

CSE458 Bayesian Networks

Lecture 3

Ann E. Nicholson

Faculty of Information Technology
Monash University
Clayton, VIC 3168
AUSTRALIA

{annn}@csse.monash.edu.au
HTTP://WWW.CSSE.MONASH.EDU.AU/BAI

Text: *Bayesian Artificial Intelligence*, Kevin B. Korb and Ann E. Nicholson, Chapman & Hall/CRC, 2004.

CSE458 2007

Introduction to Bayesian Networks

- Nodes, structure and probabilities
- Reasoning with BNs
- Understanding BNs

CSE458 2007

Nicholson & Korb

3

Bayesian Networks

- Data Structure which represents the dependence between variables.
- Gives concise specification of the joint probability distribution.
- A Bayesian Network is a graph in which the following holds:
 1. A set of random variables makes up the nodes in the network.
 2. A set of directed links or arrows connects pairs of nodes.
 3. Each node has a conditional probability table that *quantifies* the effects the parents have on the node.
 4. Directed, acyclic graph (DAG), i.e. no directed cycles.

CSE458 2007

Nicholson & Korb

4

Example: Lung Cancer Diagnosis

A patient has been suffering from shortness of breath (called dyspnoea) and visits the doctor, worried that he has lung cancer. The doctor knows that other diseases, such as tuberculosis and bronchitis are possible causes, as well as lung cancer. She also knows that other relevant information includes whether or not the patient is a smoker (increasing the chances of cancer and bronchitis) and what sort of air pollution he has been exposed to. A positive XRay would indicate either TB or lung cancer.

CSE458 2007

Nodes and Values

Q: What are the nodes to represent and what values can they take?

Nodes can be discrete or continuous; will focus on discrete for now.

- Boolean nodes: represent propositions, taking binary values true (*T*) and false (*F*).

Example: *Cancer* node represents proposition “the patient has cancer”.

- Ordered values..

Example: *Pollution* node with values {*low*, *medium*, *high*}.

- Integral values.

Example: *Age* node with possible values from 1 to 120.

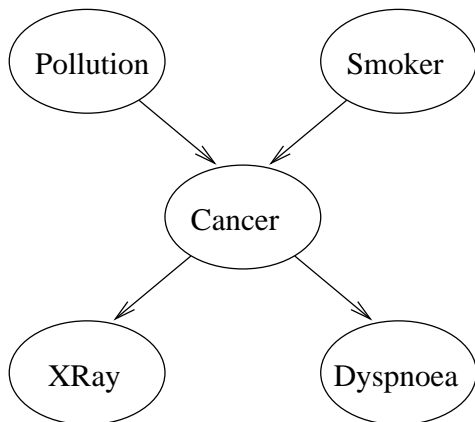
CSE458 2007

Lung cancer example: nodes and values

Node name	Type	Values
<i>Pollution</i>	Binary	{ <i>low</i> , <i>high</i> }
<i>Smoker</i>	Boolean	{ <i>T</i> , <i>F</i> }
<i>Cancer</i>	Boolean	{ <i>T</i> , <i>F</i> }
<i>Dyspnoea</i>	Boolean	{ <i>T</i> , <i>F</i> }
<i>XRay</i>	Binary	{ <i>pos</i> , <i>neg</i> }

CSE458 2007

Lung cancer example: network structure



Note: No explicit representation of other causes of cancer, or other causes of symptoms.

CSE458 2007

Structure terminology and layout

- Family metaphor:
Parent \Rightarrow *Child*
Ancestor $\Rightarrow \dots \Rightarrow$ *Descendant*
- Markov Blanket = parents + children + children's parents
- Tree analogy:
 - **root** node: no parents
 - **leaf** node: no children
 - **intermediate** node: non-leaf, non-root
- Layout convention: root notes at top, leaf nodes at bottom, arcs point down the page.

