Introduction to Bayes Nets

CS 486/686: Introduction to Artificial Intelligence
Fall 2013

Introduction

- Review probabilistic inference, independence and conditional independence
- Bayesian Networks
  - What they are
  - What they mean

Example: Joint Distribution

sunny

	cold	~cold
headache	0.108	0.012
~headache	0.016	0.064

~sunny

	cold	~cold
headache	0.072	0.008
~headache	0.144	0.576

P(headache^sunny^cold)=0.108 P(~headache^sunny^~cold)=0.064

P(headache V sunny) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28

P(headache)=0.108+0.012+0.072+0.008=0.2

marginalization

Example: Joint Distribution

sunny

	cold	~cold
headache	0.108	0.012
~headache	0.016	0.064

~sunny

	cold	~cold
headache	0.072	0.008
~headache	0.144	0.576

P(headache ^ cold| sunny)= P(headache ^ cold ^ sunny)/P(sunny)

= 0.108/(0.108+0.012+0.016+0.064)

= 0. 54

P(headache ^ cold| ~sunny)= P(headache ^ cold ^ ~sunny)/P(~sunny)

= 0.072/(0.072+0.008+0.144+0.576)

= 0.09

Bayes Rule

- Note:
  - $P(A|B)P(B)=P(A \wedge B)=P(B \wedge A)=P(B|A)P(A)$
- Bayes Rule:
  - $P(B|A)=[P(A|B)P(B)]/P(A)$

Memorize this!

General Forms of Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

$$P(A|B \wedge X) = \frac{P(B|A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

$$P(A = v_i|B) = \frac{P(B|A = v_i)P(A = v_i)}{\sum_{k=1}^{n} P(B|A = v_k)P(A = v_k)}$$

Using Bayes Rule for Inference

- Often we want to form a hypothesis about the world based on what we have observed
- Bayes rule is vitally important when viewed in terms of stating the belief given to hypothesis **H,** given evidence **e**

Likelihood

Prior probability

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)}$$

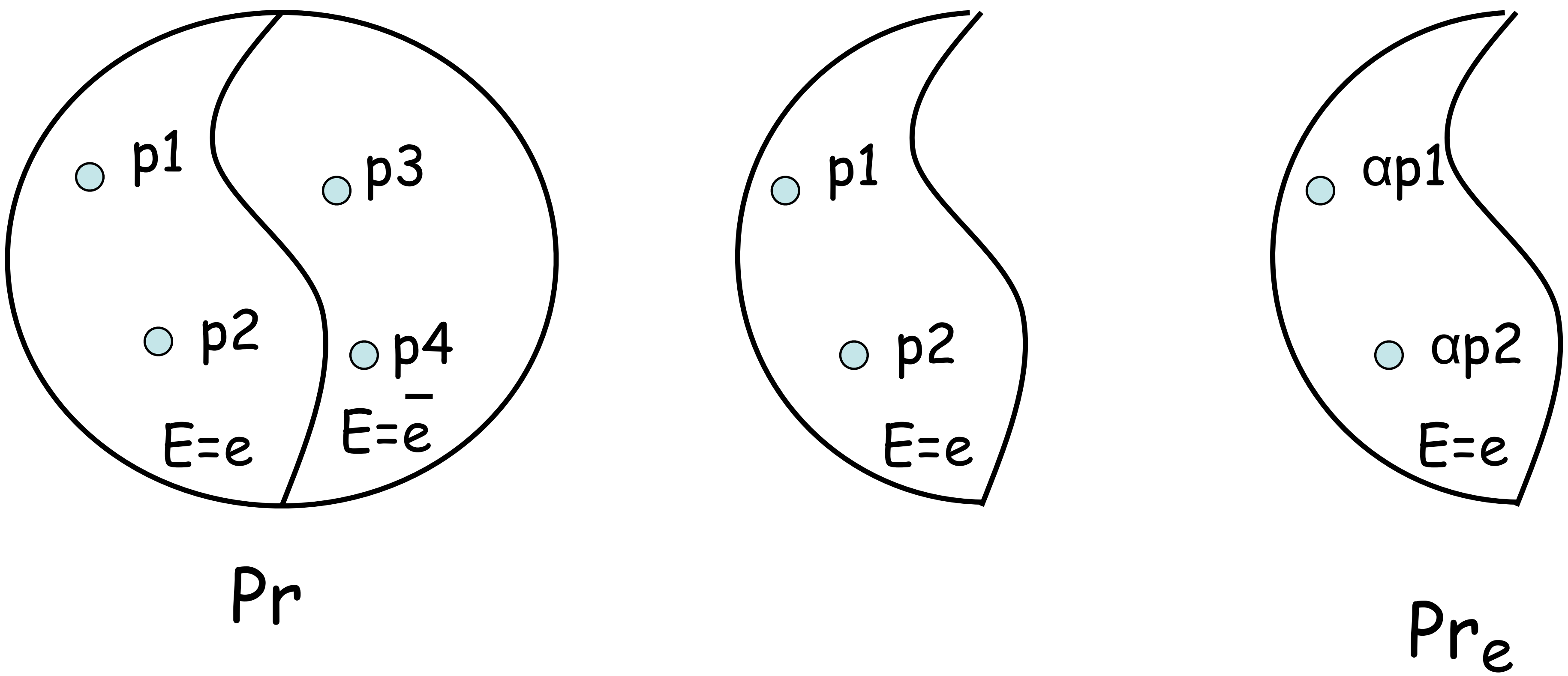
Posterior probability

Normalizing constant

Conditioning

- We define $P_e(x)=P(x|e)$
  - Produce P_e by conditioning prior distribution on observed evidence
- Semantically we take original measure μ
  - Set $\mu=0$ for any world where e was false
  - Set $\mu=\mu(w)/P(e)$ for any e-world
    - Normalization

Semantics of Conditioning



$\alpha = 1/(p1+p2)$
normalizing constant

Inference

- Semantically/conceptually, the picture is clear
- But several issues must be addressed

Issue 1

- How do we specify the full joint distribution over a set of random variables $X_1, X_2, \ldots, X_n$?
  - What are the difficulties?

Issue 2

- Inference in this representation is very slow

Independence

- Two variables A and B are **independent** if knowledge of A does not change uncertainty of B (and vice versa)
  - $P(A|B)=P(A)$
  - $P(B|A)=P(B)$
  - $P(A \wedge B)=P(A)P(B)$
  - In general: $P(X_1, X_2, \ldots, X_n)=\prod_i P(X_i)$

Variable Independence

- Two **variables** X and Y are conditionally independent given variable Z iff x, y are conditionally independent given z for all x in Dom(X), y in Dom(Y) and z in Dom(Z)
  - Also applies to sets of variables **X, Y, Z**
- If you know the value of Z (whatever it is) nothing you learn about Y will influence your beliefs about X

What good is independence?

