CMT311 Principles of Machine Learning

Bayesian Networks

Angelika Kimmig
KimmigA@cardiff.ac.uk

15.11.2019

- Last week:
  - basics of discrete probability
  - Naive Bayes: imposing structure on joint distribution by making strong conditional independence assumptions
- Today:
  - Bayesian Networks: general, graphical representation of conditional independence assumptions
- Later:
  - efficient reasoning with Bayesian networks, learning Bayesian networks from data

days described by

< Sky, AirTemp, Humidity, Wind, Water, Forecast > + label enjoy

- Sky: *Sunny / Cloudy / Rainy*
- AirTemp: *Warm / Cold*
- Humidity: *Normal / High*
- Wind: *Strong / Weak*
- Water: *Warm / Cool*
- Forecast: *Same / Change*
- enjoy: *0 / 1*

full joint distribution: 192 parameters

Sky	AirTemp	Humidity	Wind	Water	Forecast	Enjoy	P(ω)
Sunny	Warm	Normal	Strong	Warm	Same	0	0.0007875
Sunny	Warm	Normal	Strong	Warm	Same	1	0.00648
Sunny	Warm	Normal	Strong	Warm	Change	0	0.0070875
Sunny	Warm	Normal	Strong	Warm	Change	1	0.00648
Sunny	Warm	Normal	Strong	Cool	Same	0	0.0018375
Sunny	Warm	Normal	Strong	Cool	Same	1	0.00432
Sunny	Warm	Normal	Strong	Cool	Change	0	0.0165375
Sunny	Warm	Normal	Strong	Cool	Change	1	0.00432
Sunny	Warm	Normal	Weak	Warm	Same	0	0.0003375
			…				
			…				
Rainy	Cold	High	Weak	Cool	Change	1	0.00448

days described by

< Sky, AirTemp, Humidity, Wind, Water, Forecast > + label enjoy

Sunny / Cloudy / Rainy

Warm / Cold

Normal / High

Strong / Weak

Warm / Cool

Same / Change

0 / 1

Let's assume the attributes are independent given the label:

$$P(S,A,H,Wi,Wa,F,E) = P(S\,|\,E)P(A\,|\,E)P(H\,|\,E)P(Wi\,|\,E)P(Wa\,|\,E)P(F\,|\,E)P(E)$$

E=0	E=1
10/20	10/20

P(S|E)	S=Sunny	S=Cloudy	S=Rainy
E=0	3/10	3/10	4/10
E=1	4/10	2/10	4/10

P(A|E)	A=Warm	A=Cold
E=0	5/10	5/10
E=1	6/10	4/10

P(H|E)	H=Normal	H=High
E=0	6/10	4/10
E=1	5/10	5/10

P(Wi|E)	Wi=Strong	Wi=Weak
E=0	7/10	3/10
E=1	3/10	7/10

P(Wa|E)	Wa=Warm	Wa=Cool
E=0	3/10	7/10
E=1	6/10	4/10

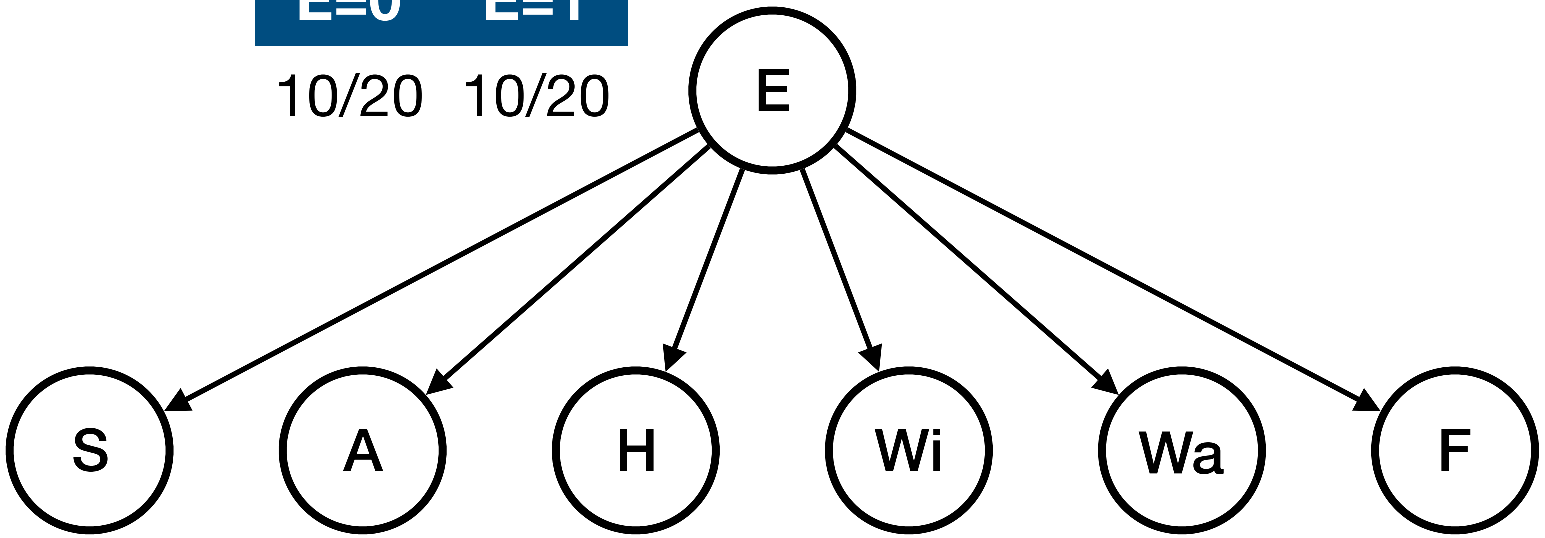
P(F|E)	F=Same	F=Change
E=0	1/10	9/10
E=1	5/10	5/10

exploiting conditional independence: 28 parameters

Let's assume the attributes are independent given the label:

$$P(S, A, H, Wi, Wa, F, E) = P(S \mid E)P(A \mid E)P(H \mid E)P(Wi \mid E)P(Wa \mid E)P(F \mid E)P(E)$$

E=0	E=1
10/20	10/20



P(S|E)	S=Sunny	S=Cloudy	S=Rainy
E=0	3/10	3/10	4/10
E=1	4/10	2/10	4/10

P(Wi|E)	Wi=Strong	Wi=Weak
E=0	7/10	3/10
E=1	3/10	7/10

P(A|E)	A=Warm	A=Cold
E=0	5/10	5/10
E=1	6/10	4/10

P(Wa|E)	Wa=Warm	Wa=Cool
E=0	3/10	7/10
E=1	6/10	4/10

P(H|E)	H=Normal	H=High
E=0	6/10	4/10
E=1	5/10	5/10

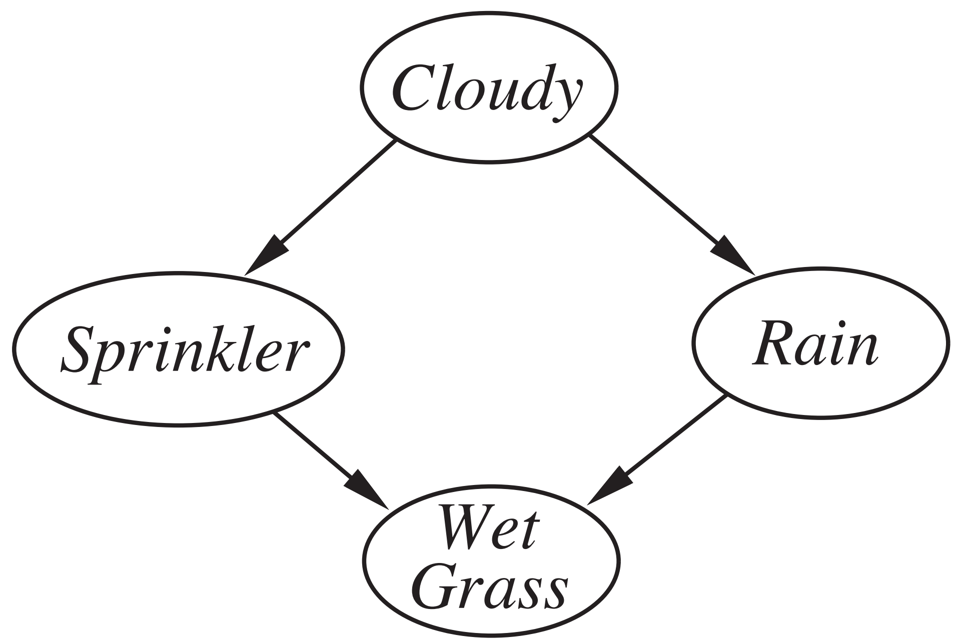
P(F|E)	F=Same	F=Change
E=0	1/10	9/10
E=1	5/10	5/10

Example: Bayesian Network

C=t	C=f
0.5	0.5

$P(C) = .5$

P(R|C)	R=t	R=f
C=t	0.8	0.2
C=f	0.2	0.8



C	$P(S)$
t	.10
f	.50

C	$P(R)$
t	.80
f	.20

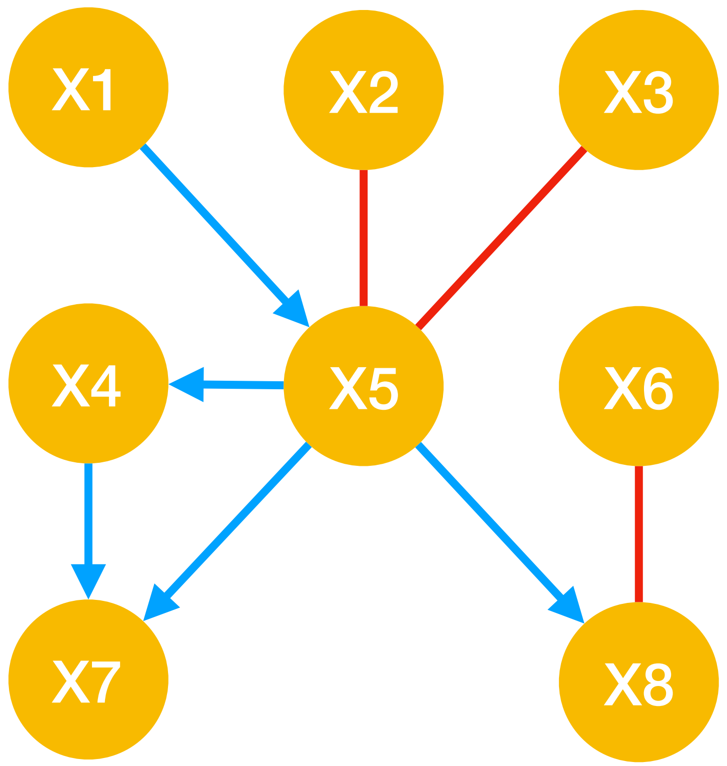
P(S|C)	S=t	S=f
C=t	0.1	0.9
C=f	0.5	0.5

S	R	$P(W)$
t	t	.99
t	f	.90
f	t	.90
f	f	.00

P(W|S,R)		W=t	W=f
S=t	R=t	0.99	0.01
S=t	R=f	0.90	0.10
S=f	R=t	0.90	0.10
S=f	R=f	0.00	1.00

[figure: Russell& Norvig]

