Conditional Independence: Representation and Learning

INTELLIGENT SYSTEMS FOR PATTERN RECOGNITION (ISPR)

RICCARDO MASSIDDA, DAVIDE BACCIU – DIPARTIMENTO DI INFORMATICA - UNIVERSITA’ DI PISA

RICCARDO.MASSIDDA@DI.UNIPI.IT DAVIDE.BACCIU@UNIPI.IT

Probabilistic and Causal Learning

- Bayesian Networks (Tuesday 4th, **today**!)
  - Compact representation of joint probabilities
  - Plate Notation
  - Local Markov Property
  - Ancestral Sampling
- d-separation, Markov blankets (Thursday 6th)
- Graphical Causal Models (Tuesday 11th)
- Structure Learning and Causal Discovery (Wednesday 12th)

Representing Joint Distributions

- The main goal of **probabilistic modeling** is to define models able to represent the **joint distribution** of a set of variables.
- Probabilistic models enable
  - **Sampling** new instances
  - Inferencing values of **hidden** variables
  - Estimating the **likelihood** of a configuration
  - ...

Representing Joint Distributions

- Assume N discrete random variables with k distinct values.
- How many parameters in the **joint probability distribution**?

Y_1	Y_2	Y_3	$P(Y_1, Y_2, Y_3)$
0	0	0	0.03
0	0	1	0.12
0	1	0	0.31
⋮	⋮	⋮	⋮
1	1	1	0.04

$k^N - 1$

UNIVERSITÀ DI PISA

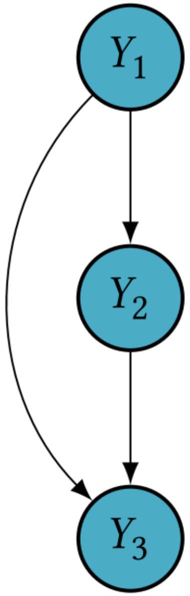
Representing Joint Distributions

- What if we compute the probability **one variable** at the time?
- We can exploit the **chain rule** to decompose the joint.

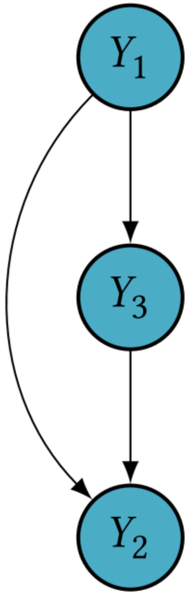
$$
\begin{aligned}
P(Y_1, Y_2, Y_3) &= P(Y_1)P(Y_2 \mid Y_1)P(Y_3 \mid Y_1, Y_2) \\
&= P(Y_2)P(Y_1 \mid Y_2)P(Y_3 \mid Y_1, Y_2) \\
&= \ldots \\
&= P(Y_3)P(Y_2 \mid Y_3)P(Y_1 \mid Y_2, Y_3).
\end{aligned}
$$

Representing Joint Distributions

- The **order** of the variables can be represented by **directed graphs**.

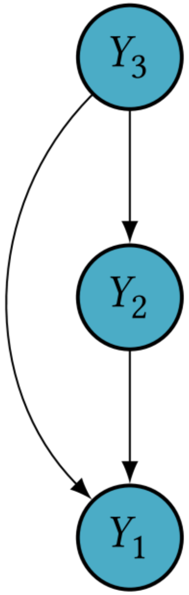


$P(Y_1)P(Y_2 \mid Y_1)P(Y_3 \mid Y_1, Y_2)$



$P(Y_1)P(Y_3 \mid Y_1)P(Y_2 \mid Y_1, Y_3)$

...