- Suppose (boolean) random variables $X_1, X_2, \ldots, X_n$ are mutually independent
  - Specify the full joint using only n parameters instead of 2^n-1
- How? Specify $P(x_1)$, $P(x_2)$,..., $P(x_n)$
  - Can now recover probability for any query
    - $P(x,y)=P(x)P(y)$ and $P(x|y)=P(x)$ and $P(y|x)=P(y)$

Value of Independence

- Complete independence reduce both **representation of the joint** and **inference** from $O(2^n)$ to $O(n)$!
- Unfortunately, rarely have complete mutual independence

Conditional Independence

- Full independence is often too strong a requirement
- Two variables A and B are **conditionally independent** given C if
  - $P(a|b,c)=P(a|c)$ for all a,b,c
  - i.e. knowing the value of B does not change the prediction of A ***if the value of C is known***

Conditional Independence

- Diagnosis problem
  - Fl=Flu, Fv=Fever, C=Cough
- Full joint dist. has $2^3-1=7$ independent entries
- If someone has the flu, we can assume that the probability of a cough does not depend on having a fever (P(C | Fl,Fv)=P(C | Fl))
- If the same condition holds if the patient does not have the Flu then C and Fv are **conditionally independent** given FL (P(C | ~Fl, Fv)=P(C | ~Fl))

Conditional Independence

- Full distribution can be written as

$$
\begin{aligned}
P(C, Fl, Fv) &= P(C, FV|Fl)P(Fl) \\
&= P(C|Fl)P(Fv|Fl)P(Fl)
\end{aligned}
$$

- We only need 5 numbers!
- Huge savings if there are lots of variables

Conditional Independence

- Such a probability distribution is sometimes called a **Naive Bayes model**

- In practice they work well - even when the independence assumption is not true

Value of Independence

- Fortunately, most domains do exhibit a fair amount of conditional independence
  - Exploit conditional independence for both representation and inference

- **Bayesian networks** do just this

Notation

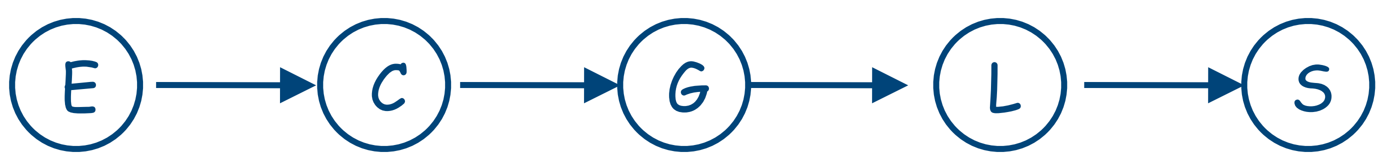
- P(X) for variable X (or set of variables) refers to (marginal) distribution over X
- P(X|Y) is the **family** of conditional distributions over X (one for each y in Dom(Y)
- Distinguish between P(X) (distribution) and P(x) (numbers)
  - Think of P(X) as a function that accepts any x_i in Dom(X) and returns a number

Notation

- Think of $P(X|Y)$ as a function that accepts any x_i and y_k and returns $P(x_i|y_k)$
- Note (again) that $P(X|Y)$ is not a single distribution

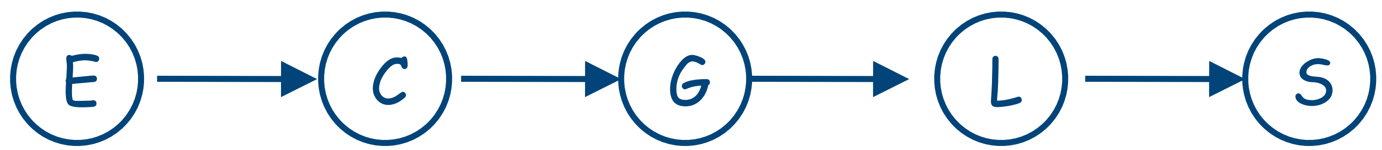
Exploiting Conditional Independence

- Consider the following story
  - If Kate woke up too early (E), she probably needs coffee (C); if Kate needs coffee (C), she is likely to be grumpy (G). If she is grumpy, then it's possible that the lecture won't go smoothly (L). If the lecture does not go smoothly, then the students will likely be sad (S).



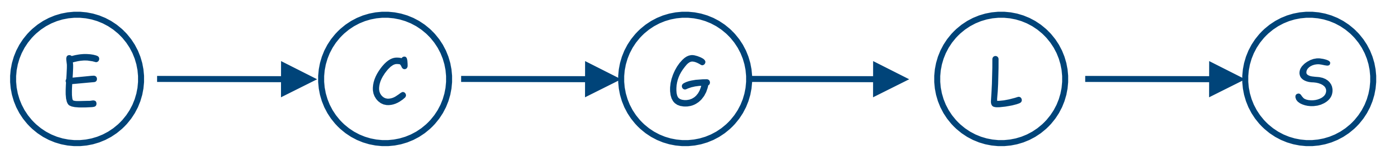
E – Kate woke too early G – Kate is grumpy S – Students are sad
C – Kate needs coffee L– The lecture did not go smoothly

Conditional Independence



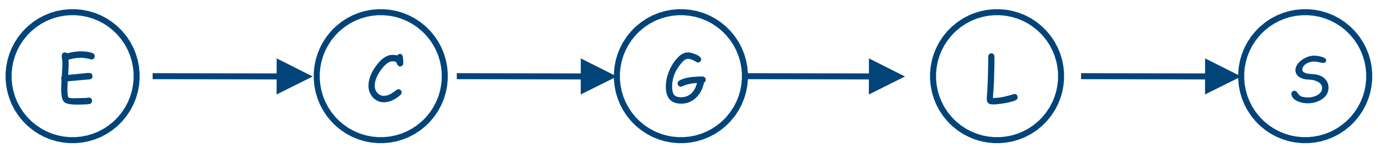
- If you learned any of E, C, G, or L then your assessment of P(S) would change
  - if any of these are seen to be true, you would increase P(s) and decrease P(~s)
  - So S is **not independent** of E, C, G, or L

Conditional Independence



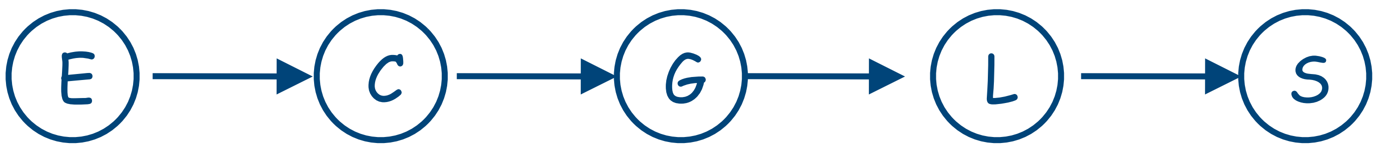
- But if you knew the value of L (true or false) then learning the values of E, C, or G would not influence P(S)
  - Students are not sad because Kate did not have a coffee, they are sad because of the lecture
  - **So S is independent of E, C, and G, given L**