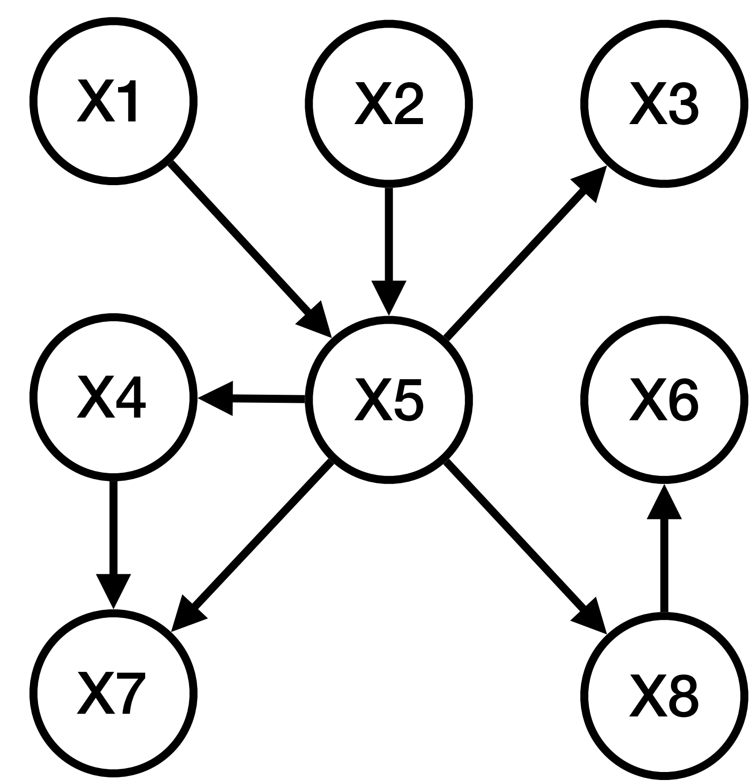
Background: Graphs



- A **graph** consists of **nodes** (vertices) and **directed** or **undirected** **edges**
- **directed graph:** all edges are directed
- **undirected graph:** all edges are undirected
- a **path** from node A to node B is a sequence of nodes connected by edges starting at A and ending at B
- **directed path**: path following the direction of arrows
- **cycle**: directed path that starts and ends at the same node
- **loop**: path with more than 2 nodes that starts and ends at the same node (ignoring edge directions)
- **Directed acyclic graph** (DAG): directed graph with no cycles

Relationships in DAGs

- X is a **parent** of Y if there is a directed edge from X to Y.
- X is a **child** of Y if there is a directed edge from Y to X.
- X is an **ancestor** of Y if there is a directed path from X to Y.
- X is a **descendant** of Y if there is a directed path from Y to X.
- **Markov blanket** of X = parents of X + children of X + parents of children of X (excluding X itself)



Bayesian Networks

Bayesian Network

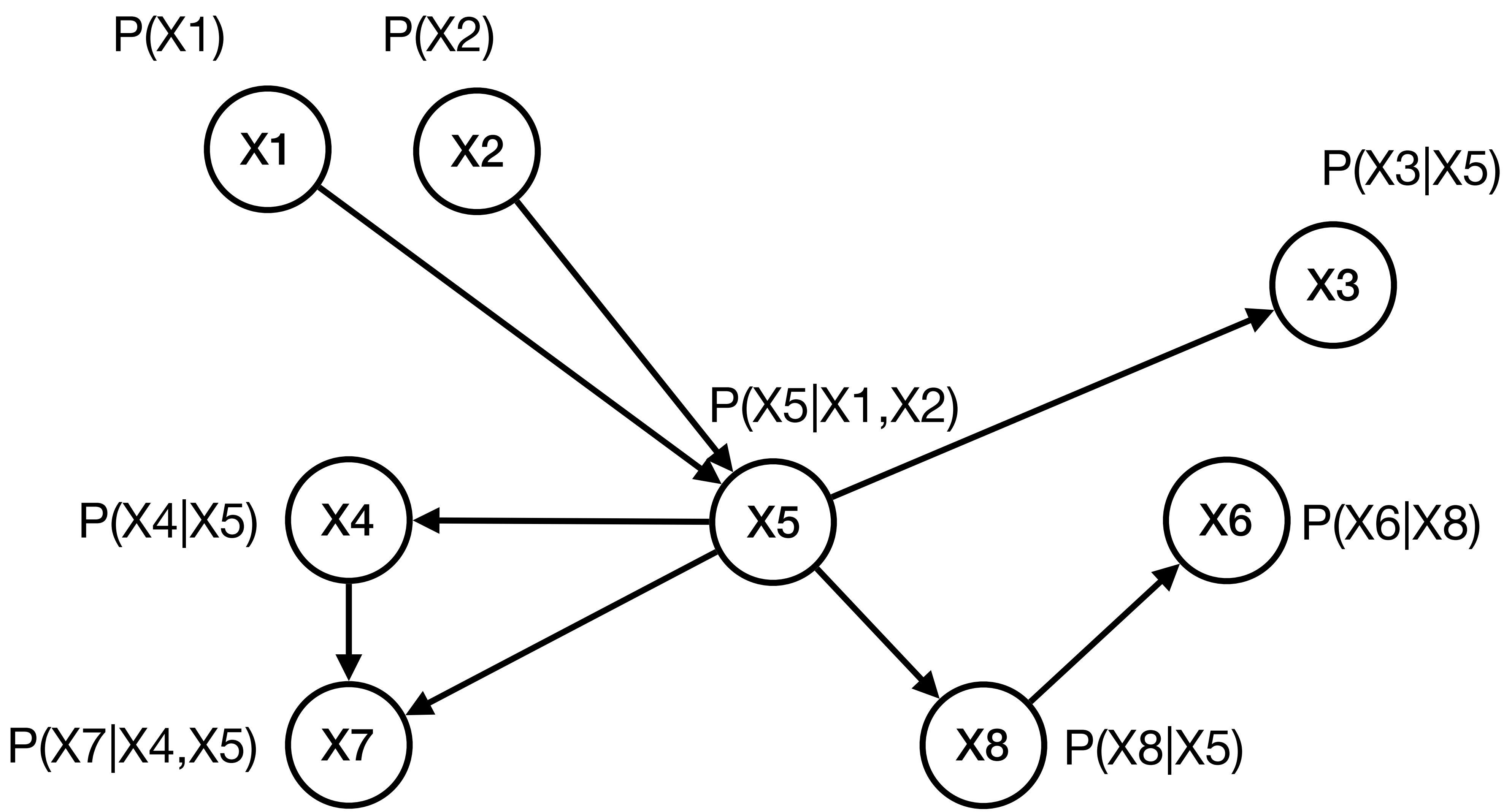
- A **Bayesian network** (BN, also called **belief network**) is a DAG in which each node corresponds to a random variable with an associated conditional probability of the node given its parents.

- Structured factorisation of the joint distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i))$$

- Factors $P(X_i \mid parents(X_i))$ often written as conditional probability tables (CPTs)

Example



$$P(X1,X2,X3,X4,X5,X6,X7,X8)=P(X1)*P(X2)*P(X3|X5)*P(X4|X5)*P(X5|X1,X2)*P(X6|X8)*P(X7|X4,X5)*P(X8|X5)$$

Example

- Sally's burglary **A**larm is sounding. Was there a **B**urglary, or was the alarm triggered by an **E**arthquake? She turns on the **R**adio for news of an earthquake.

- From the chain rule:

$$
\begin{aligned}
P(A, R, E, B) &= P(A \mid R, E, B) \cdot P(R, E, B) \\
&= P(A \mid R, E, B) \cdot P(R \mid E, B) \cdot P(E, B) \\
&= P(A \mid R, E, B) \cdot P(R \mid E, B) \cdot P(E \mid B) \cdot P(B)
\end{aligned}
$$

Example

$$P(A, R, E, B) = P(A \mid R, E, B) \cdot P(R \mid E, B) \cdot P(E \mid B) \cdot P(B)$$

Assumptions:

- the alarm does not directly depend on reports on the radio: $P(A \mid R, E, B) = P(A \mid E, B)$

- reports on the radio do not directly depend on burglaries: $P(R \mid E, B) = P(R \mid E)$

- earthquakes do not directly depend on burglaries: $P(E \mid B) = P(E)$

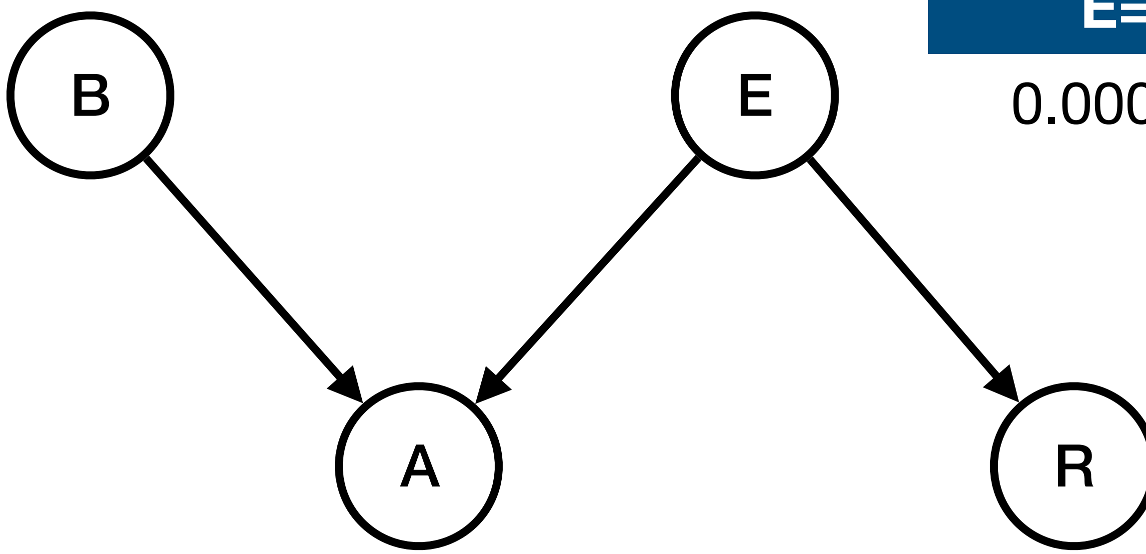
$$P(A, R, E, B) = P(A \mid E, B) \cdot P(R \mid E) \cdot P(E) \cdot P(B)$$

Example

$$P(A, R, E, B) = P(A \mid E, B) \cdot P(R \mid E) \cdot P(E) \cdot P(B)$$

B=1
0.01

E=1
0.000001

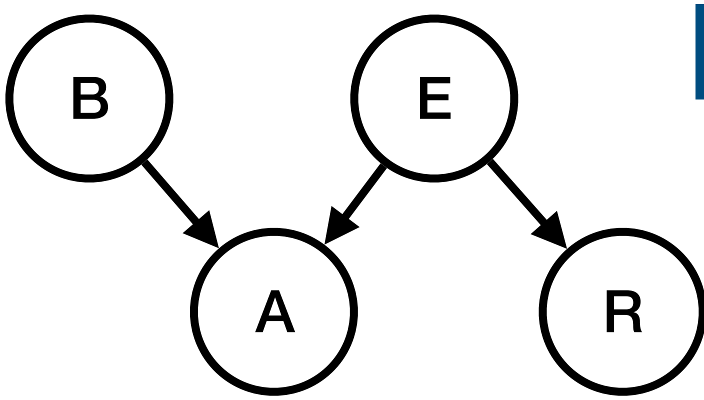


P(A|B,E)		A=1
B=1	E=1	0.9999
B=1	E=0	0.99
B=0	E=1	0.99
B=0	E=0	0.0001

P(R|E)	R=1
E=1	1
E=0	0

Example

B=1
0.01



E=1
0.000001

P(A|B,E)		A=1
B=1	E=1	0.9999
B=1	E=0	0.99
B=0	E=1	0.99
B=0	E=0	0.0001

P(R|E)	R=1
E=1	1
E=0	0

What is the probability that there was a burglary if the alarm sounds?

$$P(B=1\,|\,A=1) = \frac{P(B=1,A=1)}{P(A=1)} = \frac{\sum_{E,R} P(A=1,R,E,B=1)}{\sum_{E,R,B} P(A=1,R,E,B)}$$

$$= \frac{\sum_{E,R} P(A=1\,|\,E,B=1)P(B=1)P(E)P(R\,|\,E)}{\sum_{E,R,B} P(A=1\,|\,E,B)P(B)P(E)P(R\,|\,E)} = 0.99$$

What is the probability that there was a burglary if the alarm sounds and the radio reports on an earthquake?

