

CS5340

Uncertainty Modeling in AI

Lecture 2: Bayesian Networks (Directed Graphical Models)

Asst. Prof. Lee Gim Hee

AY 2020/21

Semester 1

Course Schedule

Week	Date	Topic	Remarks
1	12 Aug	Introduction to probabilistic reasoning	1830hrs: MS Teams (Live Introduction)
2	19 Aug	Bayesian networks (Directed graphical models)	
3	26 Aug	Markov random Fields (Undirected graphical models)	1830hrs: Zoom discussions
4	02 Sep	Variable elimination and belief propagation	Assignment 1: Belief propagation and maximal probability (15%)
5	09 Sep	Factor graph and the junction tree algorithm	
6	16 Sep	Parameter learning with complete data	Assignment 1: Due Assignment 2: Junction tree and parameter learning (15%) 1830hrs: Zoom discussions
-	23 Sep	Recess week	No lecture
7	30 Sep	Mixture models and the EM algorithm	Assignment 2: Due Online quiz 1 (20%)
8	07 Oct	Hidden Markov Models (HMM)	Assignment 3: Hidden Markov model (15%)
9	14 Oct	Monte Carlo inference (Sampling)	1830hrs: Zoom discussions
10	21 Oct	Variational inference	Assignment 3: Due Assignment 4: MCMC Sampling (15%)
11	28 Oct	Variational Auto-Encoder and Mixture Density Networks	
12	04 Nov	Graph-cut and alpha expansion	Assignment 4: Due 1830hrs: Zoom discussions
-	11 Nov	--	Online quiz 2 (20%)

Acknowledgements

- A lot of slides and content of this lecture are adopted from:
 1. "An introduction to probabilistic graphical models", Michael I. Jordan, 2002
<http://people.eecs.berkeley.edu/~jordan/prelims/chapter2.pdf> (Section 2.1)
 2. "Pattern recognition and machine learning", Christopher Bishop (Chapter 8, Section 8.1 and 8.2).
 3. "Machine learning - a probabilistic approach", Kevin Murphy (Chapter 10)
 4. "Probabilistic graphical models", Koller and Friedman (Chapter 3)

Learning Outcomes

- Students should be able to:
 1. Explain the concepts of **conditional independence**.
 2. Use the **Bayesian network** to represent conditional independence in joint distributions.
 3. Describe **d-separation** using the **three canonical 3-node graph**.
 4. Deduce all conditional independence in a Bayesian network using the **Bayes ball algorithm**.
 5. Explain the concepts of **Markov Blanket**.

Why the Need for Graphical Models?

It difficult to work with **joint probabilities**
 $p(x_1, \dots, x_N | \theta)$ with fully correlated random variables.

Why?

Why the Need for Graphical Models?

- Let's illustrate this with N **discrete** random variables X_1, \dots, X_N , where $x_i \in \{1, \dots, K\}$.
- We need $O(K^N)$ **parameters** θ to represent the joint distribution $p(x_1, \dots, x_N | \theta)$.
- **Inference becomes intractable** when N is large, and **a huge amount of data** is needed to learn all parameters.

Why the Need for Graphical Models?

Easy solution?

- **Fully independent assumption** on all random variables reduces the number of parameters to $O(NK)$.

$$p(x_1, \dots, x_N | \theta) = \prod_{i=1}^N p(x_i | \theta_i)$$

- Inference becomes **tractable** products of $p(x_i | \theta_i)$, and **smaller amount of data** is needed to learn all parameters.

Why the Need for Graphical Models?

- Fully independent assumption:

$$p(x_1, \dots, x_N | \theta) = \prod_{i=1}^N p(x_i | \theta_i)$$

- **Is it always correct** to assume that all random variables are fully independent?

Examples:

Probability of the next letter in a word:

“ probabili  ”

Is  independent of the letters that are already known in the word?

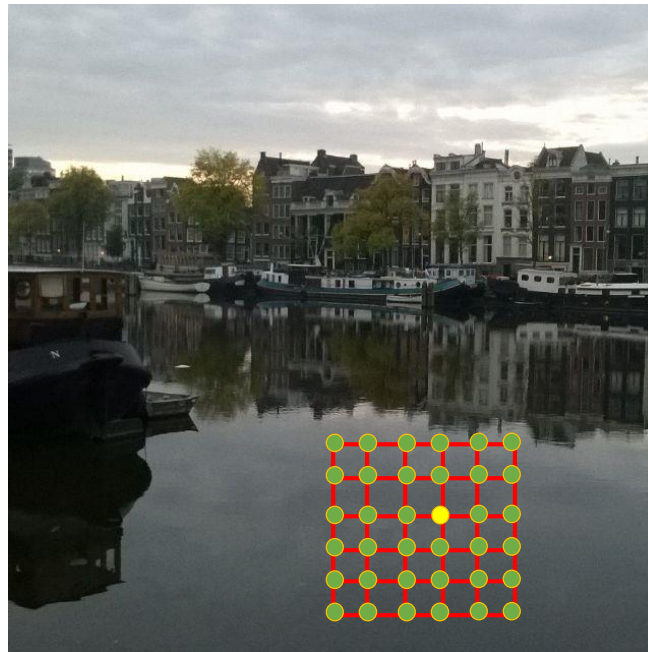
Why the Need for Graphical Models?

- Fully independent assumption :

$$p(x_1, \dots, x_N | \theta) = \prod_{i=1}^N p(x_i | \theta_i)$$

- Is it always correct** to assume that all random variables are fully independent?

Examples:



● : pixel is labeled as “water”

● : is this pixel more likely to be “water” or “sky”?

Photo Source:
G.H. Lee “Amsterdam”

Why the Need for Graphical Models?

Random variables are often NOT fully independent, how can we:

- Compactly **represent the joint distribution** $p(x_1, \dots, x_N | \theta)$ of multiple correlated variables?
- Use the joint distribution to **infer** one set of variables given another in a reasonable amount of computation time?
- **Learn** the parameters of the joint distribution with a reasonable amount of data?

Use Graphical Models!!!

Conditional Independence

- We have seen that:
 - Fully independent assumption is insufficient to model real-world random variables which are unlikely to be fully independent.
 - Fully correlated joint distributions can become intractable.
- A good compromise is by assuming an intermediate degree of dependency among the random variables.
- This is conditional independence.

Conditional Independence

- More formally, two random variables X_A and X_C are **conditionally independent given X_B** if:

$$p(x_A, x_C | x_B) = p(x_A | x_B) p(x_C | x_B)$$

- Or alternatively:

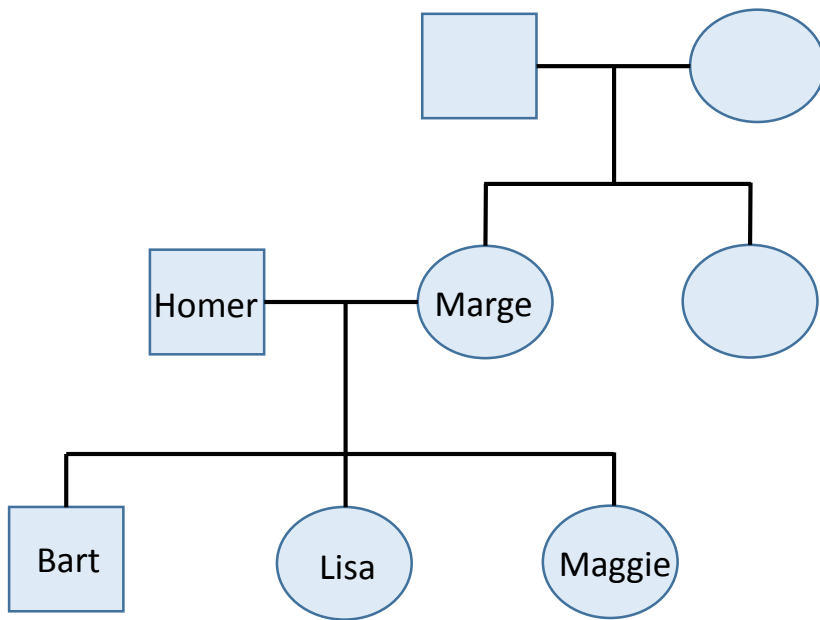
$$p(x_A | x_B, x_C) = p(x_A | x_B), \quad \forall x_B: p(x_B) > 0$$

- That is, learning the values of X_C **does not change** prediction of X_A once we know the value of X_B .
- Written as $X_A \perp X_C \mid X_B$.

Conditional Independence

Example: Family Trees (Pedigree)

A node represents an individual's genotype.



Conditional Independence:

$$X_{Bart} \perp (X_{nonDesc} \setminus X_{Parent}) \mid X_{Parent}$$

Non-descendants

Any random variable is locally dependent on only its parent nodes, also known as the **Markov Assumption**.

Bayesian Networks: Definitions

- We use a **directed acyclic graph (DAG)** to represent conditional independence.
- A DAG is a pair $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where \mathcal{V} is a set nodes and \mathcal{E} a set of **oriented edges**.

Example of a **Directed Graphical Model (DGM)**, i.e.
Bayesian Network:

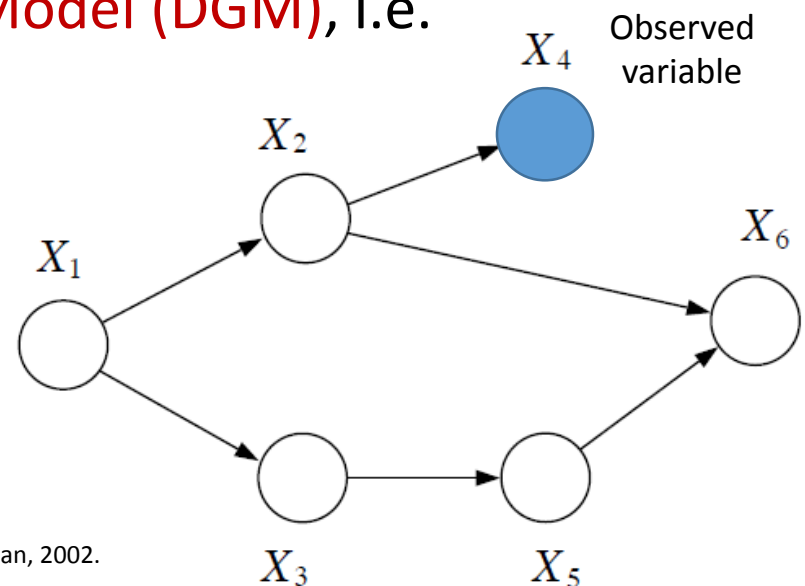


Image modified from: "An introduction to probabilistic graphical models", Michael I. Jordan, 2002.

Bayesian Networks: Definitions

- \mathcal{G} **does not** contain any cycles.
- **Shaded node** refers to observed variable.
- **Unshaded node** refers to latent/hidden (i.e. unobserved) variables.

