Probabilistic Graphical Models
The Big Objective(s)

In a wide variety of application fields two main problems need to be addressed over and over:

1. **How can (expert) knowledge of complex domains be efficiently represented?**

2. **How can inferences be carried out within these representations?**

3. **How can such representations be (automatically) extracted from collected data?**

We will deal with all three questions during the lecture.
Example 1: Planning in car manufacturing

Available information

“Engine type $e_1$ can only be combined with transmission $t_2$ or $t_5$.”

“Transmission $t_5$ requires crankshaft $c_2$.”

“Convertibles have the same set of radio options as SUVs.”

Possible questions/inferences:

“Can a station wagon with engine $e_4$ be equipped with tire set $y_6$?”

“Supplier $S_8$ failed to deliver on time. What production line has to be modified and how?”

“Are there any peculiarities within the set of cars that suffered an aircondition failure?”
Example 2: Medical reasoning

Available information:

“Malaria is much less likely than flu.”

“Flu causes cough and fever.”

“Nausea can indicate malaria as well as flu.”

“Nausea never indicated pneumonia before.”

Possible questions/inferences

“The patient has fever. How likely is he to have malaria?”

“How much more likely does flu become if we can exclude malaria?”
Both scenarios share some severe problems:

**Large Data Space**
It is intractable to store all value combinations, i.e. all car part combinations or inter-disease dependencies.

(Example: VW Bora has $10^{200}$ theoretical value combinations*)

**Sparse Data Space**
Even if we could handle such a space, it would be extremely sparse, i.e. it would be impossible to find good estimates for all the combinations.

(Example: with 100 diseases and 200 symptoms, there would be about $10^{62}$ different scenarios for which we had to estimate the probability.*)

* The number of particles in the observable universe is estimated to be between $10^{78}$ and $10^{85}$. 
**Given:** A large (high-dimensional) distribution $\delta$ representing the domain knowledge.

**Desired:** A set of smaller (lower-dimensional) distributions $\{\delta_1, \ldots, \delta_s\}$ (maybe overlapping) from which the original $\delta$ *could* be reconstructed with no (or as few as possible) errors.

With such a decomposition we can draw any conclusions from $\{\delta_1, \ldots, \delta_s\}$ that could be inferred from $\delta$ — without, however, actually reconstructing it.
Let us consider a car configuration is described by three attributes:

- Engine $E$, $\text{dom}(E) = \{ e_1, e_2, e_3 \}$
- Breaks $B$, $\text{dom}(B) = \{ b_1, b_2, b_3 \}$
- Tires $T$, $\text{dom}(T) = \{ t_1, t_2, t_3, t_4 \}$

Therefore the set of all (theoretically) possible car configurations is:

$$\Omega = \text{dom}(E) \times \text{dom}(B) \times \text{dom}(T)$$

Since not all combinations are technically possible (or wanted by marketing) a set of rules is used to cancel out invalid combinations.
Example: Car Manufacturing

Possible car configurations

Every cube designates a valid value combination.

10 car configurations in our model.

Different colors are intended to distinguish the cubes only.
Is it possible to reconstruct $\delta$ from the $\delta_i$?
Example: Reconstruction of $\delta$ with $\delta_{BE}$ and $\delta_{ET}$
Example: Reconstruction of $\delta$ with $\delta_{BE}$ and $\delta_{ET}$
Example: Reconstruction of $\delta$ with $\delta_{BE}$ and $\delta_{ET}$
Is it possible to exploit local constraints (wherever they may come from — both structural and expert knowledge-based) in a way that allows for a decomposition of the large (intractable) distribution $P(X_1, \ldots, X_n)$ into several sub-structures $\{C_1, \ldots, C_m\}$ such that:

The collective size of those sub-structures is much smaller than that of the original distribution $P$.

