Reasoning Under Uncertainty: Bayesian networks intro

Alan Mackworth

UBC CS 322 – Uncertainty 4

March 18, 2013

Textbook §6.3, 6.3.1, 6.5, 6.5.1, 6.5.2

Lecture Overview

- Recap: marginal and conditional independence
- Bayesian Networks Introduction
- Hidden Markov Models

Marginal Independence

Definition (Marginal independence)
Random variable X is (marginally) independent of random variable Y, written $X \perp\!\!\!\perp Y$, if for all $x \in \mathrm{dom}(X)$, $y_j \in \mathrm{dom}(Y)$ and $y_k \in \mathrm{dom}(Y)$, the following equation holds:

$$
\begin{aligned}
& P(X = x | Y = y_j) \\
= \; & P(X = x | Y = y_k) \\
= \; & P(X = x)
\end{aligned}
$$

- Intuitively: if $X \perp\!\!\!\perp Y$, then
  - learning that Y=y does not change your belief in X
  - and this is true for all values y that Y could take

- For example, weather is marginally independent of the result of a coin toss

Marginal Independence

Definition (Marginal independence)
Random variable X is (marginally) independent of random variable Y, written $X \perp\!\!\!\perp Y$, if for all $x \in \mathrm{dom}(X)$, $y_j \in \mathrm{dom}(Y)$ and $y_k \in \mathrm{dom}(Y)$, the following equation holds:

$$P(X = x | Y = y_j)$$
$$= P(X = x | Y = y_k)$$
$$= P(X = x)$$

- Recall the product rule:
  - $P(X = x \wedge Y = y) = P(X = x | Y = y) \times P(Y = y)$
- If $X \perp\!\!\!\perp Y$, we have:
  - $P(X = x \wedge Y = y) = P(X = x) \times P(Y = y)$
  - In distribution form: $P(X, Y) = P(X) \times P(Y)$
- If $X_i \perp\!\!\!\perp X_j$ for all i, j: $P(X_1, \dots, Xn) = \prod_{i=1}^{n} P(Xi)$

Conditional Independence

Definition (Conditional independence)
Random variable X is (conditionally) independent of random variable Y given random variable Z, written $X \perp\!\!\!\perp Y \mid Z$ if, for all $x \in dom(X)$, $y_j \in dom(Y)$, $y_k \in dom(Y)$ and $z \in dom(Z)$ the following equation holds:

$$\begin{aligned} & P(X = x | Y = y_j, Z = z) \\ = & P(X = x | Y = y_k, Z = z) \\ = & P(X = x | Z = z) \end{aligned}$$

- Intuitively: if $X \perp\!\!\!\perp Y \mid Z$, then
  - learning that Y=y does not change your belief in X when we already know Z=z
  - and this is true for all values y that Y could take and all values z that Z could take
- For example,
  ExamGrade $\perp\!\!\!\perp$ AssignmentGrade | UnderstoodMaterial

Conditional Independence

Definition (Conditional independence)
Random variable X is (conditionally) independent of random variable Y given random variable Z, written X ⫫ Y | Z if, for all $x \in dom(X)$, $y_j \in dom(Y)$, $y_k \in dom(Y)$ and $z \in dom(Z)$ the following equation holds:

$$\begin{aligned} &P(X = x|Y = y_j, Z = z) \\ &= P(X = x|Y = y_k, Z = z) \\ &= P(X = x|Z = z) \end{aligned}$$

- Definition of X ⫫ Y | Z in distribution form: $P(X|Y,Z) = P(X|Z)$
- Product rule still holds when every term is conditioned on Z=z:
  - $P(X = x \wedge Y = y|Z = z) = P(X = x|Y = y, Z = z) \times P(Y = y|Z = z)$
- Thus, if X ⫫ Y | Z :
  - $P(X = x \wedge Y = y|Z = z) = P(X = x|Z = z) \times P(Y = y|Z = z)$
  - In distribution form: $P(X,Y|Z) = P(X|Z) \times P(Y|Z)$

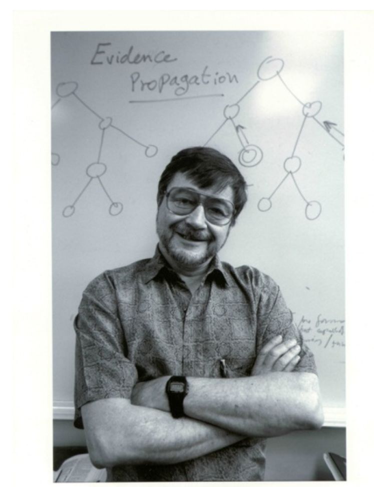
Lecture Overview

- Recap: marginal and conditional independence
- Bayesian Networks Introduction
- Hidden Markov Models

Bayesian Network Motivation

- We want a representation and reasoning system that is based on conditional (and marginal) independence
  - Compact yet expressive representation
  - Efficient reasoning procedures
- Bayes[ian] (Belief) Net[work]s are such a representation
  - Named after Thomas Bayes (ca. 1702 –1761)
  - Term coined in 1985 by Judea Pearl (1936 –)
  - Their invention changed the *primary* focus of AI from logic to probability!