Example of a **Directed Graphical Model (DGM)**, i.e. **Bayesian Network**:

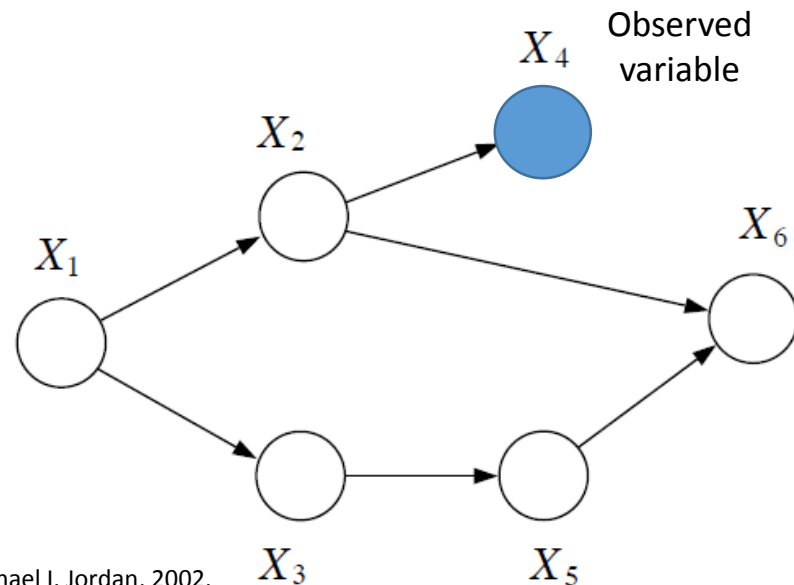
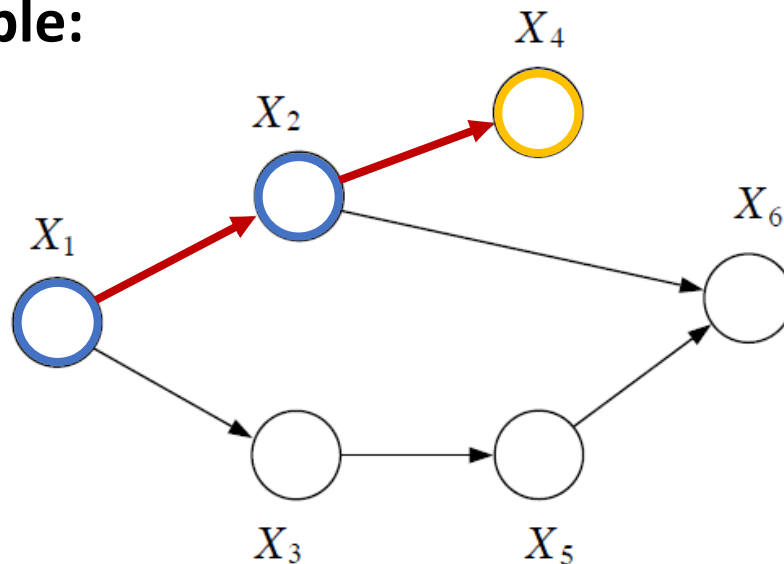


Image modified from: "An introduction to probabilistic graphical models", Michael I. Jordan, 2002.

Bayesian Networks: Definitions

- **Ancestors** are the parents, grand-parents, etc of a node.
- The ancestors of t is the set of node s that connect to t via a trail: $\text{anc}(t) \triangleq \{s: s \rightsquigarrow t\}$.

Example:

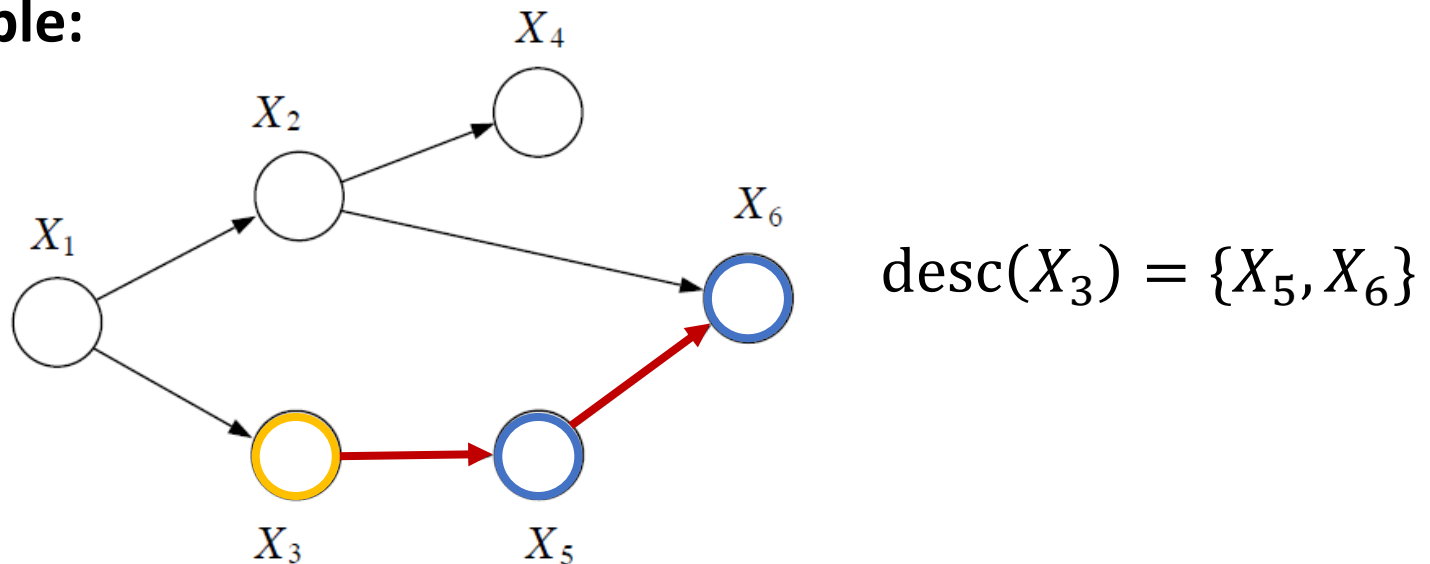


$$\text{anc}(X_4) = \{X_1, X_2\}$$

Bayesian Networks: Definitions

- **Descendants** are the children, grand-children, etc of a node.
- The descendants of s is the set of nodes that can be reached via trials from s : $\text{desc}(s) \triangleq \{t: s \rightsquigarrow t\}$.

Example:



Bayesian Networks: Definitions

- There is an associated random variable X_i for each $i \in \mathcal{V}$.
- Each node $i \in \mathcal{V}$ has a set of **parent nodes** π_i , which can be the empty set.
- Let X_{π_i} represent all the random variables that are parents to the random variable X_i .

Example:

$$\begin{aligned} X_{\pi_1} &= \emptyset, & X_{\pi_2} &= X_1 \\ X_{\pi_3} &= X_1, & X_{\pi_4} &= X_2 \\ X_{\pi_5} &= X_3, & X_{\pi_6} &= \{X_2, X_5\} \end{aligned}$$

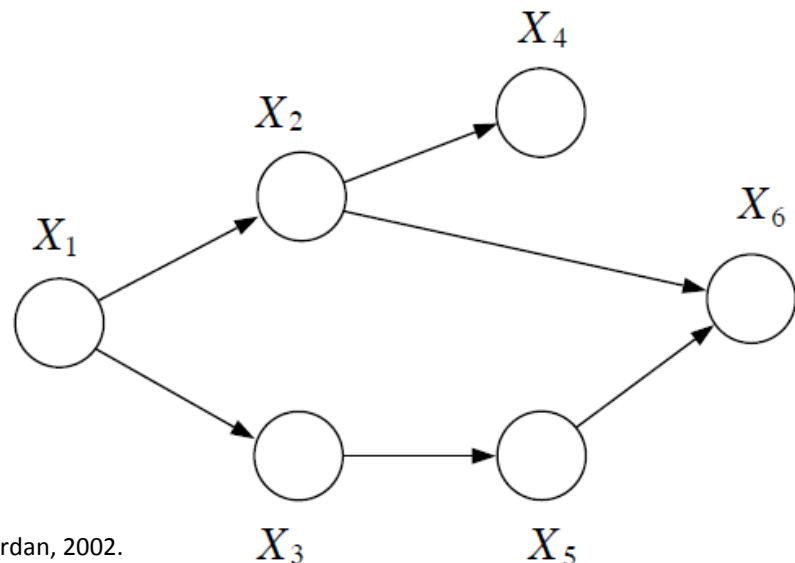


Image source: "An introduction to probabilistic graphical models", Michael I. Jordan, 2002.

Markov Assumption

- **Markov assumption:** Each random variable X_i is independent of its non-descendants $X_{\text{nonDesc}(X_i)}$ given its parents X_{π_i} .
- The following set of **basic conditional independence statements** can be associated to the DGM:

$$\{X_i \perp (X_{\text{nonDesc}(X_i)} \setminus X_{\pi_i}) \mid X_{\pi_i}\}$$

Example:

Non-descendants of X_2 are outlined blue.

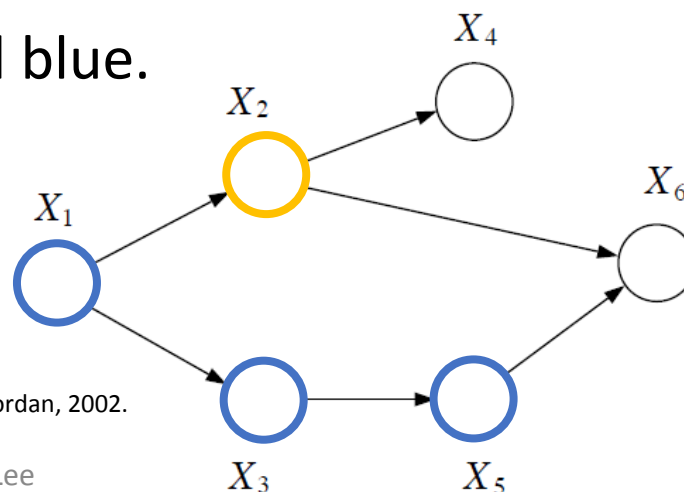


Image modified from: “An introduction to probabilistic graphical models”, Michael I. Jordan, 2002.

Markov Assumption

Example:

We have the following set of basic conditional independence from the given Bayesian network.

$$X_1 \perp \emptyset \mid \emptyset,$$

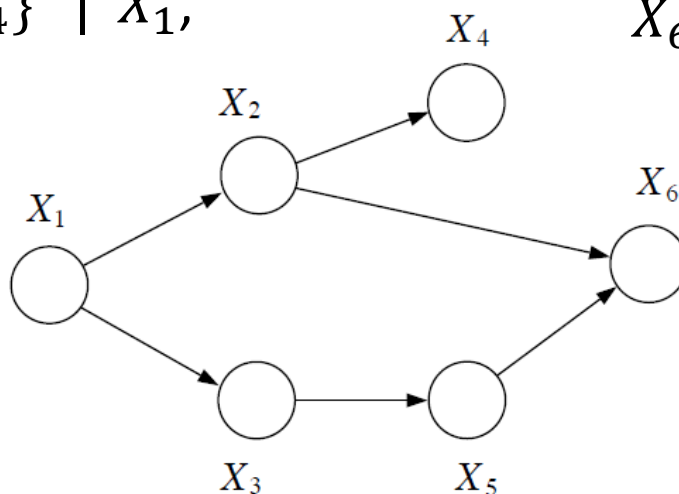
$$X_2 \perp \{X_3, X_5\} \mid X_1,$$

$$X_3 \perp \{X_2, X_4\} \mid X_1,$$

$$X_4 \perp \{X_1, X_3, X_5, X_6\} \mid X_2,$$

$$X_5 \perp \{X_1, X_2, X_4\} \mid X_3,$$

$$X_6 \perp \{X_1, X_3, X_4\} \mid \{X_2, X_5\}$$



Markov Assumption

- The basic conditional independence statements in the DGM give rise to a set of **conditional probabilities**:

$$\{X_i \perp (X_{\text{nonDesc}(x_i)} \setminus X_{\pi_i}) \mid X_{\pi_i}\}$$



$$p(x_i | x_{\pi_i}), \quad i = 1, \dots, N$$

- $p(x_i | x_{\pi_i})$ is **defined locally** according to the parent-child relationship specified by the DGM.