Conditional Independence



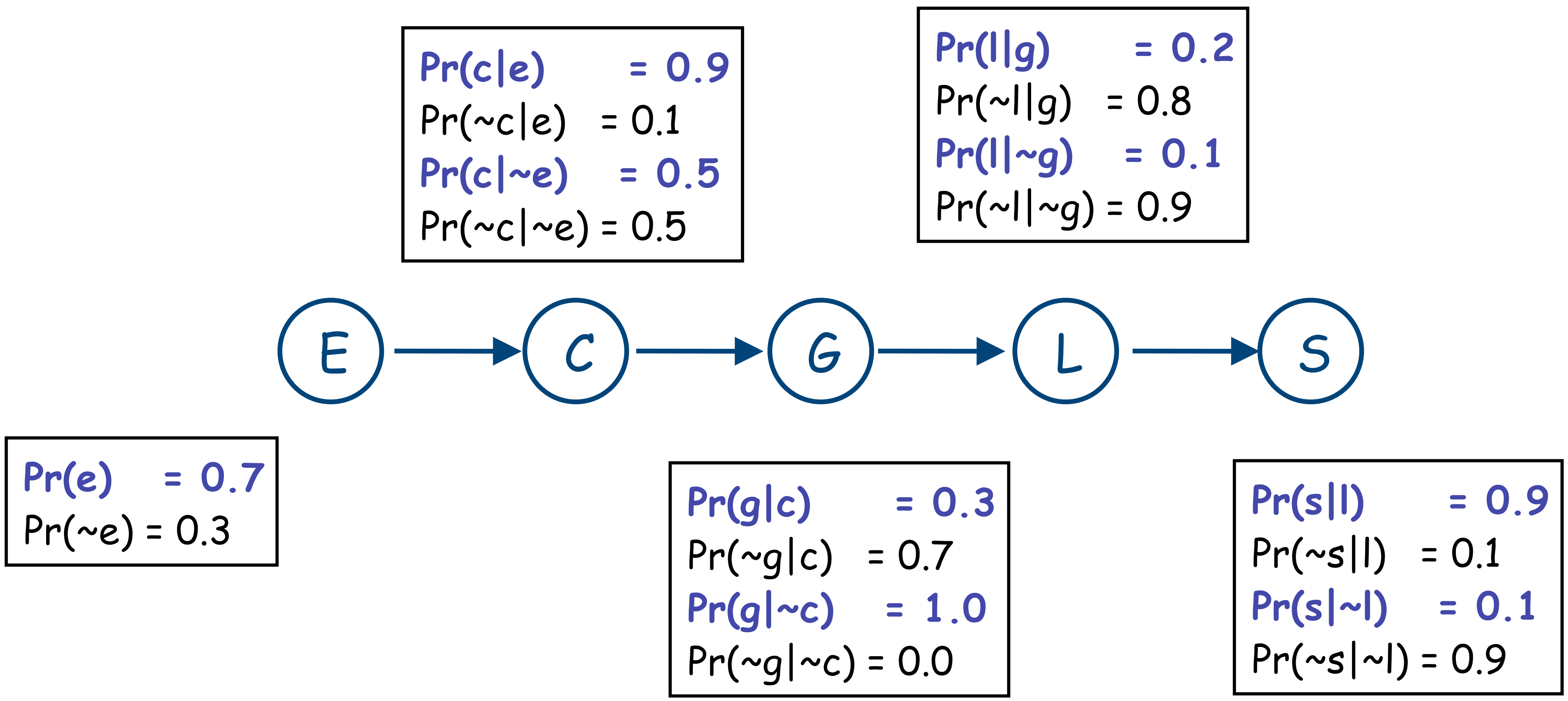
- S is independent of E, and C and G given L
- Similarly
  - L is independent of E and C, given G
  - G is independent of E given C
- This means that
  - P(S|L,{G,C,E})=
  - P(L|G, {C,E})=
  - P(G|C,{E})=
  - P(C|E)=
  - P(E)=

Conditional Independence



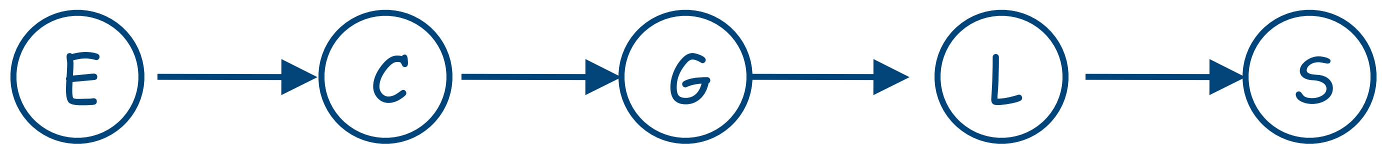
- By the chain rule
  - P(S,L,G,C,E)=?
- By our independence assumptions
  - P(S,L,G,C,E)=?
- We can specify the full joint by specifying five **conditional distributions**: P(S|L), P(L|G), P(G|C), P(C|E) and P(E)

Example Quantification



- Specifying the joint requires only 9 parameters instead of 31 for explicit representation
  - linear in number of vars instead of exponential
  - linear in general if dependence has a chain structure

Inference is easy

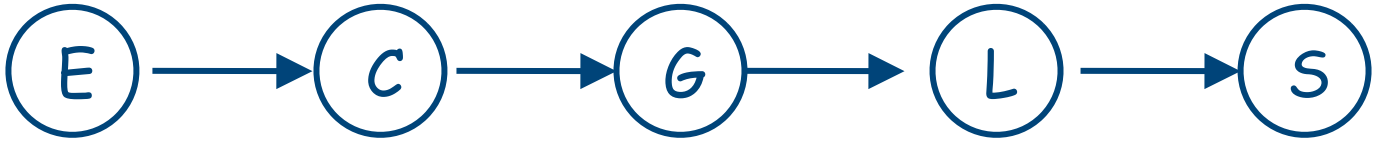


- Want to know P(g)? Use marginalization!

$$
\begin{aligned}
P(g) &= \sum_{c_i \in Dom(C)} \Pr(g \mid c_i)\Pr(c_i) \\
&= \sum_{c_i \in Dom(C)} \Pr(g \mid c_i) \sum_{e_i \in Dom(E)} \Pr(c_i \mid e_i)\Pr(e_i)
\end{aligned}
$$

These are all terms specified in our local distributions!