CSE458 2007

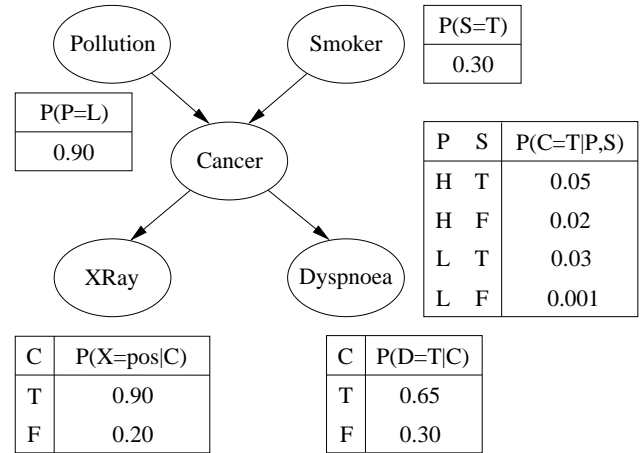
Conditional Probability Tables

Once specified topology, need to specify **conditional probability table (CPT)** for each node.

- Each row contains the conditional probability of each node value for a each possible combination of values of its parent nodes.
- Each row must sum to 1.
- A table for a Boolean var with n Boolean parents contain 2^{n+1} probs.
- A node with no parents has one row (the prior probabilities)

CSE458 2007

Lung cancer example: CPTs



CSE458 2007

The Markov Property

- Modelling with BNs requires the assumption of the **Markov Property**:
there are no direct dependencies in the system being modeled which are not already explicitly shown via arcs.
- Example: there is no way for smoking to influence dyspnoea except by way of causing cancer.
- BNs which have the Markov property are called Independence-Maps (I-Maps).
- Note: existence of arc does not have to correspond to real dependency in the system being modelled - can be nullified in the CPT.

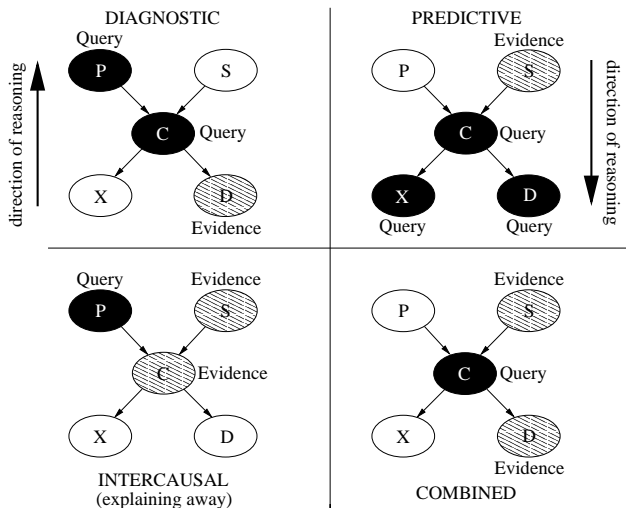
CSE458 2007

Reasoning with Bayesian Networks

- Basic task for any probabilistic inference system:
Compute the posterior probability distribution for a set of **query variables**, given new information about some **evidence variables**.
- Also called *conditioning* or *belief updating* or *inference*.

CSE458 2007

Types of Reasoning



CSE458 2007

- Specific evidence: a definite finding that a node X has a particular value, x .

Example: $Smoker=T$

- Negative evidence: a finding that node Y is *not* in state y_1 (but may take any other values).
- “Virtual” or “likelihood” evidence: source of information is not sure about it.

Example:

- $e =$ Radiologist is 80% sure that $Xray=pos$
- Want e.g.:

$$P(Cancer|e) = P(Cancer|Xray, e)P(Xray|e) + P(Cancer|\neg Xray, e)P(\neg Xray|e)$$

- **Jeffrey Conditionalization**

CSE458 2007

Reasoning with numbers

- Reasoning with lung cancer example using Netica BN software.

(See Table 2.2 in *Bayesian AI* text.)

CSE458 2007

Understanding of Bayesian Networks (Semantics)

- A (more compact) representation of the joint probability distribution.
 - helpful in understanding how to construct network
- Encoding a collection of conditional independence statements.
 - helpful in understanding how to design inference procedures
 - via *Markov property / I-map*:
 - Each conditional independence implied by the graph is present in the probability distribution

CSE458 2007

probability distribution

Write $P(X_1 = x_1, \dots, X_n = x_n)$ as $P(x_1, x_2, \dots, x_n)$.