The original distribution $P$ is recomposable (with no or at least as few as possible errors) from these sub-structures in the following way:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{m} \Psi_i(c_i)$$

where $c_i$ is an instantiation of $C_i$ and $\Psi_i(c_i) \in \mathbb{R}^+$ a factor potential.
The Big Picture / Lecture Roadmap

Graphical Model

- Bayesian Network
- Markov Network

Major part of lecture
Last part of lecture

Evidence Propagation

Data

Learning/Induction

Domain Knowledge

Knowledge/Empirical Data

Expert
(Semi-)Graphoid Axioms

**Definition:** Let \( V \) be a set of (mathematical) objects and \((\cdot \perp \cdot \mid \cdot)\) a three-place relation of subsets of \( V \). Furthermore, let \( W, X, Y, \) and \( Z \) be four disjoint subsets of \( V \). The four statements

symmetry: \((X \perp Y \mid Z) \Rightarrow (Y \perp X \mid Z)\)

decomposition: \((W \cup X \perp Y \mid Z) \Rightarrow (W \perp Y \mid Z) \land (X \perp Y \mid Z)\)

weak union: \((W \cup X \perp Y \mid Z) \Rightarrow (X \perp Y \mid Z \cup W)\)

contraction: \((X \perp Y \mid Z \cup W) \land (W \perp Y \mid Z) \Rightarrow (W \cup X \perp Y \mid Z)\)

are called the **semi-graphoid axioms**. A three-place relation \((\cdot \perp \cdot \mid \cdot)\) that satisfies the semi-graphoid axioms for all \( W, X, Y, \) and \( Z \) is called a **semi-graphoid**.

The above four statements together with

intersection: \((W \perp Y \mid Z \cup X) \land (X \perp Y \mid Z \cup W) \Rightarrow (W \cup X \perp Y \mid Z)\)

are called the **graphoid axioms**. A three-place relation \((\cdot \perp \cdot \mid \cdot)\) that satisfies the graphoid axioms for all \( W, X, Y, \) and \( Z \) is called a **graphoid**.
Illustration of the (Semi-)Graphoid Axioms

- **decomposition:**
  \[
  \begin{array}{c}
  W \\
  X \\
  Z \\
  Y
  \end{array} \quad \Rightarrow \quad \begin{array}{c}
  W \\
  Z \\
  Y
  \end{array} \land \begin{array}{c}
  X \\
  Z \\
  Y
  \end{array}
  \]

- **weak union:**
  \[
  \begin{array}{c}
  W \\
  X \\
  Z \\
  Y
  \end{array} \quad \Rightarrow \quad \begin{array}{c}
  W \\
  Z \\
  Y
  \end{array}
  \]

- **contraction:**
  \[
  \begin{array}{c}
  W \\
  X \\
  Z \\
  Y
  \end{array} \land \begin{array}{c}
  W \\
  Z \\
  Y
  \end{array} \quad \Rightarrow \quad \begin{array}{c}
  W \\
  X \\
  Z \\
  Y
  \end{array}
  \]

- **intersection:**
  \[
  \begin{array}{c}
  W \\
  X \\
  Z \\
  Y
  \end{array} \land \begin{array}{c}
  W \\
  Z \\
  Y
  \end{array} \quad \Rightarrow \quad \begin{array}{c}
  W \\
  X \\
  Z \\
  Y
  \end{array}
  \]

Similar to the properties of separation in graphs. Idea: **Represent conditional independence by separation in graphs.**
Separation in Directed Acyclic Graphs

Example Graph:

Valid Separations:
\[\langle \{A_1\} \mid \{A_3\} \mid \{A_4\} \rangle\]
\[\langle \{A_3\} \mid \{A_4, A_6\} \mid \{A_7\} \rangle\]
\[\langle \{A_8\} \mid \{A_7\} \mid \{A_9\} \rangle\]
\[\langle \{A_1\} \mid \emptyset \mid \{A_2\} \rangle\]