$P(Y_3)P(Y_2 \mid Y_3)P(Y_1 \mid Y_2, Y_3)$

Representing Joint Distributions

- Decomposing the joint with the **chain rule** reduces the **number of parameters**?
- No! 🥹

$$P(Y_1, Y_2, Y_3) = \underbrace{P(Y_1)}_{1}\underbrace{P(Y_2 \mid Y_1)}_{2}\underbrace{P(Y_3 \mid Y_1, Y_2)}_{4}$$

$$\sum_{i=0}^{N-1}(k-1)k^i = k^N - 1$$

UNIVERSITÀ DI PISA

Marginal and Conditional Independence

- Two random variables X and Y are **independent** if knowledge about X does not change the uncertainty about Y and vice versa

$$I(X,Y) \iff X \perp Y \iff P(X,Y) = P(X \mid Y)P(Y)$$
$$= P(Y \mid X)P(X) = P(X)P(Y).$$

Representing Joint Distributions

- When variables are **independent**, we only need Nk parameters.

$$
\begin{aligned}
P(Y_1, Y_2, Y_3) &= P(Y_1)P(Y_2 \mid Y_1)P(Y_3 \mid Y_1, Y_2) \\
&= \underbrace{P(Y_1)}_{\mathbf{1}}\underbrace{P(Y_2)}_{\mathbf{1}}\underbrace{P(Y_3)}_{\mathbf{1}}
\end{aligned}
$$

Università di Pisa

Marginal and Conditional Independence

- Two random variables X and Y are **conditionally independent** given Z if knowledge about X does not change the uncertainty about Y and vice versa on the conditional distribution

$$
\begin{aligned}
I(X,Y \mid Z) \iff X \perp Y \mid Z \iff P(X,Y \mid Z) &= P(X \mid Y \mid Z)P(Y \mid Z) \\
&= P(Y \mid X \mid Z)P(X \mid Z) \\
&= P(X \mid Z)P(Y \mid Z).
\end{aligned}
$$

Representing Joint Distributions

- Conditional independences reduce the **number of parameters**
- Yes! 🥳

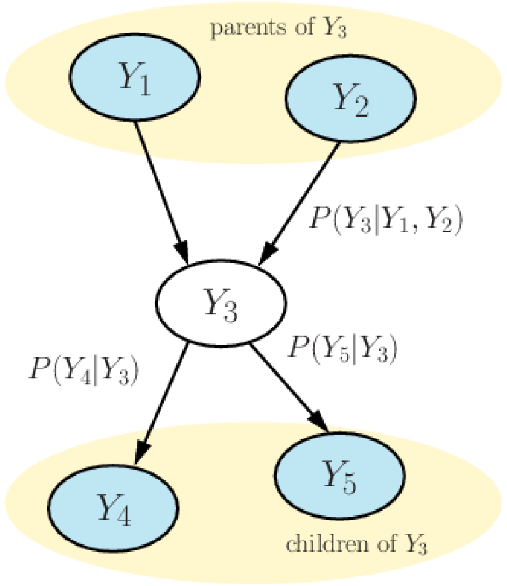
$$Y_1 \perp Y_3 \mid Y_2$$

$$\Longrightarrow \quad P(Y_1, Y_2, Y_3) = P(Y_1)P(Y_2 \mid Y_1)P(Y_3 \mid Y_1, Y_2)$$

$$= \underbrace{P(Y_1)}_{1}\underbrace{P(Y_2 \mid Y_1)}_{2}\underbrace{P(Y_3 \mid Y_2)}_{2}$$

UNIVERSITÀ DI PISA

Bayesian Network



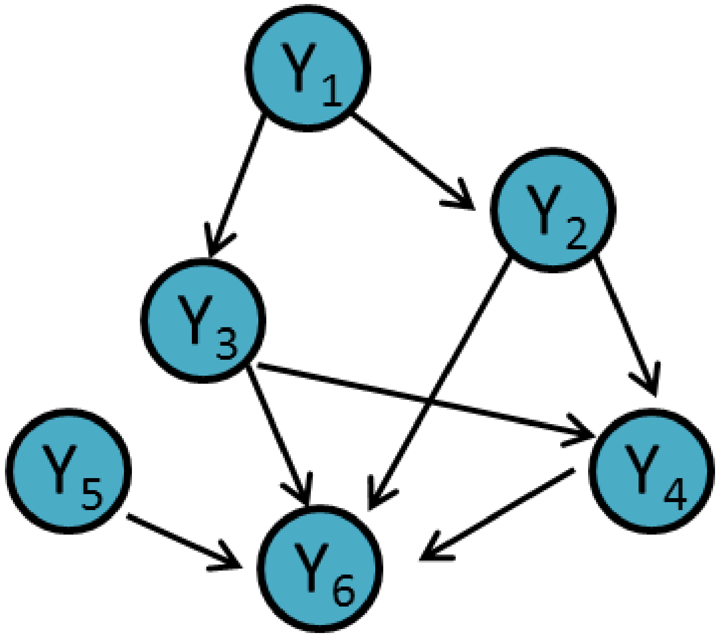
- Directed Acyclic Graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
- Nodes $v \in \mathcal{V}$ represent random variables
  - Shaded $\Rightarrow$ observed
  - Empty $\Rightarrow$ un-observed
- Edges $e \in \mathcal{E}$ describe the conditional independence relationships