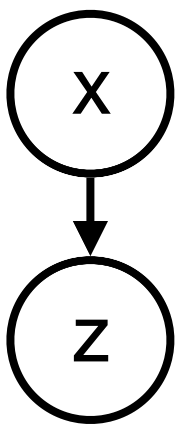
$$P(B=1\,|\,A=1,R=1) = \frac{P(B=1,A=1,R=1)}{P(A=1,R=1)} = \frac{\sum_{E} P(A=1,R=1,E,B=1)}{\sum_{E,B} P(A=1,R=1,E,B)} = 0.01$$

What have we gained?

- Here: 1+1+2+4=8 parameters instead of $2^4 - 1 = 15$
- In general, a distribution over n Boolean variables needs $2^n - 1$ probability values
- If using a BN with at most k parents per node, only $n \times 2^k$
- e.g., for $n = 20$ and $k = 5$ reduction from 1048575 to 640
- number of values depends on skill of designer (and problem)
- fewer parameters means faster inference and learning

Logic as a special case of BNs

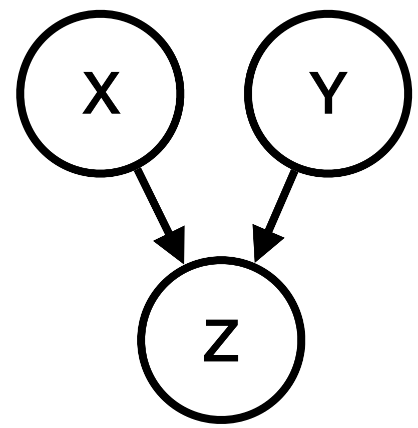
$Z = \neg X$



	Z=1
X=1	0
X=0	1

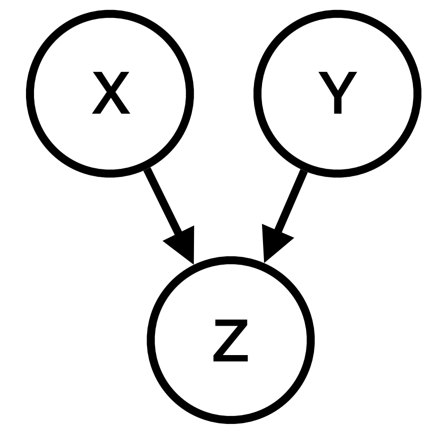
We write such CPTs compactly using the logic notation

$Z = X \wedge Y$



P(Z|X,Y)		Z=1
X=1	Y=1	1
X=1	Y=0	0
X=0	Y=1	0
X=0	Y=0	0

$Z = X \vee Y$



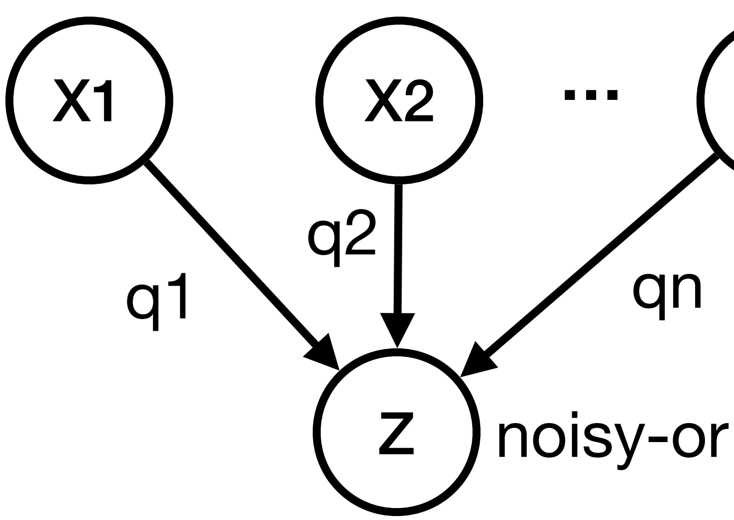
P(Z|X,Y)		Z=1
X=1	Y=1	1
X=1	Y=0	1
X=0	Y=1	1
X=0	Y=0	0

similarly for $Z = X_1 \wedge \ldots \wedge X_n$ and $Z = X_1 \vee \ldots \vee X_n$

Noisy-OR

If Z is a disjunction (**OR**-gate), $Z = X_1 \vee \ldots \vee X_n$, each event $X_i = 1$ causes the event $Z = 1$

In a **noisy-OR**, each event $X_i = 1$ causes the event $Z = 1$ **unless** an inhibitor prevents it, which happens with probability q_i (independently for each i)



for fixed values $X_1 = v_1, \ldots, X_n = v_n$ of the parents, when do we get Z=1?

any X_i with $v_i = 0$ never causes Z=1

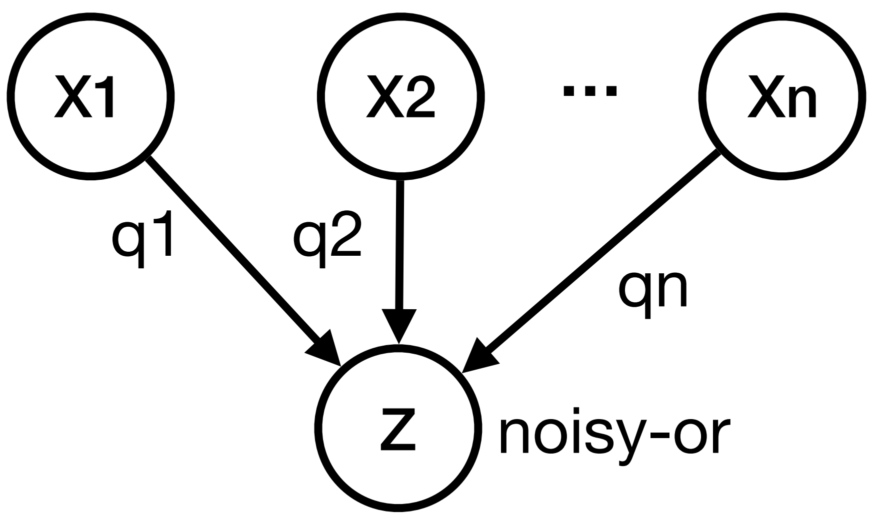
any X_i with $v_i = 1$ causes Z=1 unless inhibited

we get Z=1 if at least one of the X_i with $v_i = 1$ is not inhibited

conversely, we get Z=0 if all of the X_i with $v_i = 1$ are inhibited

$$P(Z = 0 \,|\, X_1 = v_1, \ldots, X_n = v_n) = \prod_{\{i | v_i = 1\}} q_i \qquad P(Z = 1 \,|\, X_1 = v_1, \ldots, X_n = v_n) = 1 - \prod_{\{i | v_i = 1\}} q_i$$

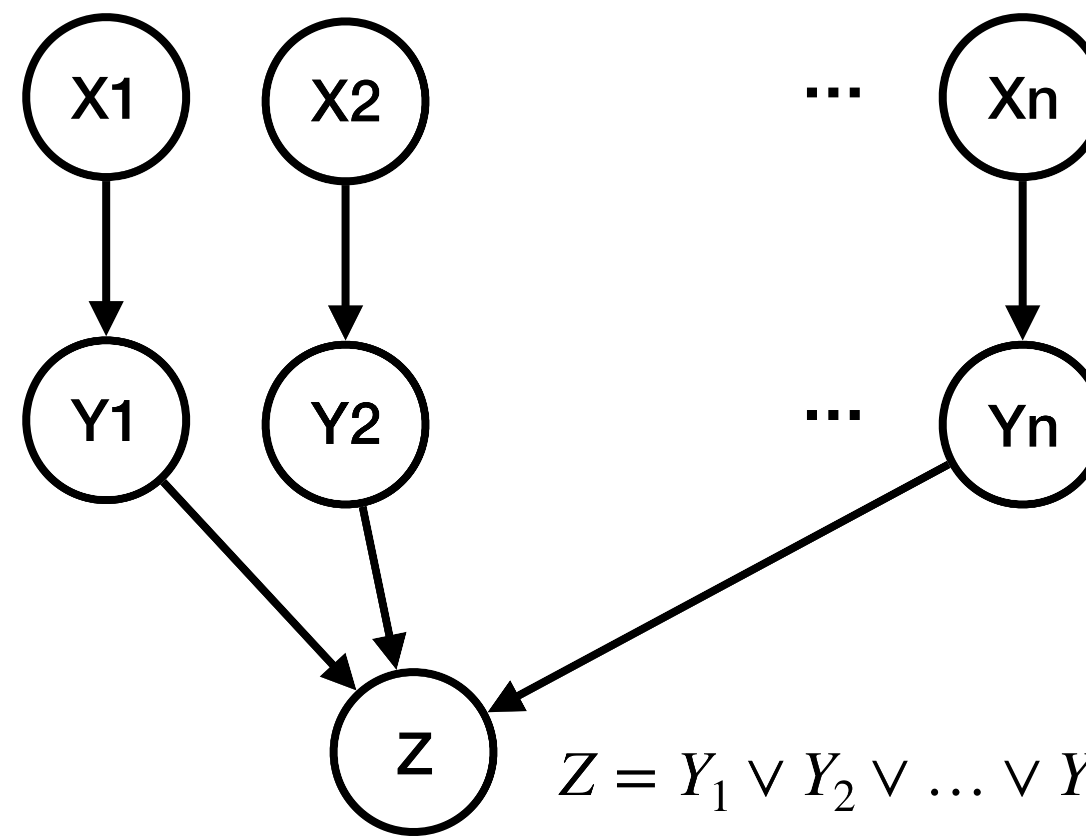
Noisy-OR



explicit encoding of noisy-or:

for each i:

