence

A and B are conditionally independent given C if any of the following hold:

- $p(A, B|C) = p(A|C) * p(B|C)$
- $p(A|B, C) = p(A|C)$
- $p(B|A, C) = p(B|C)$

Thus knowing C means we gain no information about A from B or vice-versa.

Examples

- Is your car *red*?
- Do you have a *toothache*?

Examples

- Is your car *red*?
- Do you have a *toothache*?
- Do you have a spot on an *X-ray*?
- Do you have a *Cavity*?

Combining Evidence

- We can now make predictions from information.

$$p(C|T,X) = \frac{p(T,X|C) * p(C)}{p(T,X)}$$

- Assuming that T and X are CI given C.

$$p(C|T,X) = \frac{p(T|C) * p(X|C) * p(C)}{p(T,X)}$$

- We can calculate this and even account for the *normalizing factor* $p(T,X)$.

$$\begin{aligned} p(C|T,X) + p(\neg C|T,X) &= 1 \\ \frac{p(T|C) * p(X|C) * p(C)}{p(T,X)} + \frac{p(T|\neg C) * p(X|\neg C) * p(\neg C)}{p(T,X)} &= 1 \\ p(T|C) * p(X|C) * p(C) + p(T|\neg C) * p(X|\neg C) * p(\neg C) &= p(T,X) \end{aligned}$$

Chaining

Full joint distributions

C	F	M	V	p
0	0	0	0	0.1
0	0	0	1	0.00612
0	0	1	0	0.02
0	0	1	1	0.08
0	1	0	0	0.012
0	1	0	1	0.0098
1	0	0	0	0.06
1	0	0	1	0.04
1	0	1	0	0.16
1	0	1	1	0.094
1	1	0	0	0.112
1	1	0	1	0.088
1	1	1	0	0.12
1	1	1	1	0.00988

Full joint distributions

C	F	M	V	p
0	0	0	0	0.1
0	0	0	1	0.00612
0	0	1	0	0.02
0	0	1	1	0.08
0	1	0	0	0.012
0	1	0	1	0.0098
1	0	0	0	0.06
1	0	0	1	0.04
1	0	1	0	0.16
1	0	1	1	0.094
1	1	0	0	0.112
1	1	0	1	0.088
1	1	1	0	0.12
1	1	1	1	0.00988

$$p(x_0, \ldots, x_n) = \prod_{i=0}^{n} p(x_i | parents(X_i))$$

The Chain Rule

- Given a set of variables we can define an arbitrary ordering over them.
- We then calculate the probabilities using the *chain rule* shown above.
- This yields a set of probability tables with a graph representation.

Full joint distributions

C	F	M	V	p
0	0	0	0	0.1
0	0	0	1	0.00612
0	0	1	0	0.02
0	0	1	1	0.08
0	1	0	0	0.012
0	1	0	1	0.0098
1	0	0	0	0.06
1	0	0	1	0.04
1	0	1	0	0.16
1	0	1	1	0.094
1	1	0	0	0.112
1	1	0	1	0.088
1	1	1	0	0.12
1	1	1	1	0.00988

$$p(C, F, V, M)$$
$$= p(C)p(F|C)p(V|F, C)p(M|V, F, C)$$

Independence(2)

- What if F is independent of C?

$$\begin{aligned} p(C,F,V,M) &= p(C)\ p(F|C)\ p(V|F,C)\ p(M|V,F,C) \\ &= p(C)\ p(F)\ p(V|F,C)\ p(M|V,F,C) \\ O(2^{10}) &\rightarrow O(2^9) \end{aligned}$$

- Or what if M is conditionally independent of C and F given V?

$$\begin{aligned} p(C,F,V,M) &= p(C)\ p(F|C)\ p(V|F,C)\ p(M|V,F,C) \\ &= p(C)\ p(F|C)\ p(V|F,C)\ p(M|V) \\ O(2^{10}) &\rightarrow O(2^8) \end{aligned}$$

Why does this matter?

- Remember full joint tables are exponential in the number of variables.
- Even if a lot of the entries don't matter.
- And even with the chain rule calculation of probabilities is exponential *irrespective of ordering*.

Why does this matter?

- Remember full joint tables are exponential in the number of variables.
- Even if a lot of the entries don't matter.
- And even with the chain rule calculation of probabilities is exponential *irrespective of ordering*.
- But! In some cases conditionally independent variables can just be removed.
- Reductions through CI can thus reduce computation time or even allow us to *isolate* variables entirely.
- Real problems usually have lots of CI

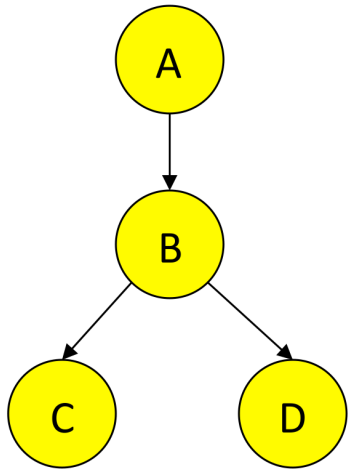
Bayesian Networks

- A probabilistic graphical model that represents a set of **random variables** and their **conditional dependencies** via a **directed acyclic graph** (dag).
- A DAG is a collection of vertices (random vars) and directed edges (the conditional dependencies).
- Each edge connects one vertex to another such that there is no way to start at at some vertex v and follow a path that loops back to v again.
- This is the *graph form of the chain rule*.