Inference is Easy



- Computing P(g) in more concrete terms

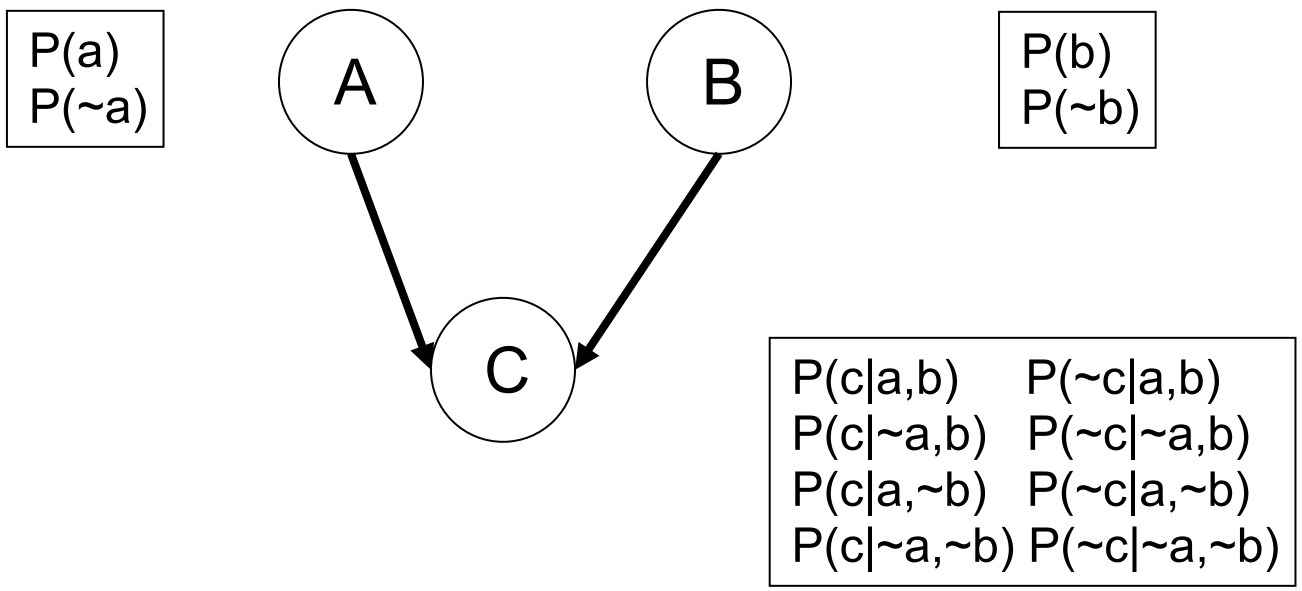
Bayesian Networks

- The structure just introduced is a Bayesian Network
  - **Graphical representation** of direct dependencies over a set of variables + a set of **conditional probability distributions** (CPTs) quantifying the strength of the influences

Bayesian Networks

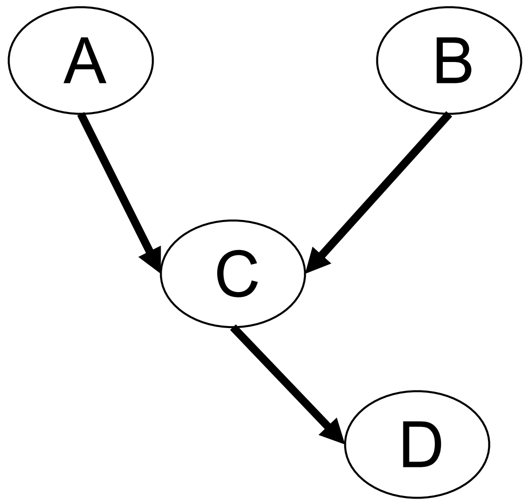
(aka belief networks, causal networks, probabilistic networks...)

- A BN over a set of variables $\{X_1,...,X_n\}$ consists of
  - A directed acyclic graph whose nodes are the variables
  - A set of CPTs $(P(X_i|Parents(X_i))$ for each X_i



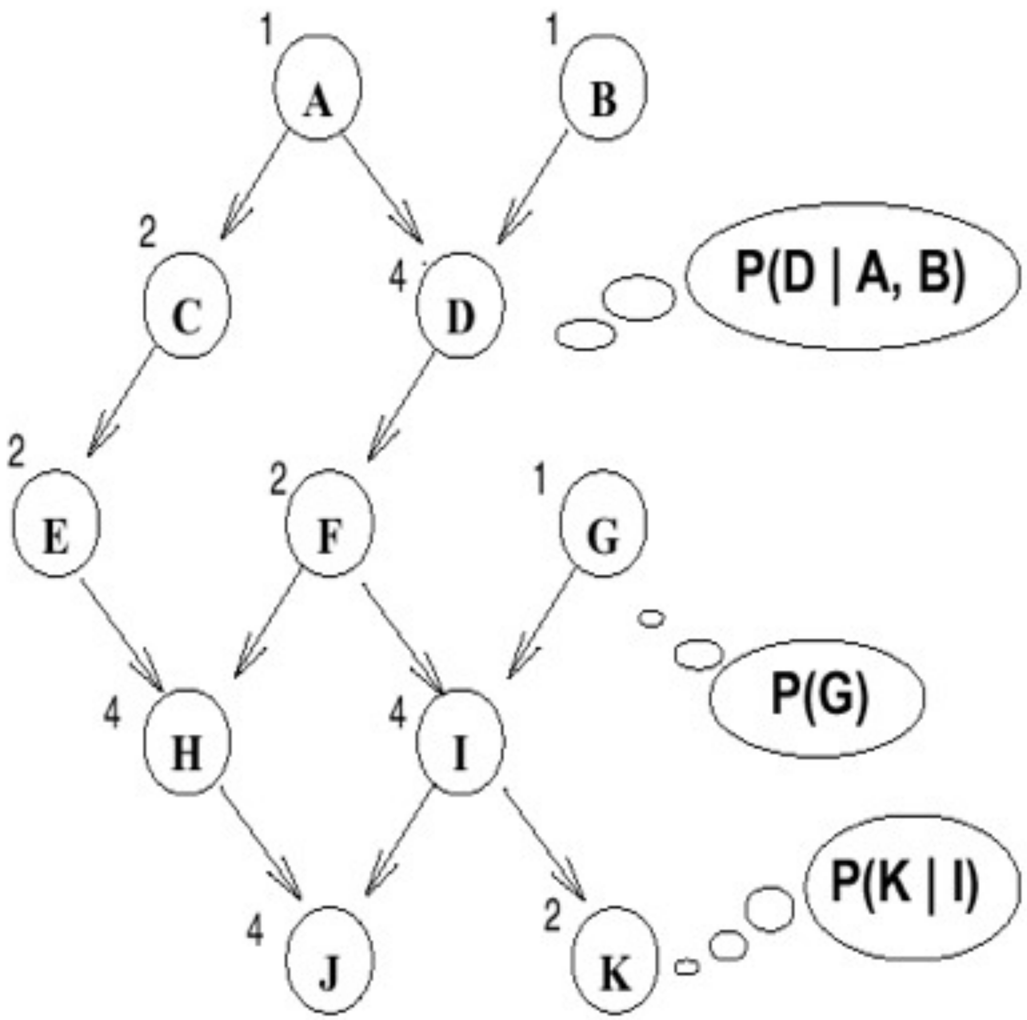
Bayesian Networks

- Key notions
  - **parents** of a node: Par(X_i)
  - **children** of a node
  - **descendents** of a node
  - **ancestors** of a node
  - **family**: set of nodes consisting of X_i and its parents
    - CPT are defined over families



Parents(C)={A,B}
Children(A)={C}
Descendents(B)={C,D}
Ancestors{D}={A,B,C}
Family{C}={C,A,B}