Invalid Separations:
\[\langle \{A_1\} \mid \{A_4\} \mid \{A_2\} \rangle\]
\[\langle \{A_4\} \mid \{A_3, A_7\} \mid \{A_6\} \rangle\]
\[\langle \{A_1\} \mid \{A_6\} \mid \{A_7\} \rangle\]
\[\langle \{A_1\} \mid \{A_4, A_9\} \mid \{A_5\} \rangle\]
Definition: Let \(( \cdot \perp \delta \cdot \mid \cdot)\) be a three-place relation representing the set of conditional independence statements that hold in a given distribution \(\delta\) over a set \(U\) of attributes. An undirected graph \(G = (U, E)\) over \(U\) is called a conditional dependence graph or a dependence map w.r.t. \(\delta\), iff for all disjoint subsets \(X, Y, Z \subseteq U\) of attributes

\[
X \perp \delta Y \mid Z \Rightarrow \langle X \mid Z \mid Y \rangle_G,
\]

i.e., if \(G\) captures by \(u\)-separation all (conditional) independences that hold in \(\delta\) and thus represents only valid (conditional) dependences. Similarly, \(G\) is called a conditional independence graph or an independence map w.r.t. \(\delta\), iff for all disjoint subsets \(X, Y, Z \subseteq U\) of attributes

\[
\langle X \mid Z \mid Y \rangle_G \Rightarrow X \perp \delta Y \mid Z,
\]

i.e., if \(G\) captures by \(u\)-separation only (conditional) independences that are valid in \(\delta\). \(G\) is said to be a perfect map of the conditional (in)dependences in \(\delta\), if it is both a dependence map and an independence map.
Perfect directed map, no perfect undirected map:

\[
\begin{align*}
\text{p}_{ABC} & \quad A = a_1 & B = b_1 & B = b_2 & \quad A = a_2 & B = b_1 & B = b_2 \\
C = c_1 & 4/24 & 3/24 & \quad & 3/24 & 2/24 \\
C = c_2 & 2/24 & 3/24 & \quad & 3/24 & 4/24 \\
\end{align*}
\]

Perfect undirected map, no perfect directed map:

\[
\begin{align*}
\text{p}_{ABCD} & \quad A = a_1 & B = b_1 & B = b_2 & \quad A = a_2 & B = b_1 & B = b_2 \\
C = c_1 & D = d_1 & 1/47 & \quad & 1/47 & \quad & 1/47 & \quad & 2/47 \\
& D = d_2 & 1/47 & \quad & 1/47 & \quad & 2/47 & \quad & 4/47 \\
C = c_2 & D = d_1 & 1/47 & \quad & 2/47 & \quad & 1/47 & \quad & 4/47 \\
& D = d_2 & 2/47 & \quad & 4/47 & \quad & 4/47 & \quad & 16/47 \\
\end{align*}
\]
Markov Properties of Undirected Graphs

**Definition:** An undirected graph $G = (U, E)$ over a set $U$ of attributes is said to have (w.r.t. a distribution $\delta$) the

**pairwise Markov property,**
iff in $\delta$ any pair of attributes which are nonadjacent in the graph are conditionally independent given all remaining attributes, i.e., iff

$$\forall A, B \in U, A \neq B : \ (A, B) \notin E \ \Rightarrow \ A \indep \delta B \mid U - \{A, B\},$$

**local Markov property,**
iff in $\delta$ any attribute is conditionally independent of all remaining attributes given its neighbors, i.e., iff

$$\forall A \in U : \ A \indep \delta U - \text{closure}(A) \mid \text{boundary}(A),$$

**global Markov property,**
iff in $\delta$ any two sets of attributes which are $u$-separated by a third are conditionally independent given the attributes in the third set, i.e., iff

$$\forall X, Y, Z \subseteq U : \ \langle X \mid Z \mid Y \rangle_G \ \Rightarrow \ X \indep \delta Y \mid Z.$$
Markov Properties of Directed Acyclic Graphs

**Definition:** A directed acyclic graph \( \vec{G} = (U, \vec{E}) \) over a set \( U \) of attributes is said to have (w.r.t. a distribution \( \delta \)) the

**pairwise Markov property,**
iff in \( \delta \) any attribute is conditionally independent of any non-descendant not among its parents given all remaining non-descendants, i.e., iff

\[
\forall A, B \in U : B \in \text{non-descs}(A) - \text{parents}(A) \Rightarrow A \perp \perp B \mid \text{non-descs}(A) - \{B\},
\]

**local Markov property,**
iff in \( \delta \) any attribute is conditionally independent of all remaining non-descendants given its parents, i.e., iff

\[
\forall A \in U : A \perp \perp \text{non-descs}(A) - \text{parents}(A) \mid \text{parents}(A),
\]