Bayesian Networks

A Bayesian network is made up of:

1. A Directed Acyclic Graph



2. A set of tables for each node in the graph

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

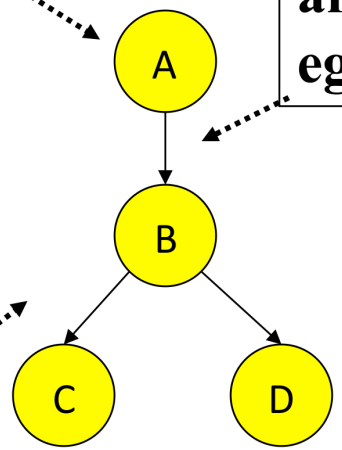
B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

Directed Acyclic Graph

Each node in the graph is a random variable

A node X is a parent of another node Y if there is an arrow from node X to node Y eg. A is a parent of B



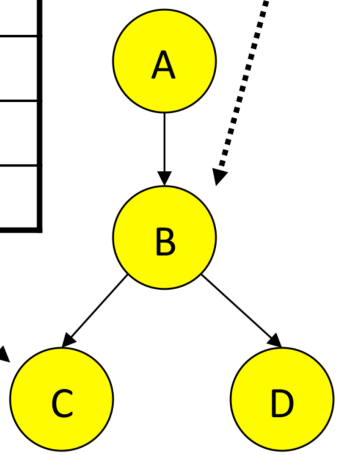
Informally, an arrow from node X to node Y means X has a direct influence on Y

Tables for each node

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1



B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

Each node X_i has a conditional probability distribution P(X_i | Parents(X_i)) that quantifies the effect of the parents on the node

The parameters are the probabilities in these conditional probability tables (CPTs)

Tables (2)

Conditional Probability
Distribution for C given B

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

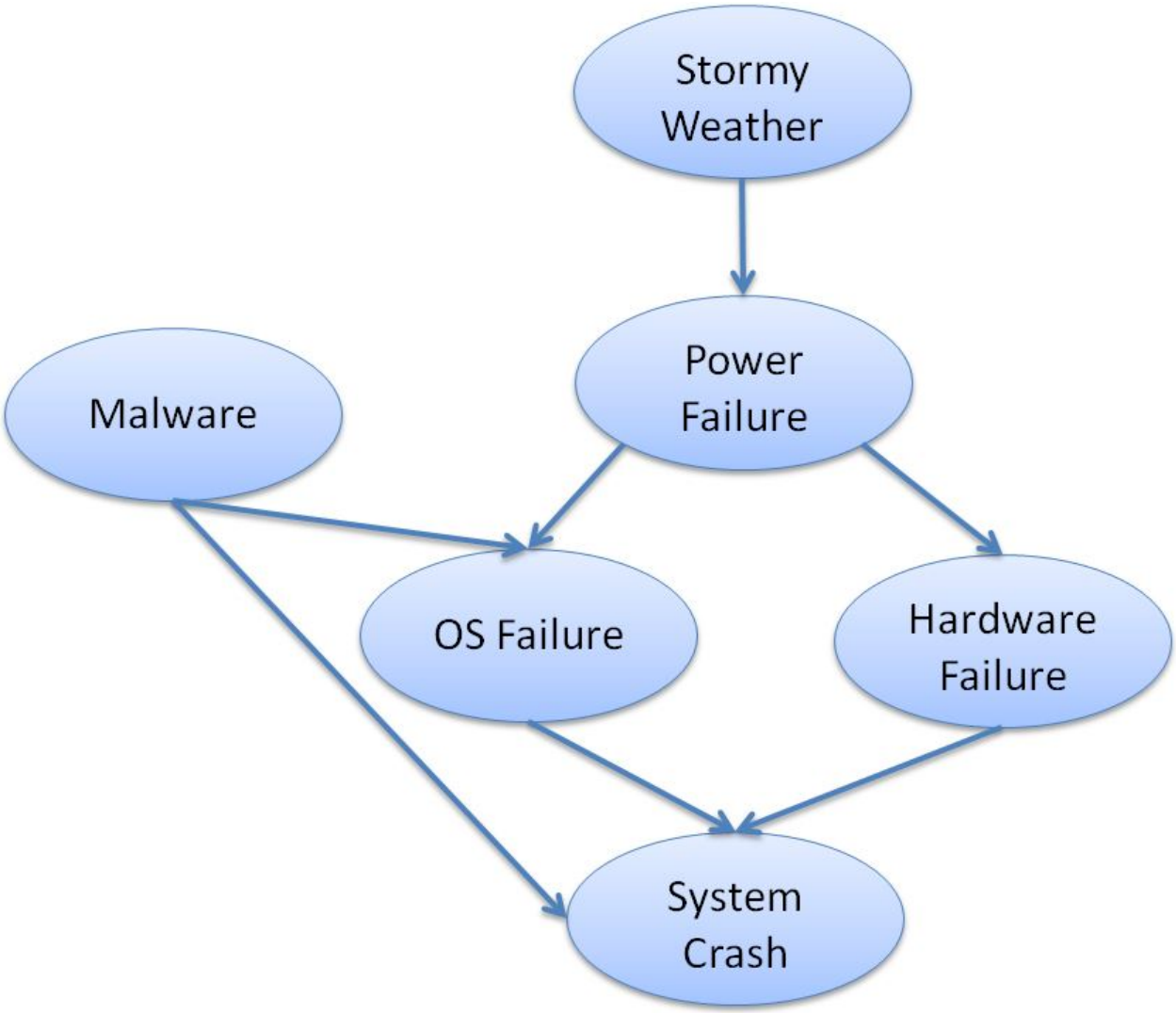
For a given combination of values of the parents (B in this example), the entries for P(C=true | B) and P(C=false | B) must add up to 1
eg. P(C=true | B=false) + P(C=false | B=false)=1