Bayes Net Example



- A couple CPTS are "shown"
- Explicit joint requires 2^{11} -1 =2047 params
- BN requires only 27 parms (the number of entries for each

Semantics

- The structure of the BN means: *every X_i is conditionally independent of all of its nondescendents given its parents*

$$\Pr(X_i \mid S \cup Par(X_i)) = \Pr(X_i \mid Par(X_i))$$

for any subset $S \subseteq NonDescendants(X_i)$

Semantics

- Imagine we make the query $P(x_1,x_2,...,x_n)$
  - $= P(x_n|x_{n-1},...,x_1)P(x_{n-1}|x_{n-2},...,x_1)...P(x_1)$
  - $= P(x_n|Par(x_n)P(x_{n-1}|Par(x_{n-1}))...P(x_1)$

- The joint is recoverable using the parameters (CPT) specified in an arbitrary BN

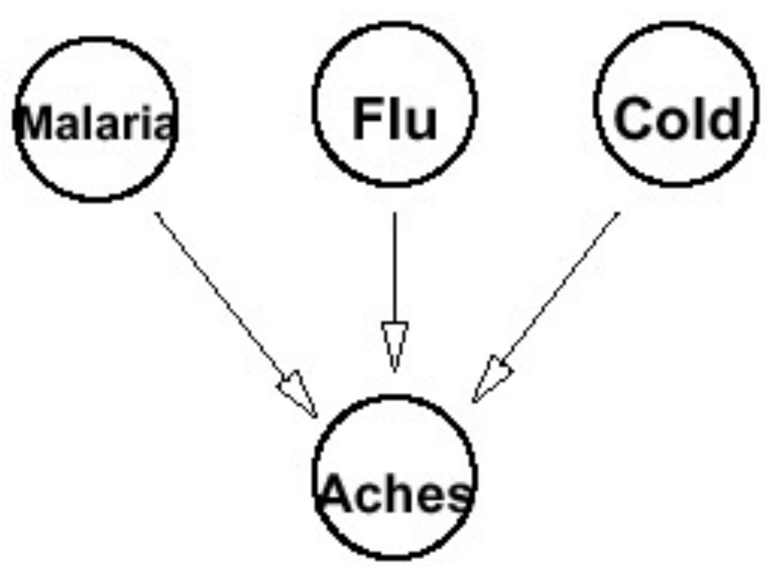
Constructing a BN

- Given any distribution over variables $X_1, X_2, \ldots, X_n$, we can construct a BN that faithfully represents that distribution

Take any ordering of the variables (say, the order given), and go through the following procedure for X_n down to X_1. Let Par(X_n) be any subset S $\subseteq$ $\{X_1, \ldots, X_{n-1}\}$ such that X_n is independent of $\{X_1, \ldots, X_{n-1}\}$ - S given S. Such a subset must exist. Then determine the parents of X_{n-1} in the same way, finding a similar S $\subseteq$ $\{X_1, \ldots, X_{n-2}\}$, and so on. In the end, a DAG is produced and the BN semantics must hold by construction.

Causal Intuitions

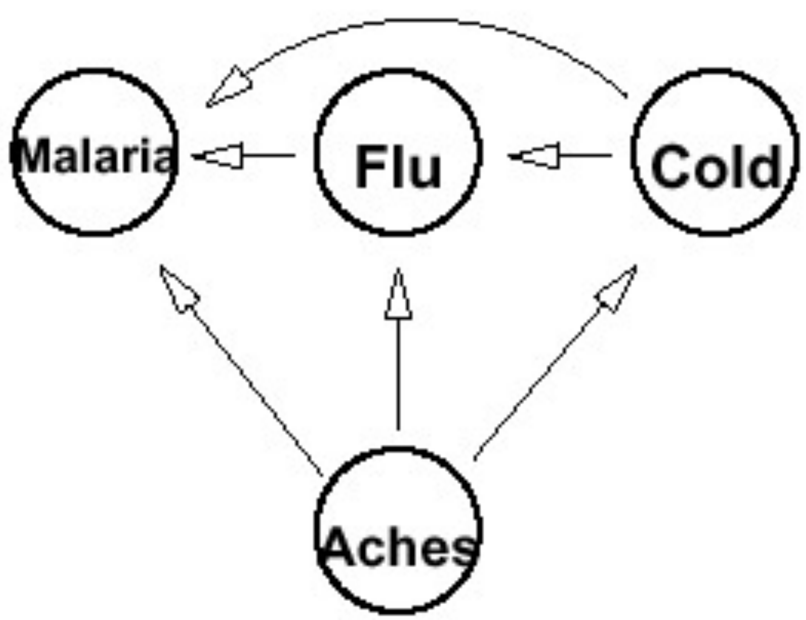
- The construction of a BN is simple
  - Works with arbitrary orderings of variable set
  - But some orderings are much better than others
  - Generally, if ordering/dependence structure reflects causal intuitions, we get a more compact BN



- In this BN, we've used the ordering Malaria, Cold, Flu, Aches to build BN for distribution P for Aches
  - Variable can only have parents that come earlier in the ordering

Causal Intuitions

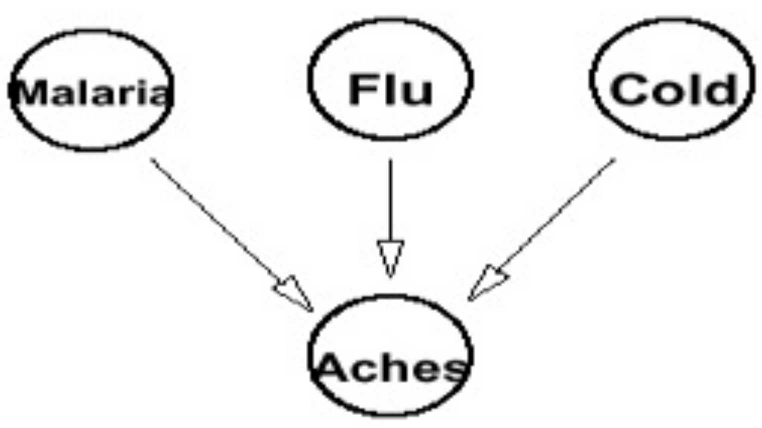
- We could have used a different ordering
  - Aches, Cold, Flu, Malaria



