

Causal Semantics of Bayesian Networks

Jirka Vomlel

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
<http://www.utia.cz/vomlel>

Salzburg, 26 February 2010

Development of Western science is based on two great achievements:

- *the invention of the formal logical system by the Greek philosophers, and*
- *the discovery of the possibility to find out causal relationship by systematic experiment during Renaissance.*

Albert Einstein (1953)

- Bayesian networks

- Bayesian networks
- Observations versus interventions

- Bayesian networks
- Observations versus interventions
- Causal semantics of Bayesian networks

- Bayesian networks
- Observations versus interventions
- Causal semantics of Bayesian networks
- Latent variables in causal models

- Bayesian networks
- Observations versus interventions
- Causal semantics of Bayesian networks
- Latent variables in causal models
- Tractable causal models with latent variables

Bayesian network (BN)

See file `fever1.net` in Hugin.

Bayesian network (BN)

See file `fever1.net` in Hugin.

- Acyclic directed graph $G = (V, E)$ where $V \subset \{1, 2, \dots, n\}$ is a set of nodes and E is a set of directed edges - formally, a subset of $V \times V$.

Bayesian network (BN)

See file [fever1.net](#) in Hugin.

- Acyclic directed graph $G = (V, E)$ where $V \subset \{1, 2, \dots, n\}$ is a set of nodes and E is a set of directed edges - formally, a subset of $V \times V$.
- $X_i, i \in V$ are discrete random variables.

Bayesian network (BN)

See file [fever1.net](#) in Hugin.

- Acyclic directed graph $G = (V, E)$ where $V \subset \{1, 2, \dots, n\}$ is a set of nodes and E is a set of directed edges - formally, a subset of $V \times V$.
- $X_i, i \in V$ are discrete random variables.
- Let $pa(i)$ denote the set of nodes that are parents of i in G - formally, $pa(i) = \{j \in V : (j \rightarrow i) \in E\}$.

Bayesian network (BN)

See file [fever1.net](#) in Hugin.

- Acyclic directed graph $G = (V, E)$ where $V \subset \{1, 2, \dots, n\}$ is a set of nodes and E is a set of directed edges - formally, a subset of $V \times V$.
- $X_i, i \in V$ are discrete random variables.
- Let $pa(i)$ denote the set of nodes that are parents of i in G - formally, $pa(i) = \{j \in V : (j \rightarrow i) \in E\}$.
- Further, let for $A \subseteq V$ symbol X_A denotes a set of variables $\{X_j\}_{j \in A}$.

Bayesian network (BN)

See file [fever1.net](#) in Hugin.

- Acyclic directed graph $G = (V, E)$ where $V \subset \{1, 2, \dots, n\}$ is a set of nodes and E is a set of directed edges - formally, a subset of $V \times V$.
- $X_i, i \in V$ are discrete random variables.
- Let $pa(i)$ denote the set of nodes that are parents of i in G - formally, $pa(i) = \{j \in V : (j \rightarrow i) \in E\}$.
- Further, let for $A \subseteq V$ symbol X_A denotes a set of variables $\{X_j\}_{j \in A}$.
- For each random variable $X_i, i \in V$ a conditional probability distribution $P(X_i | X_{pa(i)})$ is defined.

Bayesian network (BN)

See file [fever1.net](#) in Hugin.

- Acyclic directed graph $G = (V, E)$ where $V \subset \{1, 2, \dots, n\}$ is a set of nodes and E is a set of directed edges - formally, a subset of $V \times V$.
- $X_i, i \in V$ are discrete random variables.
- Let $pa(i)$ denote the set of nodes that are parents of i in G - formally, $pa(i) = \{j \in V : (j \rightarrow i) \in E\}$.
- Further, let for $A \subseteq V$ symbol X_A denotes a set of variables $\{X_j\}_{j \in A}$.
- For each random variable $X_i, i \in V$ a conditional probability distribution $P(X_i | X_{pa(i)})$ is defined.
- Then the joint probability distribution defined by a Bayesian network is

$$P(X_V) = \prod_{i \in V} P(X_i | X_{pa(i)}) .$$

Conditional independence

Let $C \subset V$ denote the set with indexes corresponding to variables whose state is known.

Conditional independence

Let $C \subset V$ denote the set with indexes corresponding to variables whose state is known.