If you have a Boolean variable with k Boolean parents, this table has 2^{k+1} probabilities (but only 2^k need to be stored)

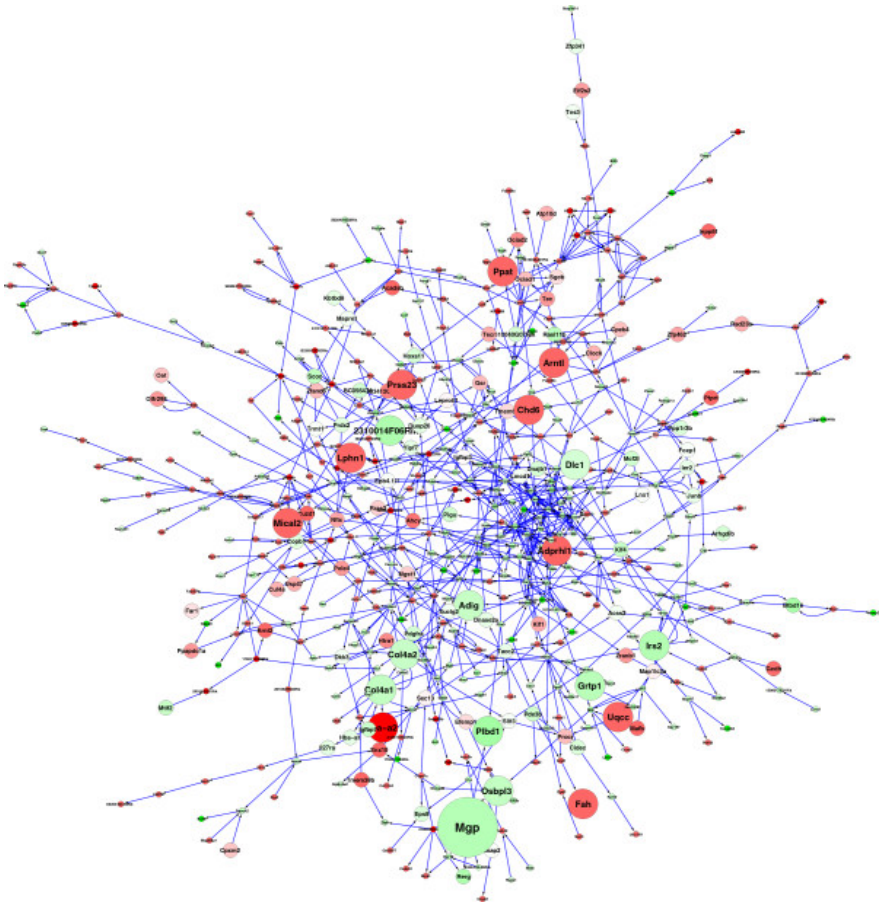
Key Points

- A BN encodes the *conditional independence relationships* between variables in the graph structure.
- It does **not necessarily** encode the **causal relationships** between variables.
- With suitable tables we can represent Sunny $\rightarrow$ Hot or Hot $\rightarrow$ Sunny.
- The representation is *compact* over the variables relative to a full table.
- The missing arrows *signal* conditional independencies.
- What makes it useful is the *independence*.

Diagnosis


Stormy Weather
Power Failure
Malware
OS Failure
Hardware Failure
System Crash

Modeling



Lionikas et al. (2012) “Resolving candidate genes of mouse skeletal muscle QTL via RNA-Seq and expression network analyses” BMC Genomics 13(1):592

Queries

Calculating Probabilities

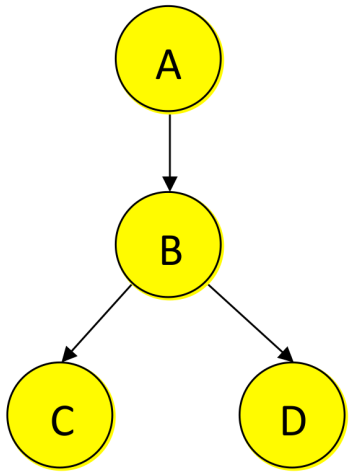
Using the network in the example, suppose you want to calculate:

$$P(A = true, B = true, C = true, D = true)$$

$$= P(A = true) * P(B = true \mid A = true) *$$

$$P(C = true \mid B = true)\, P(D = true \mid B = true)$$

$$= (0.4)*(0.3)*(0.1)*(0.95)$$



A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

Calculating Prob (2)

Using the network in the example, suppose you want to calculate:

$$P(A = true, B = true, C = true, D = true)$$

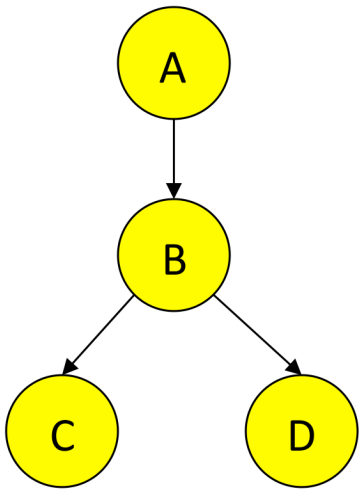
$$= P(A = true) * P(B = true \mid A = true) *$$