Factorization (chain rule):

$$\begin{aligned}
 P(x_1, x_2, \dots, x_n) &= P(x_1) \times \dots \times P(x_n | x_1, \dots, x_{n-1}) \\
 &= \prod_i P(x_i | x_1, \dots, x_{i-1})
 \end{aligned}$$

Since BN structure implies that the value of a particular node is conditional *only* on the values of its parent nodes, this reduces to

$$P(x_1, x_2, \dots, x_n) = \prod_i P(x_i | Parents(X_i))$$

provided $Parents(X_i) \subseteq \{x_1, \dots, x_{i-1}\}$.

$$\begin{aligned}
 P(X = pos \wedge D = T \wedge C = T \wedge P = lo \wedge S = F) \\
 &= P(X = pos | D = T, C = T, P = lo, S = F) \\
 &\quad \times P(D = T | C = T, P = lo, S = F) \\
 &\quad \times P(C = T | P = lo, S = F) P(P = lo | S = F) P(S = F) \\
 &= P(X = pos | C = T) P(D = T | C = T) \\
 &\quad \times P(C = T | P = lo, S = F) P(P = lo) P(S = F)
 \end{aligned}$$

CSE458 2007

Pearl's Network Construction Algorithm

1. Choose the set of relevant variables $\{X_i\}$ that describe the domain.
2. Choose an ordering for the variables, $\langle X_1, \dots, X_n \rangle$.
3. While there are variables left:
 - (a) Add the next variable X_i to the network.
 - (b) Add arcs to the X_i node from some minimal set of nodes already in the net, $Parents(X_i)$, such that the following conditional independence property is satisfied:

$$P(X_i | X'_1, \dots, X'_m) = P(X_i | Parents(X_i))$$

where X'_1, \dots, X'_m are all the variables preceding X_i , including $Parents(X_i)$.

- (c) Define the CPT for X_i .

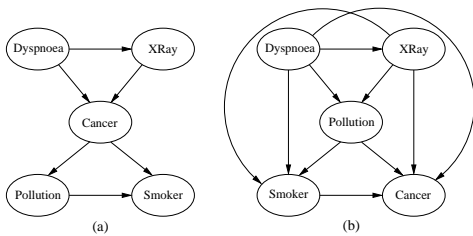
CSE458 2007

Compactness and Node Ordering

- Compactness of BN depends upon sparseness of the system.
- The best order to add nodes is to add the “root causes” first, then the variable they influence, so on until “leaves” reached.

→ Causal structure

- Alternative structures using different orderings (a) $\langle D, X, C, P, S \rangle$ (b) $\langle D, X, P, S, C \rangle$.

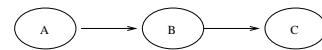


1. These BNs still represent same joint distribution.
2. Structure (b) requires as many probabilities as the full joint distribution! See below for *why*.

CSE458 2007

Conditional Independence: Causal Chains

Causal chains give rise to conditional independence:



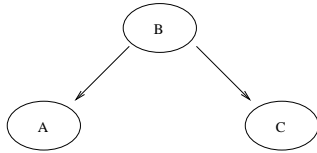
$$P(C | A \wedge B) = P(C | B)$$

Example: “smoking causes cancer which causes dyspnoea”

CSE458 2007

Conditional Independence: Common Causes

Common causes (or ancestors) also give rise to conditional independence:



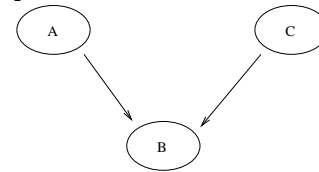
$$P(C|A \wedge B) = P(C|B) \equiv A \perp\!\!\!\perp C|B$$

Example: cancer is a common cause of the two symptoms, a positive XRay result and dyspnoea.

CSE458 2007

Conditional Dependence: Common Effects

Common effects (or their descendants) give rise to conditional *dependence*:



$$P(A|C \wedge B) \neq P(A)P(C) \equiv \neg(A \perp\!\!\!\perp C|B)$$

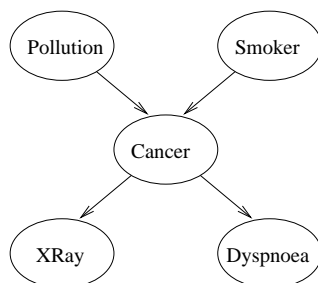
Example: Cancer is a common effect of pollution and smoking.

Given lung cancer, smoking “explains away” pollution.