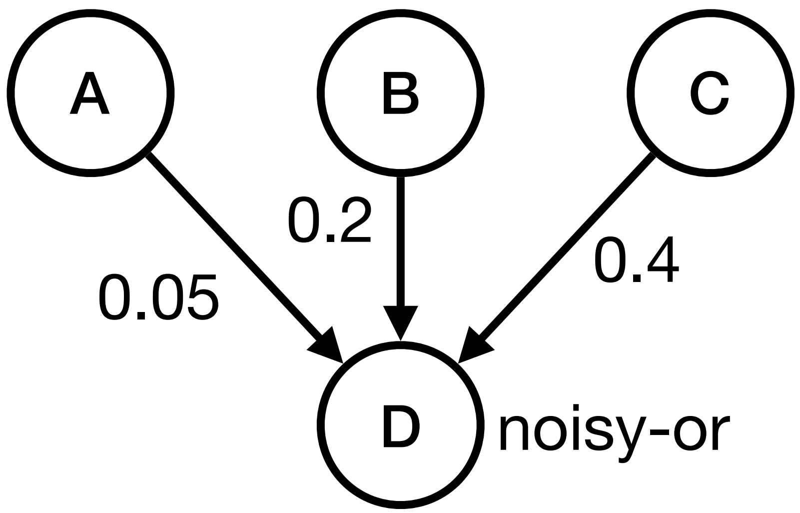
P(Yi|Xi)	Yi=1
Xi=1	1-q_i
Xi=0	0



$$Z = Y_1 \vee Y_2 \vee \ldots \vee Y_n$$

noisy-AND follows the same principle for logical **AND**: for Z to be 1, all parents need to be 1, but X_i is independently inhibited with probability q_i

Noisy-OR: example



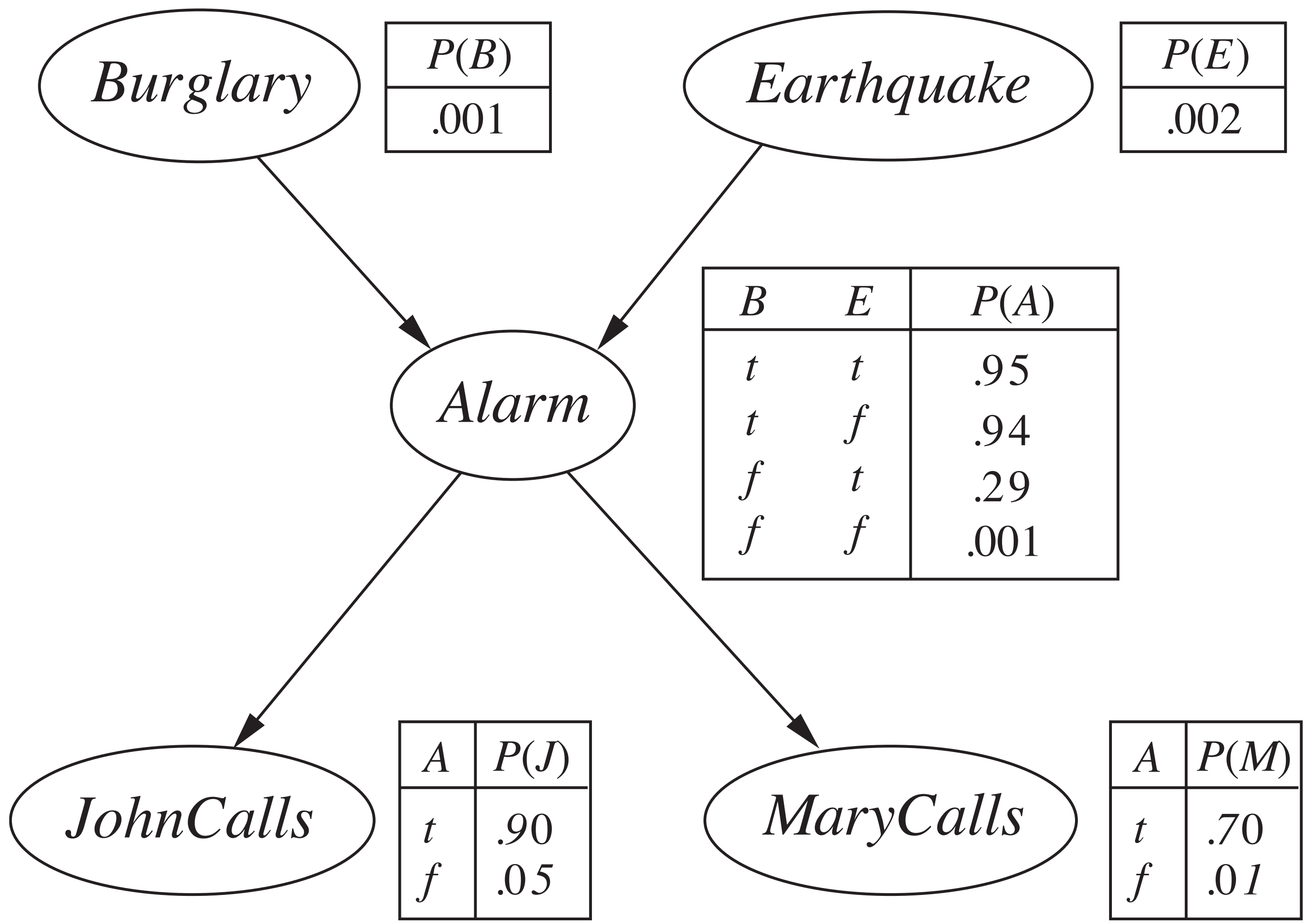
			Z=1
A=1	B=1	C=1	1-0.05*0.2*0.4=0.996
A=1	B=1	C=0	1-0.05*0.2=0.990
A=1	B=0	C=1	1-0.05*0.4=0.980
A=1	B=0	C=0	1-0.05=0.950
A=0	B=1	C=1	1-0.2*0.4=0.920
A=0	B=1	C=0	1-0.2=0.800
A=0	B=0	C=1	1-0.4=0.600
A=0	B=0	C=0	1-1=0.000

$$P(Z=0 \mid X_1 = v_1, \ldots, X_n = v_n) = \prod_{\{i \mid v_i = 1\}} q_i$$

$$P(Z=1 \mid X_1 = v_1, \ldots, X_n = v_n) = 1 - \prod_{\{i \mid v_i = 1\}} q_i$$

Judea Pearl's Alarm network

You have a new burglary alarm that is fairly reliable at detecting a burglary, but also responds to earthquakes. Your neighbours, Mary and John, promise to call you if they hear the alarm sounding.

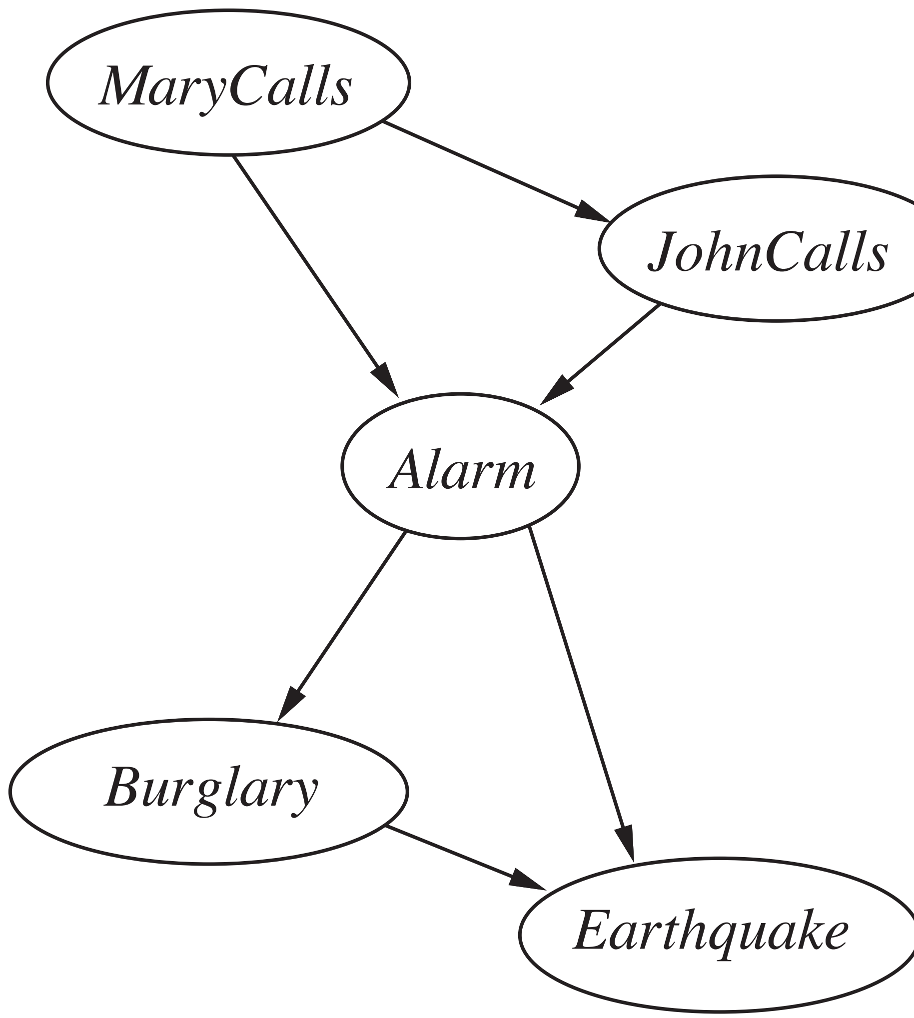


[figure: Russell& Norvig]

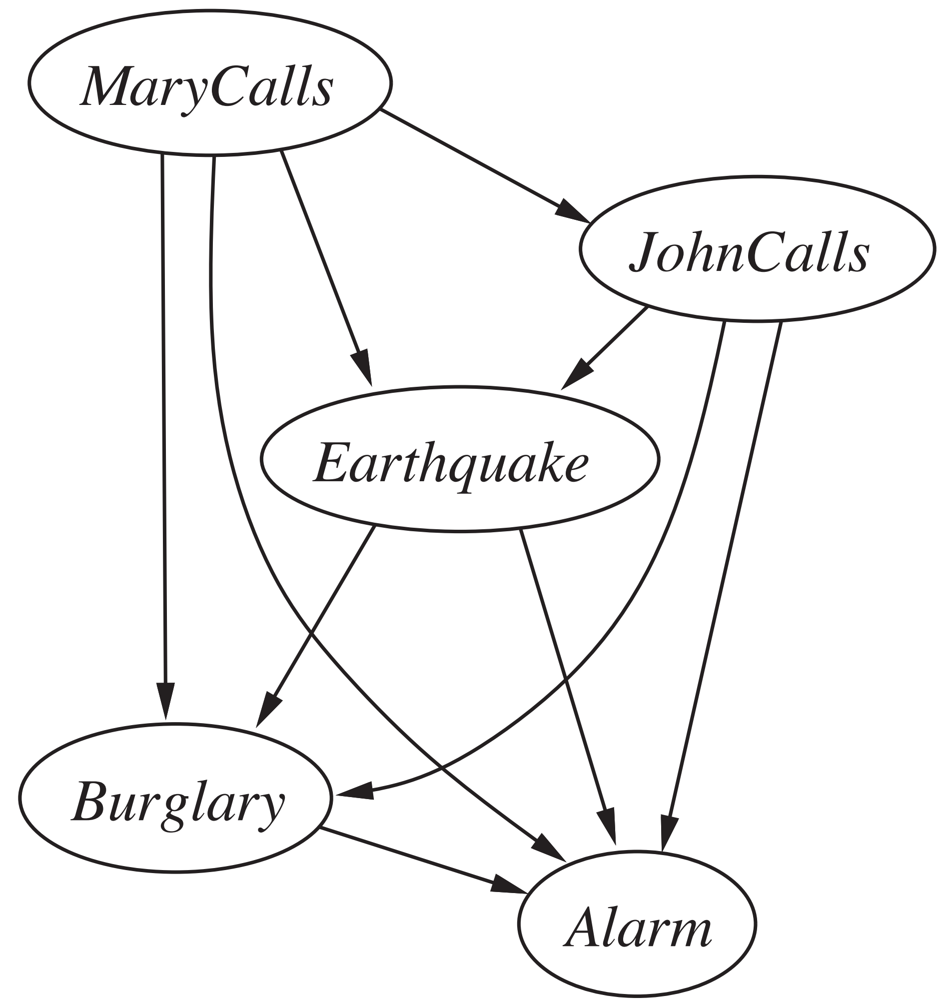
Judea Pearl's Alarm network

order: M,J,A,B,E

order: M,J,E,B,A



(a)