$$P(C = true \mid B = true)\, P(D = true \mid B = true)$$

$$= (0.4)*(0.3)*(0.1)*(0.95)$$

This is from the graph structure

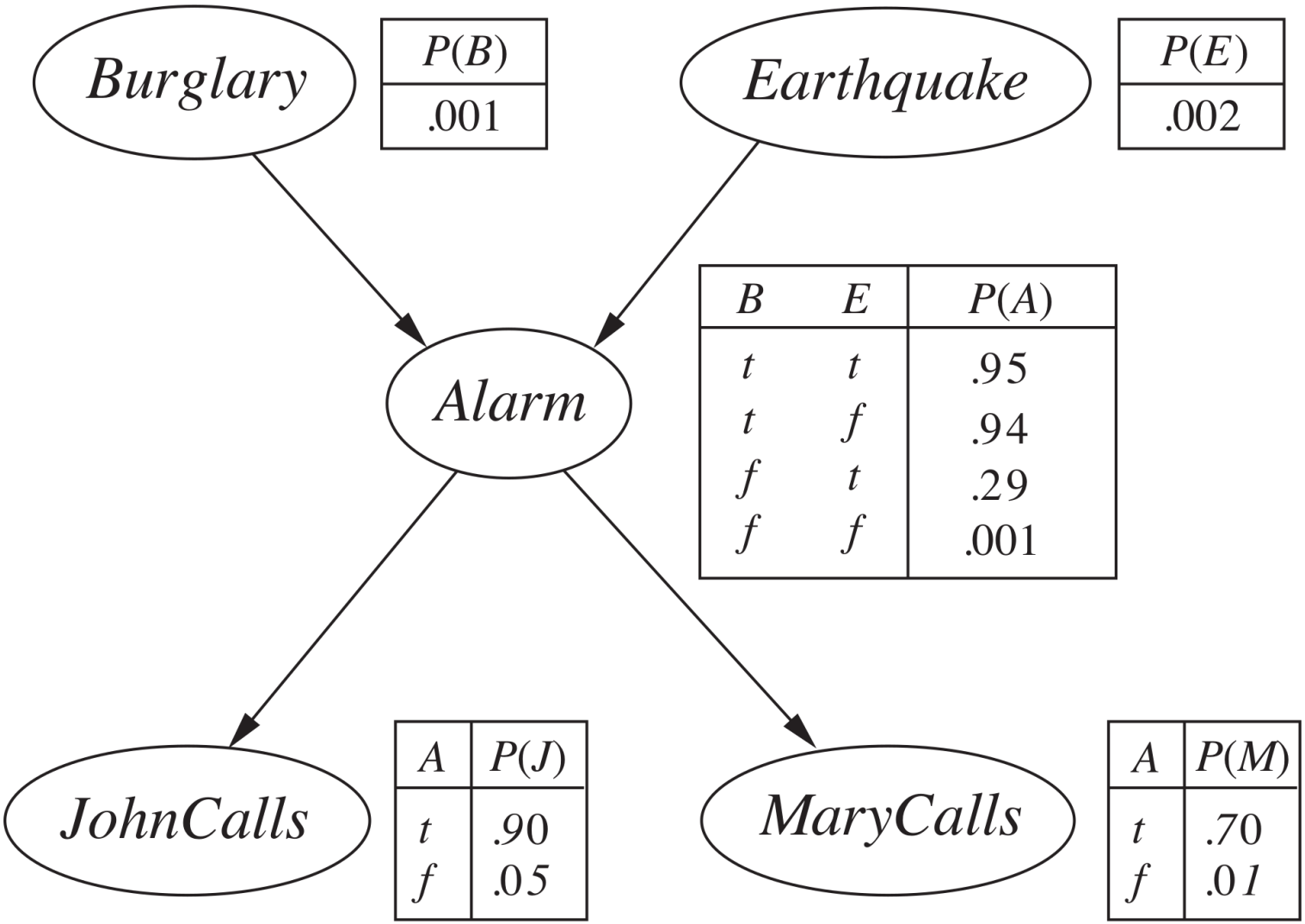
These numbers are from the conditional probability tables



Broad Inference in Bayes Nets

- Bayes nets represent probability distributions.
- Basically When new evidence comes in:
  - Update the distributions of the variables.
  - Propagate the information along the network to other nodes.
- *How* we do that is the trick.

Example Network



Naïve Inference

1. Suppose that we want to calculate the odds of some variable (e.g. Mary Calls) if there was no earthquake.

2. Separate nodes in the network into:
   - Unknown *Query Variables* (X)
   - Fixed *Evidence Variables* (V)
   - The remaining *Hidden Variables* (Y)