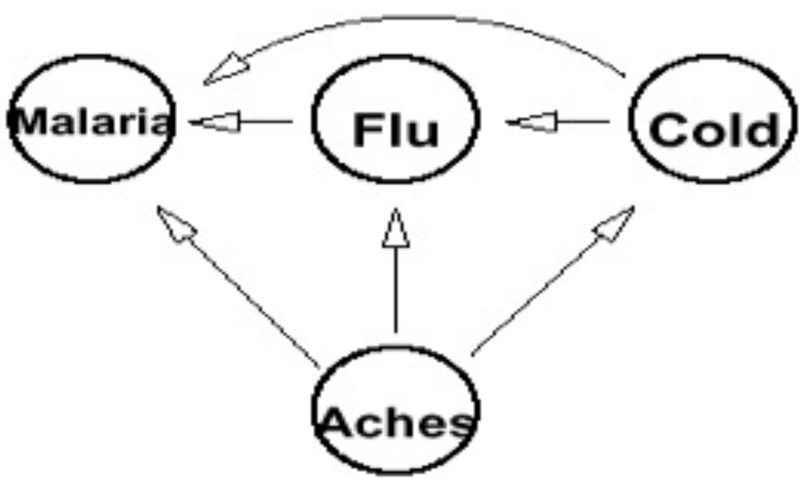
- Mal depends on Aches; but it also depends on Cold, Flu *given* Aches
  - Cold, Flu explain away Mal given Aches
- Flu depends on Aches; but also on Cold *given* Aches
- Cold depends on Aches

Compactness

- In general, if each random variable is directly influenced by at most k others then each CPT will be at most 2^k. Thus the entire network of n variables can be specified by $n2^k$



1+1+1+8=11 numbers



1+2+4+8=15 numbers

Testing Independence

- Given a BN, how we do determine if two variables X and Y are independent given evidence E?
  - We use a simple graphical property
- **D-separation**: *A set of variables E d-separates X and Y if it blocks every undirected path between X and Y*
- X and Y are conditionally independent given E if E d-separates X and Y

Blocking

- P is an undirected path from X to Y in BN. Let **E** be evidence set. **E** blocks path P iff there is some node in Z on the path such that
  - **Case 1**: one arc on P goes into Z and one goes out of Z and Z in **E**, or
  - **Case 2**: both arcs on P leave Z and Z in **E**, or
  - **Case 3:** both arcs on P enter Z and neither Z, nor any of its descendents, are in **E**

Blocking

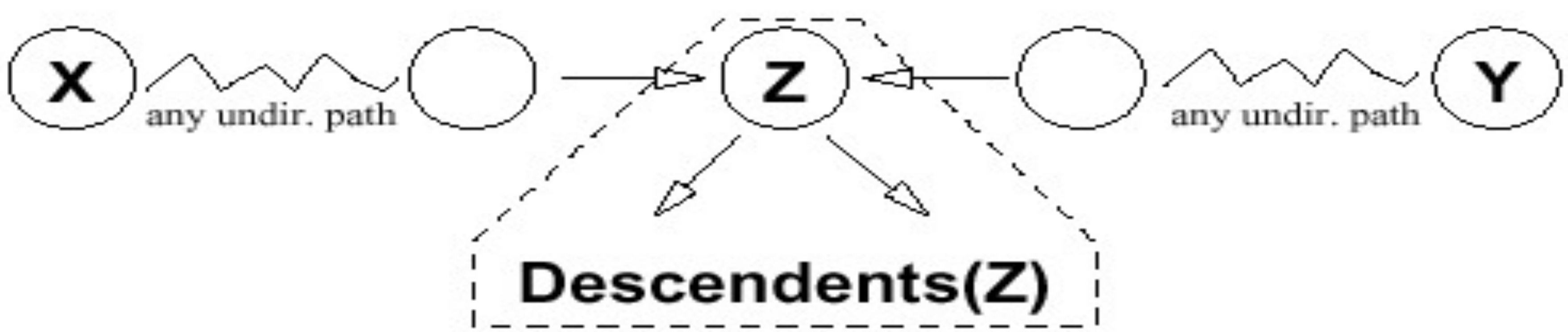
(1)

If Z in evidence, the path between X and Y blocked

(2)

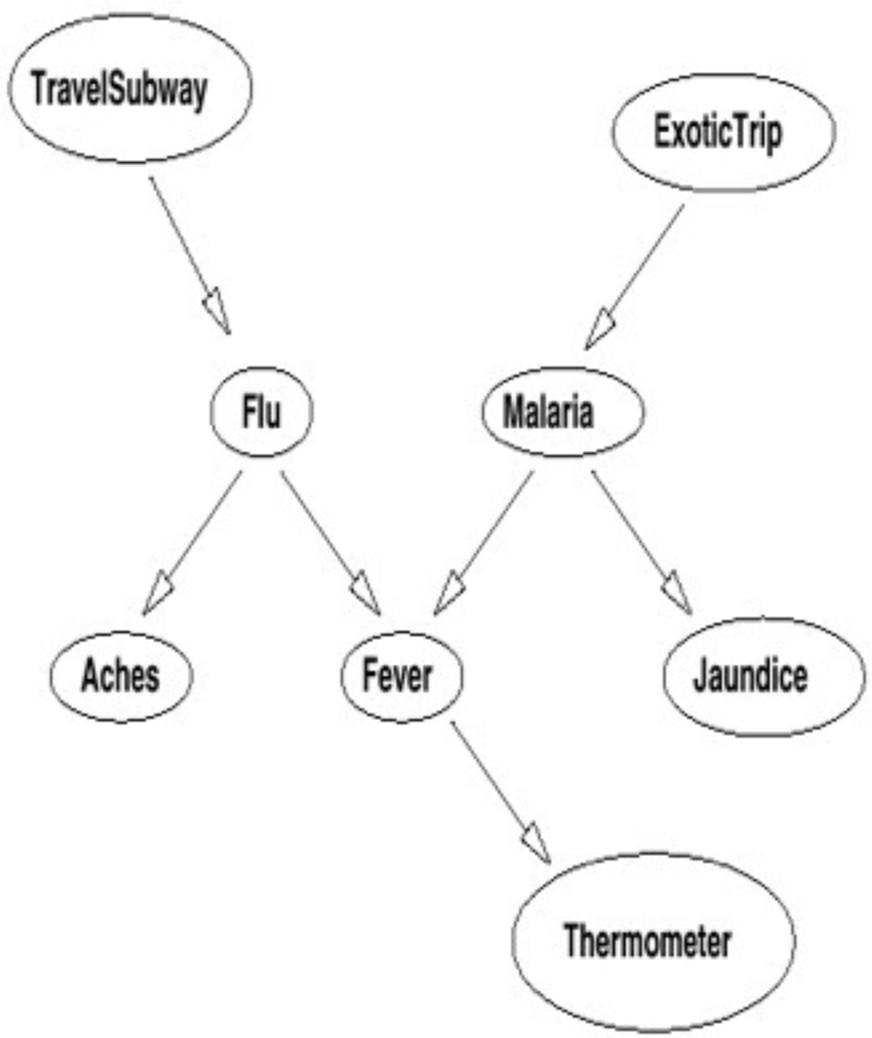
If Z in evidence, the path between X and Y blocked

(3)



If Z is ***not*** in evidence and ***no*** descendent of Z is in evidence, then the path between X and Y is blocked