Definition (Path blocked by evidence)

A path in a acyclic directed graph is **blocked** by a set C if there is a node $n \in V$ in the path such that

Conditional independence

Let $C \subset V$ denote the set with indexes corresponding to variables whose state is known.

Definition (Path blocked by evidence)

A path in a acyclic directed graph is **blocked** by a set C if there is a node $n \in V$ in the path such that

- the arrows do not meet head-to-head in n and $n \in C$ or

Conditional independence

Let $C \subset V$ denote the set with indexes corresponding to variables whose state is known.

Definition (Path blocked by evidence)

A path in a acyclic directed graph is **blocked** by a set C if there is a node $n \in V$ in the path such that

- the arrows do not meet head-to-head in n and $n \in C$ or
- the arrows meet head-to-head in n and neither n nor any of its descendants belong to C .

Conditional independence

Let $C \subset V$ denote the set with indexes corresponding to variables whose state is known.

Definition (Path blocked by evidence)

A path in a acyclic directed graph is **blocked** by a set C if there is a node $n \in V$ in the path such that

- the arrows do not meet head-to-head in n and $n \in C$ or
- the arrows meet head-to-head in n and neither n nor any of its descendants belong to C .

Definition (Conditional independence)

Let A, B, C be pairwise disjoint subsets of V .

X_A is **independent** of X_B given X_C ($X_A \perp\!\!\!\perp X_B | X_C$) iff all paths between A and B are blocked by set C .

Conditional independence

Let $C \subset V$ denote the set with indexes corresponding to variables whose state is known.

Definition (Path blocked by evidence)

A path in a acyclic directed graph is **blocked** by a set C if there is a node $n \in V$ in the path such that

- the arrows do not meet head-to-head in n and $n \in C$ or
- the arrows meet head-to-head in n and neither n nor any of its descendants belong to C .

Definition (Conditional independence)

Let A, B, C be pairwise disjoint subsets of V .

X_A is **independent** of X_B given X_C ($X_A \perp\!\!\!\perp X_B | X_C$) iff all paths between A and B are blocked by set C .

See file fever1.net in Hugin again.

BN example: Removing Kidney stones (Charig et al., 1986)

Let $V = \{X_1, X_2, X_3\}$ where

- X_1 is *Stones' size* taking values s (small) and l (large),

BN example: Removing Kidney stones (Charig et al., 1986)

Let $V = \{X_1, X_2, X_3\}$ where

- X_1 is *Stones' size* taking values s (small) and l (large),
- X_2 is *Treatment* taking values A (all open procedures) and B (percutaneous nephrolithotomy),

BN example: Removing Kidney stones (Charig et al., 1986)

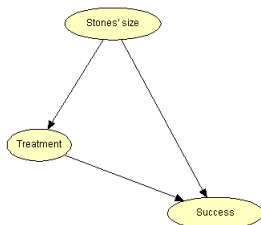
Let $V = \{X_1, X_2, X_3\}$ where

- X_1 is *Stones' size* taking values s (small) and l (large),
- X_2 is *Treatment* taking values A (all open procedures) and B (percutaneous nephrolithotomy),
- X_3 is *Success* taking values 0 (False) and 1 (True).

BN example: Removing Kidney stones (Charig et al., 1986)

Let $V = \{X_1, X_2, X_3\}$ where

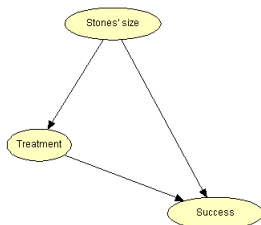
- X_1 is *Stones' size* taking values *s* (small) and *l* (large),
- X_2 is *Treatment* taking values *A* (all open procedures) and *B* (percutaneous nephrolithotomy),
- X_3 is *Success* taking values **0** (False) and **1** (True).