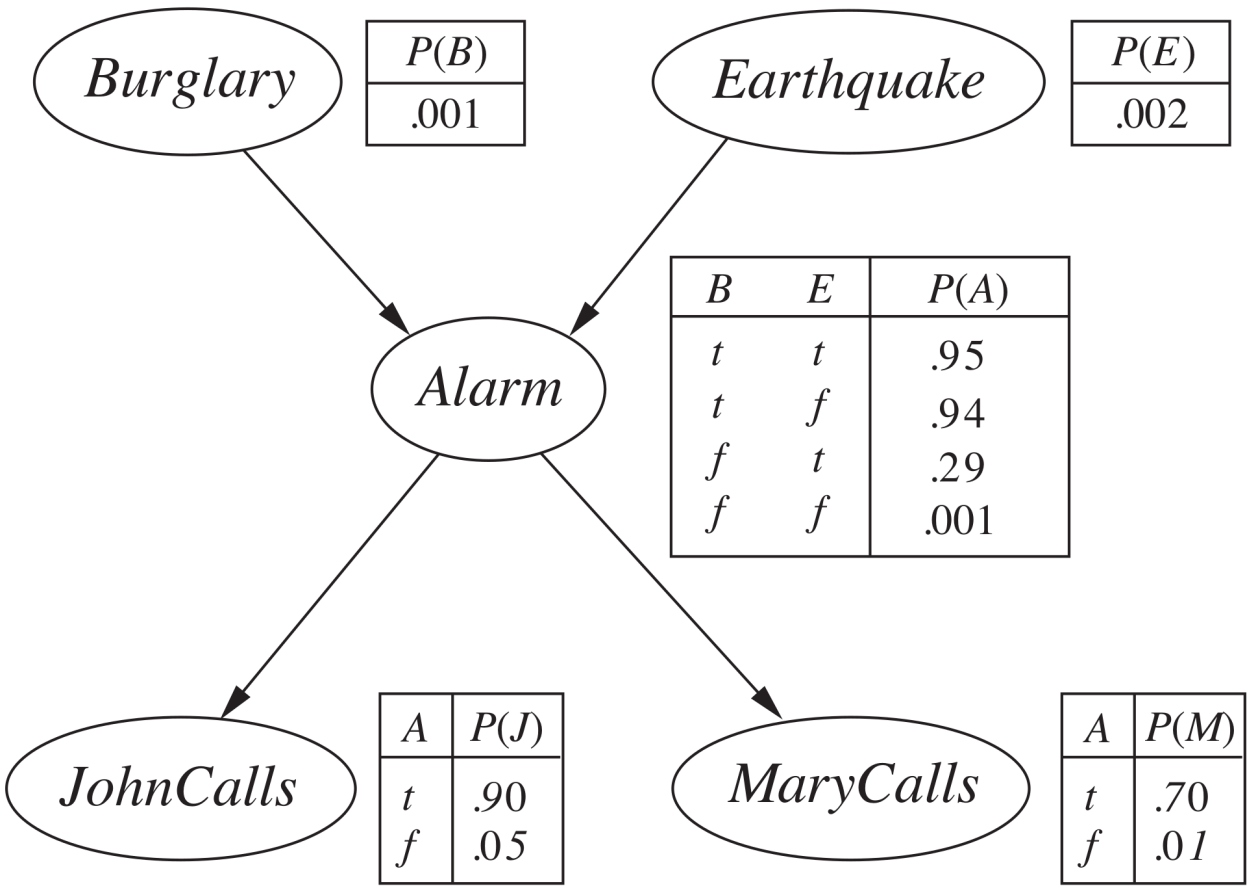
3. Compute the joint probability distribution using the chain rule:

$$p(B, E = f, A, J, M = t) =$$
$$p(B =?)p(E = f)p(A =?|B =?, E = f)p(J =?|A =?)p(M = t|A =?)$$

4. We have to consider the alternatives:

$$p(B = t)p(E = f)p(A = t|B = t, E = f)p(J = t|A = t)p(M = t|A = t)$$
$$+p(B = f)p(E = f)p(A = t|B = f, E = f)p(J = t|A = t)p(M = t|A = t)$$

Example case



X = {J=t, E=f}; V = {}; Y = {B,A, M}

Example Case (1)

- In order to calculate this we sum over the values of Y.

$$P(J = t, E = f)$$
$$= \sum_{B,A,M} p(B)p(E = f)p(A|B, E = f)p(J = t|A)p(M|A)$$

Example Case (1)

- In order to calculate this we sum over the values of Y.

$$P(J = t, E = f)$$
$$= \sum_{B,A,M} p(B)p(E = f)p(A|B, E = f)p(J = t|A)p(M|A)$$

- We can make the computation more efficient by moving summations inwards:

$$= p(E = f)\left(\sum_B p(B)\left(\sum_A p(A|B, E = f)p(J = t|A)\left(\sum_M p(M|A)\right)\right)\right)$$

Example Case (1)

- In order to calculate this we sum over the values of Y.

$$P(J = t, E = f)$$
$$= \sum_{B,A,M} p(B)p(E = f)p(A|B, E = f)p(J = t|A)p(M|A)$$

- We can make the computation more efficient by moving summations inwards:

$$= p(E = f)\left(\sum_B p(B)\left(\sum_A p(A|B, E = f)p(J = t|A)\left(\sum_M p(M|A)\right)\right)\right)$$

- Hidden leaves can be removed.

$$= p(E = f)\left(\sum_B p(B)\left(\sum_A p(A|B, E = f)p(J = t|A)\right)\right)$$

Example Case (1)

- In order to calculate this we sum over the values of Y.

$$P(J = t, E = f)$$
$$= \sum_{B,A,M} p(B)p(E = f)p(A|B, E = f)p(J = t|A)p(M|A)$$

- We can make the computation more efficient by moving summations inwards:

$$= p(E = f)\left(\sum_B p(B)\left(\sum_A p(A|B, E = f)p(J = t|A)\left(\sum_M p(M|A)\right)\right)\right)$$

- Hidden leaves can be removed.

$$= p(E = f)\left(\sum_B p(B)\left(\sum_A p(A|B, E = f)p(J = t|A)\right)\right)$$