CSE458 2007

D-separation

- Graphical criterion of conditional independence.
 $X \perp\!\!\!\perp Y|Z$
- We can determine whether a set of nodes X is independent of another set Y , given a set of evidence nodes E , via the Markov property:
 $X \perp\!\!\!\perp Y|E \rightarrow X \perp\!\!\!\perp Y|E$.
- Example



CSE458 2007

D-separation

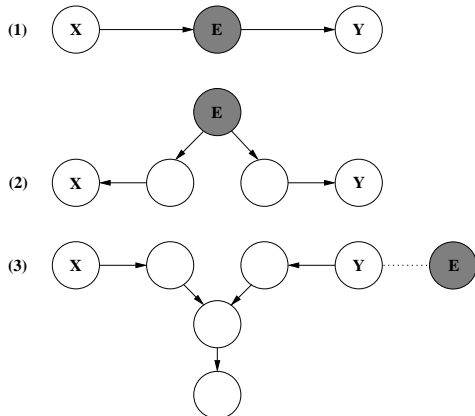
How to determine d-separation, $X \perp\!\!\!\perp Y|E$:

- If every undirected path from a node in X to a node in Y is *d-separated* by E , then X and Y are *conditionally independent* given E .
- A set of nodes E *d-separates* two sets of nodes X and Y if every undirected path from a node in X to a node in Y is *blocked* given E .
- A path is *blocked* given a set of nodes E if there is a node Z on the path for which one of three conditions holds:
 - Z is in E and Z has one arrow on the path leading in and one arrow out (chain).
 - Z is in E and Z has both path arrows leading out (common cause).
 - Neither Z nor any descendant of Z is in E , and both path arrows lead in to Z (common effect).

CSE458 2007

D-separation (cont'd)

- Evidence nodes **E** shown shaded.



CSE458 2007

Causal Ordering

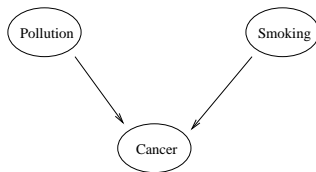
Why does variable order affect network density?

Because

- Using the causal order allows direct representation of conditional independencies
- Violating causal order requires new arcs to re-establish conditional independencies

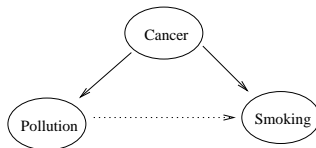
CSE458 2007

Causal Ordering (cont'd)



Pollution and *Smoking* are marginally independent.

Given the ordering: Cancer, Pollution, Smoking:



Marginal independence of *Pollution* and *Smoking* must be re-established by adding $Pollution \rightarrow Smoking$ or $Smoking \leftarrow Pollution$

CSE458 2007

Bayesian Networks: Summary

- Bayes' rule allows unknown probabilities to be computed from known ones.
- Conditional independence (due to causal relationships) allows efficient updating
- BNs are a natural way to represent conditional independence info.
 - links between nodes: qualitative aspects;
 - conditional probability tables: quantitative aspects.
- Probabilistic inference: compute the probability distribution for query variables, given evidence variables
- BN Inference is very flexible: can enter evidence about any node and update beliefs in any other nodes.

CSE458 2007

Inference Algorithms: Overview

- Exact inference
 - Trees and polytrees:
 - * message-passing algorithm
 - Multiply-connected networks:
- Approximate Inference
 - Large, complex networks:
 - * Stochastic Simulation
 - * Other approximation methods
- In the general case, both exact and approximate inference are computationally complex (“NP-hard”).
- Causal inference

Inference in chains

Two node network $X \rightarrow Y$.

- Evidence $X = x$, then $Bel(Y) = P(Y|X = x)$ straight from CPT.
- Evidence $Y = y$

$$\begin{aligned}
 Bel(X = x) &= P(X = x|Y = y) \\
 &= \frac{P(Y = y|X = x)P(X = x)}{P(y)} \\
 &= \alpha P(x)\lambda(x)
 \end{aligned}$$

where

$$\alpha = \frac{1}{P(Y = y)}$$

$P(x)$ is the prior, and $\lambda(x) = P(Y = y|X = x)$ is the **likelihood**.