BN example: Removing Kidney stones (Charig et al., 1986)

Let $V = \{X_1, X_2, X_3\}$ where

- X_1 is *Stones' size* taking values s (small) and l (large),
- X_2 is *Treatment* taking values A (all open procedures) and B (percutaneous nephrolithotomy),
- X_3 is *Success* taking values 0 (False) and 1 (True).



$$P(X_1, X_2, X_3) = P(X_3|X_1, X_2) \cdot P(X_2|X_1) \cdot P(X_1)$$

Conditioning by observation

Assume that we have observed that a person with kidney stones is going to undergo treatment A. What is the probability of the success?

$$P(X_3 = 1 | X_2 = A) = \frac{P(X_3 = 1, X_2 = A)}{P(X_2 = A)}$$

Conditioning by observation

Assume that we have observed that a person with kidney stones is going to undergo treatment A. What is the probability of the success?

$$P(X_3 = 1 | X_2 = A) = \frac{P(X_3 = 1, X_2 = A)}{P(X_2 = A)}$$

where

$$P(X_2 = A) = \sum_{x_3 \in \{0,1\}} P(X_3 = x_3, X_2 = A)$$

Conditioning by observation

Assume that we have observed that a person with kidney stones is going to undergo treatment A. What is the probability of the success?

$$P(X_3 = 1 | X_2 = A) = \frac{P(X_3 = 1, X_2 = A)}{P(X_2 = A)}$$

where

$$P(X_2 = A) = \sum_{x_3 \in \{0,1\}} P(X_3 = x_3, X_2 = A)$$

$$P(X_3 = x_3, X_2 = A) = \sum_{x_1 \in \{s,l\}} P(X_3 = x_3, X_2 = A, X_1 = x_1)$$

Conditioning by observation

Assume that we have observed that a person with kidney stones is going to undergo treatment A. What is the probability of the success?

$$P(X_3 = 1 | X_2 = A) = \frac{P(X_3 = 1, X_2 = A)}{P(X_2 = A)}$$

where

$$P(X_2 = A) = \sum_{x_3 \in \{0,1\}} P(X_3 = x_3, X_2 = A)$$

$$P(X_3 = x_3, X_2 = A) = \sum_{x_1 \in \{s,l\}} P(X_3 = x_3, X_2 = A, X_1 = x_1)$$

and using the definition of the Bayesian network

$$\begin{aligned} P(X_3 = x_3, X_2 = A, X_1 = x_1) &= \\ &P(X_3 = x_3 | X_2 = A, X_1 = x_1) \cdot P(X_2 = A | X_1 = x_1) \cdot P(X_1 = x_1) \end{aligned}$$

Conditioning by observation

Assume that we have observed that a person with kidney stones is going to undergo treatment A. What is the probability of the success?

$$P(X_3 = 1 | X_2 = A) = \frac{P(X_3 = 1, X_2 = A)}{P(X_2 = A)}$$

where

$$P(X_2 = A) = \sum_{x_3 \in \{0,1\}} P(X_3 = x_3, X_2 = A)$$

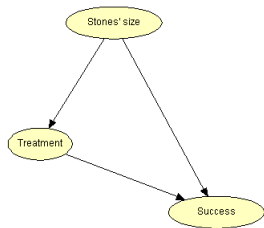
$$P(X_3 = x_3, X_2 = A) = \sum_{x_1 \in \{s,l\}} P(X_3 = x_3, X_2 = A, X_1 = x_1)$$

and using the definition of the Bayesian network

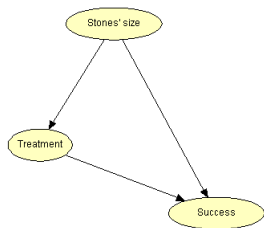
$$\begin{aligned} P(X_3 = x_3, X_2 = A, X_1 = x_1) &= \\ &P(X_3 = x_3 | X_2 = A, X_1 = x_1) \cdot P(X_2 = A | X_1 = x_1) \cdot P(X_1 = x_1) \end{aligned}$$

See file [kidney_stones.net](#) in Hugin.

Conditioning by intervention

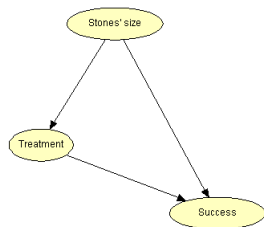


Conditioning by intervention



Now, we want to evaluate which treatment A or B is generally better.