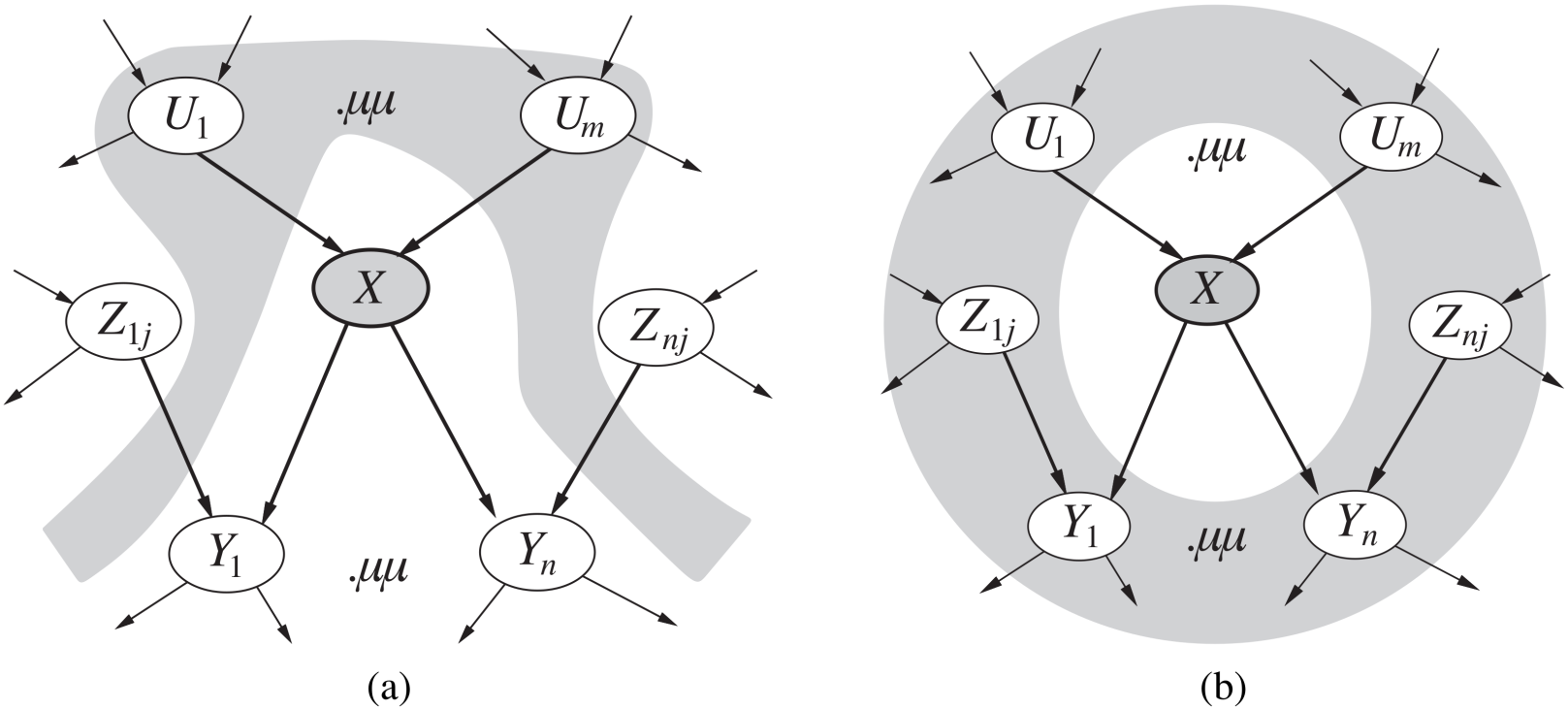
- In any *polytree* (undirected free tree), variable elimination is $O(n)$.

Independence

Causal Markov Assumption

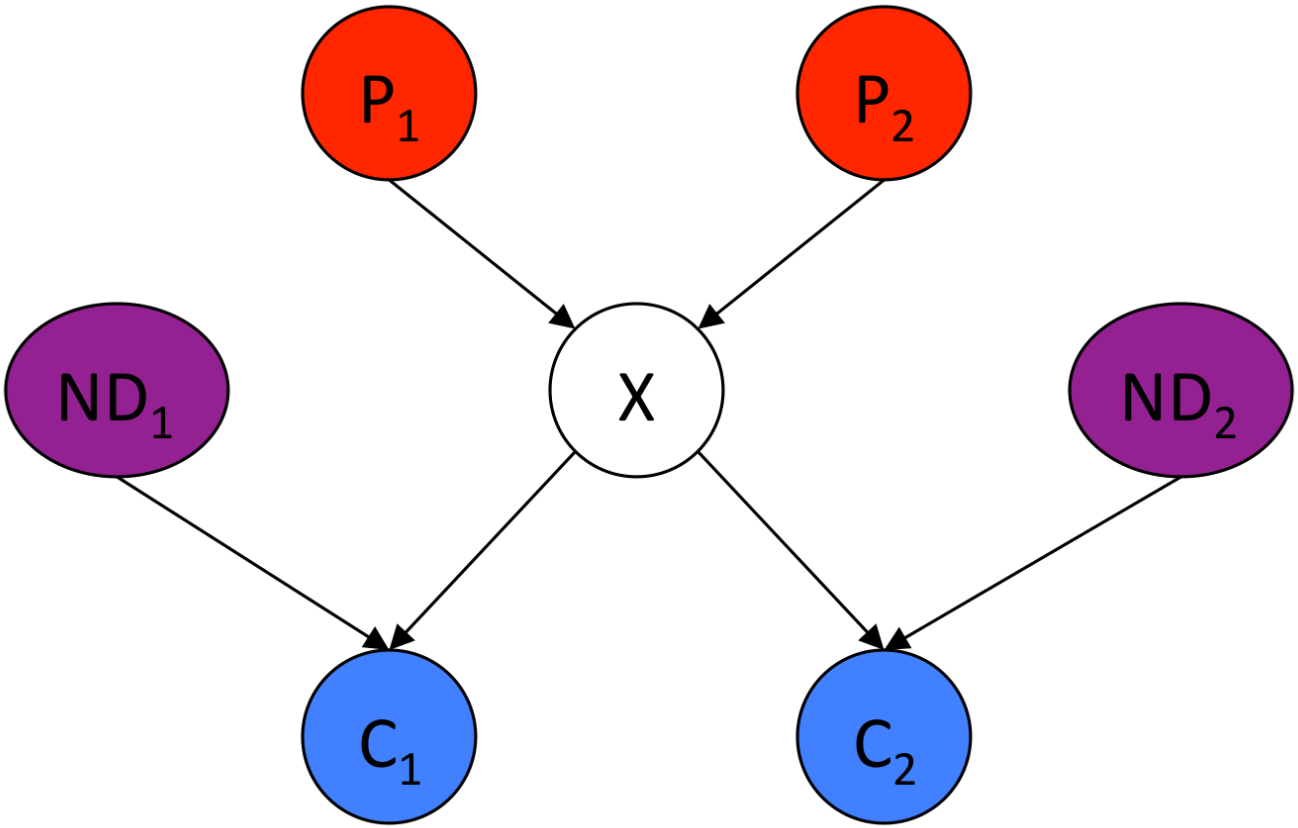
- Bayesian networks encode dependencies *and independencies* between variables.
- Under the *Causal Markov Assumption* each variable in a Bayesian Network is independent of its ancestors given the values of *its parents*.