**global Markov property,**
iff in \( \delta \) any two sets of attributes which are \( d \)-separated by a third are conditionally independent given the attributes in the third set, i.e., iff

\[
\forall X, Y, Z \subseteq U : \langle X \mid Z \mid Y \rangle_{\vec{G}} \Rightarrow X \perp \perp Y \mid Z.
\]
Theorem: If a three-place relation \((\cdot \perp_\delta \cdot \mid \cdot)\) representing the set of conditional independence statements that hold in a given joint distribution \(\delta\) over a set \(U\) of attributes satisfies the graphoid axioms, then the pairwise, the local, and the global Markov property of an undirected graph \(G = (U, E)\) over \(U\) are equivalent.

Theorem: If a three-place relation \((\cdot \perp_\delta \cdot \mid \cdot)\) representing the set of conditional independence statements that hold in a given joint distribution \(\delta\) over a set \(U\) of attributes satisfies the semi-graphoid axioms, then the local and the global Markov property of a directed acyclic graph \(\vec{G} = (U, \vec{E})\) over \(U\) are equivalent.

If \((\cdot \perp_\delta \cdot \mid \cdot)\) satisfies the graphoid axioms, then the pairwise, the local, and the global Markov property are equivalent.
**Definition:** A probability distribution \( p_V \) over a set \( V \) of variables is called **decomposable** or **factorizable w.r.t. an undirected graph** \( G = (V, E) \) iff it can be written as a product of nonnegative functions on the maximal cliques of \( G \).

That is, let \( \mathcal{M} \) be a family of subsets of variables, such that the subgraphs of \( G \) induced by the sets \( M \in \mathcal{M} \) are the maximal cliques of \( G \). Then there exist functions \( \phi_M : \mathcal{E}_M \to \mathbb{R}^+ \), \( M \in \mathcal{M} \), \( \forall a_1 \in \text{dom}(A_1) : \ldots \forall a_n \in \text{dom}(A_n) : \)

\[
p_V \left( \bigwedge_{A_i \in V} A_i = a_i \right) = \prod_{M \in \mathcal{M}} \phi_M \left( \bigwedge_{A_i \in M} A_i = a_i \right).
\]

**Example:**

\[
p_V(A_1 = a_1, \ldots, A_6 = a_6)
= \phi_{A_1A_2A_3}(A_1 = a_1, A_2 = a_2, A_3 = a_3)
\cdot \phi_{A_3A_5A_6}(A_3 = a_3, A_5 = a_5, A_6 = a_6)
\cdot \phi_{A_2A_4}(A_2 = a_2, A_4 = a_4)
\cdot \phi_{A_4A_6}(A_4 = a_4, A_6 = a_6).
\]
**Definition:** A probability distribution $p_U$ over a set $U$ of attributes is called **decomposable** or **factorizable w.r.t. a directed acyclic graph** $\tilde{G} = (U, \tilde{E})$ over $U$, iff it can be written as a product of the conditional probabilities of the attributes given their parents in $\tilde{G}$, i.e., iff

$$\forall a_1 \in \text{dom}(A_1) : \ldots \forall a_n \in \text{dom}(A_n) :$$

$$p_U \left( \bigwedge_{A_i \in U} A_i = a_i \right) = \prod_{A_i \in U} \left. P \left( A_i = a_i \right| \bigwedge_{A_j \in \text{parents}_{\tilde{G}}(A_i)} A_j = a_j \right).$$

**Example:**

$$P(A_1 = a_1, \ldots, A_7 = a_7)$$

$$= P(A_1 = a_1) \cdot P(A_2 = a_2 \mid A_1 = a_1) \cdot P(A_3 = a_3)$$

$$\cdot P(A_4 = a_4 \mid A_1 = a_1, A_2 = a_2)$$

$$\cdot P(A_5 = a_5 \mid A_2 = a_2, A_3 = a_3)$$

$$\cdot P(A_6 = a_6 \mid A_4 = a_4, A_5 = a_5)$$

$$\cdot P(A_7 = a_7 \mid A_5 = a_5).$$
Core Theorem of Graphical Models:
Let $p_V$ be a strictly positive probability distribution on a set $V$ of (discrete) variables. A directed or undirected graph $G = (V, E)$ is a conditional independence graph w.r.t. $p_V$ if and only if $p_V$ is factorizable w.r.t. $G$.