Since $\sum_i Bel(Y = y_i) = 1$, we can compute α as a **normalizing constant**.

Example: $Flu \rightarrow HighTemp$

Suppose $P(Flu = T) = 0.05$,
 $P(HighTemp = T|Flu = T) = 0.9$,
 $P(HighTemp = T|Flu = F) = 0.2$.

Given evidence $HighTemp = T$, then

$$\begin{aligned}
 Bel(Flu = T) &= \alpha P(Flu = T)\lambda(Flu = T) \\
 &= \alpha \times 0.05 \times 0.9 = \alpha 0.045 \\
 Bel(Flu = F) &= \alpha P(Flu = F)\lambda(Flu = F) \\
 &= \alpha \times 0.95 \times 0.2 = \alpha 0.19
 \end{aligned}$$

We can compute α via

$$Bel(Flu = T) + Bel(Flu = F) = 1 = \alpha 0.045 + \alpha 0.19$$

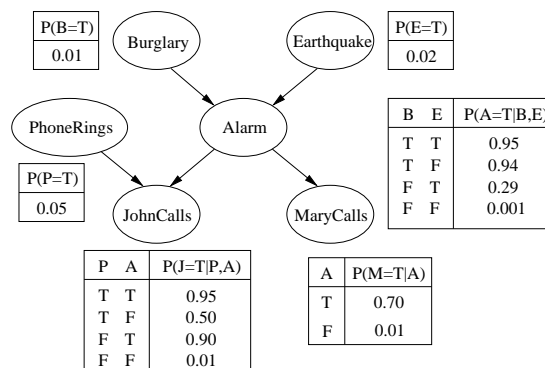
giving

$$\alpha = \frac{1}{0.19 + 0.045}$$

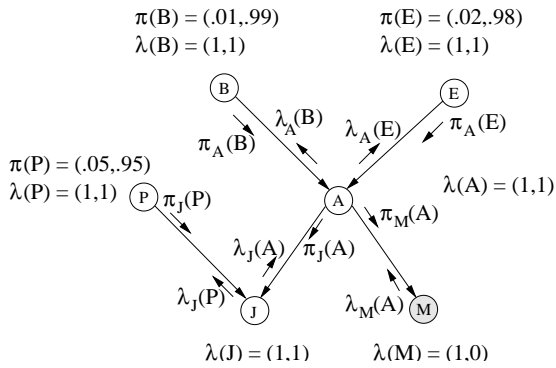
This allows us to finish the belief update:

$$\begin{aligned}
 Bel(Flu = T) &= \frac{0.045}{0.19 + 0.045} = 0.8085 \\
 Bel(Flu = F) &= \frac{0.19}{0.19 + 0.045} = 0.1915
 \end{aligned}$$

Earthquake example



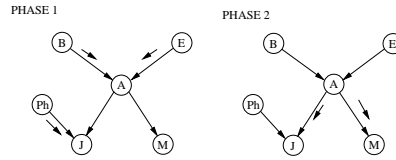
Inference in polytrees: message passing



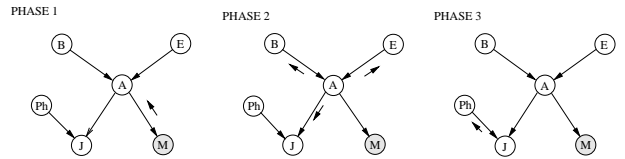
CSE458 2007

Message propagation

PROPAGATION, NO EVIDENCE



PROPAGATION, EVIDENCE for node M



CSE458 2007

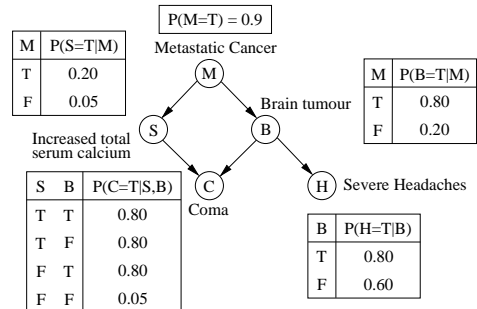
Message-passing algorithm: features

- All computations are local \Rightarrow efficient
- Requires summation over all joint instantiations of parent nodes \Rightarrow exponential in no. of parents.
- No. of propagation steps depends on length of longest path