Graphical Independence



Graphical Independence (2)

The Markov Condition: given its parents (P_1 & P_2) a node (X) is conditionally independent of its non-descendants (ND_1 & ND_2).



Back to the Chain Rule

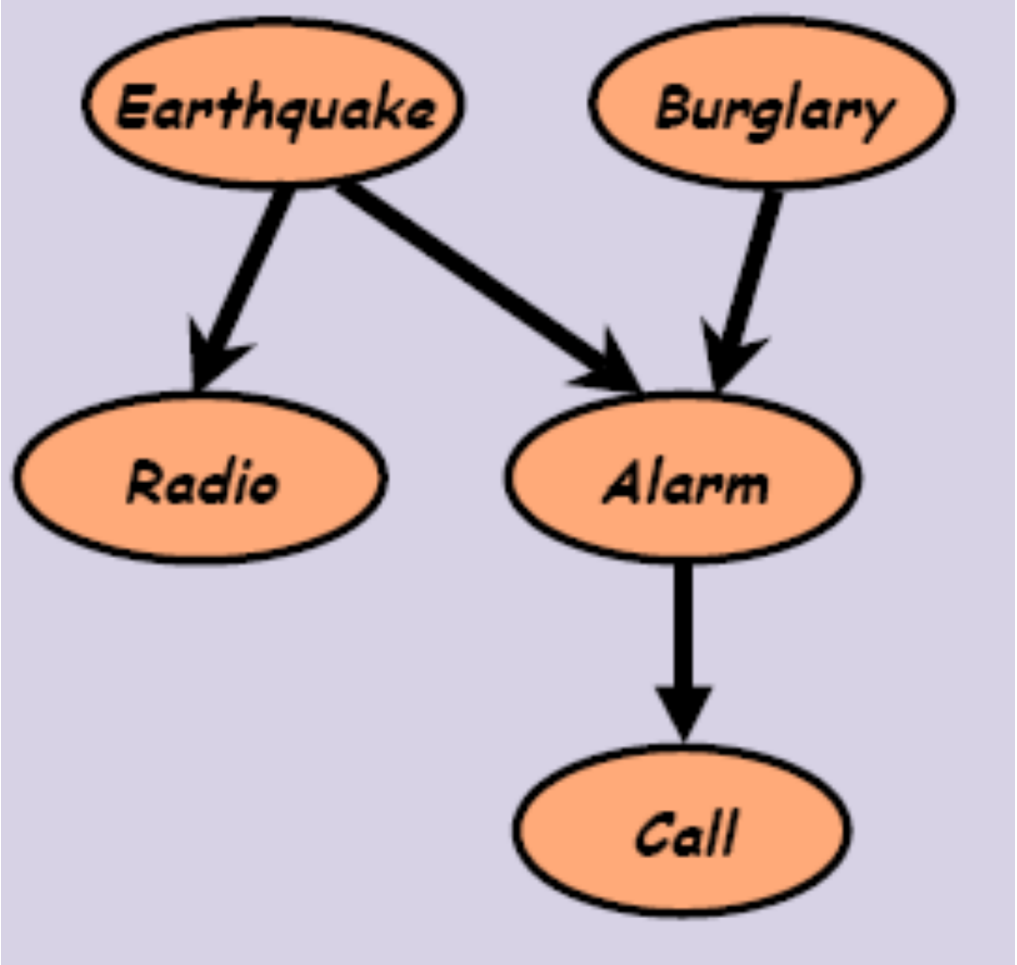
Due to the Markov condition we can compute the full-joint probability distribution using the chain rule:

$$p(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) = \prod_{i=1}^{n} p(X_i = x_i | Parents(X_i)) \quad (3)$$