Definition: A Markov network is an undirected conditional independence graph of a probability distribution $p_V$ together with the family of positive functions $\phi_M$ of the factorization induced by the graph.

Definition: A Bayesian network is a directed conditional independence graph of a probability distribution $p_U$ together with the family of conditional probabilities of the factorization induced by the graph.

Sometimes the conditional independence graph is required to be minimal, if it is to be used as the graph underlying a Markov or Bayesian network. For correct evidence propagation it is not required that the graph is minimal. Evidence propagation may just be less efficient than possible.
Bayes Networks
A *Bayes Network* $(V, E, P)$ consists of a set $V = \{X_1, \ldots, X_n\}$ of random variables and a set $E$ of directed edges between the variables.

Each variable has a finite set of mutual exclusive and collectively exhaustive states.

The variables in combination with the edges form a directed, acyclic graph.

Each variable with parent nodes $B_1, \ldots, B_m$ is assigned a table $P(A | B_1, \ldots, B_m)$.

Note, that the connections between the nodes not necessarily express a causal relationship.

For every belief network, the following equation holds:

$$P(V) = \prod_{v \in V: P(c(v)) > 0} P(v | c(v))$$

with $c(v)$ being the parent nodes of $v$. 
Probabilistic Dependency Networks

Probabilistic dependency networks are directed acyclic graphs (DAGs) where the nodes represent propositions or variables and the directed edges model a direct dependence between the connected nodes. The strength of dependence is defined by conditional probabilities.

In general (according chain rule):

\[ P(X_1, \ldots, X_6) = P(X_6 \mid X_5, \ldots, X_1) \cdot P(X_5 \mid X_4, \ldots, X_1) \cdot P(X_4 \mid X_3, X_2, X_1) \cdot P(X_3 \mid X_2, X_1) \cdot P(X_2 \mid X_1) \cdot P(X_1) \]
Probabilistic dependency networks are directed acyclic graphs (DAGs) where the nodes represent propositions or variables and the directed edges model a direct causal dependence between the connected nodes. The strength of dependence is defined by conditional probabilities.

According graph (independence structure):

\[
P(X_1, \ldots, X_6) = P(X_6 \mid X_5) \cdot P(X_5 \mid X_2, X_3) \cdot P(X_4 \mid X_2) \cdot P(X_3 \mid X_1) \cdot P(X_2 \mid X_1) \cdot P(X_1)
\]
Nomenclature for the next slides:

- \( X_1, \ldots, X_n \)  
  Variables  
  (properties, attributes, random variables, propositions)

- \( \Omega_1, \ldots, \Omega_n \)  
  respective finite domains  
  (also designated with \( \text{dom}(X_i) \))

\[
\Omega = \prod_{i=1}^{n} \Omega_i
\]

- \( \Omega_i = \{x_i^{(1)}, \ldots, x_i^{(n_i)}\} \)  
  Universe of Discourse (tuples that characterize objects described by \( X_1, \ldots, X_n \))

\[
n = 1, \ldots, n, \ n_i \in \mathbb{N}
\]
The product space \((\Omega, 2^\Omega, P)\) is unique iff \(P(\{(x_1, \ldots, x_n)\})\) is specified for all \(x_i \in \{x_i^{(1)}, \ldots, x_i^{(n_i)}\}\), \(i = 1, \ldots, n\).

When the distribution \(P(X_1, \ldots, X_n)\) is given in tabular form, then \(\prod_{i=1}^n |\Omega_i|\) entries are necessary.

For variables with \(|\Omega_i| \geq 2\) at least \(2^n\) entries.