CSE458 2007

Inference in multiply connected networks

- Networks where two nodes are connected by more than one path
 - Two or more possible causes which share a common ancestor
 - One variable can influence another through more than one causal mechanism
- Example: Cancer network

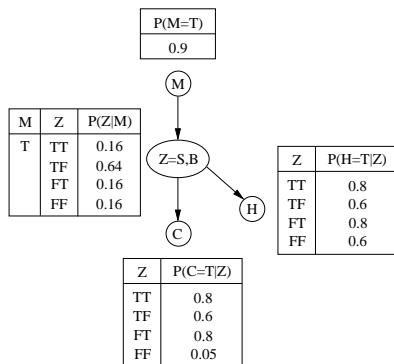


- Message passing doesn't work - evidence gets "counted twice"

CSE458 2007

Clustering methods

- Transform network into a probabilistically equivalent polytree by merging (clustering) offending nodes
- Cancer example: new node Z combining B and C



$$P(z|a) = P(b, c|a) = P(b|a)P(c|a)$$

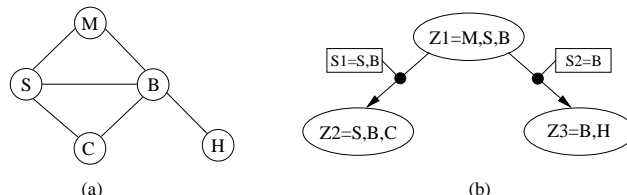
$$P(e|z) = P(e|b, c) = P(e|c)$$

$$P(d|z) = P(d|b, c)$$

CSE458 2007

Jensen join-tree method

- Jensen Join-tree (Jensen, 1996) version the current most efficient algorithm in this class (e.g. used in Hugin, Netica).



CSE458 2007

Jensen join-tree method (cont.)

Network evaluation done in two stages

- Compile into join-tree
 - May be slow
 - May require too much memory if original network is highly connected
- Do belief updating in join-tree (usually fast)

Caveat: clustered nodes have increased complexity; updates may be computationally complex

CSE458 2007

Approximate inference with stochastic simulation

- Use the network to generate a large number of cases that are consistent with the network distribution.
- Evaluation may not converge to exact values (in reasonable time).
- Usually converges to close to exact solution quickly if the evidence is not too unlikely.
- Performs better when evidence is nearer to root nodes, however in real domains, evidence tends to be near leaves (Nicholson&Jitnah, 1998)

CSE458 2007

Causal modeling

We should like to do causal modeling with our Bayesian networks.

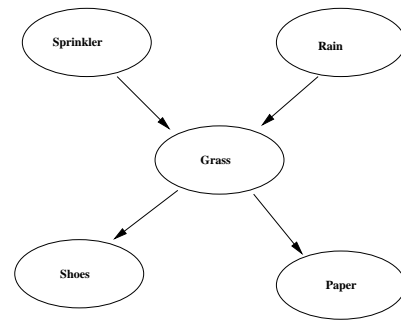
Prerequisite: arcs are truly causal (hence, nodes are properly ordered).

Reasoning about real or hypothetical interventions:

- what if we upgrade quality in manufacturing?
- what if we treat the patient with X, Y, Z?

For planning, control, prediction.

Common practice appears to be: let observation stand for intervention.



If we observe that the lawn is wet:

- We can infer in any direction; everything updates
- We get, e.g., “explaining away” between causes

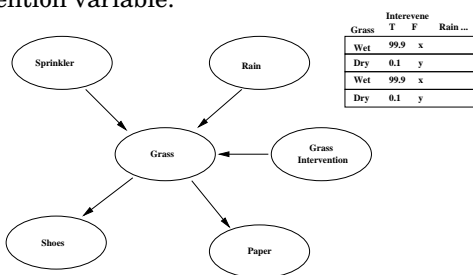
What happens if we *intervene* in a causal process?

Spirtes, et al., (1993), Pearl (2000) answer: cut links to parents and *then* update.