Bayesian Networks: Joint Probability

- **Locality of the parent-child** relationship is used to construct **economical representations** of the joint distribution.
- The **parent-child** represents conditional independence:

$$p(x_i | x_{\pi_i})$$

- Joint probability can be read off the graph as the **product** of all **local conditional independence**:

$$p(x_1, \dots, x_N) = \prod_{i=1}^N p(x_i | x_{\pi_i})$$

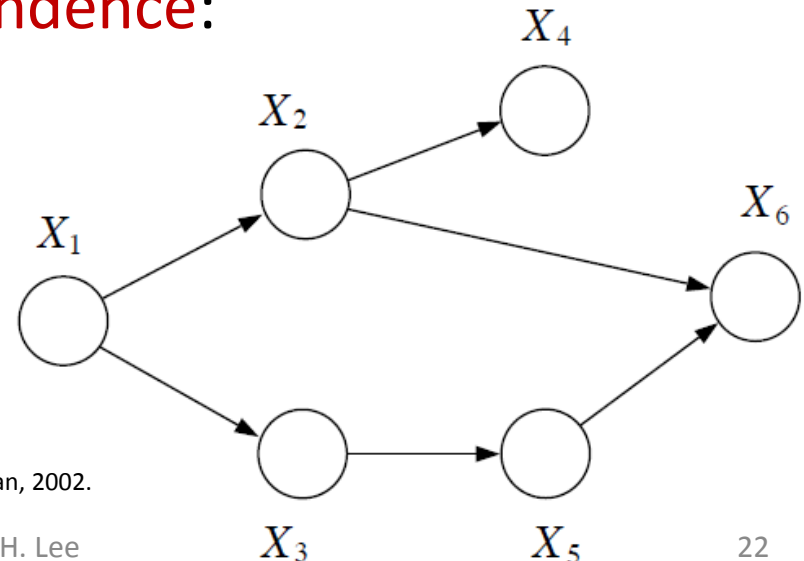


Image source: "An introduction to probabilistic graphical models", Michael I. Jordan, 2002.

Bayesian Networks: Joint Probability

Proof Sketch:

$$p(x_1, \dots, x_N) = p(x_1) \prod_{i=2}^N p(x_i | x_{x_1}, \dots, x_{i-1}) \quad (\text{chain rule})$$



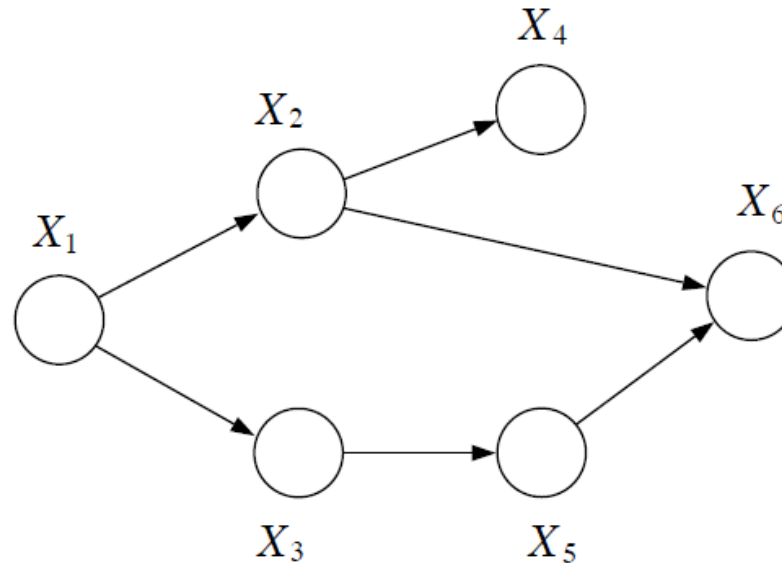
$$\{X_i \perp (X_{\text{nonDesc}(x_i)} \setminus X_{\pi_i}) \mid X_{\pi_i}\}$$

$$p(x_1, \dots, x_N) = \prod_{i=1}^N p(x_i | x_{\pi_i})$$

(Assuming topological ordering of DGM)

Bayesian Networks: Joint Probability

Example:



$$p(x_1, \dots, x_N) = \prod_{i=1}^N p(x_i | x_{\pi_i})$$

$$p(x_1, \dots, x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)$$

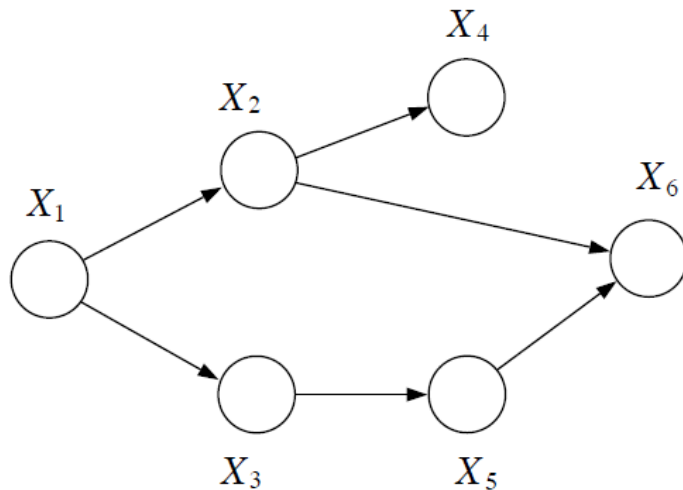
Image source: "An introduction to probabilistic graphical models", Michael I. Jordan, 2002.

Bayesian Networks: Joint Probability

Example:

Let's verify that the basic sets of conditional independence are indeed represented in the joint probability:

$$p(x_1, \dots, x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)$$



$$X_1 \perp \emptyset \mid \emptyset,$$

$$X_2 \perp \{X_3, X_5\} \mid X_1,$$

$$X_3 \perp \{X_2, X_4\} \mid X_1,$$

$$X_4 \perp \{X_1, X_3, X_5, X_6\} \mid X_2,$$

$$X_5 \perp \{X_1, X_2, X_4\} \mid X_3,$$

$$X_6 \perp \{X_1, X_3, X_4\} \mid \{X_2, X_5\}$$

Bayesian Networks: Joint Probability

Example:

Let's verify that X_1 and X_3 are independent of X_4 given X_2 .

First we compute the **marginal probability** of $\{X_1, X_2, X_3, X_4\}$:

$$\begin{aligned} p(x_1, x_2, x_3, x_4) &= \sum_{x_5} \sum_{x_6} p(x_1, x_2, x_3, x_4, x_5, x_6) \\ &= \sum_{x_5} \sum_{x_6} p(x_1) p(x_2 | x_1) p(x_3 | x_1) p(x_4 | x_2) p(x_5 | x_3) p(x_6 | x_2, x_5) \\ &= p(x_1) p(x_2 | x_1) p(x_3 | x_1) p(x_4 | x_2) \sum_{x_5} p(x_5 | x_3) \sum_{x_6} p(x_6 | x_2, x_5) \\ &= p(x_1) p(x_2 | x_1) p(x_3 | x_1) p(x_4 | x_2), \end{aligned}$$

Bayesian Networks: Joint Probability

Example:

Let's verify that X_1 and X_3 are independent of X_4 given X_2 .

Next we compute the **marginal probability** of $\{X_1, X_2, X_3\}$:

$$\begin{aligned} p(x_1, x_2, x_3) &= \sum_{x_4} p(x_1)p(x_2 | x_1)p(x_3 | x_1)p(x_4 | x_2) \\ &= p(x_1)p(x_2 | x_1)p(x_3 | x_1). \end{aligned}$$

Dividing the two marginal yields the desired conditional:

$$p(x_4 | x_1, x_2, x_3) = p(x_4 | x_2),$$

Which demonstrates the conditional independence relationship $X_4 \perp \{X_1, X_3\} \mid X_2$.

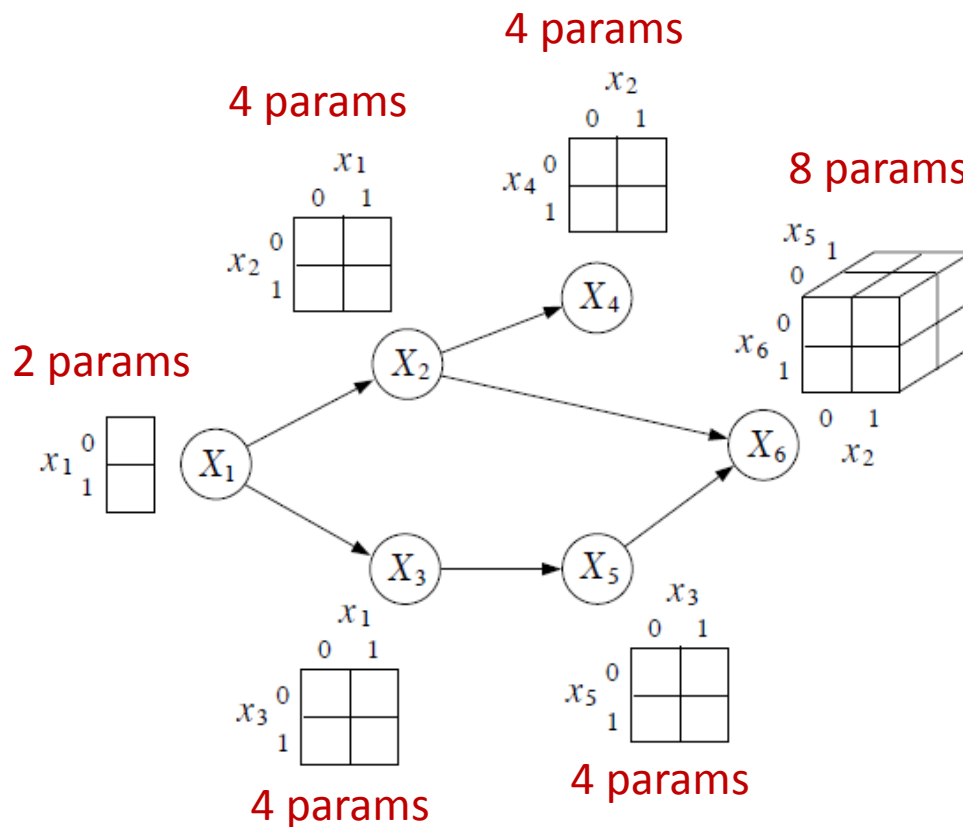
Bayesian Networks: Parameter Reduction

- Let m_i denote the number of parents of node X_i , and each node takes on K values.
- The conditional probability associated with X_i can be represented with a table of size K^{m_i+1} .
- Results in **huge reduction of parameters** needed to represent the joint probability, i.e. from $O(K^N)$ to $O(K^{m+1})$, $m \ll N$.

Bayesian Networks: Parameter Reduction

Example:

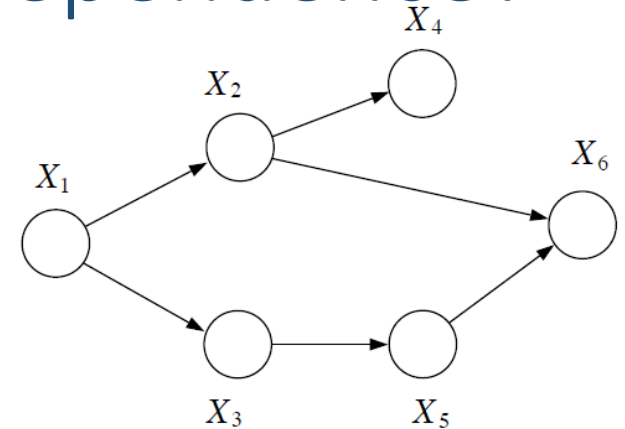
Binary state random variable $x_i \in \{0,1\}$.



- Total parameters = 26.
- Total parameters needed for fully dependent joint probability = $2^6 = 64$.
- More significant difference with higher number of nodes.

Additional Conditional Independence?

It turns out $X_1 \perp X_6 \mid \{X_2, X_3\}$ is also a conditional independence, but not directly observed from the parent-child relation.

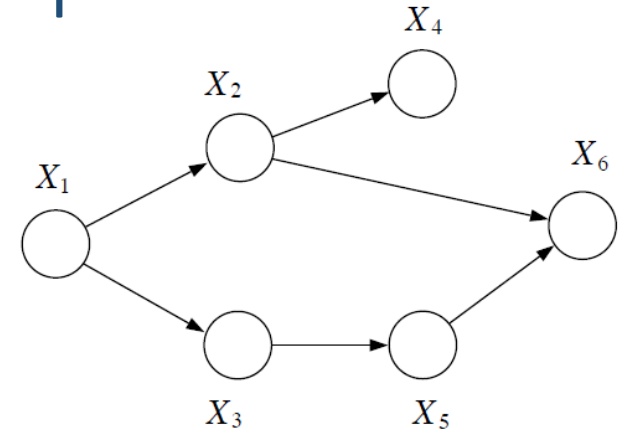


Proof:

$$\begin{aligned} p(x_1, x_2, x_3, x_6) &= \sum_{x_4} \sum_{x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1) \sum_{x_4} \cancel{p(x_4|x_2)}^1 \sum_{x_5} p(x_5|x_3)p(x_6|x_2, x_5) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1) \sum_{x_5} p(x_5|x_3)p(x_6|x_2, x_5) \\ p(x_2, x_3, x_6) &= \sum_{x_1} \sum_{x_4} \sum_{x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5) \\ &= \sum_{x_1} p(x_1)p(x_2|x_1)p(x_3|x_1) \sum_{x_4} \cancel{p(x_4|x_2)}^1 \sum_{x_5} p(x_5|x_3)p(x_6|x_2, x_5) \\ &= \sum_{x_1} p(x_1)p(x_2|x_1)p(x_3|x_1) \sum_{x_5} p(x_5|x_3)p(x_6|x_2, x_5) \end{aligned}$$

Additional Conditional Independence?