The application of DAGs allows for the representation of existing (in)dependencies.
Constructing a DAG

\textbf{input} \; P(X_1, \ldots, X_n)  \\
\textbf{output} \; \text{a DAG } G

1. Set the nodes of $G$ to $\{X_1, \ldots, X_n\}$.
2. Choose a total ordering on the set of variables (e.g. $X_1 \prec X_2 \prec \cdots \prec X_n$)
3. For $X_i$ find the smallest (uniquely determinable) set $S_i \subseteq \{X_1, \ldots, X_n\}$ sucht that $P(X_i \mid S_i) = P(X_i \mid X_1 \ldots, X_{i-1})$.
4. Connect all nodes in $S_i$ with $X_i$ and store $P(X_i \mid S_i)$ as quantization of the dependencies for that node $X_i$ (given its parents).
5. return $G$
Example

Let $a_1, a_2, a_3$ be three blood groups and $b_1, b_2, b_3$ three indications of a blood group test.

Variables: $A$ (blood group) $B$ (indication)

Domains: $\Omega_A = \{a_1, a_2, a_3\} \quad \Omega_B = \{b_1, b_2, b_3\}$

It is conjectured that there is a causal relationship between the variables.

$$P(\{(a_i, b_j)\}) \quad | \quad b_1 \quad b_2 \quad b_3 \quad \sum$$

<table>
<thead>
<tr>
<th></th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$\sum$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>0.64</td>
<td>0.08</td>
<td>0.08</td>
<td>0.8</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0.01</td>
<td>0.08</td>
<td>0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>$a_3$</td>
<td>0.01</td>
<td>0.01</td>
<td>0.08</td>
<td>0.1</td>
</tr>
<tr>
<td>$\sum$</td>
<td>0.66</td>
<td>0.17</td>
<td>0.17</td>
<td>1</td>
</tr>
</tbody>
</table>

$$P(A, B) = P(B | A) \cdot P(A)$$

We are dealing with a belief network.
Example

**Expert Knowledge**

Metastatic cancer is a possible cause of brain cancer, and an explanation for elevated levels of calcium in the blood. Both phenomena together can explain that a patient falls into a coma. Severe headaches are possibly associated with a brain tumor.

**Special Case**

The patient has severe headaches.

**Question**

Will the patient go into a coma?
### Example

#### Choice of universe of discourse

<table>
<thead>
<tr>
<th>Variable</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ metastatic cancer</td>
<td>{a_1, a_2}</td>
</tr>
<tr>
<td>$B$ increased serum calcium</td>
<td>{b_1, b_2}</td>
</tr>
<tr>
<td>$C$ brain tumor</td>
<td>{c_1, c_2}</td>
</tr>
<tr>
<td>$D$ coma</td>
<td>{d_1, d_2}</td>
</tr>
<tr>
<td>$E$ headache</td>
<td>{e_1, e_2}</td>
</tr>
</tbody>
</table>

\(\Omega = \{a_1, a_2\} \times \cdots \times \{e_1, e_2\}\)

\(|\Omega| = 32\)

#### Analysis of dependencies

![Bayesian Network Diagram]
Example

Choice of probability parameters

\[ P(a, b, c, d, e) \overset{\text{abbr.}}{=} P(A = a, B = b, C = c, D = d, E = e) = P(e | c)P(d | b, c)P(c | a)P(b | a)P(a) \]

Shorthand notation

11 values to store instead of 31

Consult experts, textbooks, case studies, surveys, etc.

Calculation of conditional probabilities

Calculation of marginal probabilities
Knowledge acquisition (Where do the numbers come from?)
→ learning strategies

Computational complexities
→ exploit independencies

**Problem:**

When does the independency of $X$ and $Y$ given $Z$ hold in $(V, E, P)$?

How to determine a decomposition based on the graph structure?
Example

\[ A \xrightarrow{} C \xrightarrow{} B \]

Meal quality

\[ A \quad \text{quality of ingredients} \\
B \quad \text{cook’s skill} \\
C \quad \text{meal quality} \]

If \( C \) is not known, \( A \) and \( B \) are independent.

If \( C \) is known, then \( A \) and \( B \) become (conditionally) dependent given \( C \).

\( A \perp\!\!\!\!\!\!\!\!\!\!\!\perp B \mid C \)
**Converging Connection:** Marginal Independence

Decomposition according to graph:

\[ P(A, B, C) = P(C \mid A, B) \cdot P(A) \cdot P(B) \]

Embedded Independence:

\[ P(A, B, C) = \frac{P(A, B, C)}{P(A, B)} \cdot P(A) \cdot P(B) \quad \text{with} \quad P(A, B) \neq 0 \]

\[ P(A, B) = P(A) \cdot P(B) \]

\[ \Rightarrow A \indep B \mid \emptyset \]
Example (cont.)