(b)

Edges in BNs do not always have a **causal** interpretation, but directing them from causes to effects often gives cleaner models

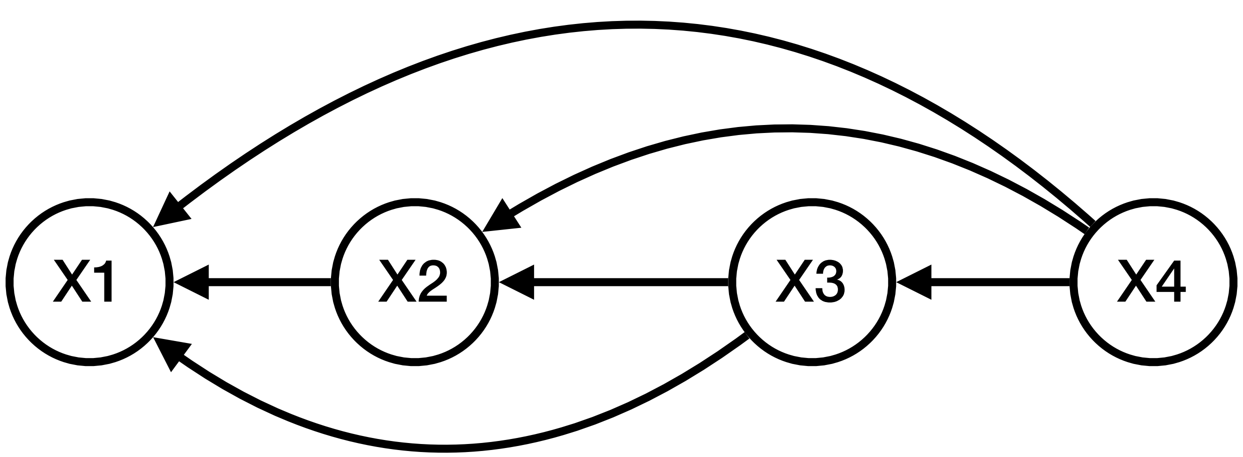
[figure: Russell& Norvig]

Factorisation in BNs

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i \,|\, parents(X_i))$$

chain rule lets us write **any** joint distribution in this form:

$$\begin{aligned} P(X_1, X_2, \dots, X_n) &= P(X_1 \,|\, X_2, \dots, X_n) \cdot P(X_2, \dots, X_n) \\ &= P(X_1 \,|\, X_2, \dots, X_n) \cdot P(X_2 \,|\, X_3, \dots, X_n) \cdot P(X_3, \dots, X_n) \\ &= P(X_n) \cdot \prod_{i=1}^{n-1} P(X_i \,|\, X_{i+1}, \dots, X_n) \end{aligned}$$



Order of variables is important if we want to gain something: determines which edges we can omit because of independence

Why graphs?

- Data structure to compactly represent a **factored** joint distribution

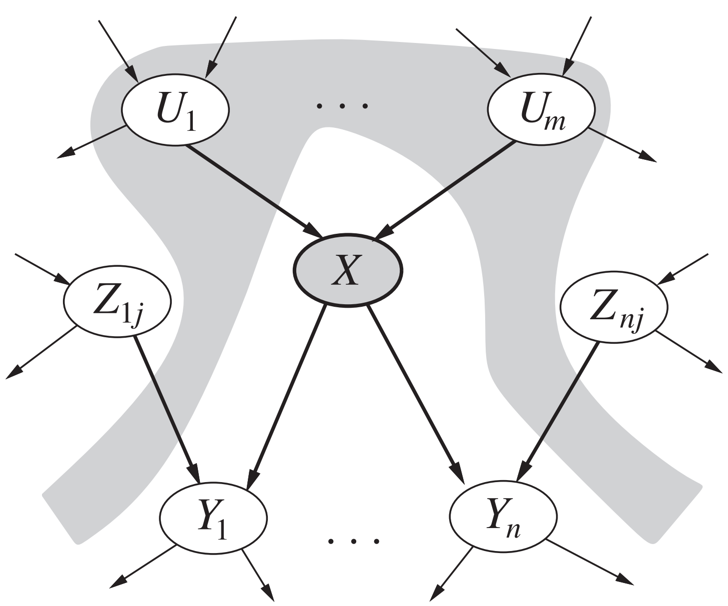
- Compact representation of a **set of conditional independence assumption** about a joint distribution

- Both views are **equivalent**: a distribution P satisfies all conditional independence assumptions in a DAG if and only if it has the factorised form.

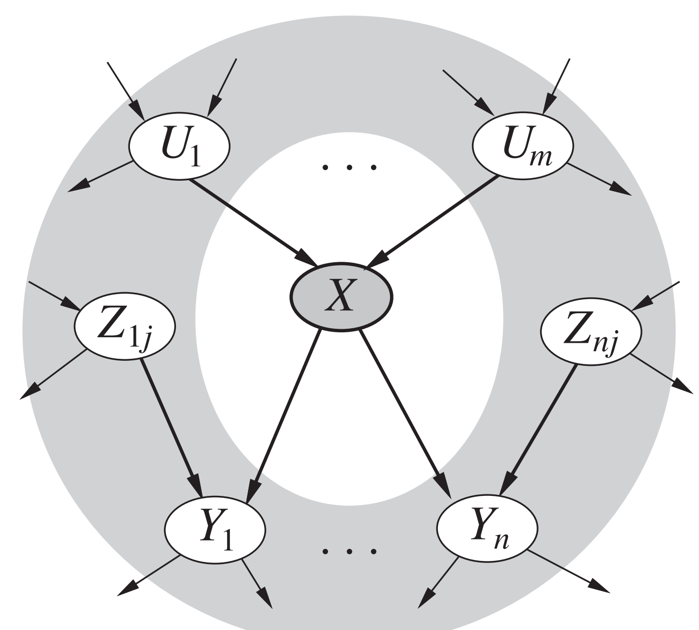
DAGs & conditional independence

Random variables X and Y are conditionally independent of each other given the state of random variable Z, written X⫫Y|Z, if $P(X, Y \mid Z) = P(X \mid Z) \cdot P(Y \mid Z)$

Each node is conditionally independent of its **non-descendants given** its **parents.**



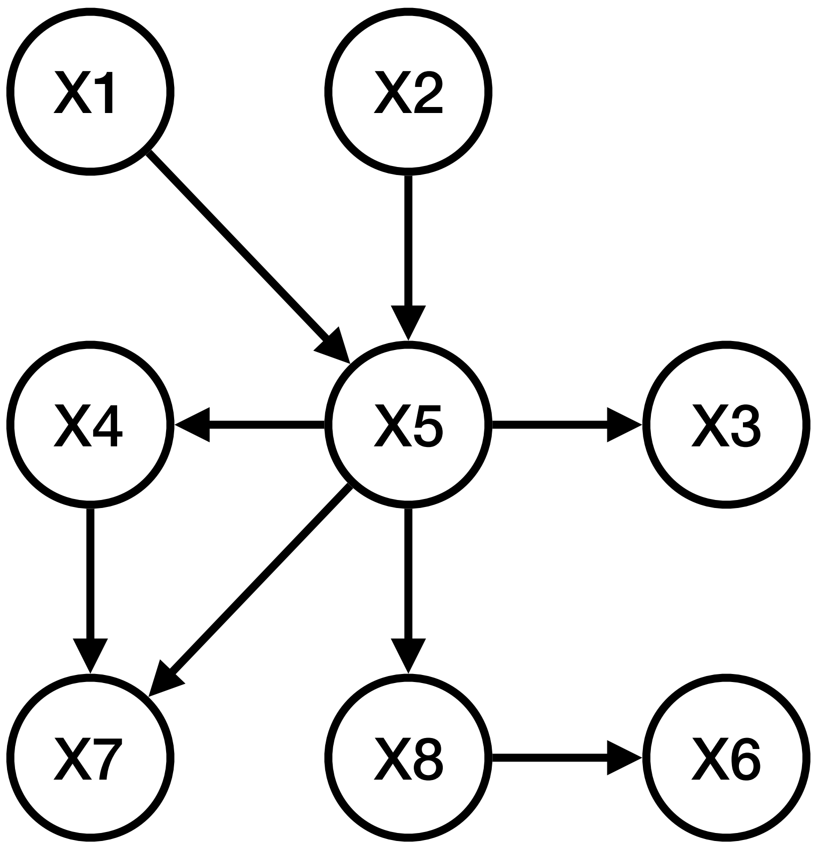
Each node is conditionally independent of **all other nodes** in the network **given** its **Markov blanket.**



Both characterisations follow from the more general notion of d-separation [proof: see exercises]

[figures: Russell& Norvig]

Example

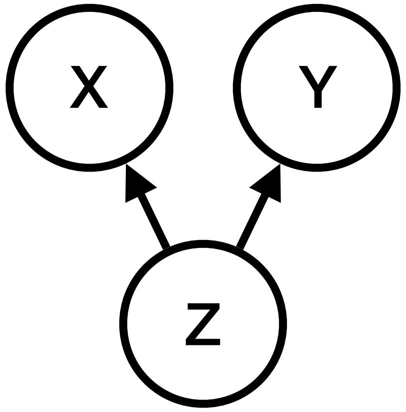


Each node is conditionally independent of its **non-descendants given** its **parents.**

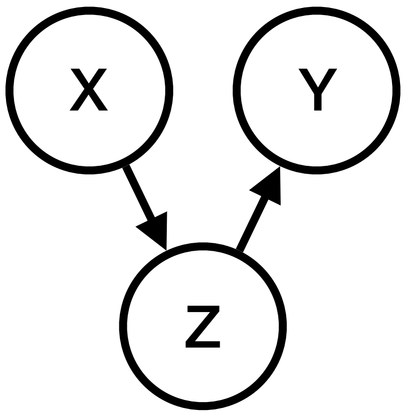
Each node is conditionally independent of **all other nodes** in the network **given** its **Markov blanket.**