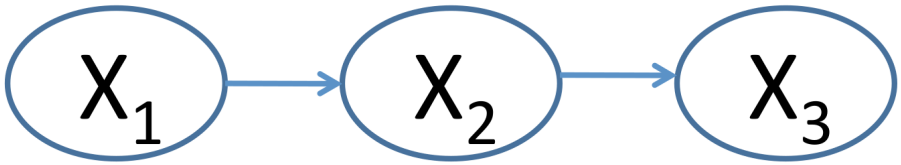
D-Separation

- We can formalize separation as *d-separation*
- Two nodes n & m are d-separated by a set Z if for all undirected paths (P) between them:
  - P contains a directed chain, $u \ldots \leftarrow m \leftarrow \ldots v$ or $u \ldots \rightarrow m \rightarrow \ldots v$, such that the middle node m is in Z,
  - P contains a fork, $u \ldots \leftarrow m \rightarrow \ldots v$, such that the middle node m is in Z, or
  - P contains an inverted fork (or collider), $u \ldots \rightarrow m \leftarrow \ldots v$, such that the middle node m is not in Z and no descendant of m is in Z.
- D-separated nodes are conditionally independent given Z.

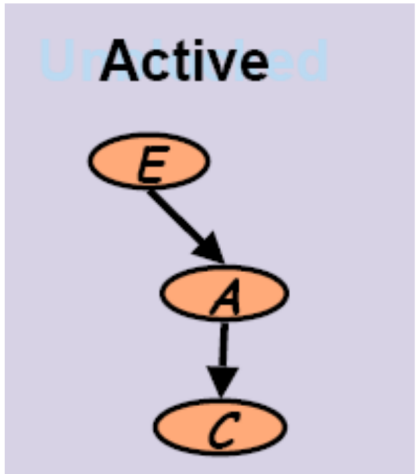
D-Separation Example

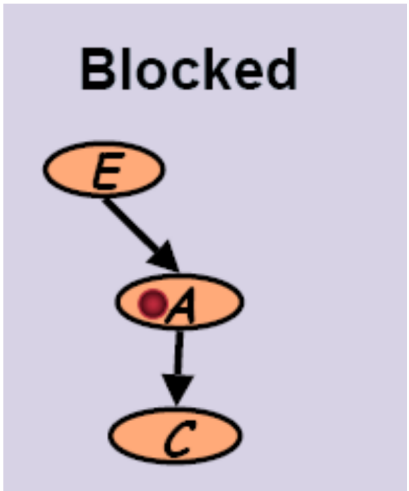


Chain

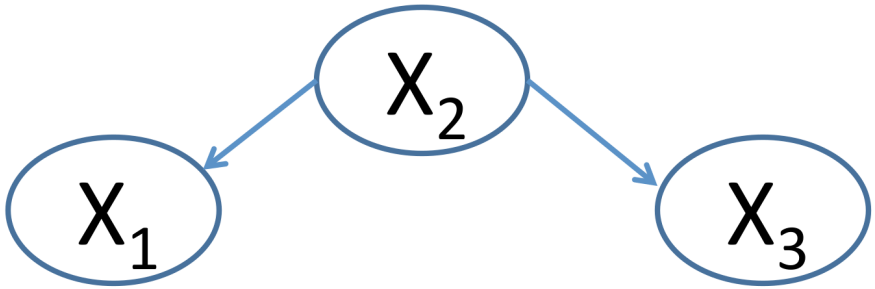


- In a serial connection from *X1 to X3 via X2, evidence from X1 to X3 is* ***blocked*** *only when we have* hard evidence about *X2.*
- Intermediate cause.

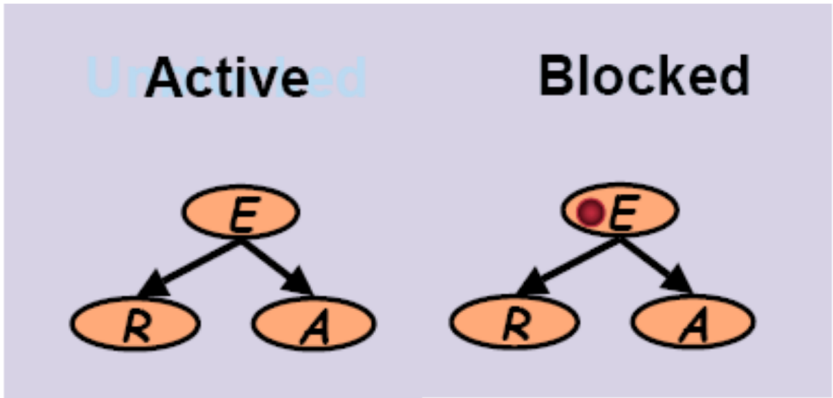




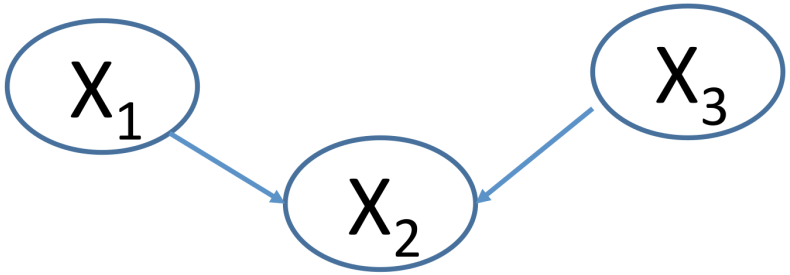
Fork



- In a diverging connection where *X1 and X3 have the common parent X2, evidence from X1 to X3 is* blocked only when we have hard evidence about *X2*.
- Common cause.



Collider (Inverted Fork)



- In a converging connection where *X2 has parents X1 and X3, any evidence about X2 results in* evidence transmitted between *X1 and X3.*
- Common Effect.