- No explaining away; parents are then unaffected
- Downstream updating is as normal

Causal inference

We prefer (conceptually) to augment the graph with an intervention variable:



- Simplistically, parent connections are severed
- With full generality, X acquires a new parent D_X
 - Allows any degree of control for intervention
 - Allows any kind of interaction with existing parents
 - Bayesian update algorithms unaffected

Inference: Summary

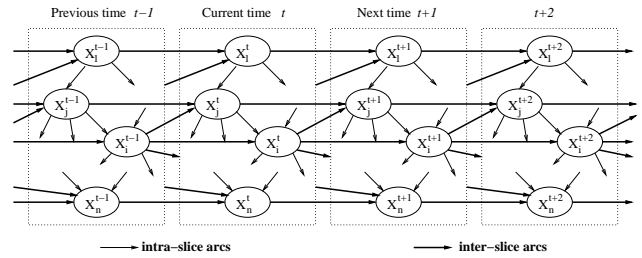
- Probabilistic inference: compute the probability distribution for query variables, given evidence variables
- Causal inference: compute the probability distribution for query variables, given intervention
- BN Inference is very flexible: can enter evidence about any node and update beliefs in any other nodes.
- The speed of inference in practice depends on the structure of the network: how many loops; numbers of parents; location of evidence and query nodes.
- BNs can be used to model causal intervention.

Extensions to Bayesian Networks

- For decision making: decision networks (David Albrecht, Lecture 4)
- For reasoning about changes over time: dynamic Bayesian networks

CSE458 2007

Dynamic Belief Networks



- One node for each variable for each time step.

- **Intra-slice** arcs $X_i^T \rightarrow X_j^T$

- **Inter-slice (temporal)** arcs

1. $X_i^T \rightarrow X_i^{T+1}$

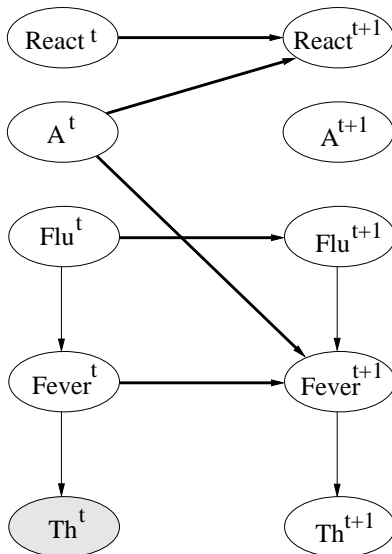
2. $X_i^T \rightarrow X_j^{T+1}$

CSE458 2007

Nicholson & Korb

47

Fever DBN



CSE458 2007

Nicholson & Korb

48

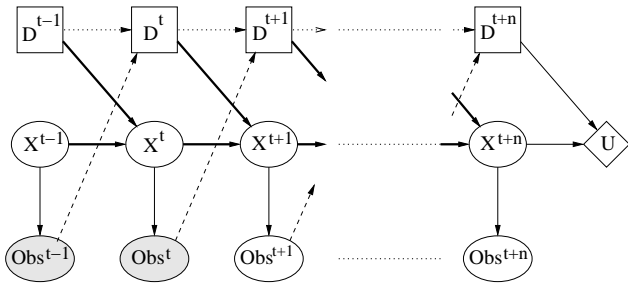
DBN reasoning

- Can calculate distributions for S_{t+1} and further: **probabilistic projection**.
- Reasoning can be done using standard BN updating algorithms
- This type of DBN gets very large, very quickly.
- Usually only keep two time slices of the network.

CSE458 2007

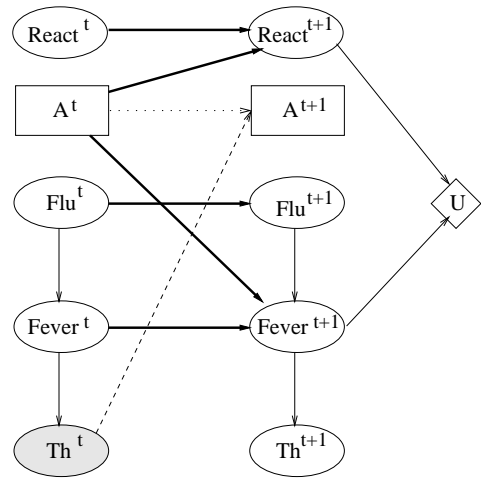
Dynamic Decision Network

- Similarly, Decision Networks can be extended to include temporal aspects.
- Sequence of decisions taken = Plan.



CSE458 2007

Fever DDN



CSE458 2007