Thomas Bayes

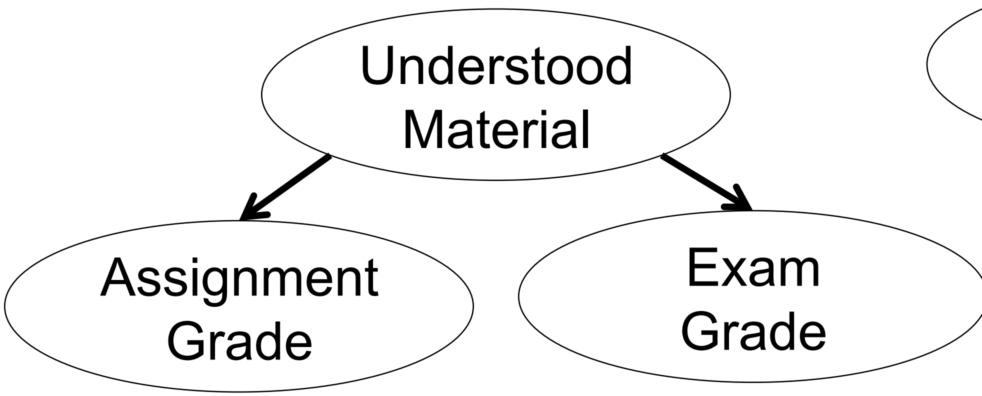


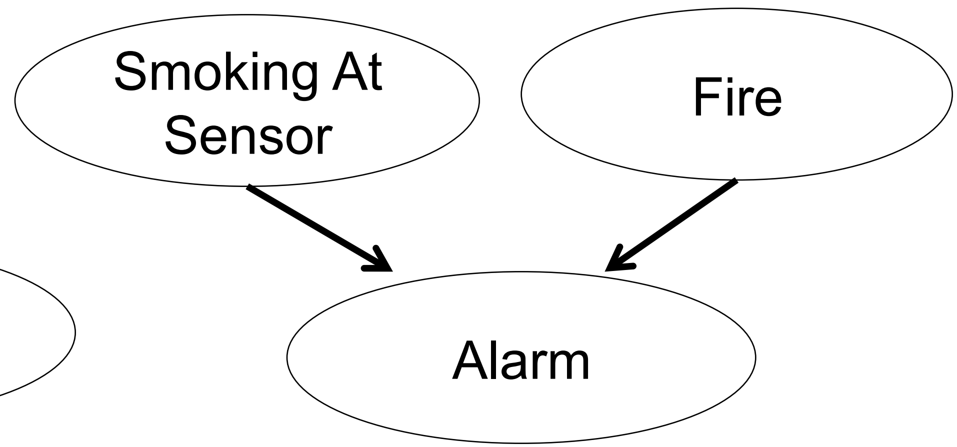
Judea Pearl

Bayesian Networks: Intuition

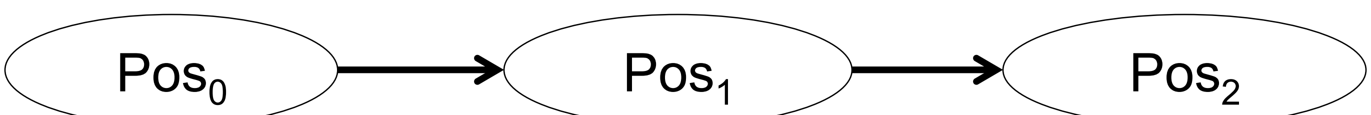
- A graphical representation for a joint probability distribution
  - Nodes are random variables
  - Directed edges between nodes reflect dependence

- Some informal examples:





Robot:



Bayesian Networks: Definition

Definition (Bayesian Network)
A Bayesian network consists of
- A directed acyclic graph (V, E) whose nodes are labeled with random variables
- A domain for each random variable
- A conditional probability distribution for each variable V
  - Specifies $P(V|Parents(V))$
  - $Parents(V)$ is the set of variables V' with $(\mathrm{V}', \mathrm{V}) \in E$
    - For nodes V without predecessors, $Parents(V) = \{\,\}$

- The parents of variable V are those V directly depends on
- A Bayesian network is a compact representation of the JPD: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i))$

Bayesian Networks: Definition

Definition (Bayesian Network)

A Bayesian network consists of

- A directed acyclic graph (V, E) whose nodes are labeled with random variables
- A domain for each random variable
- A conditional probability distribution for each variable V
  - Specifies $P(V|Parents(V))$
  - $Parents(V)$ is the set of variables V' with $(\mathrm{V}', \mathrm{V}) \in E$
    - For nodes V without predecessors, $Parents(V) = \{\,\}$

- Discrete Bayesian networks:
  - Domain of each variable is finite
  - Conditional probability distribution is a conditional probability table
  - We will assume this discrete case
    - But everything we say about independence (marginal & conditional) carries over to the continuous case

Example for BN construction: Fire Diagnosis

Bayesian networks are a compact representation of the joint probability distribution (over all variables in the network)

Encoding the joint over $\mathcal{X} = \{X_1, \ldots, X_n\}$ as a Bayesian network:

1. Totally order the variables of interest: $X_1, \ldots, X_n$
2. Use chain rule with that ordering: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_{i-1},\ldots,X_1)$
3. For every variable X_i, find the smallest set of parents $Pa(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$ such that $X_i \perp\!\!\!\perp \{X_1, \ldots, X_{i-1}\} \setminus Pa(X_i) \mid Pa(X_i)$
   - X_i is conditionally independent from its other ancestors given its parents
4. Then we can rewrite $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i))$
   - This is a compact representation of the joint probability distribution

Example for BN construction: Fire Diagnosis

5. Construct the Bayesian Net (BN)

- Nodes are the random variables
- Directed arc from each variable in $Pa(X_i)$ to X_i
- Conditional Probability Table (CPT)

  for each variable X_i: $P(X_i \mid Pa(X_i))$

Example for BN construction: Fire Diagnosis

You want to diagnose whether there is a fire in a building

- You receive a noisy report about whether everyone is leaving the building
- If everyone is leaving, this may have been caused by a fire alarm
- If there is a fire alarm, it may have been caused by a fire or by tampering
- If there is a fire, there may be smoke