If nothing is known about the restaurant success or meal quality or both, the cook’s skills and quality of the ingredients are unrelated, that is, independent.

However, if we observe that the restaurant has no success, we can infer that the meal quality might be bad.

If we further learn that the ingredients quality is high, we will conclude that the cook’s skills must be low, thus rendering both variables dependent.

\[ A \perp B \mid D \]
Diverging Connection

If \( C \) is unknown, knowledge about \( A \) is relevant for \( B \) and vice versa, i.e. \( A \) and \( B \) are marginally dependent.

However, if \( C \) is observed, \( A \) and \( B \) become conditionally independent given \( C \).

\( A \) influences \( B \) via \( C \). If \( C \) is known it in a way blocks the information from flowing from \( A \) to \( B \), thus rendering \( A \) and \( B \) (conditionally) independent.
**Diverging Connection:** Conditional Independence

Decomposition according to graph:

\[ P(A, B, C) = P(A | C) \cdot P(B | C) \cdot P(C) \]

Embedded Independence:

\[ P(A, B | C) = P(A | C) \cdot P(B | C) \]
\[ \Rightarrow A \perp\!\!\!\!\!\perp B | C \]

Alternative derivation:

\[ P(A, B, C) = P(A | C) \cdot P(B, C) \]
\[ P(A | B, C) = P(A | C) \]
\[ \Rightarrow A \perp\!\!\!\!\!\perp B | C \]
Serial Connection

![Diagram](dependencies.png)

Accidents

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>rain</td>
<td>accident risk</td>
<td>road conditions</td>
</tr>
</tbody>
</table>

Analog scenario to case 2

A influences C and C influences B. Thus, A influences B. If C is known, it blocks the path between A and B.
**Serial Connection**: Conditional Independence

Decomposition according to graph:

\[
P(A, B, C) = P(B \mid C) \cdot P(C \mid A) \cdot P(A)
\]

Embedded Independence:

\[
P(A, B, C) = P(B \mid C) \cdot P(C, A)
\]

\[
P(B \mid C, A) = P(B \mid C)
\]

\[
\Rightarrow A \perp\!\!\!\!\!\!\perp B \mid C
\]
Trivial Cases:

Marginal Independence:

\[ P(A, B) = P(A) \cdot P(B) \]

\[ A \quad B \]

Marginal Dependence:

\[ P(A, B) = P(B \mid A) \cdot P(A) \]

\[ A \rightarrow B \]
**Question:** Are $X_2$ and $X_3$ independent given $X_1$?

![Bayesian Network Diagram]
Repetition: d-Separation

Let $G = (V, E)$ a DAG and $X, Y, Z \in V$ three nodes.

a) A set $S \subseteq V\{X,Y\}$ \textit{d-separates} $X$ and $Y$, if $S$ blocks all paths between $X$ and $Y$. (paths may also route in opposite edge direction)

b) A path $\pi$ is d-separated by $S$ if at least one pair of consecutive edges along $\pi$ is blocked. There are the following blocking conditions:

1. $X \leftarrow Y \rightarrow Z$ \hspace{1cm} \text{tail-to-tail}
2. $X \leftarrow Y \leftarrow Z$ \hspace{1cm} \text{head-to-tail}
3. $X \rightarrow Y \leftarrow Z$ \hspace{1cm} \text{head-to-head}

c) Two edges that meet tail-to-tail or head-to-tail in node $Y$ are blocked if $Y \in S$.

d) Two edges meeting head-to-head in $Y$ are blocked if neither $Y$ nor its successors are in $S$. 