It turns out $X_1 \perp X_6 \mid \{X_2, X_3\}$ is also a conditional independence, but not directly observed from the parent-child relation.



Proof:

$$p(x_1, x_2, x_3, x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1) \sum_{x_5} p(x_5|x_3)p(x_6|x_2, x_5)$$

$$p(x_2, x_3, x_6) = \sum_{x_1} p(x_1)p(x_2|x_1)p(x_3|x_1) \sum_{x_5} p(x_5|x_3)p(x_6|x_2, x_5)$$

$$p(x_1|x_2, x_3, x_6) = \frac{p(x_1, x_2, x_3, x_6)}{p(x_2, x_3, x_6)}$$

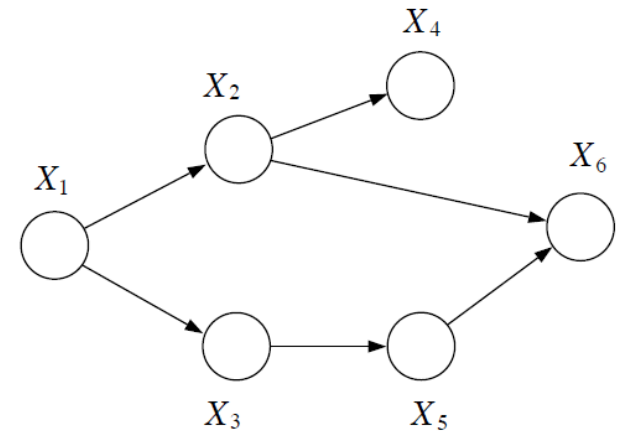
$$\begin{aligned} &= \frac{p(x_1)p(x_2|x_1)p(x_3|x_1) \sum_{x_5} p(x_5|x_3)p(x_6|x_2, x_5)}{\sum_{x_1} p(x_1)p(x_2|x_1)p(x_3|x_1) \sum_{x_5} p(x_5|x_3)p(x_6|x_2, x_5)} \\ &= \frac{p(x_1, x_2, x_3)}{\sum_{x_1} p(x_1, x_2, x_3)} = \frac{p(x_1, x_2, x_3)}{p(x_2, x_3)} = p(x_1|x_2, x_3) \end{aligned}$$

Additional Conditional Independence?

It turns out $X_1 \perp X_6 \mid \{X_2, X_3\}$ is also a conditional independence, but not directly observed from the parent-child relation.

Proof:

$$p(x_1|x_2, x_3, x_6) = p(x_1|x_2, x_3)$$



Question: Can we write all other conditional independencies by just **inspecting the DGM** without going through the complicated mathematics?

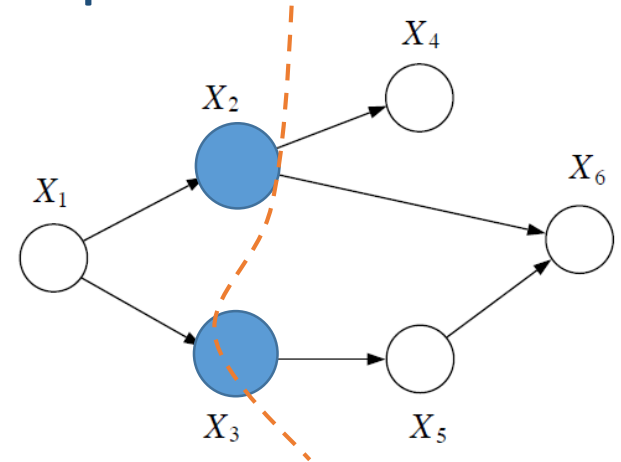
Additional Conditional Independence?

It turns out $X_1 \perp X_6 \mid \{X_2, X_3\}$ is also a conditional independence, but not directly observed from the parent-child relation.

Proof:

$$p(x_1|x_2, x_3, x_6) = p(x_1|x_2, x_3)$$

The nodes $\{X_2, X_3\}$ “block” X_1 from X_6 .

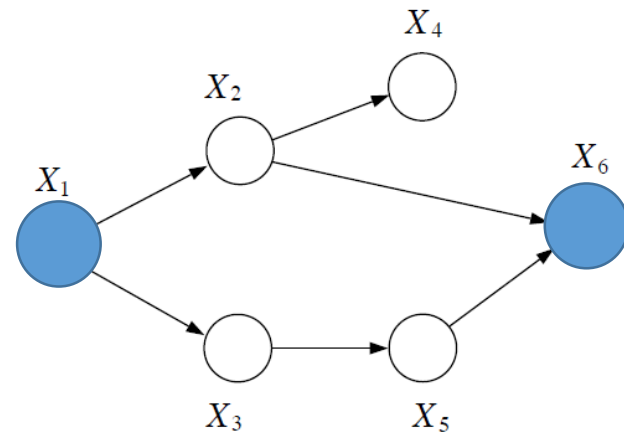
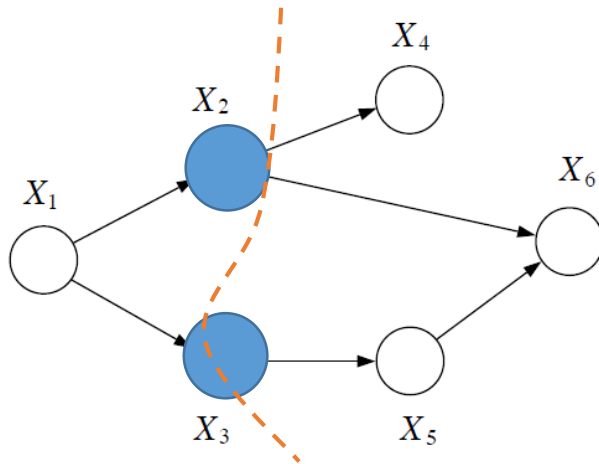


Question: Can we write all other conditional independencies by just **inspecting the DGM** without going through the complicated mathematics?

Answer: Yes, observe that the nodes $\{X_2, X_3\}$ “**block**” all paths from X_1 to X_6 . This suggests the notion of **graph separation** for inferring conditional independence.

Additional Conditional Independence?

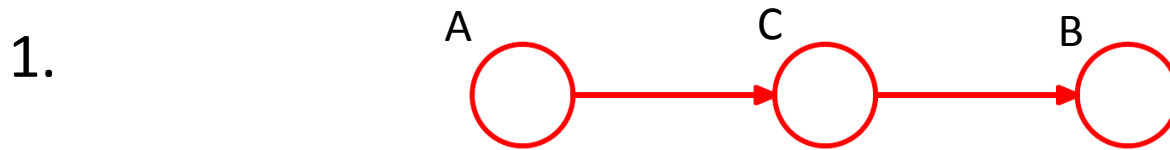
- We have to **be careful** in making the notion of “blocking”.
- For example, X_2 is **NOT independent** of X_3 given X_1 and X_6 as would be suggested by a naïve interpretation of “blocking”.
- Precise definition of “blocking” has to be done through the **“three canonical 3-node graphs”**, and **“d-separation”**.



The nodes $\{X_2, X_3\}$ **“block”** X_1 from X_6 .

X_2 is **NOT independent** of X_3 given $\{X_1, X_6\}$.

Three Canonical 3-Node Graphs



Joint distribution corresponding to this graph:

$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

If **none** of the variables are observed, we can see that A and B are **NOT independent** by marginalizing over C :

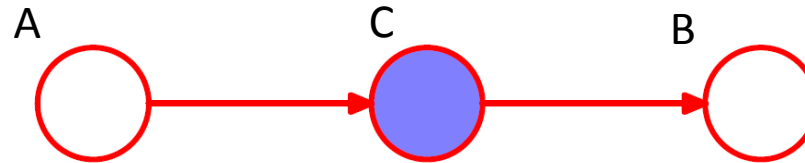
$$p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a)p(b|a)$$

$$\sum_c p(c|a)p(b|c) = \sum_c \frac{p(a)p(c|a)p(b|c)}{p(a)} = \sum_c \frac{p(a,b,c)}{p(a)} = \frac{p(a,b)}{p(a)} = p(b|a)$$

which in general **does not** factorize into $p(a)p(b)$, and so

$$A \not\perp B \mid \emptyset$$

Three Canonical 3-Node Graphs



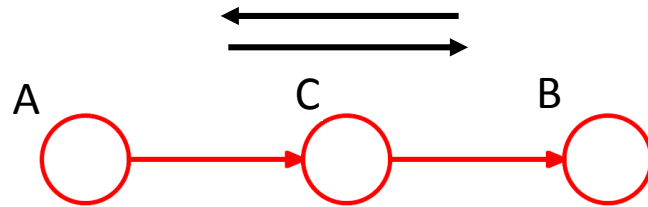
If we **condition on** node C, using Bayes' theorem together with the joint distribution, we get:

$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} && \text{(Bayes rule)} \\ &= p(a|c)p(b|c) && p(a)p(c|a) = p(a|c)p(c) \end{aligned}$$

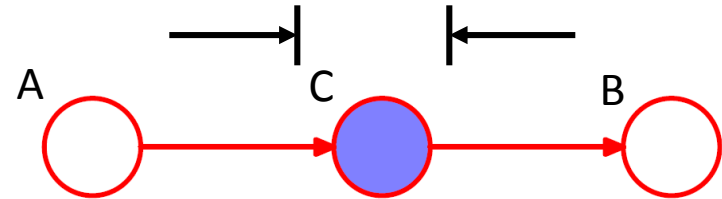
which shows the conditional independence property:

$$A \perp B \mid C$$

Three Canonical 3-Node Graphs



$$A \not\perp B \mid \emptyset$$

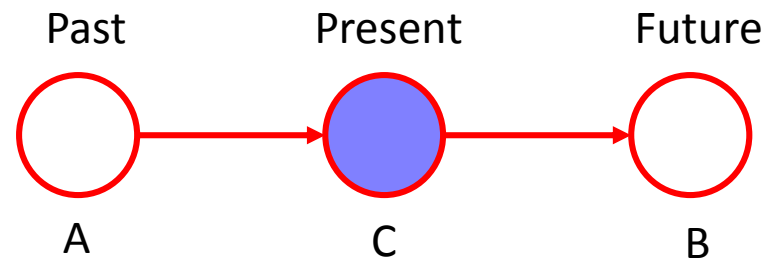


$$A \perp B \mid C$$

- The node C is said to be *head-to-tail* with respect to the path from node A to node B .
- Such a path **connects** nodes A and B and renders them dependent.
- The observation of C **'blocks'** the path from A to B and so we obtain the conditional independence property.

Three Canonical 3-Node Graphs

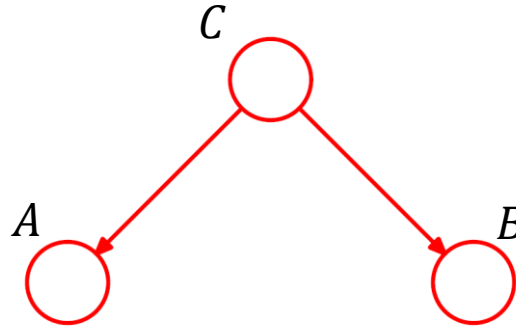
Intuitive interpretation:



- The conditional independence $A \perp B \mid C$ translates into the statement: “the past is independent of the future given the present”.
- This is an example of a simple classical **Markov Chain**.