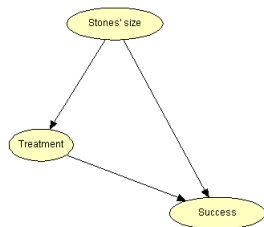
Conditioning by intervention



Now, we want to evaluate which treatment A or B is generally better. Shall we check whether

$$P(X_3 = 1 | X_2 = A) > P(X_3 = 1 | X_2 = B) ?$$

Conditioning by intervention

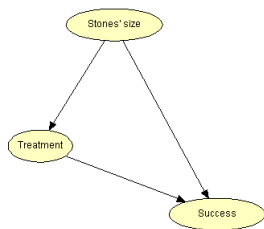


Now, we want to evaluate which treatment A or B is generally better. Shall we check whether

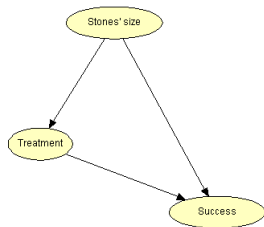
$$P(X_3 = 1 | X_2 = A) > P(X_3 = 1 | X_2 = B) ?$$

No! We have a covariate X_1 in the model that is influenced by X_2 and has an impact on X_3 , i.e., there is an open path $X_2 \leftarrow X_1 \rightarrow X_3$.

Conditioning by intervention

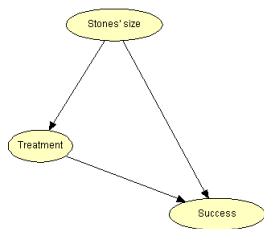


Conditioning by intervention



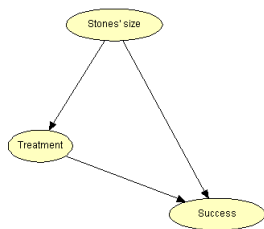
- When we use a treatment we intervene in the world.

Conditioning by intervention



- When we use a treatment we intervene in the world.
- We manipulate the value of the corresponding variable.

Conditioning by intervention



- When we use a treatment we intervene in the world.
- We manipulate the value of the corresponding variable.
- We break the causal mechanism therefore the manipulated variable should be disconnected from all variables that causally influence it.

Conditioning by intervention



- When we use a treatment we intervene in the world.
- We manipulate the value of the corresponding variable.
- We break the causal mechanism therefore the manipulated variable should be disconnected from all variables that causally influence it.

Conditioning by intervention



- When we use a treatment we intervene in the world.
- We manipulate the value of the corresponding variable.
- We break the causal mechanism therefore the manipulated variable should be disconnected from all variables that causally influence it.

See file [simpsons_paradox_kidney_stones.net](#) in Hugin.

Causal Bayesian networks

- Each causal variable X_i is governed by a causal mechanism that stochastically determines its value based on the value of its parents.

Causal Bayesian networks

- Each causal variable X_i is governed by a causal mechanism that stochastically determines its value based on the value of its parents.
- The causal mechanism takes the same form as the conditional probability distribution $P(X_i|X_{pa(i)})$.

Causal Bayesian networks

- Each causal variable X_i is governed by a causal mechanism that stochastically determines its value based on the value of its parents.
- The causal mechanism takes the same form as the conditional probability distribution $P(X_i|X_{pa(i)})$.
- Causality flows in the direction of edges.

Causal Bayesian networks

- Each causal variable X_i is governed by a causal mechanism that stochastically determines its value based on the value of its parents.
- The causal mechanism takes the same form as the conditional probability distribution $P(X_i|X_{pa(i)})$.
- Causality flows in the direction of edges.
- Intervention on a variable $X_i \leftarrow x_i^*$ results in replacing its causal mechanism with one that dictates X_i takes value x_i^* .

Causal Bayesian networks

- Each causal variable X_i is governed by a causal mechanism that stochastically determines its value based on the value of its parents.
- The causal mechanism takes the same form as the conditional probability distribution $P(X_i|X_{pa(i)})$.
- Causality flows in the direction of edges.
- Intervention on a variable $X_i \leftarrow x_i^*$ results in replacing its causal mechanism with one that dictates X_i takes value x_i^* .

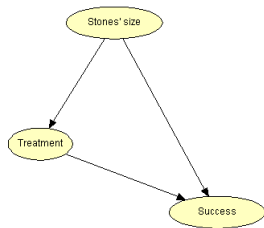
Definition

Probability after intervention $X_A \leftarrow x_A^*$ is defined as

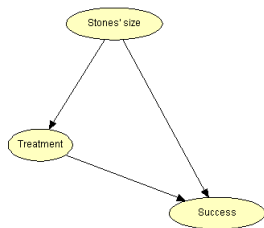
$$P(X_B = x_B | X_A \leftarrow x_A^*) = \prod_{i \in B} P(X_i | X_{pa(i)} = \bar{x}_{pa(i)})$$

where $B = V \setminus A$ and for $C \subseteq V$ the symbol \bar{x}_C denotes vector $x_C = (x_i)_{i \in C}$ with values x_i corresponding to $i \in A \cap C$ substituted by x_i^* .