Conditional Probability Tables (CPT) local to each node describe the probability distribution given its parents

$$P(Y_1, \dots, Y_N) = \prod_{i=1}^{N} P(Y_i \,|pa(Y_i))$$

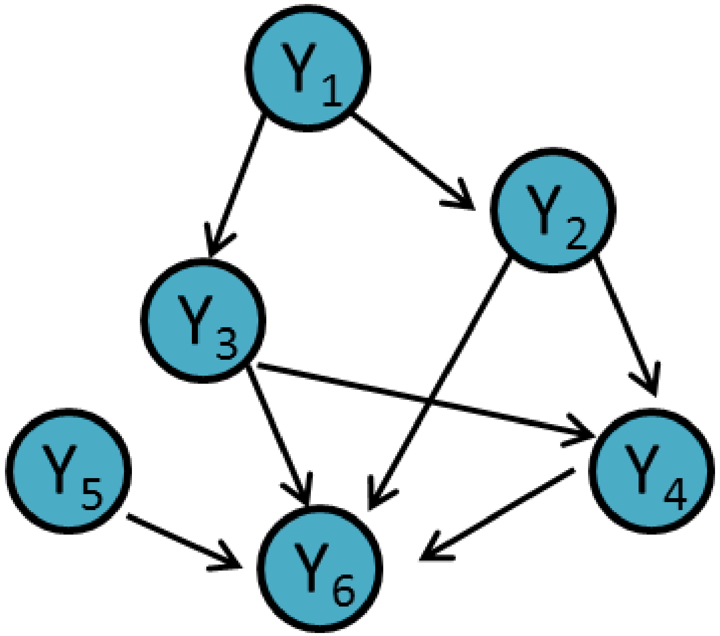
UNIVERSITÀ DI PISA

Bayesian Networks



- Let L be the **maximum number of ingoing edges** in a Bayes Net.
- Then, the number of parameters is **at most** $N \cdot (k\text{-}1)^L$
- $\Rightarrow$ The **sparser** the network, the less "complex" the parameters.

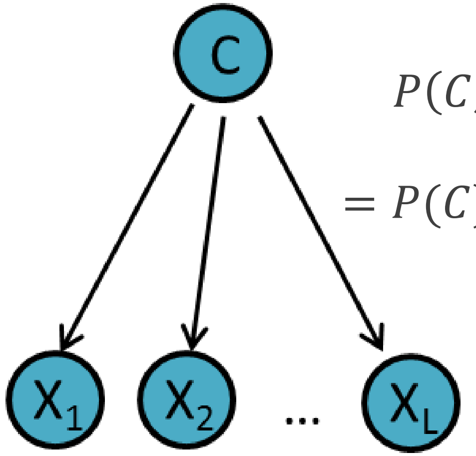
Bayesian Networks



- Are these relations **causal**?
- In general **no**, a Bayesian Network represent **statistical dependence** relations.
- However, they **might** coincide with causal dependence under further **assumptions**.

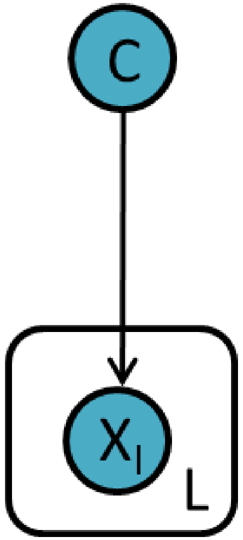
Compact Representation of Bayes Nets

If the same **dependencies are replicated** over different variables, we can compactly represent it by **plate notation**.