What do these statements tell us about the following nodes?
X6
X4
X5
X1

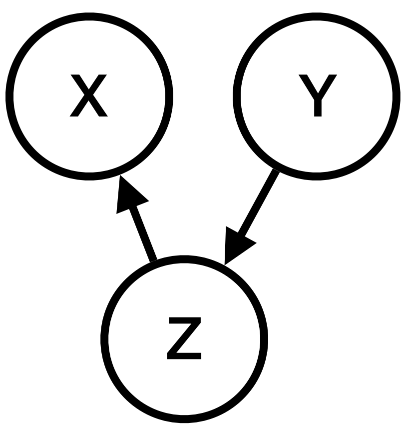
Conditional independence



$$P(X,Y|Z) = \frac{P(X|Z)P(Y|Z)P(Z)}{P(Z)} = P(X|Z)P(Y|Z)$$



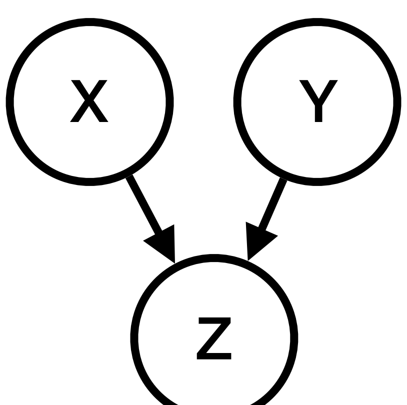
$$P(X,Y|Z) = \frac{P(X)P(Z|X)P(Y|Z)}{P(Z)} = \frac{P(X,Z)P(Y|Z)}{P(Z)} = P(X|Z)P(Y|Z)$$



$$P(X,Y|Z) = \frac{P(Y)P(Z|Y)P(X|Z)}{P(Z)} = \frac{P(Y,Z)P(X|Z)}{P(Z)} = P(Y|Z)P(X|Z)$$

$X \perp\!\!\!\perp Y|Z$ holds

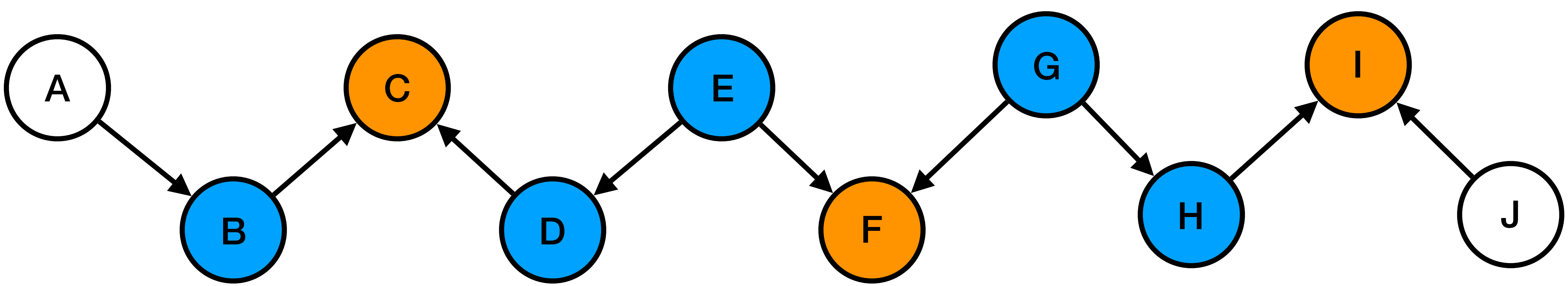
$X \perp\!\!\!\perp Y|Z$ does not hold



$$P(X,Y|Z) = \frac{P(X)P(Y)P(Z|X,Y)}{P(Z)} \text{ in general } \neq P(X|Z)P(Y|Z)$$

Collider

- Consider an acyclic path between two nodes.
- An intermediate node on the path is a **collider** if it has incoming edges from both its neighbours on the path.
- An intermediate node on the path is a **non-collider** if it is not a collider.



Observations blocking paths in a BN

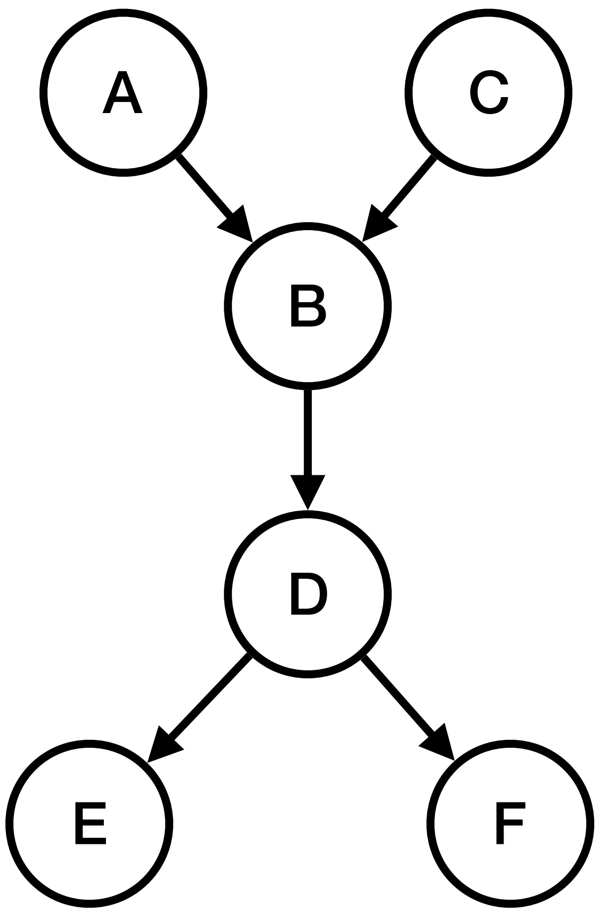
- Let $\mathscr{Z}$ be the set of nodes in a BN whose values are observed, and X and Y distinct nodes that are not in $\mathscr{Z}$

- We say a path from X to Y is **blocked** by $\mathscr{Z}$ if at least one of the following holds:

  - there is a collider on the path such that neither the collider nor any of its descendants is in $\mathscr{Z}$

  - there is a non-collider on the path that is in $\mathscr{Z}$

Example



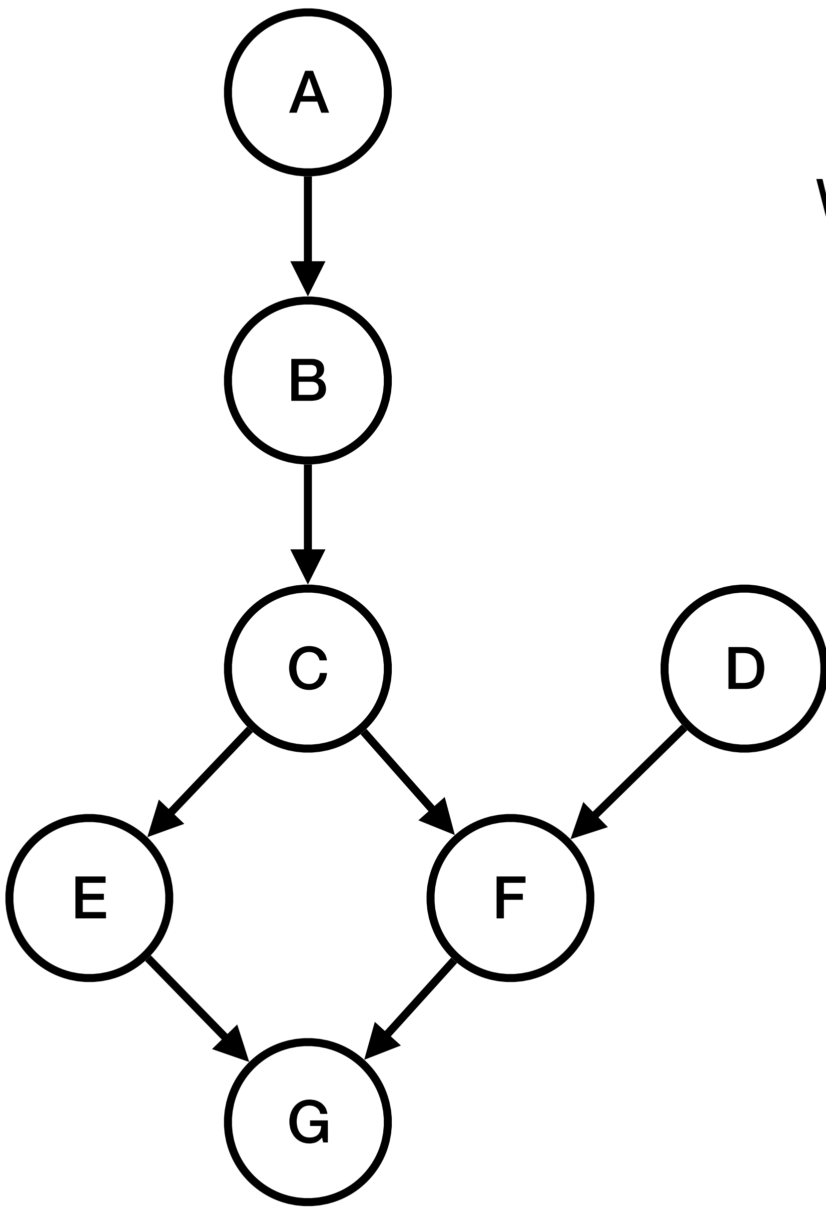
Which paths are blocked by each of the following sets?

- $\mathscr{Z} = \emptyset$
- $\mathscr{Z} = \{B\}$
- $\mathscr{Z} = \{D\}$
- $\mathscr{Z} = \{F\}$
- $\mathscr{Z} = \{D, F\}$

d-separation

- Let $\mathscr{Z}$ be the set of nodes in a BN whose values are observed, and X and Y distinct nodes that are not in $\mathscr{Z}$
- X and Y are **d-separated** (by $\mathscr{Z}$) if every path from X to Y is blocked by $\mathscr{Z}$
- X and Y are **d-connected** if they are not d-separated

Example

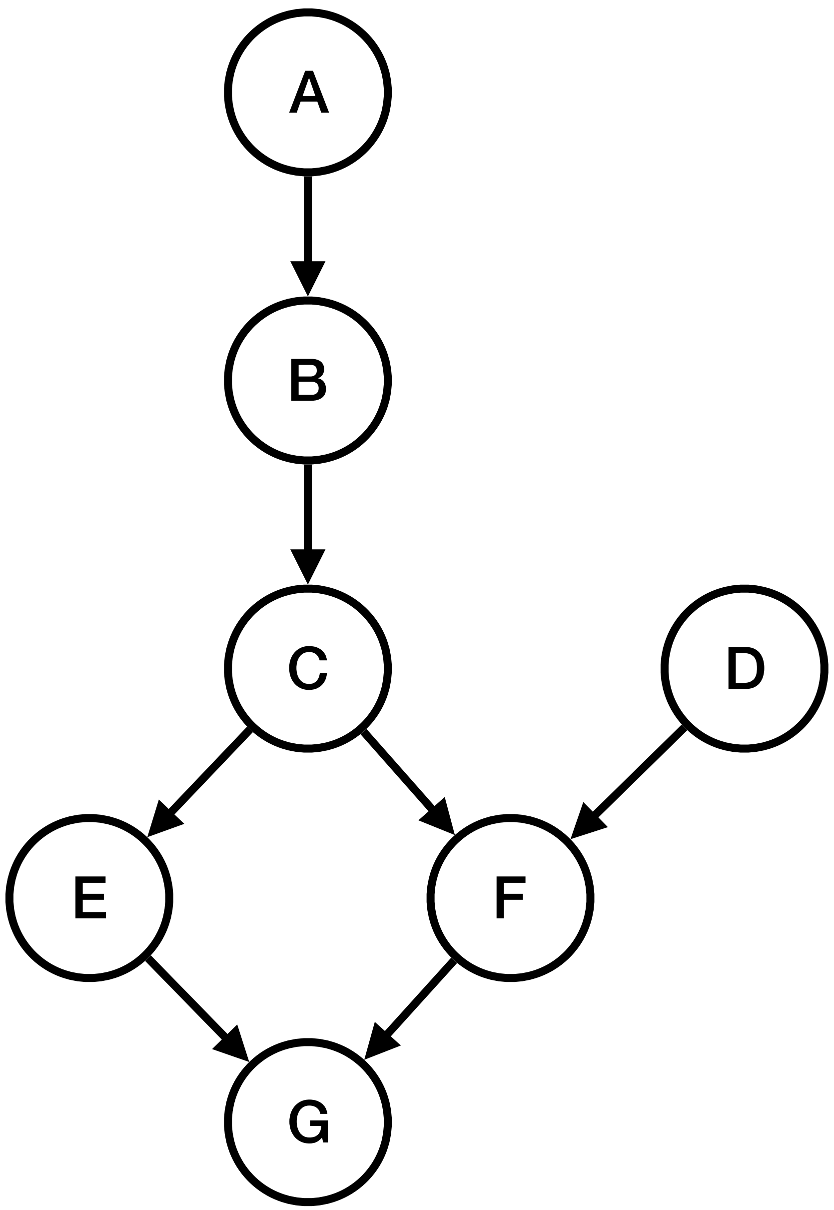


Which sets $\mathscr{Z} \subseteq \{B, C, E, F, G\}$ d-separate A and D?