Probability after intervention - example



Probability after intervention - example



$$\begin{aligned} &P(X_1 = x_1, X_3 = x_3 | X_2 \leftarrow x_2^*) \\ &= P(X_3 = x_3 | X_1 = x_1, X_2 = x_2^*) \cdot P(X_1 = x_1) \end{aligned}$$

Probability after intervention - example



$$\begin{aligned} P(X_1 = x_1, X_3 = x_3 | X_2 \leftarrow x_2^*) \\ = P(X_3 = x_3 | X_1 = x_1, X_2 = x_2^*) \cdot P(X_1 = x_1) \end{aligned}$$

Definition (General Problem)

Assume variables are partitioned by $\{t\} \cup R \cup C \cup U = V$ as:

Definition (General Problem)

Assume variables are partitioned by $\{t\} \cup R \cup C \cup U = V$ as:

- a treatment variable X_t ,

Definition (General Problem)

Assume variables are partitioned by $\{t\} \cup R \cup C \cup U = V$ as:

- a treatment variable X_t ,
- a set of response variables X_R ,

Definition (General Problem)

Assume variables are partitioned by $\{t\} \cup R \cup C \cup U = V$ as:

- a treatment variable X_t ,
- a set of response variables X_R ,
- a set of observed covariates X_C , and

Definition (General Problem)

Assume variables are partitioned by $\{t\} \cup R \cup C \cup U = V$ as:

- a treatment variable X_t ,
- a set of response variables X_R ,
- a set of observed covariates X_C , and
- a set of unobserved (latent) variables X_U .

Definition (General Problem)

Assume variables are partitioned by $\{t\} \cup R \cup C \cup U = V$ as:

- a treatment variable X_t ,
- a set of response variables X_R ,
- a set of observed covariates X_C , and
- a set of unobserved (latent) variables X_U .

Probability distribution $P(t, R, C)$ is known.

Definition (General Problem)

Assume variables are partitioned by $\{t\} \cup R \cup C \cup U = V$ as:

- a treatment variable X_t ,
- a set of response variables X_R ,
- a set of observed covariates X_C , and
- a set of unobserved (latent) variables X_U .

Probability distribution $P(t, R, C)$ is known.

$P(t, R, C, U)$ is unknown but its underlying structure is given by an acyclic directed graph $G = (V, E)$.

Definition (General Problem)

Assume variables are partitioned by $\{t\} \cup R \cup C \cup U = V$ as:

- a treatment variable X_t ,
- a set of response variables X_R ,
- a set of observed covariates X_C , and
- a set of unobserved (latent) variables X_U .

Probability distribution $P(t, R, C)$ is known.

$P(t, R, C, U)$ is unknown but its underlying structure is given by an acyclic directed graph $G = (V, E)$.

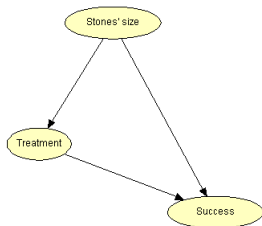
Can $P(X_R = x_R | X_t \leftarrow x_t^*)$ be determined from this information?

Definition

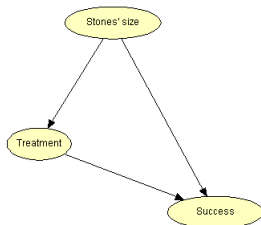
X_C identifies causal effect of X_t on X_R if for any pair of causal Bayesian networks P_1, P_2 with graph G it holds that

$$\begin{aligned} P_1(X_t = x_t, X_R = x_R, X_C = x_C) &\equiv P_2(X_t = x_t, X_R = x_R, X_C = x_C) \\ \implies P_1(X_R = x_r | X_t \leftarrow x_t^*) &\equiv P_2(X_R = x_r | X_t \leftarrow x_t^*) . \end{aligned}$$

Identifiability of causal effects



Identifiability of causal effects



Again, assume $X_t = \textit{Treatment}$, and $X_R = \textit{Success}$. If *Stones' size* is observable then it identifies causal effect of X_t on X_R , if it is latent then causal effect of X_t on X_R is not identifiable.