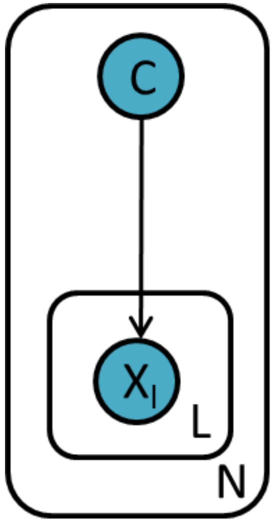


$$P(C, Y_1, \dots, Y_N) = P(C) \prod_{i=1}^{N} P(Y_i|C)$$

The Naive Bayes Classifier

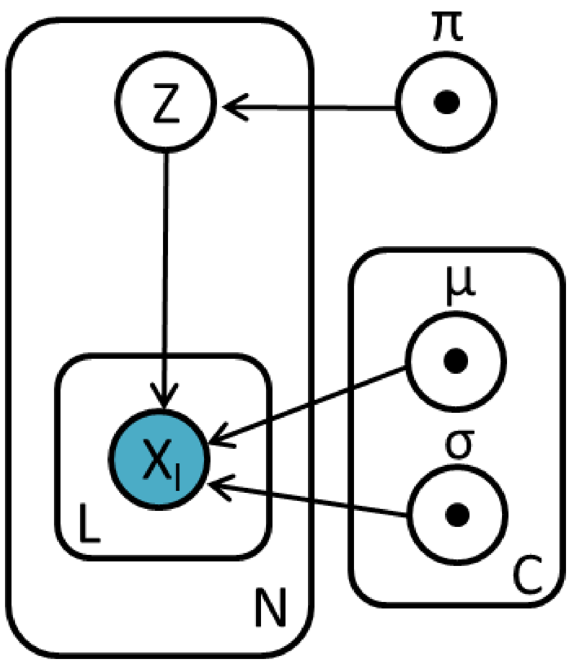


Replication for L attributes



Replication for N data samples

Full Plate Notation



Gaussian Mixture Model

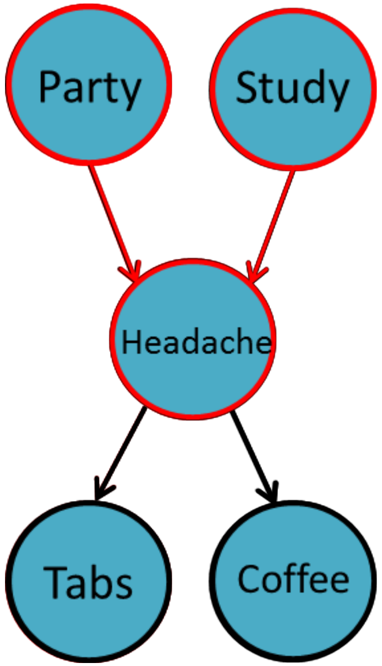
- Boxes denote replication for a number of times denoted by the letter in the corner
- Shaded nodes are observed variables
- Empty nodes denote un-observed latent variables
- Black seeds (optional) identify model parameters
  - π → multinomial prior distribution
  - μ → means of the C Gaussians
  - σ → std of the C Gaussians

Local Markov Property

Definition (Local Markov property)

Each node / random variable is conditionally independent of all its non-descendants given a joint state of its parents

$$Y_v \perp Y_{V \setminus \mathrm{ch}(v)} | Y_{pa(v)} \text{ for all } v \in V$$



$Party$ and $Study$ are marginally independent

- $Party \perp Study$

However, local Markov property does not support

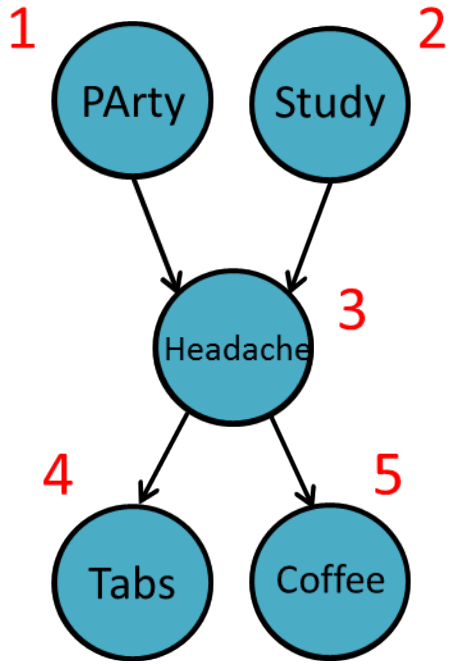
- $Party \perp Study \mid Headache$
- $Tabs \perp Party$