Example for BN construction: Fire Diagnosis

First you choose the variables. In this case, all are Boolean:

- Tampering is true when the alarm has been tampered with
- Fire is true when there is a fire
- Alarm is true when there is an alarm
- Smoke is true when there is smoke
- Leaving is true if there are lots of people leaving the building
- Report is true if the sensor reports that lots of people are leaving the building

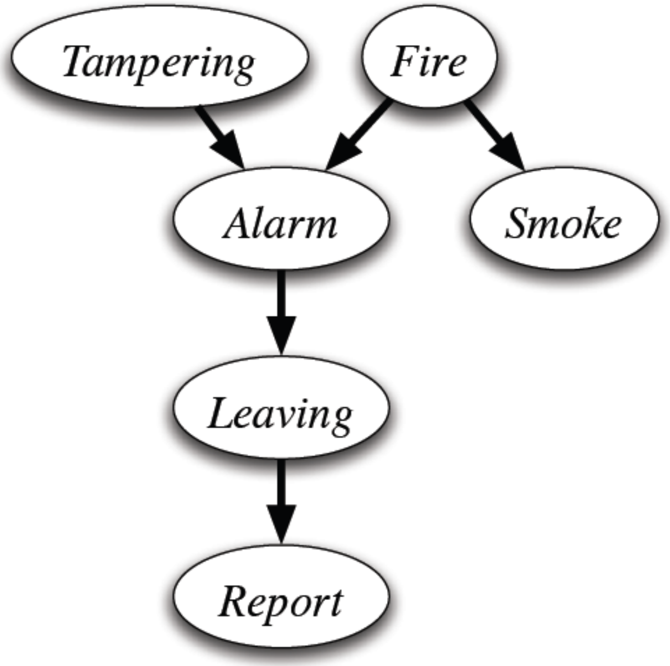
- Let's construct the Bayesian network for this (whiteboard)
  - First, you choose a total ordering of the variables, let's say: Fire; Tampering; Alarm; Smoke; Leaving; Report.

Example for BN construction: Fire Diagnosis

- Using the total ordering of variables:
  - Let’s say Fire; Tampering; Alarm; Smoke; Leaving; Report.
- Now choose the parents for each variable by evaluating conditional independencies
  - Fire is the first variable in the ordering, X_1. It does not have parents.
  - Tampering independent of fire (learning that one is true would not change your beliefs about the probability of the other)
  - Alarm depends on both Fire and Tampering: it could be caused by either or both
  - Smoke is caused by Fire, and so is independent of Tampering and Alarm given whether there is a Fire
  - Leaving is caused by Alarm, and thus is independent of the other variables given Alarm
  - Report is caused by Leaving, and thus is independent of the other variables given Leaving

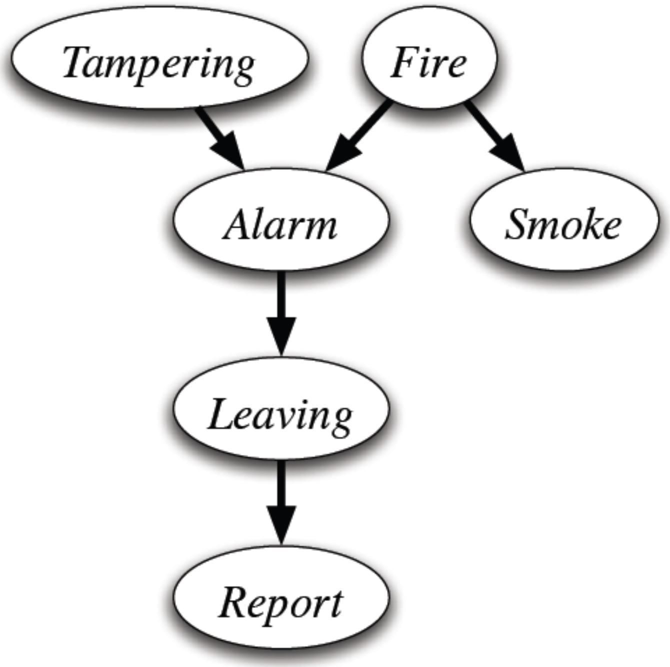
Example for BN construction: Fire Diagnosis

- This results in the following Bayesian network



- P(Tampering, Fire, Alarm, Smoke, Leaving, Report)
  = P(Tampering) × P(Fire) × P(Alarm|Tampering,Fire)
  × P(Smoke|Fire) × P(Leaving|Alarm) × P(Report|Leaving)

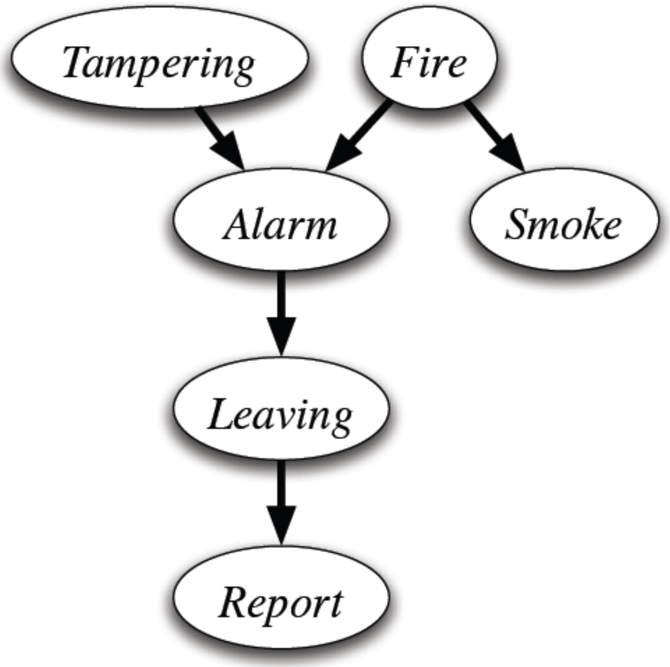
Example for BN construction: Fire Diagnosis