Intervention graph

We can decide the identifiability from the graph - we don't need to resort to probability distributions.

Intervention graph

We can decide the identifiability from the graph - we don't need to resort to probability distributions.

Definition

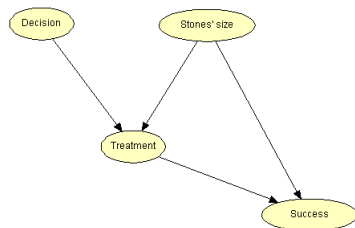
Intervention graph is obtained from the graph of considered causal Bayesian network by augmenting each node t corresponding to an intervention with an additional parent t' .

Intervention graph

We can decide the identifiability from the graph - we don't need to resort to probability distributions.

Definition

Intervention graph is obtained from the graph of considered causal Bayesian network by augmenting each node t corresponding to an intervention with an additional parent t' .

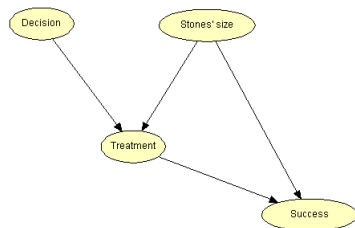


Intervention graph

We can decide the identifiability from the graph - we don't need to resort to probability distributions.

Definition

Intervention graph is obtained from the graph of considered causal Bayesian network by augmenting each node t corresponding to an intervention with an additional parent t' .



See file [kidney_stones_augmented.net](#) in Hugin.

Transformation of interventional to observational probabilities

To decide whether a set C identifies causal effect and to compute $P(X_R = x_r | X_t \leftarrow x_t^*)$ three inference rules (based on testing conditional independence in the intervention graph) can be used (for $A \subseteq C$):

Transformation of interventional to observational probabilities

To decide whether a set C identifies causal effect and to compute $P(X_R = x_r | X_t \leftarrow x_t^*)$ three inference rules (based on testing conditional independence in the intervention graph) can be used (for $A \subseteq C$):

① neutral observation

$$X_R \perp\!\!\!\perp X_t | X_A \implies P(X_R = x_r | X_A = x_A, X_t = x_t) = P(X_R = x_r | X_A = x_A)$$



Transformation of interventional to observational probabilities

To decide whether a set C identifies causal effect and to compute $P(X_R = x_r | X_t \leftarrow x_t^*)$ three inference rules (based on testing conditional independence in the intervention graph) can be used (for $A \subseteq C$):

2 neutral intervention

$$X_R \perp\!\!\!\perp X_{t'} | X_A \implies P(X_R = x_r | X_A = x_A, X_t \leftarrow x_t) \\ = P(X_R = x_r | X_A = x_A)$$

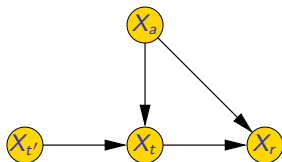


Transformation of interventional to observational probabilities

To decide whether a set C identifies causal effect and to compute $P(X_R = x_r | X_t \leftarrow x_t^*)$ three inference rules (based on testing conditional independence in the intervention graph) can be used (for $A \subseteq C$):

- 3 equivalence of observation and intervention

$$X_R \perp\!\!\!\perp X_{t'} | X_A, X_t \implies P(X_R = x_r | X_A = x_A, X_t \leftarrow x_t) = P(X_R = x_r | X_A = x_A, X_t = x_t)$$



Theorem

Assume $A \subseteq C$ such that A satisfies

- $X_A \perp\!\!\!\perp X_t$

Theorem

Assume $A \subseteq C$ such that A satisfies

- $X_A \perp\!\!\!\perp X_{t'}$
- $X_R \perp\!\!\!\perp X_{t'} | X_A, X_t$ (all back door trails are blocked by X_A).

Theorem

Assume $A \subseteq C$ such that A satisfies

- $X_A \perp\!\!\!\perp X_{t'}$
- $X_R \perp\!\!\!\perp X_{t'} | X_A, X_t$ (all back door trails are blocked by X_A).