But $Party$ and $Tabs$ are independent given $Headache$

Joint Probability Factorization

An application of Chain rule and Local Markov Property

1. Pick a topological ordering of nodes
2. Apply chain rule following the order
3. Use the conditional independence assumptions



$$P(PA,S,H,T,C) = P(PA)\cdot P(S|PA)\cdot P(H|S,PA)\cdot P(T|H,S,PA)\cdot P(C|T,H,S,PA)$$
$$= P(PA)\cdot P(S)\cdot P(H|S,PA)\cdot P(T|H)\cdot P(C|H)$$

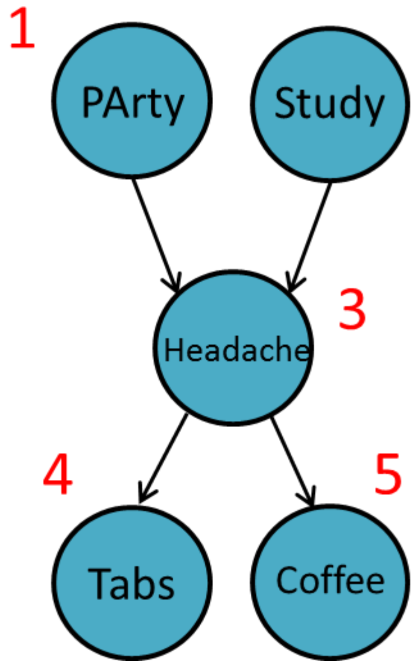
(Ancestral) Sampling of a BN

A BN describes a generative process for observations

1. Pick a topological ordering of nodes
2. Generate data by sampling from the local conditional probabilities following this order

Generate i-th sample for each variable PA, S, H, T, C

1. $pa_i \sim P(PA)$
2. $s_i \sim P(S)$
3. $h_i \sim P(H|S = s_i, PA = pa_i)$
4. $t_i \sim P(T|H = h_i)$
5. $c_i \sim P(C|H = h_i)$

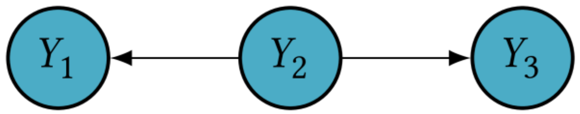


UNIVERSITÀ DI PISA

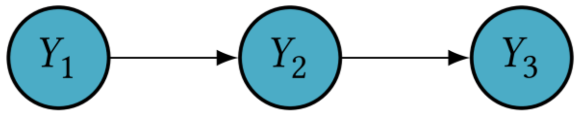
Fundamental BN structures

There exist **three fundamental substructures** that determine the conditional independence relationships in a Bayesian Network.

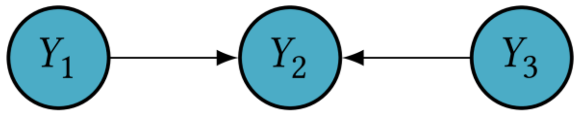
- **Tail-to-Tail** (Fork, “Common Cause”)