Examples



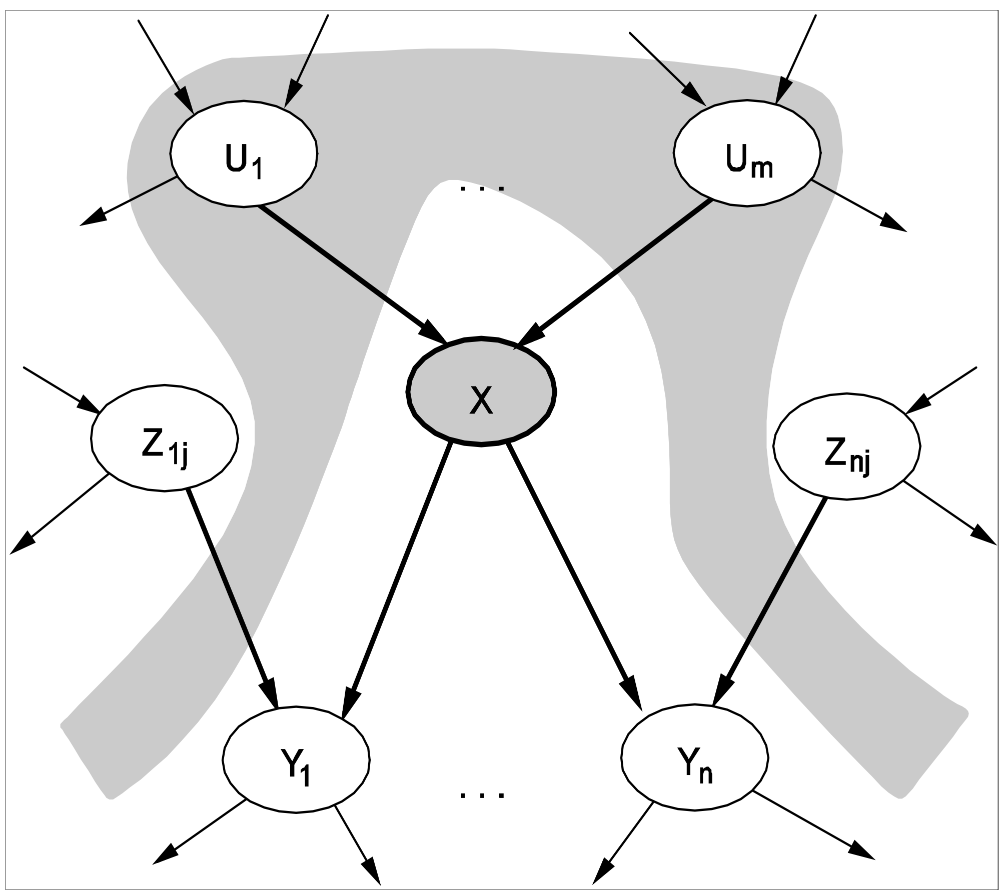
1. Subway and Thermometer?
2. Aches and Fever?
3. Aches and Thermometer?
4. Flu and Malaria?
5. Subway and ExoticTrip?

D-Separation

- Can be computed in linear time with a depth-first search like algorithm
- Useful since now have a linear time algorithm for automatically inferring whether learning the value of one variables might given us any additional info about some other variable, given when we already know
  - "Might" since vars might be conditionally independent but **not** d-seperated

Other ways of determining conditional independence

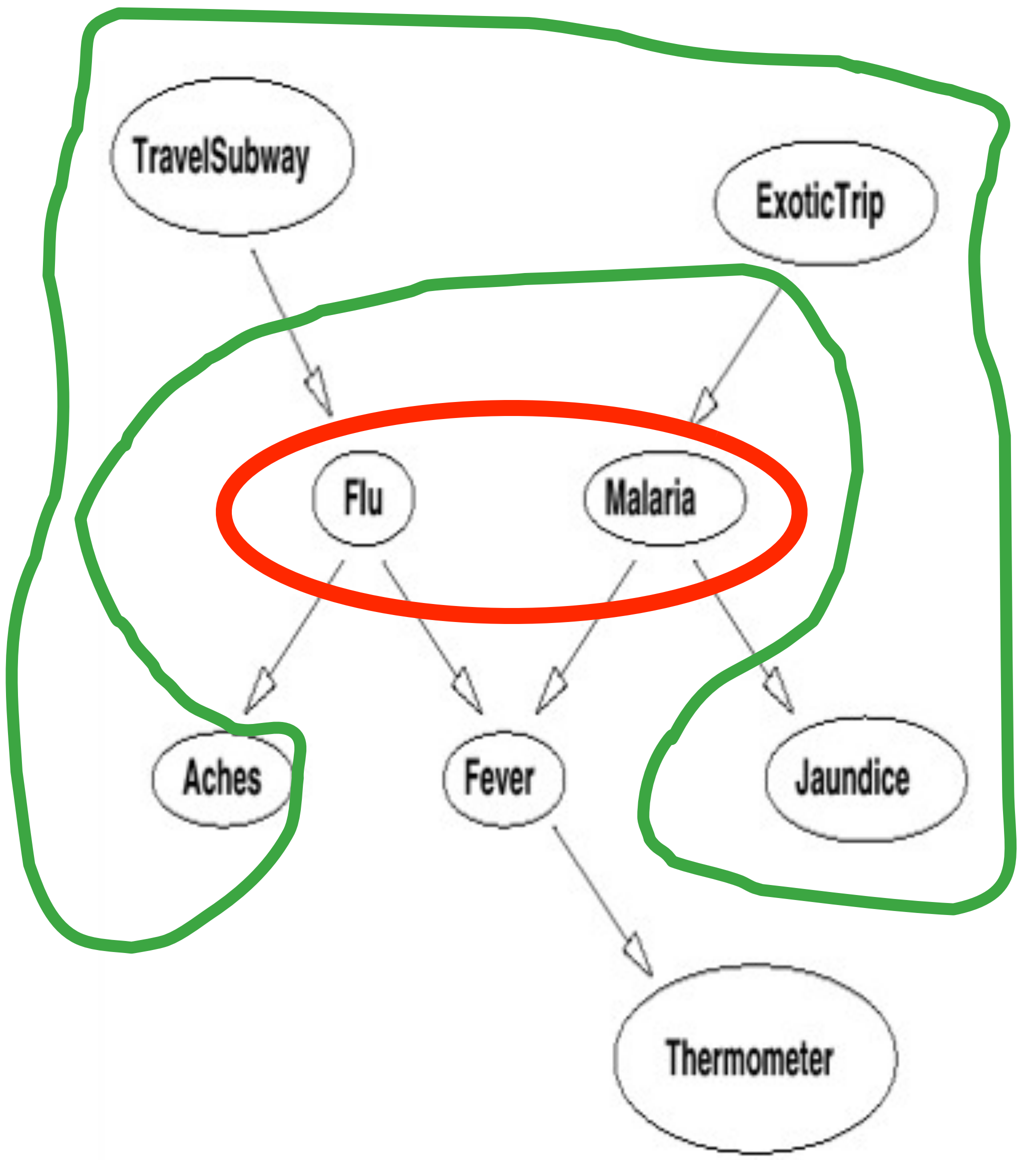
- Non-descendents



A node is conditionally independent of its non-descendents, given its parents.