- We are not done yet: must specify the Conditional Probability Table (CPT) for each variable. All variables are Boolean.
- How many probabilities do we need to specify for this Bayesian network?
  - This time taking into account that probability tables have to sum to 1

6 12 20 2^6-1

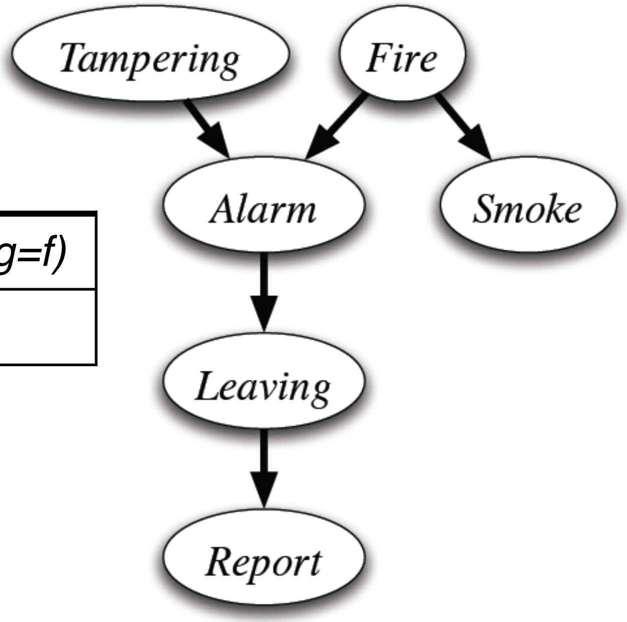
Example for BN construction: Fire Diagnosis



- We are not done yet: must specify the Conditional Probability Table (CPT) for each variable. All variables are Boolean.
- How many probabilities do we need to specify for this Bayesian network?
  - P(Tampering): 1 probability
  - P(Alarm|Tampering, Fire): 4 (independent)

    1 probability for each of the 4 instantiations of the parents
  - In total: 1+1+4+2+2+2 = 12 (compared to 2^6 -1= 63 for full JPD!)

Example for BN construction: Fire Diagnosis

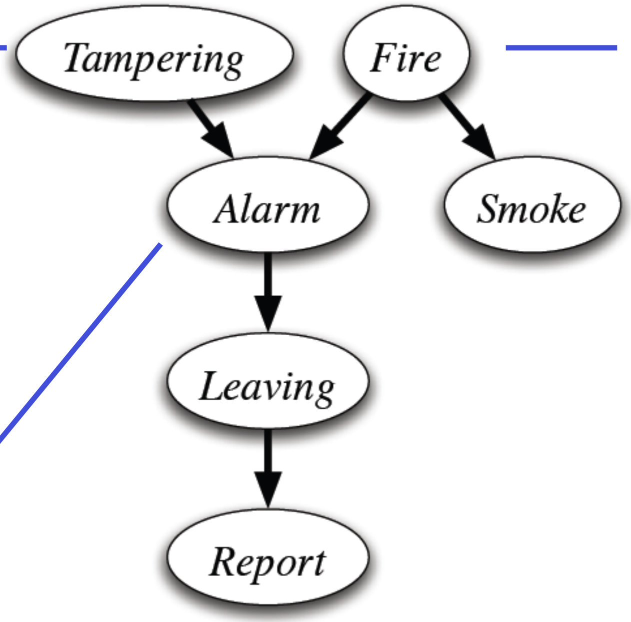


P(Tampering=t)	*P(Tampering=f)*
0.02	0.98

We don't need to store P(Tampering=f) since probabilities sum to 1

Example for BN construction: Fire Diagnosis

P(Tampering=t)
0.02



P(Fire=t)
0.01

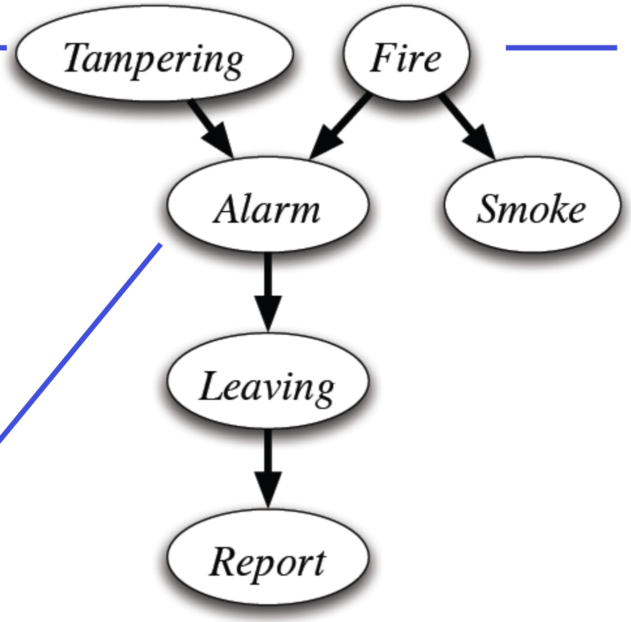
Tampering T	*Fire F*	$P(Alarm=t \mid T,F)$	$P(Alarm=f \mid T,F)$
t	t	0.5	0.5
t	f	0.85	0.15
f	t	0.99	0.01
f	f	0.0001	0.9999