Conditional independence in BNs

- Let $\mathscr{X}$, $\mathscr{Y}$ and $\mathscr{Z}$ be disjoint sets of nodes in a BN.
- $\mathscr{X}$ and $\mathscr{Y}$ are d-separated by $\mathscr{Z}$ if every pair of nodes $X \in \mathscr{X}, Y \in \mathscr{Y}$ is d-separated by $\mathscr{Z}$.
- If $\mathscr{X}$ and $\mathscr{Y}$ are d-separated by $\mathscr{Z}$, then $\mathscr{X} \perp\!\!\!\perp \mathscr{Y} \mid \mathscr{Z}$, i.e., $\mathscr{X}$ and $\mathscr{Y}$ are conditionally independent given $\mathscr{Z}$

Example



Are {E} and {D} conditionally independent given {A}?

Are {E} and {D} conditionally independent given {A,G}?

Are {A,B} and {D} conditionally independent given {C}?

Are {A,B} and {D} conditionally independent given {E}?

Are {A,B} and {D} conditionally independent given {G}?

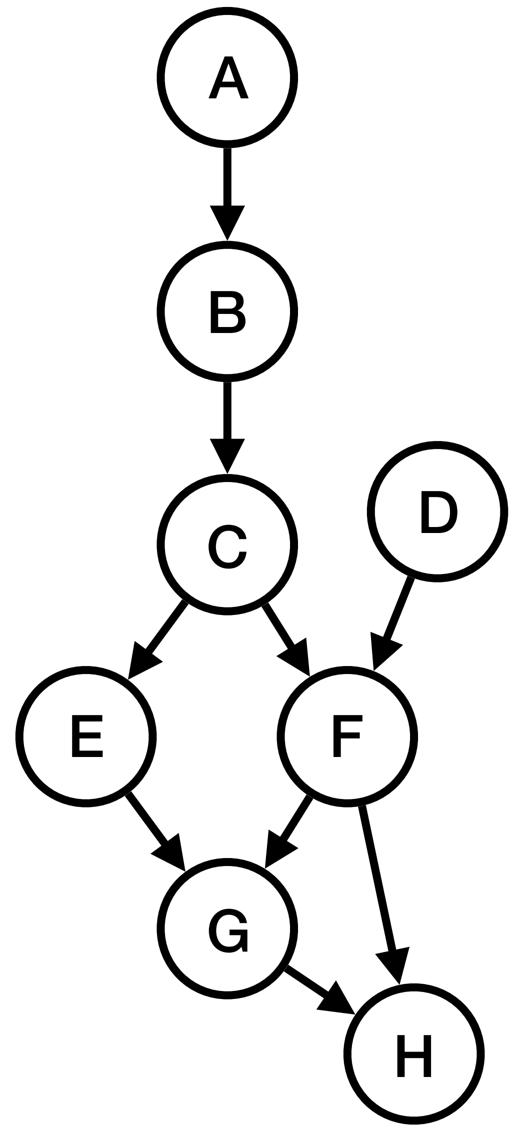
DAGs and independencies

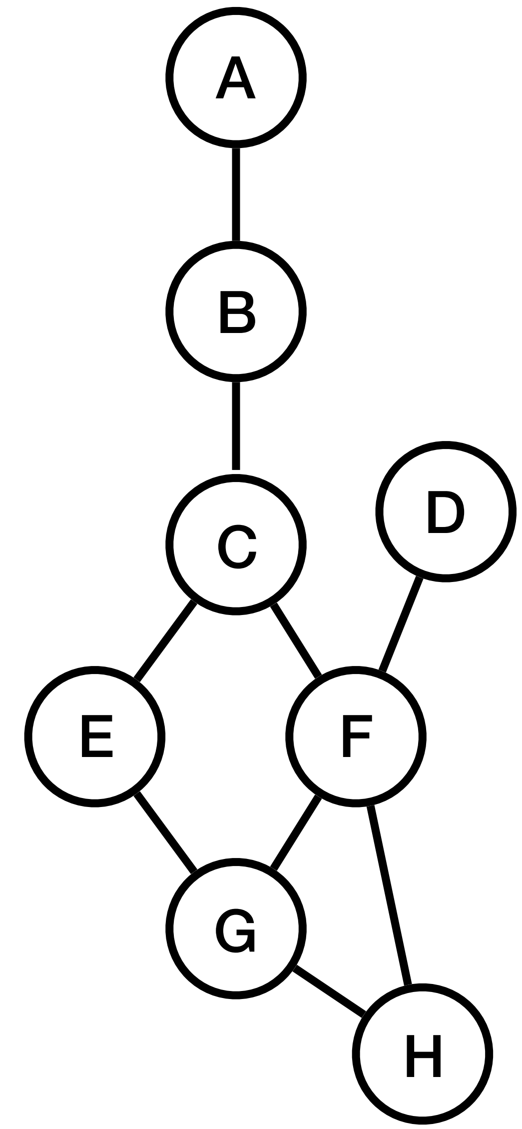
- We’ve seen earlier that different graphs can represent the same conditional independence assumptions.
- Given two DAGs, can we tell whether this is the case, without figuring out all the conditional independencies?
- YES: Markov equivalence

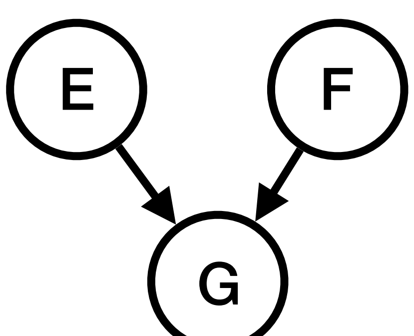
Markov Equivalence

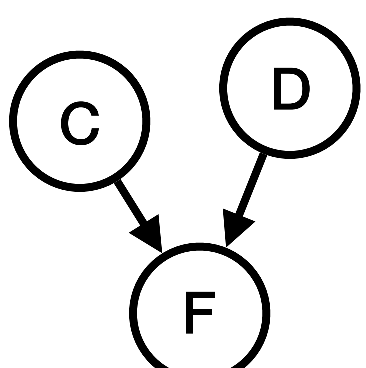
- The **skeleton** of a DAG is the undirected graph obtained by removing the direction of edges.

- An **immorality** in a DAG consists of three nodes X,Y,Z such that X and Z are parents of Y, but there is no edge between X and Z





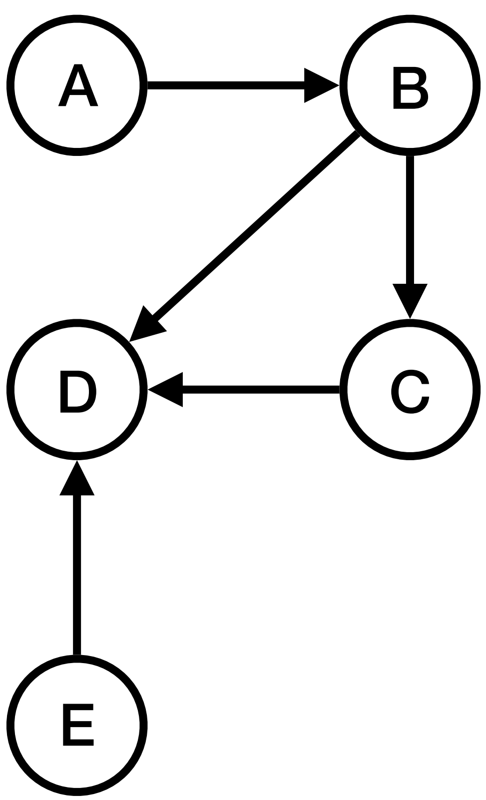




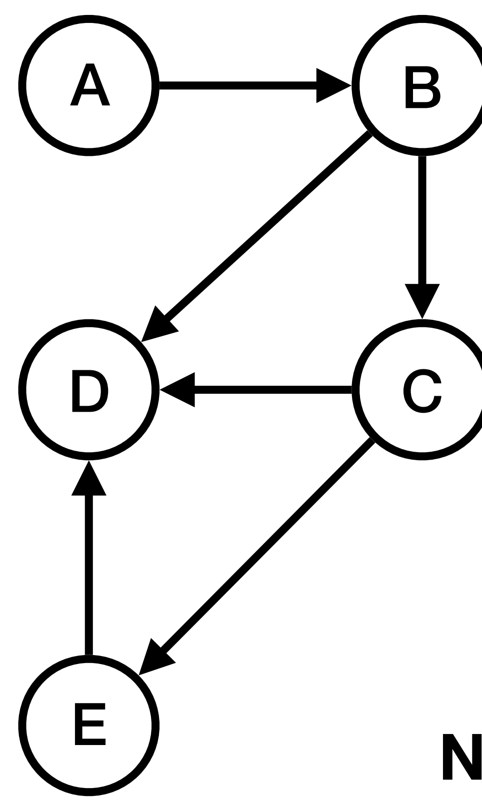
Markov Equivalence

Two DAGs represent the same set of conditional independence assumptions if and only if they have the same skeleton and the same set of immoralities.