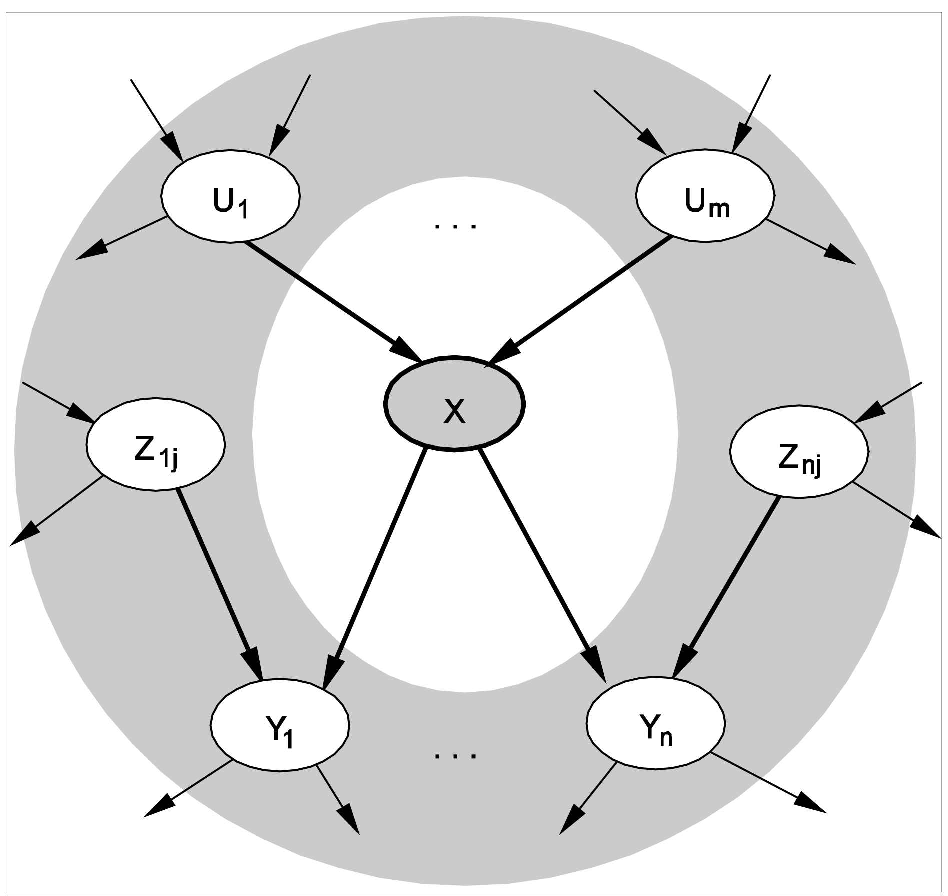
X is conditionally independent of the Z_{ij}s given U_is

Example



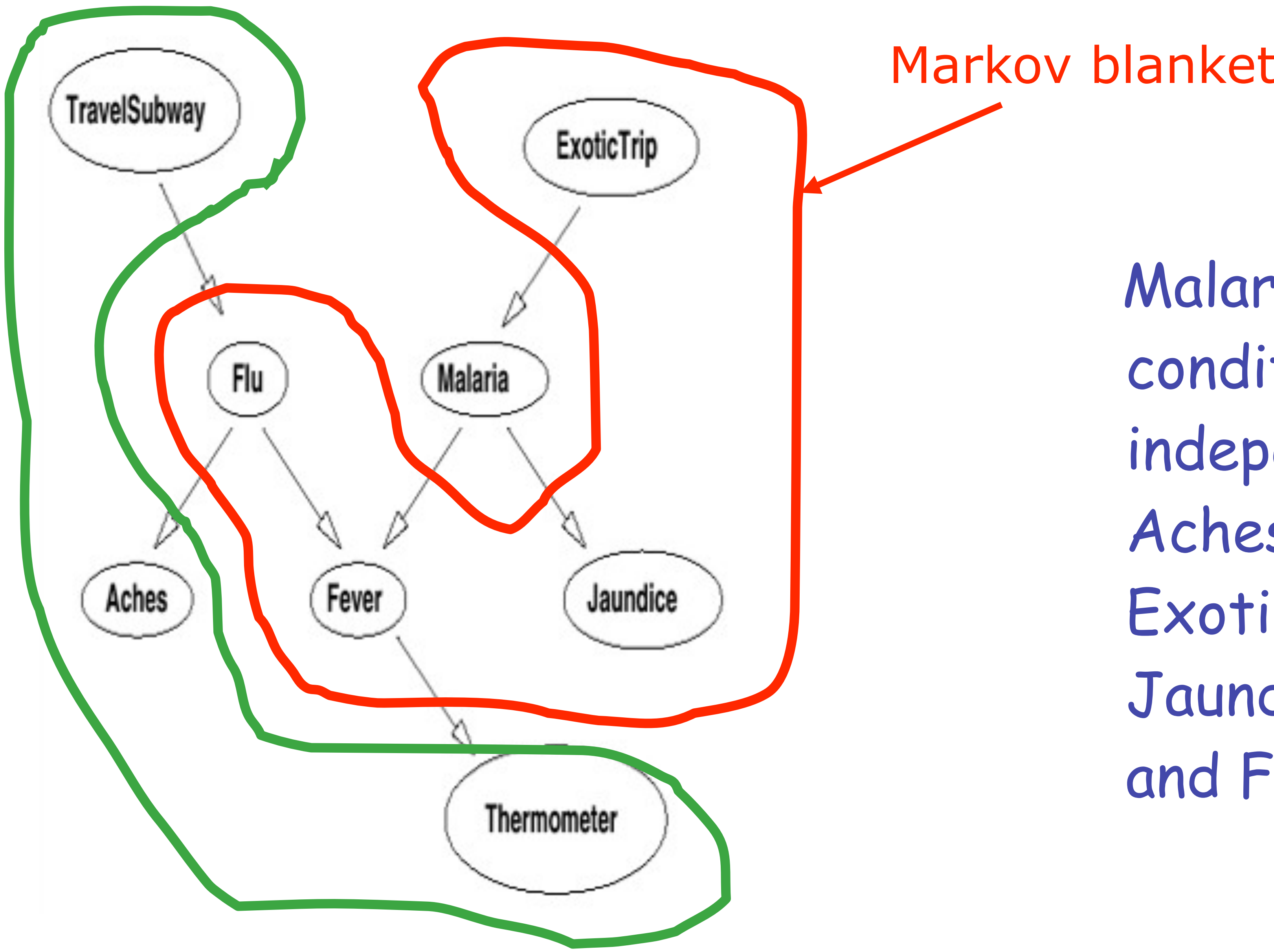
Fever is conditionally independent of Jaundice given Malaria and Flu

Markov Blanket



A node is conditionally independent of all other nodes in the network, given its parents, children and children's parents (Markov blanket).

Markov Blanket



Malaria is conditionally independent of Aches given ExoticTrip, Jaundice, Fever and Flu

Next Class

- Inference in Bayes Nets!