We don't need to store $P(Alarm=f \mid T,F)$ since probabilities sum to 1

Each row of this table is a conditional probability distribution

Example for BN construction: Fire Diagnosis

P(Tampering=t)
0.02



P(Fire=t)
0.01

Tampering T	*Fire F*	*P(Alarm=t|T,F)*
t	t	0.5
t	f	0.85
f	t	0.99
f	f	0.0001

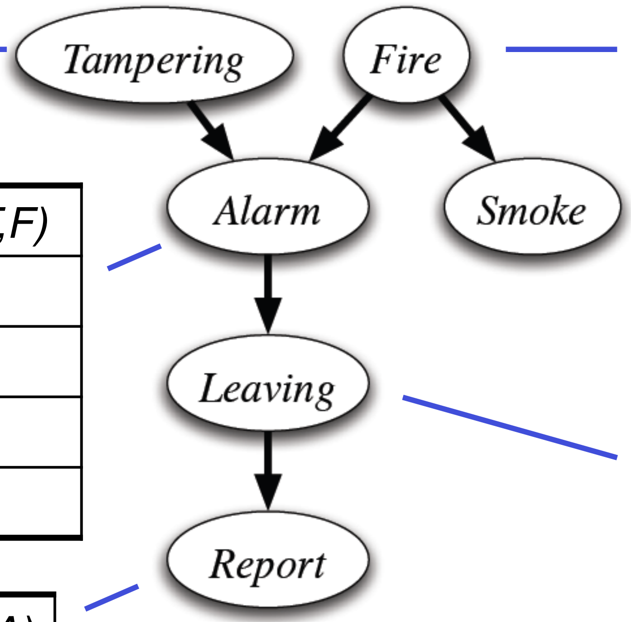
We don't need to store P(Alarm=f|T,F) since probabilities sum to 1

Each row of this table is a conditional probability distribution

Example for BN construction: Fire Diagnosis

P(Tampering=t)
0.02

P(Fire=t)
0.01



Tampering T	Fire F	$P(Alarm=t \mid T,F)$
t	t	0.5
t	f	0.85
f	t	0.99
f	f	0.0001

Fire F	$P(Smoke=t \mid F)$
t	0.9
f	0.01

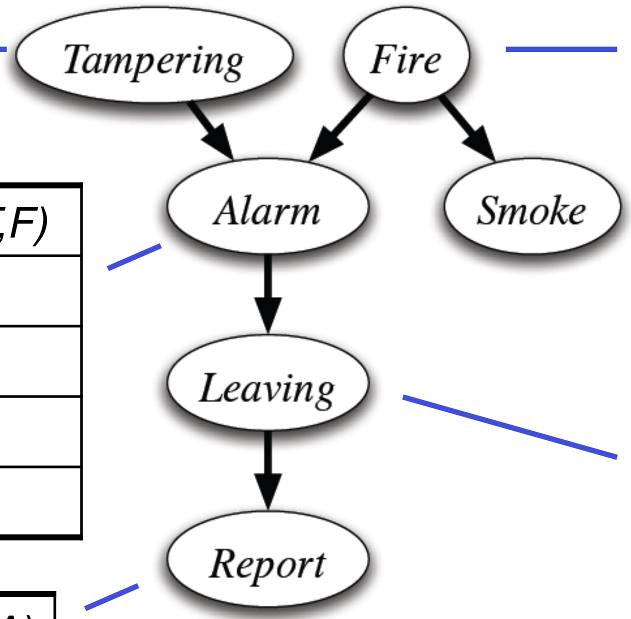
Alarm	$P(Leaving=t \mid A)$
t	0.88
f	0.001

Leaving	$P(Report=t \mid A)$
t	0.75
f	0.01

P(Tampering=t, Fire=f, Alarm=t, Smoke=f, Leaving=t, Report=t)