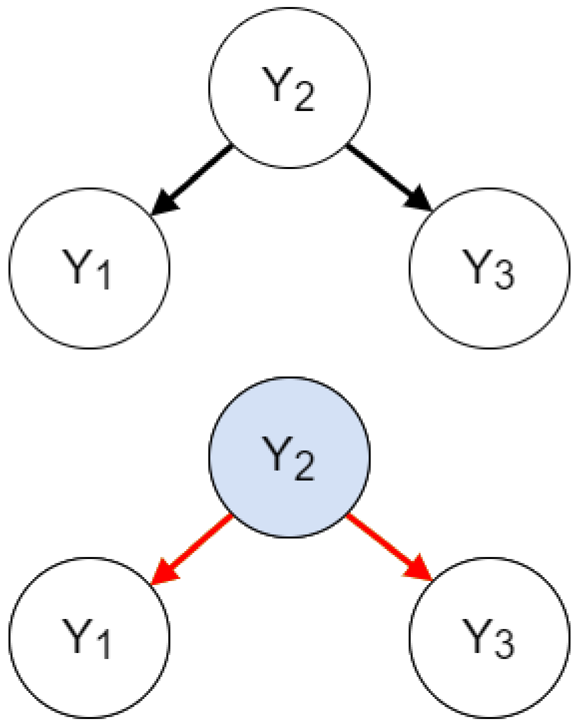
- **Head-to-Tail** (Chain, “Causal Effect”)



- **Head-to-Head** (Collider, “Common Effect”)



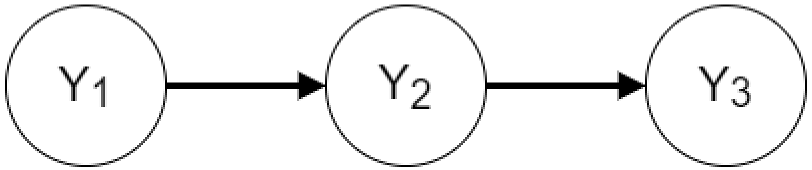
Tail-to-Tail Connections

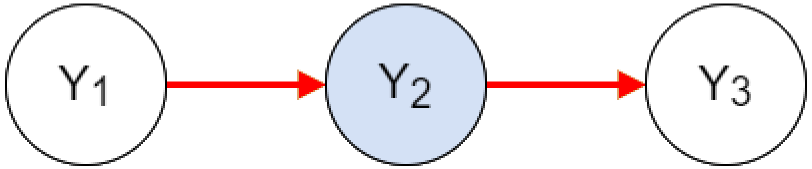


- Corresponds to
  $$P(Y_1, Y_3|Y_2)P(Y_2) = P(Y_1|Y_2)P(Y_3|Y_2)P(Y_2)$$
- If Y_2 is unobserved then Y_1 and Y_3 are marginally dependent
  $$Y_1 \not\perp Y_3$$
- If Y_2 is observed then Y_1 and Y_3 are conditionally independent
  $$Y_1 \perp Y_3 | Y_2$$

When Y_2 in observed is said to **block the path** from Y_1 to Y_3

Head-to-Tail Connections





Observed Y_2 **blocks the path** from Y_1 to Y_3

- Corresponds to

$$P(Y_1, Y_2, Y_3) = P(Y_1)P(Y_2|Y_1)P(Y_3|Y_2)$$
$$= P(Y_1|Y_2)P(Y_3|Y_2)P(Y_2)$$

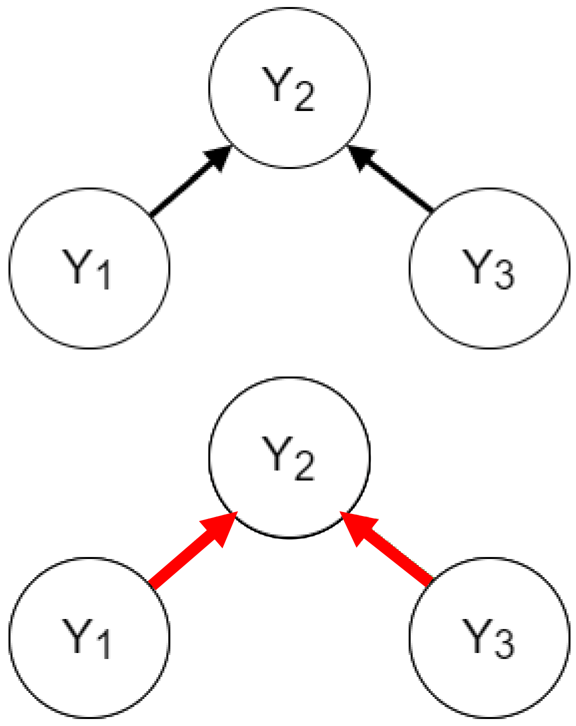
- If Y_2 is unobserved then Y_1 and Y_3 are marginally dependent Type equation here.