Relation to Conditional independence

If \( S \subseteq V \setminus \{X, Y\} \) d-separates \( X \) and \( Y \) in a Belief network \((V, E, P)\) then \( X \) and \( Y \) are conditionally independent given \( S \):

\[
P(X, Y \mid S) = P(X \mid S) \cdot P(Y \mid S)
\]

Application to the previous example:

Paths:

\[
\begin{align*}
\pi_1 &= \langle X_2 - X_1 - X_3 \rangle, \\
\pi_2 &= \langle X_2 - X_5 - X_3 \rangle, \\
\pi_3 &= \langle X_2 - X_4 - X_1 - X_3 \rangle, \\
S &= \{X_1\}
\end{align*}
\]

\( \pi_1 \) \( X_2 \leftarrow X_1 \rightarrow X_3 \) tail-to-tail
\( X_1 \in S \Rightarrow \pi_1 \) is blocked by \( S \)

\( \pi_2 \) \( X_2 \rightarrow X_5 \leftarrow X_3 \) head-to-head
\( X_5, X_6 \notin S \Rightarrow \pi_2 \) is blocked by \( S \)

\( \pi_3 \) \( X_4 \leftarrow X_1 \rightarrow X_3 \) tail-to-tail
\( X_2 \rightarrow X_4 \leftarrow X_1 \) head-to-head
both connections are blocked \( \Rightarrow \pi_3 \) is blocked
Example (cont.)

Answer: $X_2$ and $X_3$ are d-separated via $\{X_1\}$. Therefore $X_2$ and $X_3$ become conditionally independent given $X_1$.

$S = \{X_1, X_4\} \Rightarrow X_2$ and $X_3$ are d-separated by $S$

$S = \{X_1, X_6\} \Rightarrow X_2$ and $X_3$ are not d-separated by $S$
Question: Is it possible to use a formal scheme to infer new conditional independence (CI) statements from a set of initial CIs?

Repetition

Let \((\Omega, \mathcal{E}, P)\) be a probability space and \(W, X, Y, Z\) disjoint subsets of variables. If \(X\) and \(Y\) are conditionally independent given \(Z\) we write:

\[
X \perp \!\!\!\!\! \perp P Y \mid Z
\]

Often, the following (equivalent) notation is used:

\[
I_P(X \mid Z \mid Y) \quad \text{or} \quad I_P(X, Y \mid Z)
\]

If the underlying space is known the index \(P\) is omitted.
Let $(\Omega, \mathcal{E}, P)$ be a probability space and $W, X, Y$ and $Z$ four disjoint subsets of random variables (over $\Omega$). Then the propositions

a) Symmetry: $(X \perp\!\!\!\!\perp Y \mid Z) \Rightarrow (Y \perp\!\!\!\!\perp X \mid Z)$

b) Decomposition: $(W \cup X \perp\!\!\!\!\perp Y \mid Z) \Rightarrow (W \perp\!\!\!\!\perp Y \mid Z) \land (X \perp\!\!\!\!\perp Y \mid Z)$

c) Weak Union: $(W \cup X \perp\!\!\!\!\perp Y \mid Z) \Rightarrow (X \perp\!\!\!\!\perp Y \mid Z \cup W)$

d) Contraction: $(X \perp\!\!\!\!\perp Y \mid Z \cup W) \land (W \perp\!\!\!\!\perp Y \mid Z) \Rightarrow (W \cup X \perp\!\!\!\!\perp Y \mid Z)$

are called the **Semi-Graphoid Axioms**. The above propositions and

e) Intersection: $(W \perp\!\!\!\!\perp Y \mid Z \cup X) \land (X \perp\!\!\!\!\perp Y \mid Z \cup W) \Rightarrow (W \cup X \perp\!\!\!\!\perp Y \mid Z)$

are called the **Graphoid Axioms**.
Example

Proposition: \( B \perp \perp C \mid A \)

Proof:

\[
D \perp \perp A, C \mid \emptyset \quad \land \quad B \perp \perp C \mid A, D
\]

w. union \( \implies \) \( D \perp \perp C \mid A \quad \land \quad B \perp \perp C \mid A, D \)

symm. \( \iff \) \( C \perp \perp D \mid A \quad \land \quad C \perp \perp B \mid A, D \)

contr. \( \implies \) \( C \perp \perp B, D \mid A \)

decomp. \( \implies \) \( C \perp \perp B \mid A \)

symm. \( \iff \) \( B \perp \perp C \mid A \)