Example for BN construction: Fire Diagnosis

P(Tampering=t)
0.02



P(Fire=t)
0.01

Tampering T	Fire F	P(Alarm=t|T,F)
t	t	0.5
t	f	0.85
f	t	0.99
f	f	0.0001

Fire F	P(Smoke=t |F)
t	0.9
f	0.01

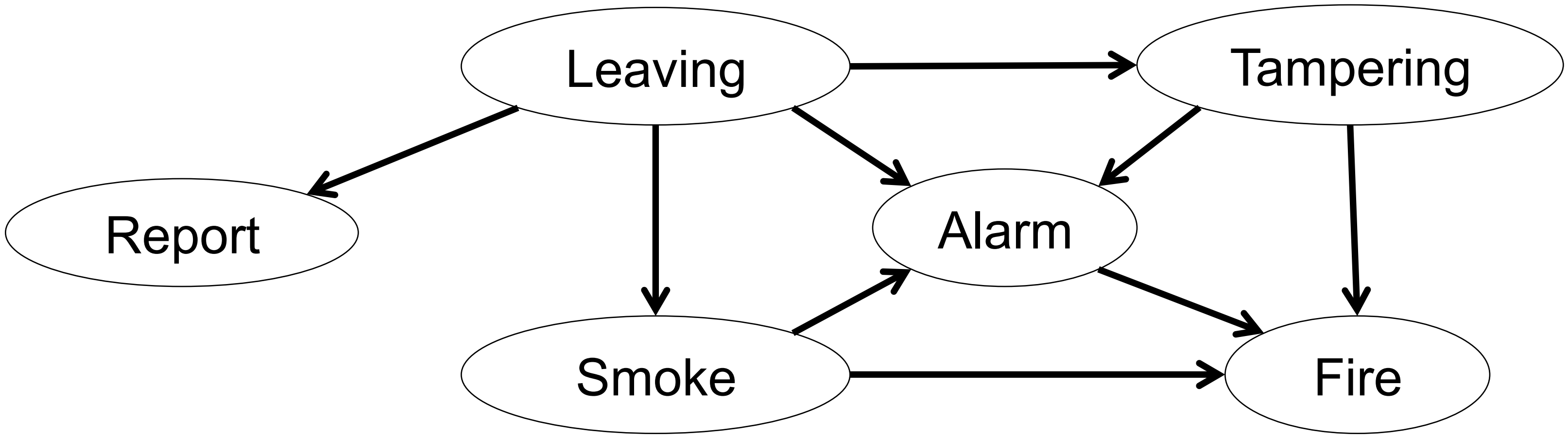
Alarm	P(Leaving=t|A)
t	0.88
f	0.001

Leaving	P(Report=t|A)
t	0.75
f	0.01

$$
\begin{aligned}
&P(\text{Tampering=t, Fire=f, Alarm=t, Smoke=f, Leaving=t, Report=t}) \\
&= P(\text{Tampering=t}) \times P(\text{Fire=f}) \times P(\text{Alarm=t}|\text{Tampering=t,Fire=f}) \\
&\quad \times P(\text{Smoke=f}|\text{Fire=f}) \times P(\text{Leaving=t}|\text{Alarm=t}) \\
&\quad \times P(\text{Report=t}|\text{Leaving=t}) \\
&= 0.02 \times (1-0.01) \times 0.85 \times (1-0.01) \times 0.88 \times 0.75 \;\; = 0.126
\end{aligned}
$$

What if we use a different ordering?

- Important for assignment 4, question 4:
- Say, we use the following order:
  - Leaving; Tampering; Report; Smoke; Alarm; Fire.



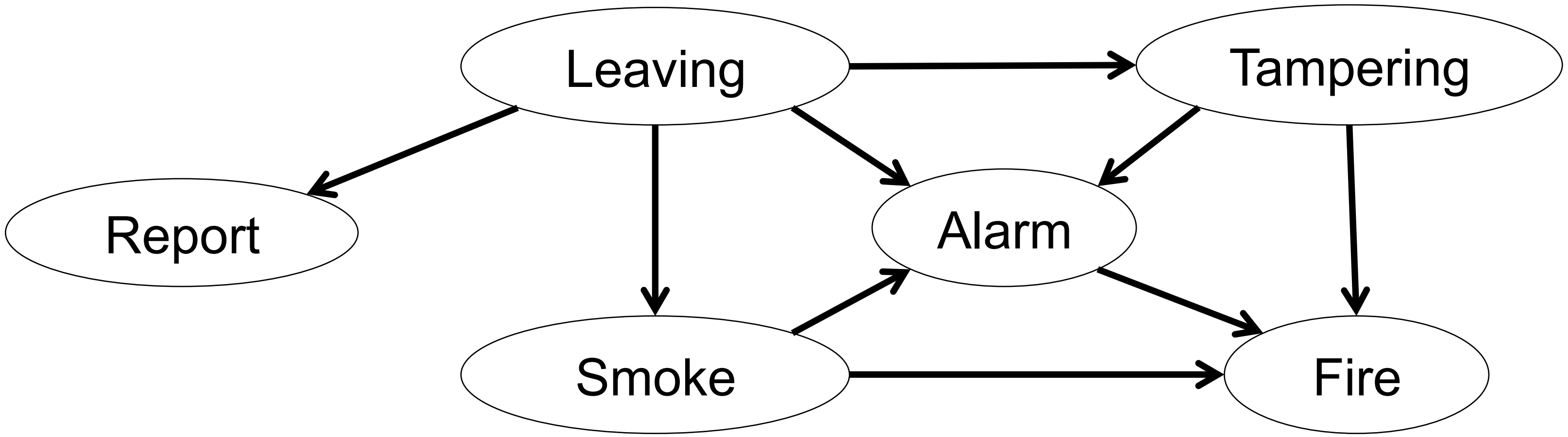
- We end up with a completely different network structure!
- Which of the two structures is better (think computationally)?

The previous structure

This structure

What if we use a different ordering?