Three Canonical 3-Node Graphs

2.



Joint distribution corresponding to this graph:

$$p(a, b, c) = p(a|c)p(b|c)p(c)$$

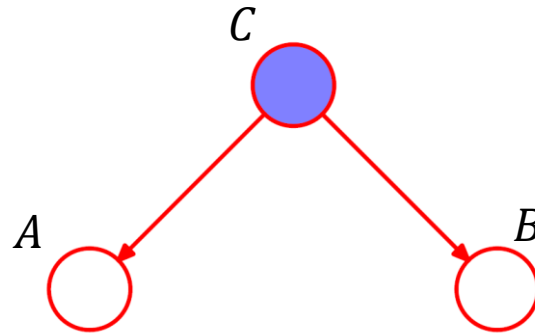
If **none** of the variables are observed, we can see that A and B are **NOT independent** by marginalizing both sides over C :

$$p(a, b) = \sum_c p(a|c)p(b|c)p(c)$$

which in general **does not** factorize into $p(a)p(b)$, and so

$$A \not\perp B \mid \emptyset$$

Three Canonical 3-Node Graphs



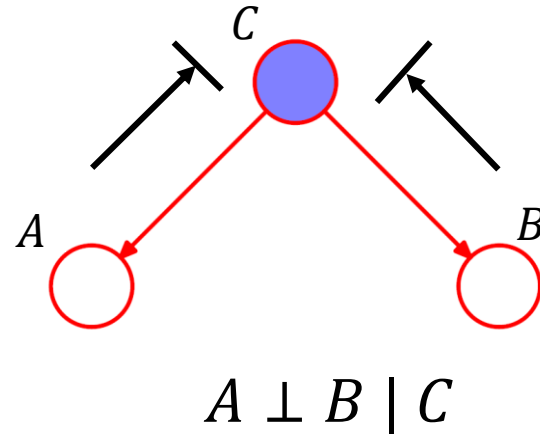
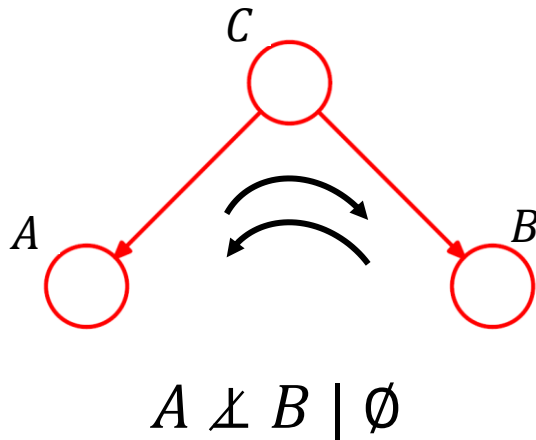
If we **condition on** node C , we can easily write down the conditional distribution of A and B , given C , in the form:

$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} & p(a, b, c) &= p(a | c)p(b | c)p(c) \\ &= p(a|c)p(b|c) \end{aligned}$$

which shows the conditional independence property:

$$A \perp B \mid C$$

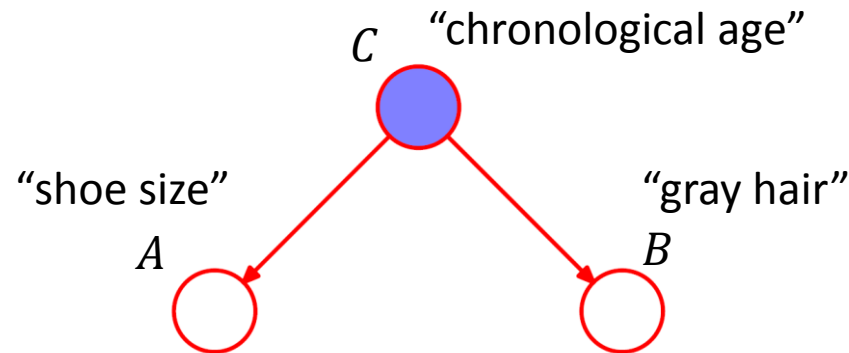
Three Canonical 3-Node Graphs



- The node C is said to be *tail-to-tail* with respect to the path from node A to B .
- Such a path **connects** nodes A and B and renders them dependent.
- The observation of C **'blocks'** the path from A to B , and we obtain the conditional independence property.

Three Canonical 3-Node Graphs

Intuitive interpretation:

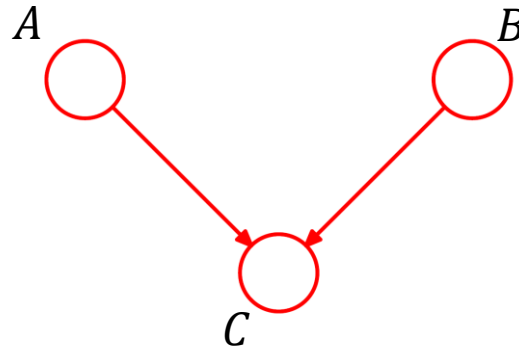


- Given the age of a person, there is **no further relationship** between the size of his feet and the amount of gray hair he has.
- We say that the variable C **“explains”** all of the observed dependence between A and B .

Image Source: “Pattern Recognition and Machine Learning”, Christopher Bishop

Three Canonical 3-Node Graphs

3.



Joint distribution corresponding to this graph:

$$p(a, b, c) = p(a)p(b)p(c|a, b)$$

If **none** of the variables are observed, marginalizing both sides over c we obtain:

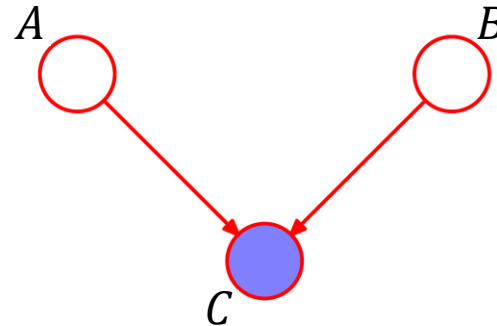
$$\begin{aligned} p(a, b) &= \sum_c p(a)p(b)p(c | a, b) \\ &= p(a)p(b) \end{aligned}$$

A and B **are independent with no variables observed**, in contrast to the two cases:

$$A \perp B \mid \emptyset$$

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Three Canonical 3-Node Graphs



If we **condition on** node C , the conditional distribution of A and B is given by:

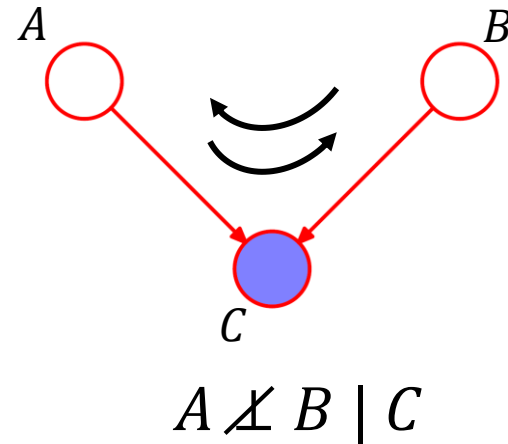
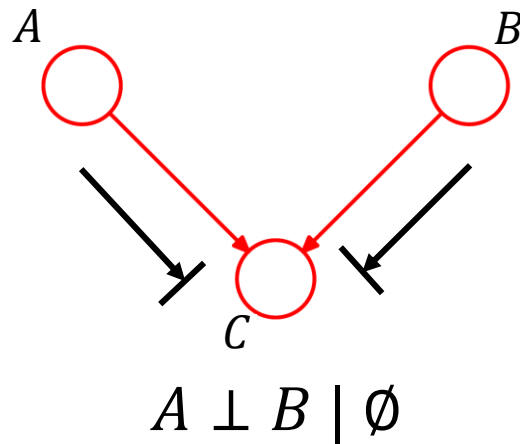
$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(b)p(c|a, b)}{p(c)} \end{aligned}$$

which in general does not factorize into the product $p(a)p(b)$, and so

$$A \not\perp B \mid C$$

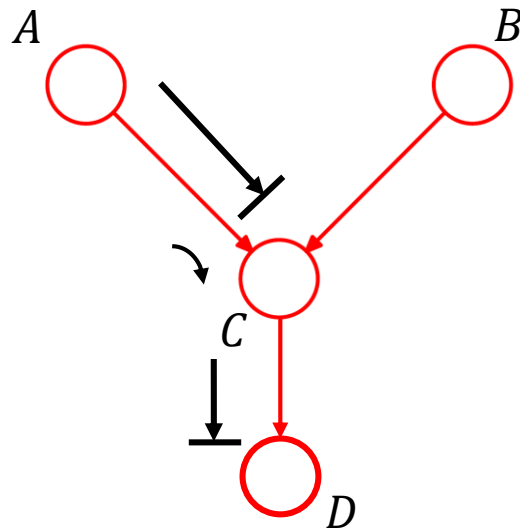
Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Three Canonical 3-Node Graphs

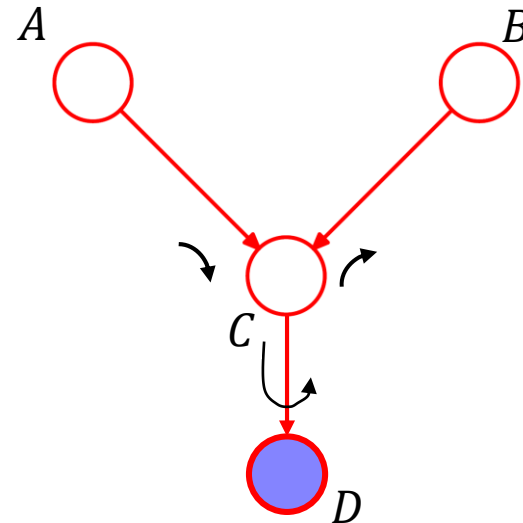


- Node C is *head-to-head* with respect to the path from A to B , also known as the “*v-structure*”.
- When node C is unobserved, it “*blocks*” the path, and the variables A and B are independent.
- However, conditioning on C “*unblocks*” the path and renders A and B dependent.

Three Canonical 3-Node Graphs



$$A \perp B \mid \emptyset$$

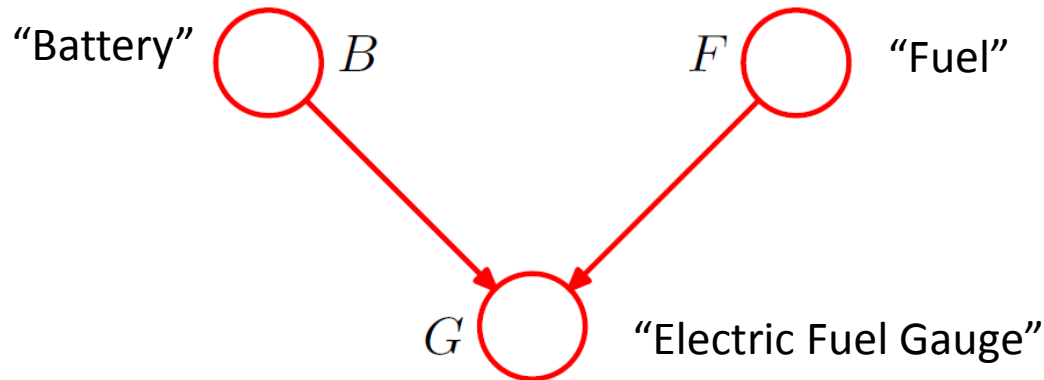


$$A \not\perp B \mid D$$

- The observation of any descendent node of C “**unblocks**” the path from A to B.