Example 1



?
=

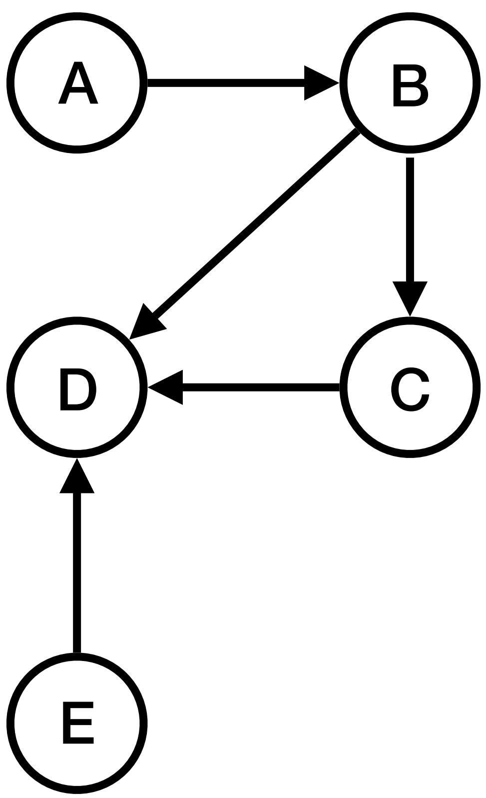


NO: different skeleton

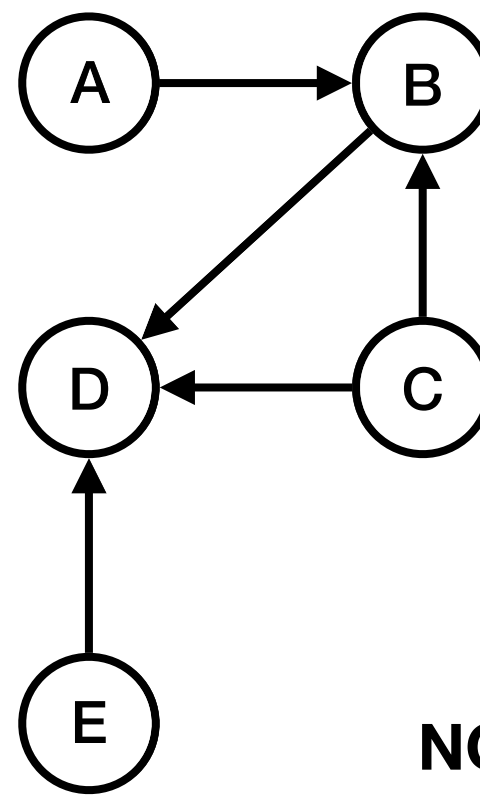
Markov Equivalence

Two DAGs represent the same set of conditional independence assumptions if and only if they have the same skeleton and the same set of immoralities.

Example 2



?
=

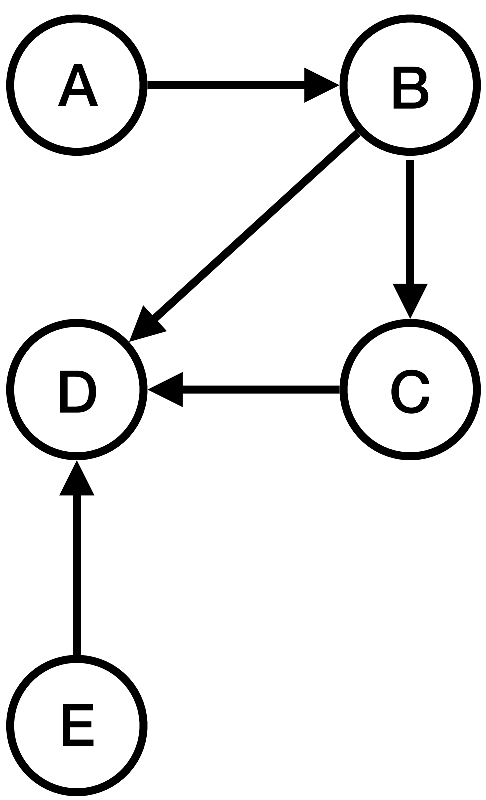


NO: same skeleton, but different immoralities

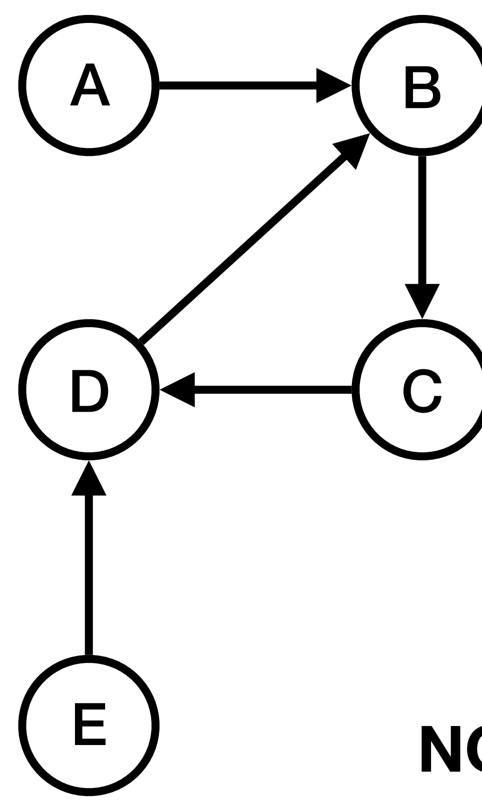
Markov Equivalence

Two DAGs represent the same set of conditional independence assumptions if and only if they have the same skeleton and the same set of immoralities.

Example 3



?
=

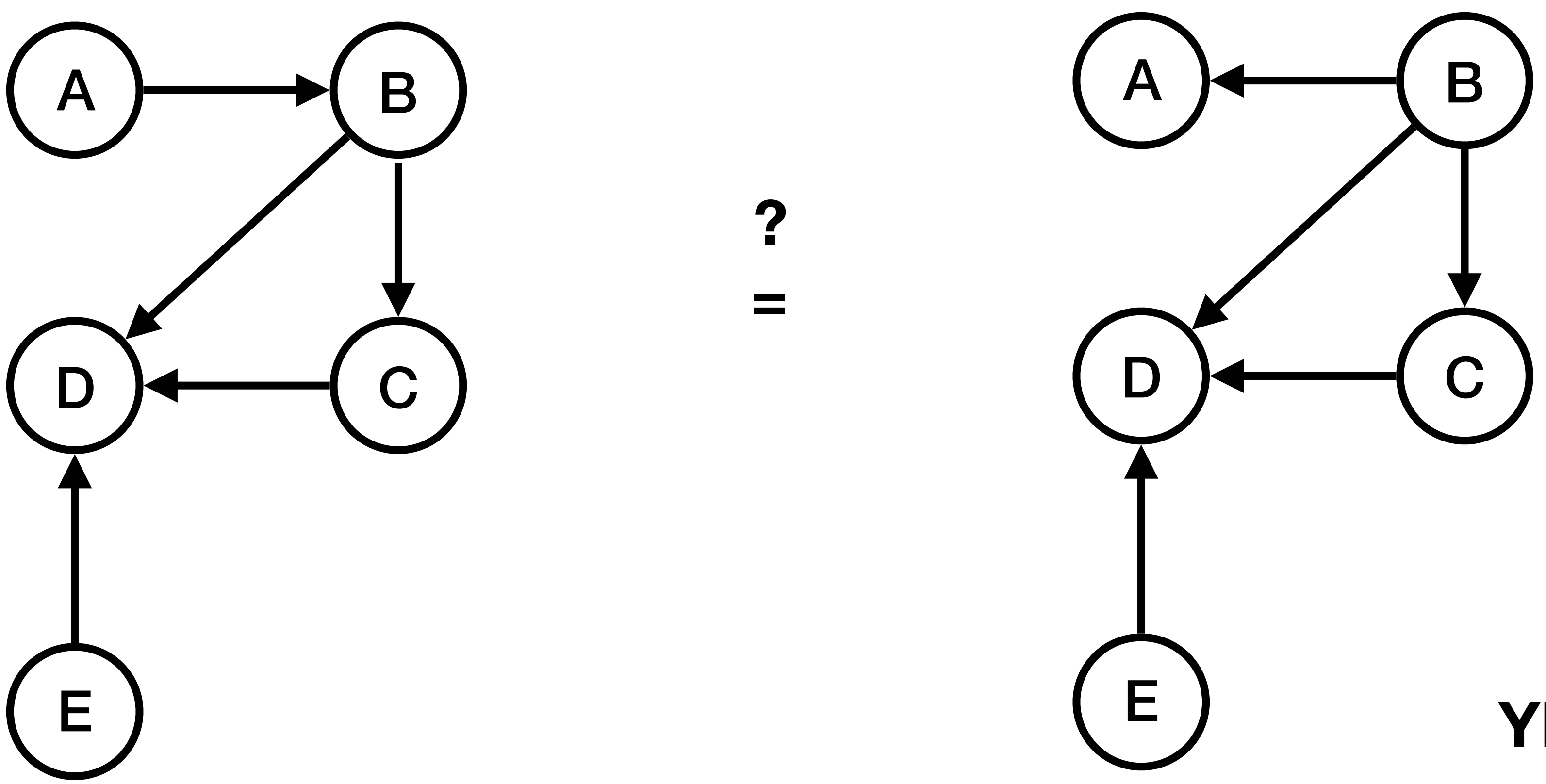


NO: same skeleton, but different immoralities

Markov Equivalence

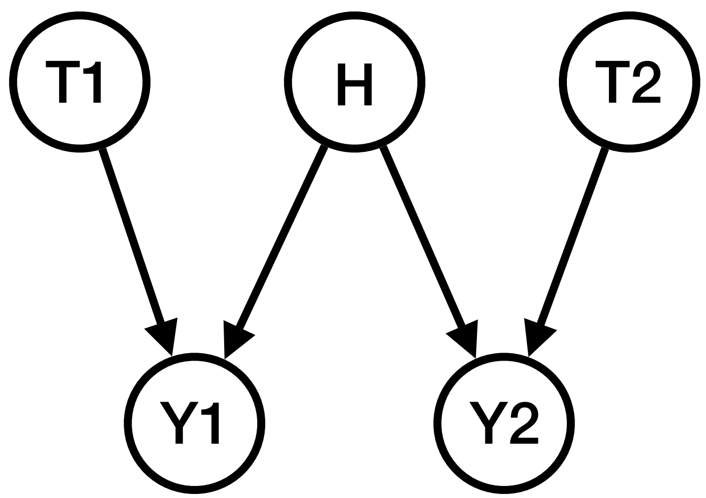
Two DAGs represent the same set of conditional independence assumptions if and only if they have the same skeleton and the same set of immoralities.

Example 4



YES: same skeleton, same immoralities

Limits of expressibility



Health of patient, two treatments **T1** and **T2** with outcomes **Y1** and **Y2**

{T1}⫫{T2,Y2} and {T2}⫫{T1,Y1}   (why?)

summing out H:

$$P(T1,Y1,T2,Y2) = \sum_H P(H)P(T1)P(T2)P(Y1 \mid H, T1)P(Y2 \mid H, T2)$$

$$= P(T1)P(T2) \sum_H P(H)P(Y1 \mid H, T1)P(Y2 \mid H, T2)$$

{T1}⫫{T2,Y2} and {T2}⫫{T1,Y1} still hold for P(T1,Y1,T2,Y2), but there is no BN over these four variables that precisely encodes these independence assumptions

Exercises:
start here, finish at home

(solutions will be on learning central later)

Reading Material

- Today:
  - Russell & Norvig: sections 14.1 & 14.2
  - Barber: chapters 2 & 3
- Next week:
  - Russell & Norvig: 14.4
  - Barber: chapters 4 & 5

- Parts of slides based on
    - David Barber’s slides for the BRML book
    - Tinne De Laet & Luc De Raedt’s slides for the UAI course at KU Leuven