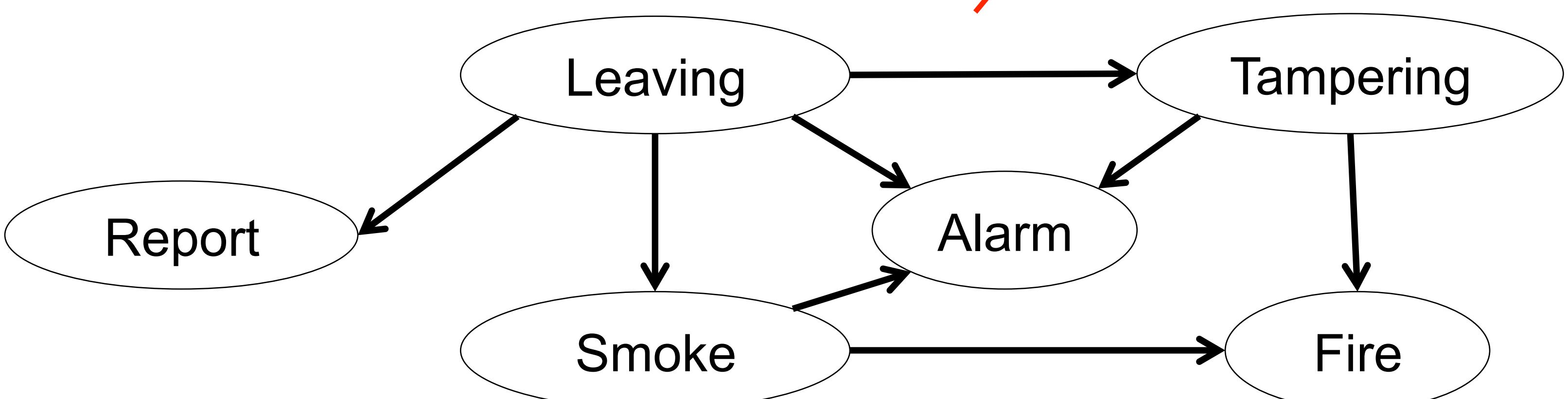
- Important for assignment 4, question 4:
- Say, we use the following order:
  – Leaving; Tampering; Report; Smoke; Alarm; Fire.



- We end up with a completely different network structure!
- Which of the two structures is better (think computationally)?
  – In the last network, we had to specify 12 probabilities
  – Here? 1 + 2 + 2 + 2 + 8 + 8 = 23
  – The causal structure typically leads to the most compact network
    - Compactness typically enables more efficient reasoning

Are there wrong network structures?

- Important for assignment 4, question 4
- Some variable orderings yield more compact, some less compact structures
  - Compact ones are better
  - But all representations resulting from this process are correct
  - One extreme: the fully connected network is always correct but rarely the best choice
- How can a network structure be wrong?
  - If it misses directed edges that are required
  - E.g. an edge is missing below: Fire $\not\perp\!\!\!\perp$ Alarm | {Tampering, Smoke}

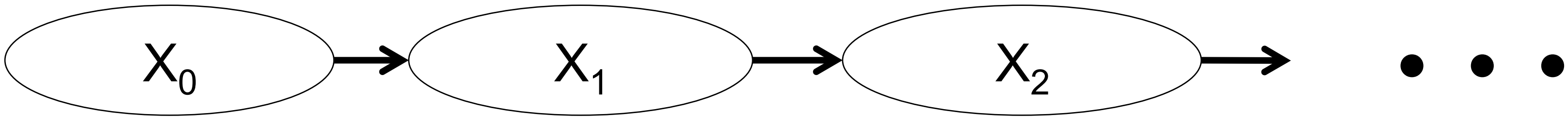


Lecture Overview

- Recap: marginal and conditional independence
- Bayesian Networks Introduction
- Hidden Markov Models

Markov Chains

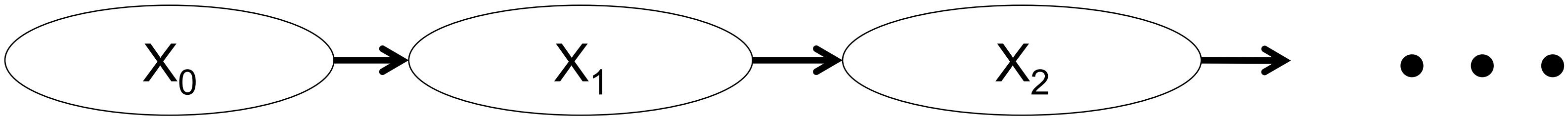
- A Markov chain is a special kind of belief network:



- X_t represents a state at time t.
- Its dependence structure yields: $P(X_t|X_1, \ldots, X_{t-1}) = P(X_t|X_{t-1})$
  - This conditional probability distribution is called the state transition probability
  - Intuitively X_t conveys all of the information about the history that can affect the future states:
    "The past is independent of the future given the present."

- JPD of a Markov Chain: $P(X_0,\ldots,X_T) = P(X_0) \times \prod_{t=1}^{T} P(X_t|X_{t-1})$

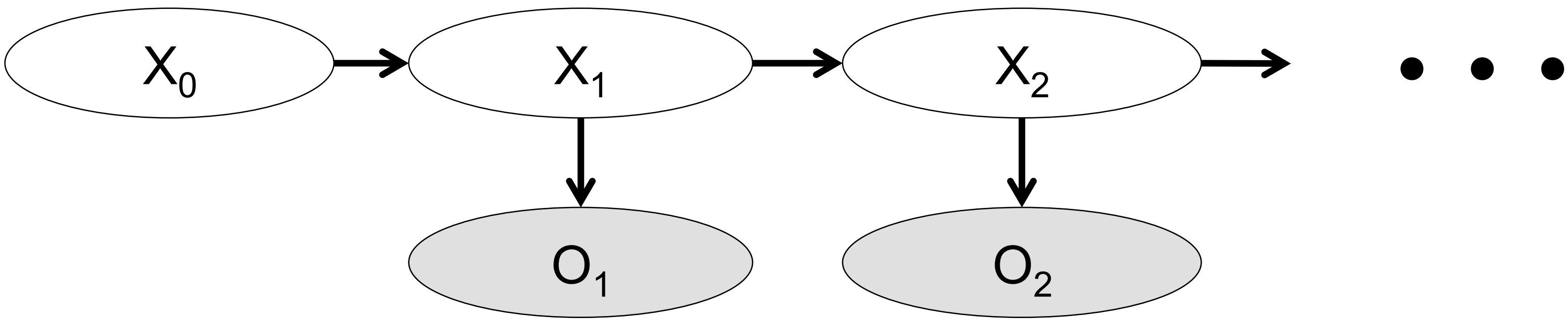
Stationary Markov Chains



- A stationary Markov chain is when
  - All state transition probability tables are the same
  - I.e., for all t > 0, t' > 0: $P(X_t|X_{t-1}) = P(X_{t'}|X_{t'-1})$
- We only need to specify $P(X_0)$ and $P(X_t|X_{t-1})$.
  - Simple model, easy to specify
  - Often the natural model
  - The network can extend indefinitely in time
- Example: Drunkard's walk, robot random motion