Then

$$P(X_R = x_R | X_t \leftarrow x_t) = \sum_{x_A} P(X_R = x_R | X_A = x_A, X_t = x_t) \cdot P(X_A = x_A)$$

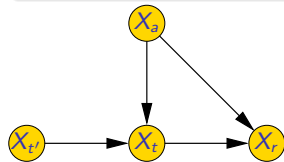
Theorem

Assume $A \subseteq C$ such that A satisfies

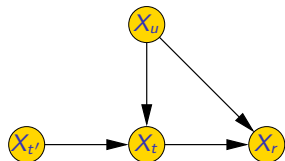
- $X_A \perp\!\!\!\perp X_{t'}$
- $X_R \perp\!\!\!\perp X_{t'} | X_A, X_t$ (all back door trails are blocked by X_A).

Then

$$P(X_R = x_R | X_t \leftarrow x_t) = \sum_{x_A} P(X_R = x_R | X_A = x_A, X_t = x_t) \cdot P(X_A = x_A)$$

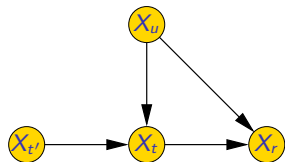


Randomization

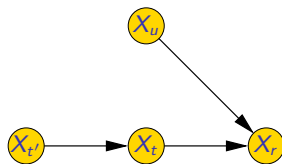


non-identifiable

Randomization

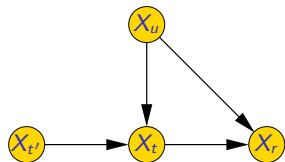


non-identifiable

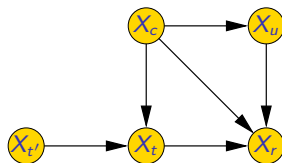


identifiable

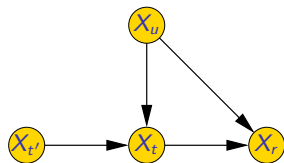
Randomization



non-identifiable

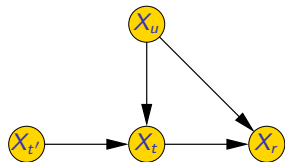


identifiable

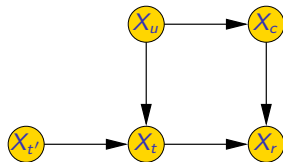


non-identifiable

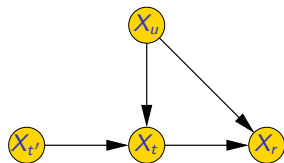
Sufficient covariate



non-identifiable

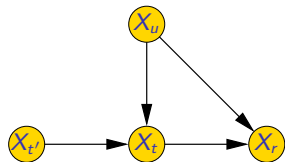


identifiable

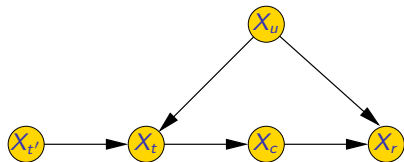


non-identifiable

Front door



non-identifiable



identifiable

- We have described a way how a causal interpretation can be assigned to Bayesian networks.

- We have described a way how a causal interpretation can be assigned to Bayesian networks.
- However, there are many other ways to do this (e.g. Shafer, 1996).

- We have described a way how a causal interpretation can be assigned to Bayesian networks.
- However, there are many other ways to do this (e.g. Shafer, 1996).
- Also, we need to check carefully that a Causal Bayesian network is a proper model for a given problem.

- Steffen Lauritzen. Causal Inference from Graphical Models. Research Report R-99-2021, Department of Mathematics, Aalborg University. Available at:

http://www.math.aau.dk/fileadmin/user_upload/www.math.aau.dk/Forskning/Rapportserien/R-99-2021.ps

- Steffen Lauritzen. Causal Inference from Graphical Models. Research Report R-99-2021, Department of Mathematics, Aalborg University. Available at:

http://www.math.aau.dk/fileadmin/user_upload/www.math.aau.dk/Forskning/Rapportserien/R-99-2021.ps

- Daphne Koller and Nir Friedman. Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009. Chapter 21.

- Steffen Lauritzen. Causal Inference from Graphical Models. Research Report R-99-2021, Department of Mathematics, Aalborg University. Available at:
http://www.math.aau.dk/fileadmin/user_upload/www.math.aau.dk/Forskning/Rapportserien/R-99-2021.ps
- Daphne Koller and Nir Friedman. Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009. Chapter 21.
- Judea Pearl. Causality. Cambridge University Press, 2000.