Three Canonical 3-Node Graphs

Intuitive interpretation:

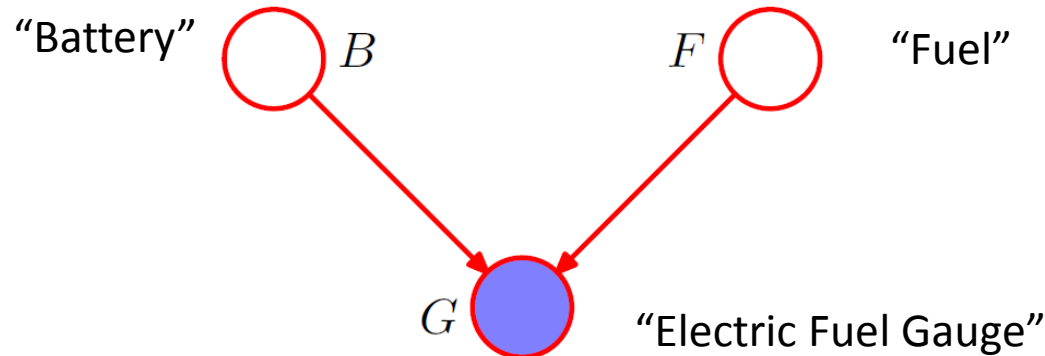


- **No relation** between the battery and fuel status if fuel gauge is not read.
- This implies **conditional independence** of "battery" and "fuel" when "gauge" is not observed.

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Three Canonical 3-Node Graphs

Intuitive interpretation:

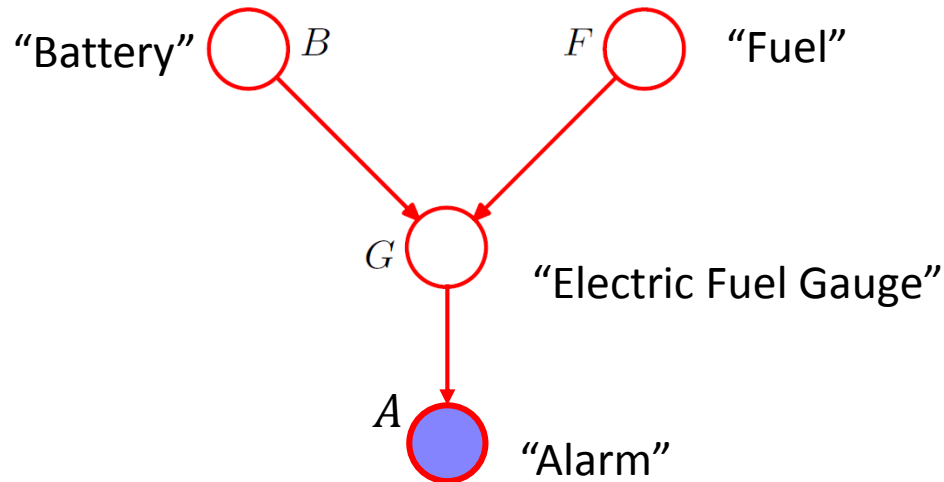


- Suppose the fuel gauge shows “empty”, knowing that the battery is flat **lowers our belief** that the fuel tank is empty.
- Battery and fuel status are now **no longer independent**.
- This is known as the **“explaining-away”** effect.

Image Source: “Pattern Recognition and Machine Learning”, Christopher Bishop

Three Canonical 3-Node Graphs

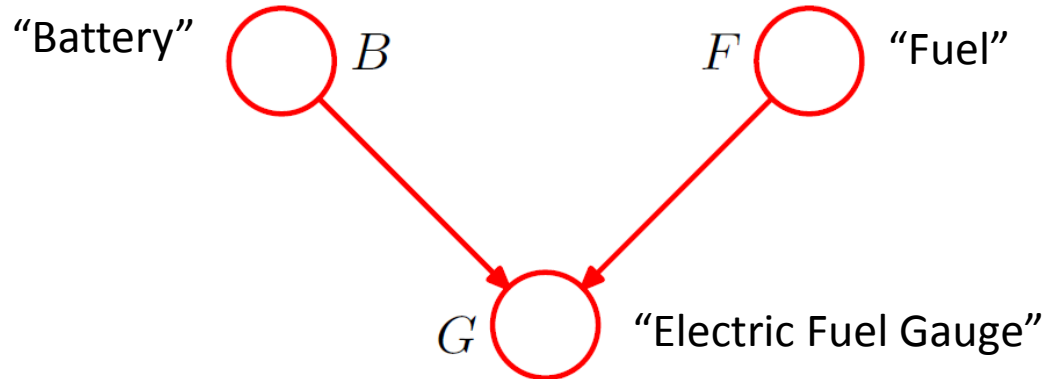
Intuitive interpretation:



- Alarm goes off when fuel gauge is empty.
- Suppose alarm goes off, we know that the fuel gauge shows "empty".
- Knowing that the battery is flat **lowers our belief** that the fuel tank is empty, i.e. battery and fuel status are now **no longer independent**.

Three Canonical 3-Node Graphs

Numerical Example:

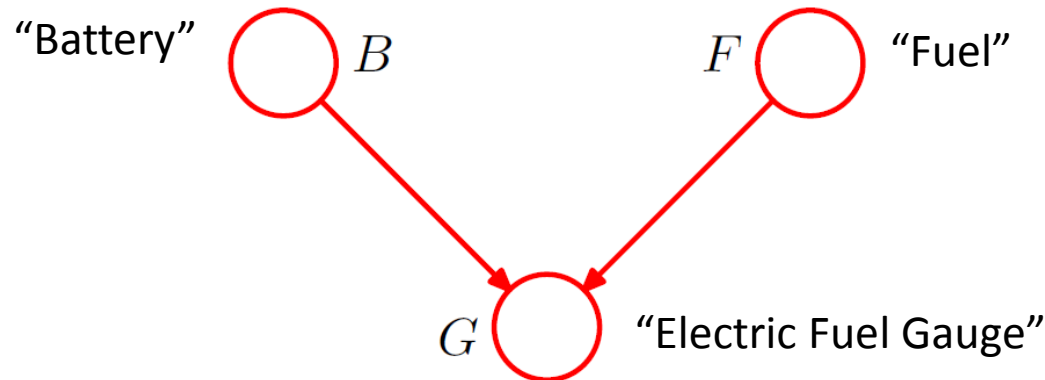


- B : battery state that is either charged ($B = 1$) or flat ($B = 0$).
- F : fuel tank state that is either full of fuel ($F = 1$) or empty ($F = 0$).
- G : electric fuel gauge state which indicates either full ($G = 1$) or empty ($G = 0$).

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Three Canonical 3-Node Graphs

Numerical Example:



Given:

$$p(B = 1) = 0.9$$

$$p(F = 1) = 0.9$$

$$p(G = 1 | B = 1, F = 1) = 0.8$$

$$p(G = 1 | B = 1, F = 0) = 0.2$$

$$p(G = 1 | B = 0, F = 1) = 0.2$$

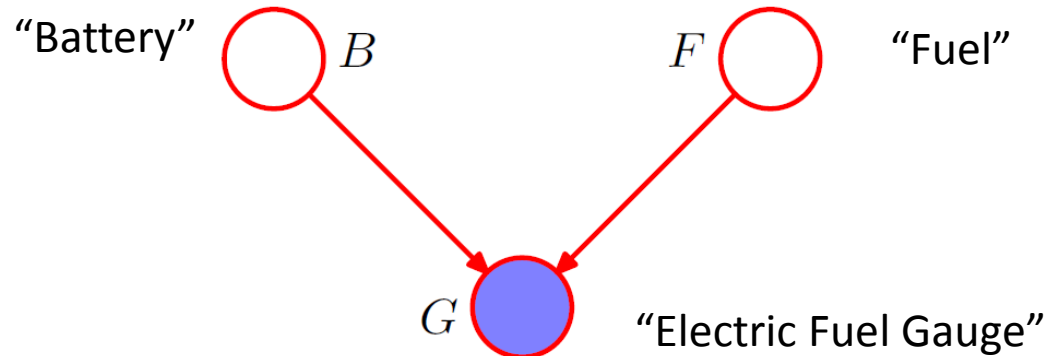
$$p(G = 1 | B = 0, F = 0) = 0.1$$

Before we observe any data, the **prior probability** of the fuel tank being empty is $p(F = 0) = 0.1$.

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Three Canonical 3-Node Graphs

Numerical Example:



Suppose that we observe the fuel gauge and it reads empty, i.e., $G = 0$, we have:

$$p(F = 0|G = 0) = \frac{p(G = 0|F = 0)p(F = 0)}{p(G = 0)} \simeq 0.257$$

where

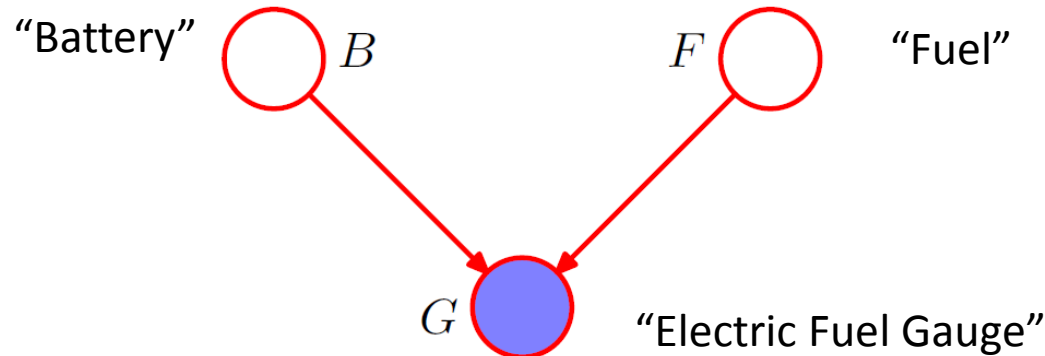
$$p(G = 0) = \sum_{b \in \{0,1\}} \sum_{f \in \{0,1\}} p(G = 0|B, F)p(B)p(F) = 0.315$$

$$p(G = 0|F = 0) = \sum_{b \in \{0,1\}} p(G = 0|B, F = 0)p(B) = 0.81$$

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Three Canonical 3-Node Graphs

Numerical Example:



Hence,

$$p(F = 0|G = 0) > p(F = 0)$$

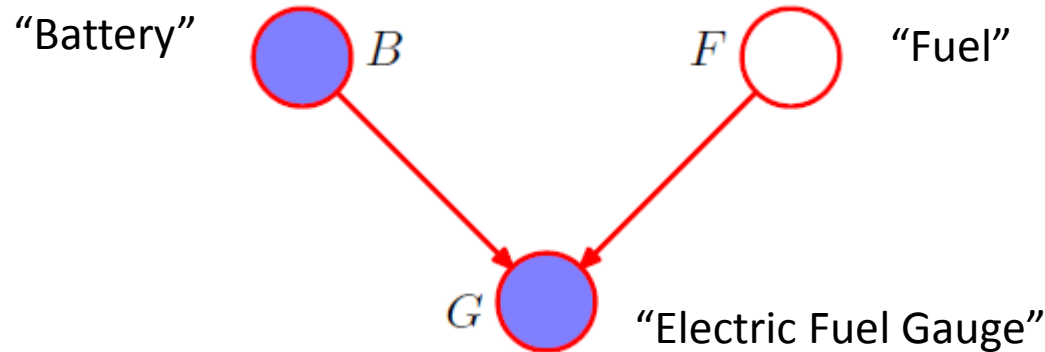
0.257 0.1

Observing that the gauge reads empty makes it **more likely** that the tank is indeed empty, as we would intuitively expect.

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Three Canonical 3-Node Graphs

Numerical Example:

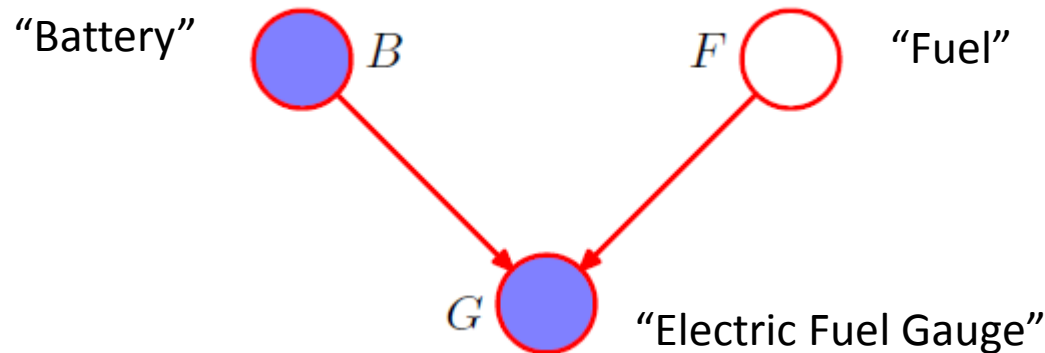


- If we also check the state of the battery and find that it is flat, i.e., $B = 0$.
- We have now observed the states of **both** fuel gauge and battery.