Hidden Markov Models (HMMs)

- A Hidden Markov Model (HMM) is a Markov chain plus a noisy observation about the state at each time step:

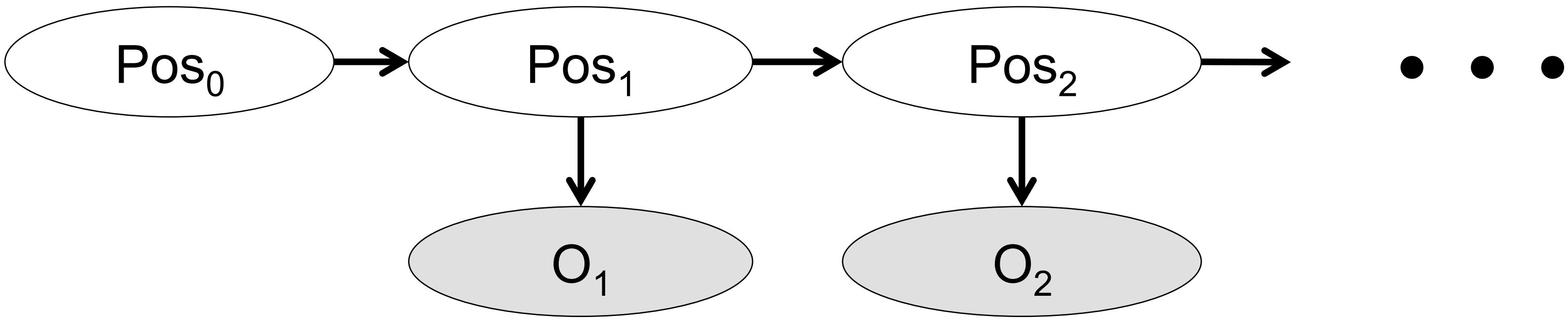


- Same conditional probability tables at each time step
  - The state transition probability $P(X_t|X_{t-1})$
    - also called the system dynamics
  - The observation probability $P(O_t|X_t)$
    - also called the sensor model

- JPD of an HMM: $P(X_0, \ldots, X_T, O_1, \ldots, O_T)$
  $= P(X_0) \times \prod_{t=1}^{T} P(X_t|X_{t-1}) \times \prod_{t=1}^{T} P(O_t|X_t)$

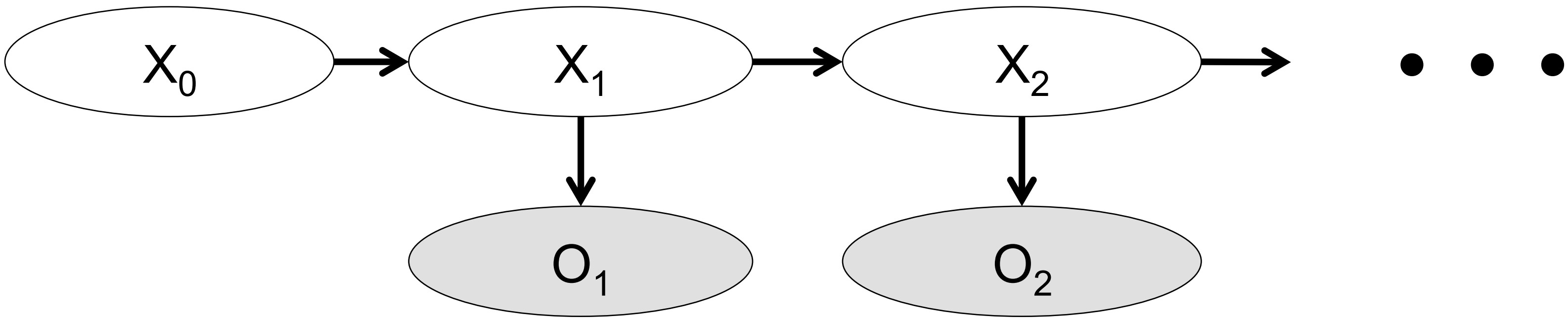
Example HMM: Robot Tracking

- Robot tracking as an HMM:



- Robot is moving at random: $P(Pos_t|Pos_{t-1})$
- Sensor observations of the current state $P(O_t|Pos_t)$

Filtering in Hidden Markov Models (HMMs)



- Filtering problem in HMMs:
  at time step t, we would like to know $P(X_t|O_1, \ldots, O_t)$

- Can derive simple update equations:
  - Compute $P(X_t|O_1, \ldots, O_t)$ if we already know $P(X_{t-1}|O_1, \ldots, O_{t-1})$

Learning Goals For Today's Class

- Build a Bayesian Network for a given domain
- Compute the representational savings in terms of number of probabilities required
- Understand basics of Markov Chains and Hidden Markov Models

---

- Assignment 4 available on Connect
  - Due Wednesday, April 4.
  - You should now be able to solve questions 1, 2, 3 and 4. Do them!
  - Material for question 5 (variable elimination): later this week