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Three Canonical 3-Node Graphs

Numerical Example:



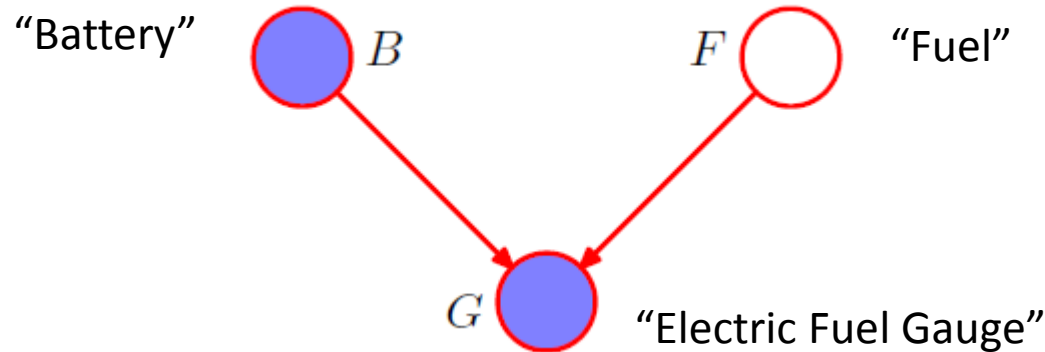
- **Posterior probability** that fuel tank is empty given observations of both fuel gauge and battery state is:

$$p(F = 0 | G = 0, B = 0) = \frac{p(G = 0 | B = 0, F = 0)p(F = 0)}{\sum_{F \in \{0,1\}} p(G = 0 | B = 0, F)p(F)} \simeq 0.111$$

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Three Canonical 3-Node Graphs

Numerical Example:



0.257

0.111

$$p(F = 0|G = 0) > p(F = 0|G = 0, B = 0)$$

- Finding out that battery is flat *explains away* observation that the fuel gauge reads empty!

Image Source: “Pattern Recognition and Machine Learning”, Christopher Bishop

Graph Separation

- We have seen earlier that $A \perp B \mid C$ if all paths from nodes in set A are “**blocked**” from nodes in set B when all nodes from set C are observed.
- A is said to be **d-separated** from B by C , and the joint distribution over all of the variables in the graph will satisfy $A \perp B \mid C$.

Graph Separation

- From the **three canonical 3-node graphs**, any path is said to be “**blocked**” / **d-separated** if it includes a node such that either:
 - a) The arrows on the path meet either **head-to-tail** or **tail-to-tail** at the node, and the node is in the set C , or
 - b) The arrows meet **head-to-head** at the node, and neither the node, nor any of its descendants, is in the set C .

Bayes Ball Algorithm

- This is a “reachability” algorithm:
 1. Shade the nodes in set C .
 2. Place a ball at each of the nodes in set A .
 3. Let the balls bounce around the graph according to the **d-separation rules**:

IF none of the balls reach B THEN

$$A \perp B \mid C,$$

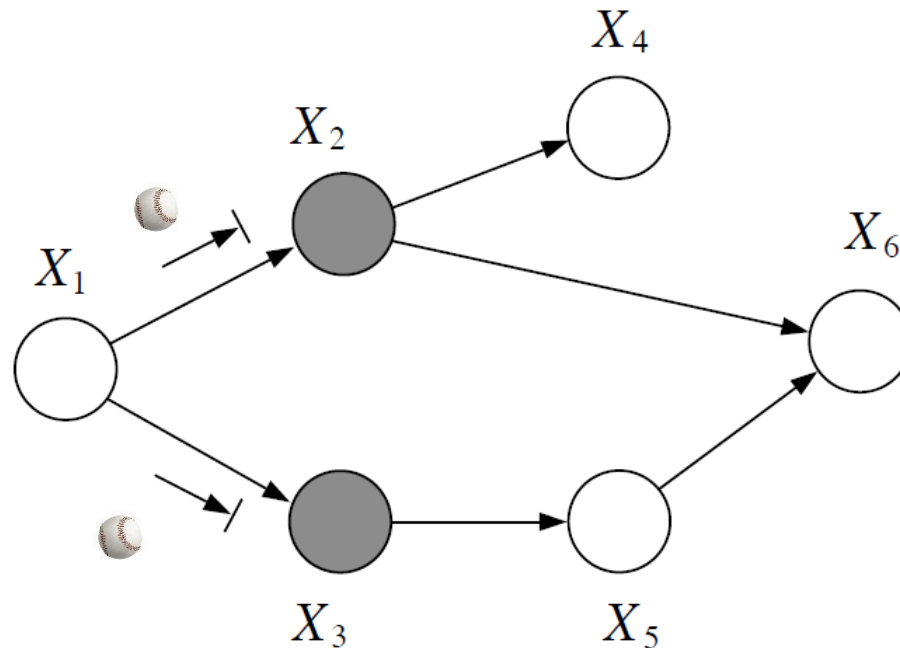
ELSE

$$A \not\perp B \mid C,$$

- Can be implemented as a **breadth-first search**.

Bayes Ball Algorithm

Example 1:

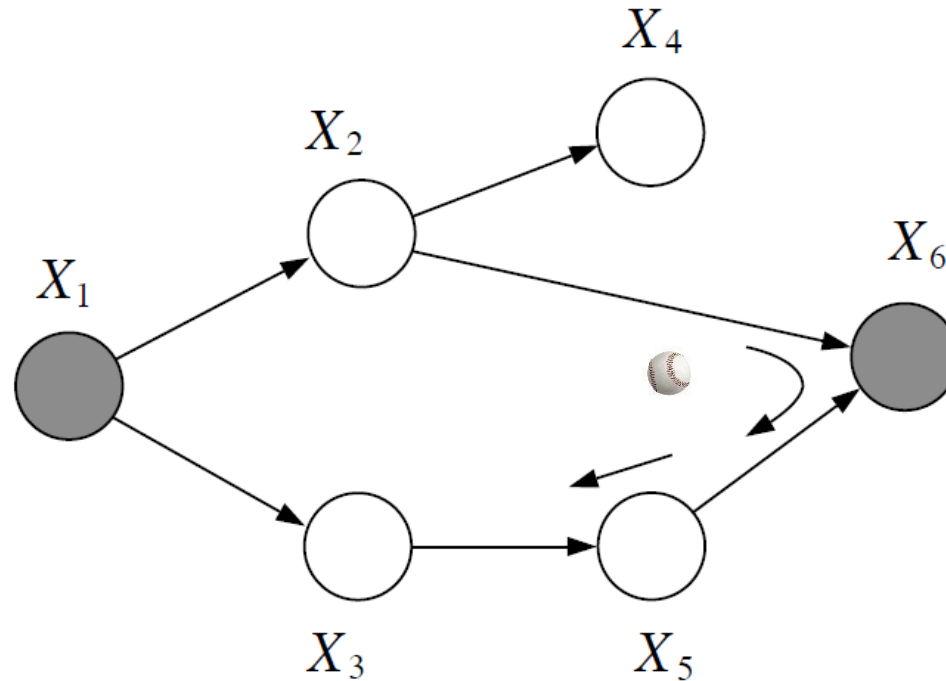


- A ball **cannot pass** through X_2 to X_6 nor through X_3 i.e. $X_1 \perp X_6 \mid \{X_2, X_3\}$.

Image modified from: "An introduction to probabilistic graphical models", Michael I. Jordan, 2002.

Bayes Ball Algorithm

Example 2:



- A ball **can pass** through X_2 to X_6 through X_5 and X_3 i.e. $X_2 \not\perp\!\!\!\perp X_3 \mid \{X_1, X_6\}$.

Image modified from: "An introduction to probabilistic graphical models", Michael I. Jordan, 2002.

Bayes Ball Algorithm

Example 3: Naïve Bayes Classifier

- **Classify e-mails** as spam ($Y = 1$) or not spam ($Y = 0$).
- Let $1:N$ index the words in our vocabulary (e.g., English)
- $$X_n = \begin{cases} 0 & \text{word } n \text{ appears in an e-mail} \\ 1 & \text{otherwise} \end{cases}$$
- E-mails are drawn according to some distribution $p(x_1, \dots, x_N, Y)$

Bayes Ball Algorithm

Example 3: Naïve Bayes Classifier

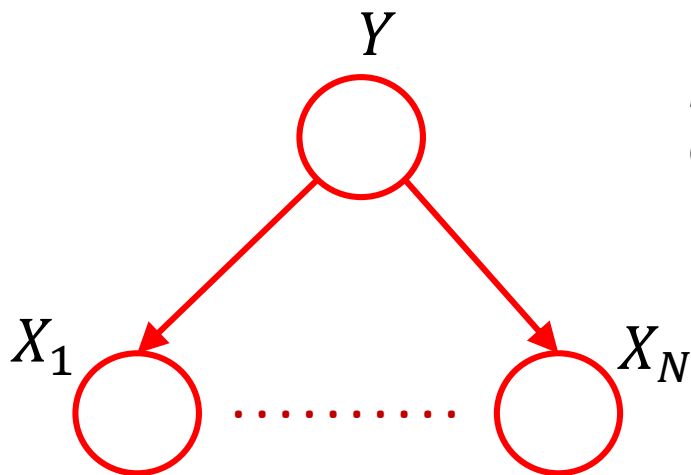
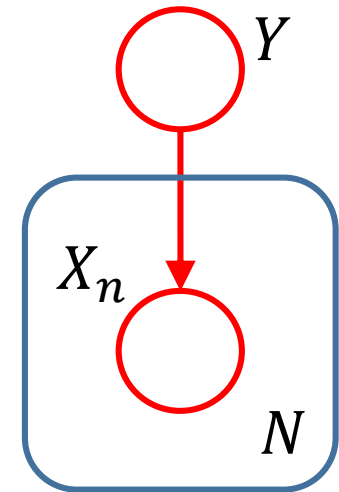


Plate notation that represents N nodes of which only a single example X_n is shown explicitly

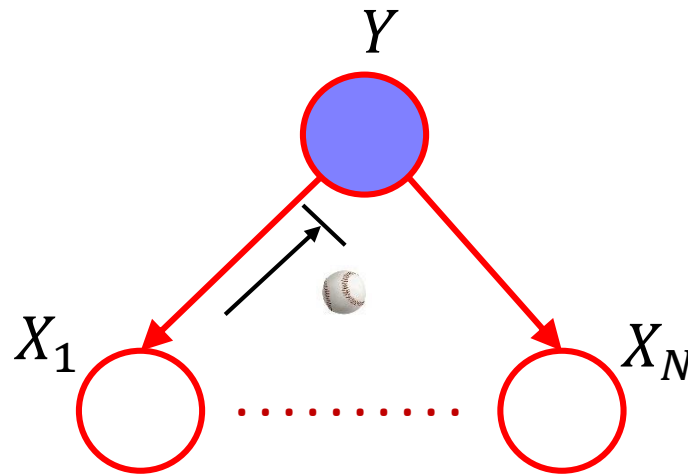


Joint distribution:

$$p(x_1, \dots, x_N, Y) = p(x_1, \dots, x_N | Y) p(Y)$$

Bayes Ball Algorithm

Example 3: Naïve Bayes Classifier

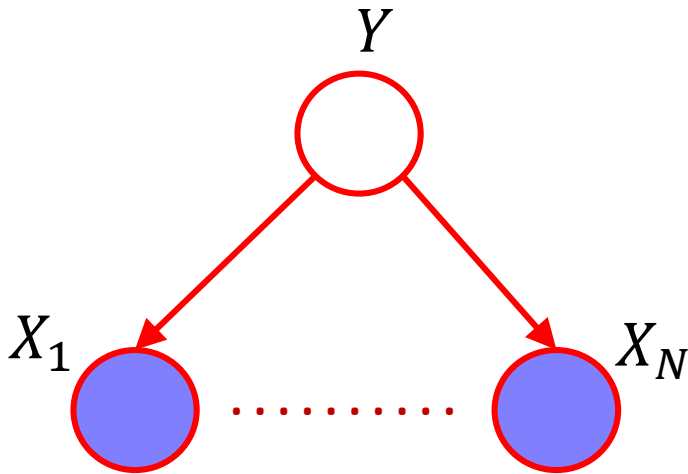


Joint distribution:

$$\begin{aligned} p(x_1, \dots, x_N, Y) &= p(x_1, \dots, x_N | Y) p(Y) \\ &= p(x_1 | Y) \dots p(x_N | Y) p(Y) \\ &= \prod_{n=1}^N p(x_n | Y) p(Y) \end{aligned}$$

Bayes Ball Algorithm

Example 3: Naïve Bayes Classifier



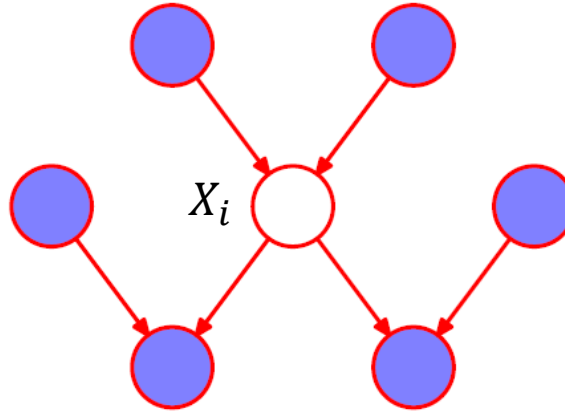
Joint distribution:

$$p(x_1, \dots, x_N, Y) = \prod_{n=1}^N p(x_n | Y) p(Y)$$

Predict with Bayes rule:

$$p(Y = 1 | x_1, \dots, x_N) = \frac{p(y = 1) \prod_{n=1}^N p(x_n | Y = 1)}{\sum_{y=\{0,1\}} p(Y = y) \prod_{n=1}^N p(x_n | Y = y)}$$

Markov Blanket

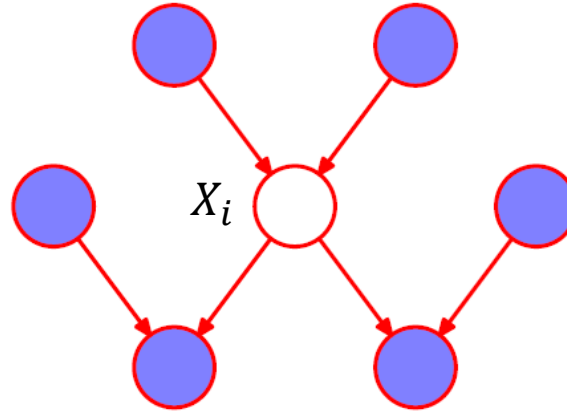


- The Markov blanket of a node X_i comprises the **set of parents, children and co-parents** of the node.
- Conditional distribution of X_i , conditioned on all the remaining variables in the graph, is **dependent only** on the variables in the Markov blanket.

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Markov Blanket

Proof:

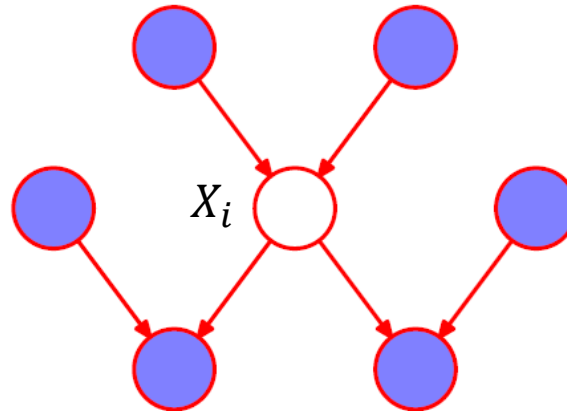


$$\begin{aligned}
 p(x_i | x_{\{j \neq i\}}) &= \frac{p(x_1, \dots, x_D)}{\int p(x_1, \dots, x_D) dx_i} = \frac{\prod_k p(x_k | x_{\pi_k})}{\int \prod_k p(x_k | x_{\pi_k}) dx_i} \\
 &= \frac{\cancel{\prod_{l \neq \{m, i\}} p(x_l | x_{\pi_l})} \prod_m p(x_m | x_{\pi_m} : x_i \in x_{\pi_m}) p(x_i | x_{\pi_i})}{\cancel{\prod_{l \neq \{m, i\}} p(x_l | x_{\pi_l})} \int \prod_m p(x_m | x_{\pi_m} : x_i \in x_{\pi_m}) p(x_i | x_{\pi_i}) dx_i} \\
 &= \frac{\prod_m p(x_m | x_{\pi_m} : x_i \in x_{\pi_m}) p(x_i | x_{\pi_i})}{\int \prod_m p(x_m | x_{\pi_m} : x_i \in x_{\pi_m}) p(x_i | x_{\pi_i}) dx_i}
 \end{aligned}$$

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Markov Blanket

Proof:



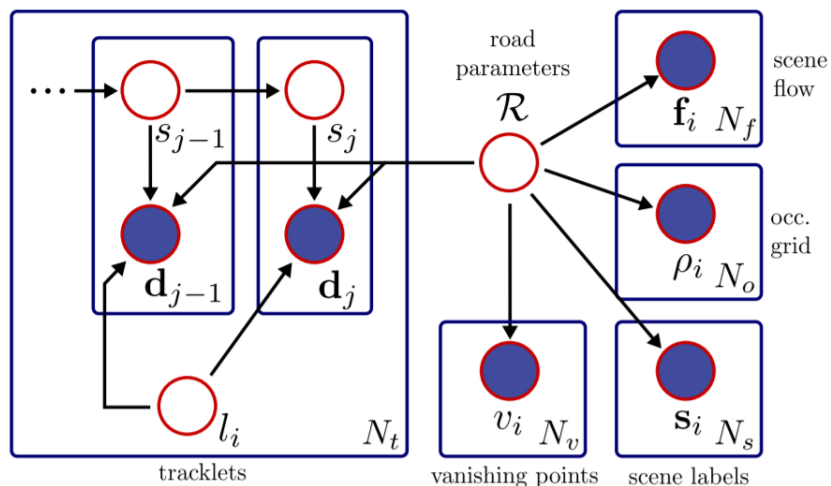
$$p(x_i | x_{\{j \neq i\}}) = \frac{\prod_m p(x_m | x_{\pi_m} : x_i \in x_{\pi_m}) p(x_i | x_{\pi_i})}{\int \prod_m p(x_m | x_{\pi_m} : x_i \in x_{\pi_m}) p(x_i | x_{\pi_i}) dx_i}$$

children and co-parents of X_i parents of X_i

Image Source: "Pattern Recognition and Machine Learning", Christopher Bishop

Examples of Bayesian Network

Urban Scene Understanding:



$$\begin{aligned}
 p(\mathcal{E}, \mathcal{R} | \Theta) &= \underbrace{p(\mathcal{R} | \Theta)}_{\text{Prior}} \underbrace{\left[\prod_{i=1}^{N_t} p(\mathbf{t}_i | \mathcal{R}, \Theta) \right]}_{\text{Vehicle Tracklets}} \\
 &\times \underbrace{\prod_{i=1}^{N_v} p(v_i | \mathcal{R}, \Theta)}_{\text{Vanishing Points}} \underbrace{\prod_{i=1}^{N_s} p(s_i | \mathcal{R}, \Theta)}_{\text{Scene Labels}} \\
 &\times \underbrace{\prod_{i=1}^{N_f} p(\mathbf{f}_i | \mathcal{R}, \Theta)}_{\text{Scene Flow}} \underbrace{\prod_{i=1}^{N_o} p(\rho_i | \mathcal{R}, \Theta)}_{\text{Occupancy Grid}}
 \end{aligned}$$

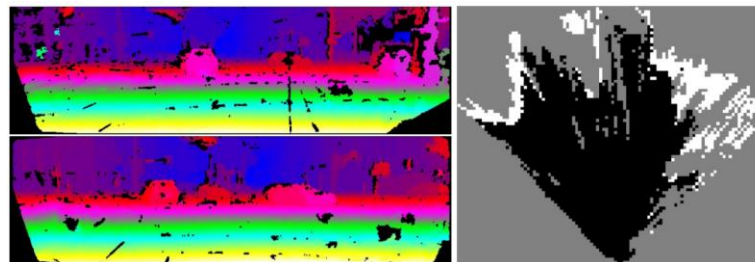
Vanishing Points



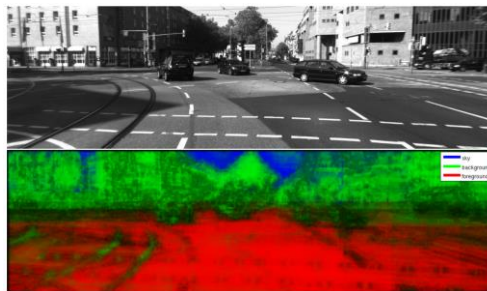
Tracklets



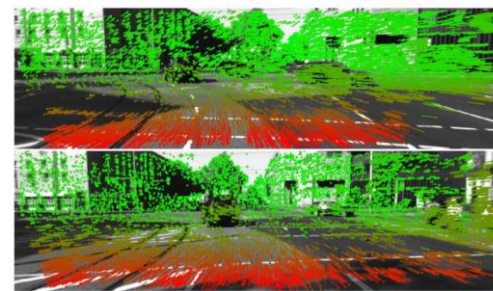
Occupancy Grid



Semantic Image Segmentation



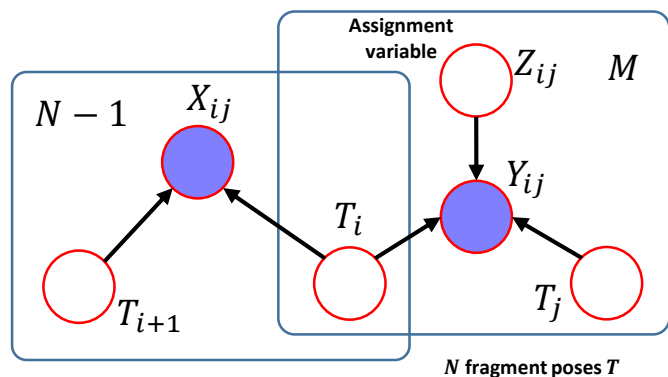
Scene Flow



Andreas Geiger, PhD Thesis 2013

Examples of Bayesian Network

Robust 3D Point Cloud Reconstruction:

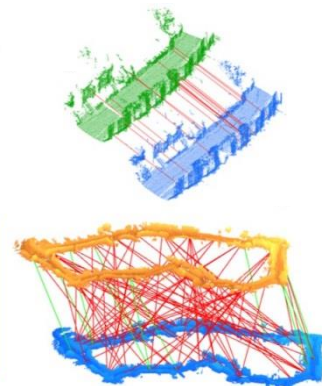


front-end
back-end

Odometry Constraints $X_{i,i+1}$

Loop-Closure Constraints Y_{ij}

Robust Reconstruction with EM



Expectation step:

$$Q_{EM} = \sum_Z p(Z | Y, T^{old}) \ln p(X, Y, Z | T)$$

Maximization step:

$$\operatorname{argmax}_T Q_{EM}$$



Summary

- You have learned how to:
 1. Explain the concepts of **conditional independence**.
 2. Use the **Bayesian network** to represent conditional independence in joint distributions.
 3. Describe **d-separation** using the **three canonical 3-node graph**.
 4. Deduce all conditional independence in a Bayesian network using the **Bayes ball algorithm**.
 5. Explain the concepts of **Markov Blanket**.