$$Y_1 \not\perp Y_3$$

- If Y_2 is observed then Y_1 and Y_3 are conditionally independent

$$Y_1 \perp Y_3 | Y_2$$

Head-to-Head Connections



- Corresponds to
$$P(Y_1, Y_2, Y_3) = P(Y_1)P(Y_3)P(Y_2|Y_1, Y_3)$$
- If Y_2 is observed then Y_1 and Y_3 are conditionally dependent
$$Y_1 \not\perp Y_3 | Y_2$$
- If Y_2 is unobserved then Y_1 and Y_3 are marginally independent
$$Y_1 \perp Y_3$$

If any Y_2 **descendants** is observed it **unlocks the path**

Probabilistic and Causal Learning

- Bayesian Networks (Tuesday 4th)
- **Reminder: no lecture tomorrow!**
- Bayesian Networks (Thursday 6th, **next**!)
  - d-separation
  - Markov Property and Faithfulness
  - Markov Blanket
  - Introduction to Markov Random Fields
- Graphical Causal Models (Tuesday 11th)
- Structure Learning and Causal Discovery (Wednesday 12th)

Conditional Independence: Representation and Learning

INTELLIGENT SYSTEMS FOR PATTERN RECOGNITION (ISPR)

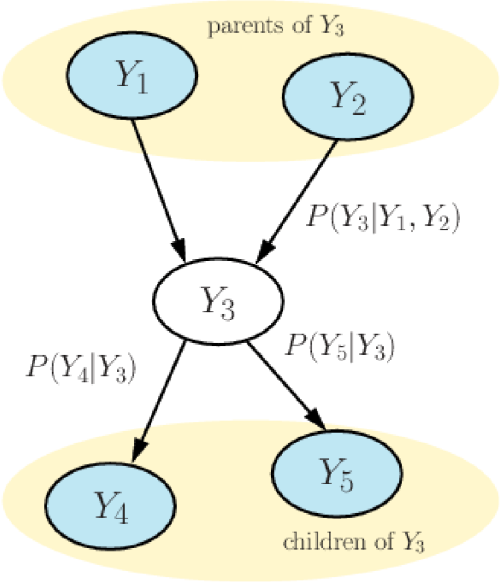
RICCARDO MASSIDDA, DAVIDE BACCIU – DIPARTIMENTO DI INFORMATICA - UNIVERSITA’ DI PISA

RICCARDO.MASSIDDA@DI.UNIPI.IT DAVIDE.BACCIU@UNIPI.IT

Probabilistic and Causal Learning

- Bayesian Networks (Tuesday 4th)
- Bayesian Networks (Thursday 6th, **today**!)
  - d-separation
  - Markov Property and Faithfulness
  - Markov Blanket
  - Introduction to Markov Random Fields
- Graphical Causal Models (Tuesday 11th)
- Structure Learning and Causal Discovery (Wednesday 12th)

Bayesian Network



- Directed Acyclic Graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
- Nodes $v \in \mathcal{V}$ represent random variables
  - Shaded $\Rightarrow$ observed
  - Empty $\Rightarrow$ un-observed
- Edges $e \in \mathcal{E}$ describe the conditional independence relationships

Conditional Probability Tables (CPT) local to each node describe the probability distribution given its parents

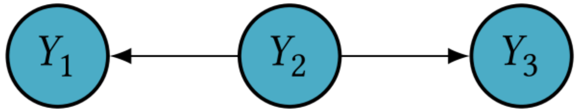
$$P(Y_1, \ldots, Y_N) = \prod_{i=1}^{N} P(Y_i \,|pa(Y_i))$$

IN SUPREMÆ DIGNITATIS 1343
UNIVERSITÀ DI PISA

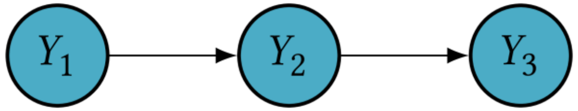
Fundamental BN structures

There exist **three fundamental substructures** that determine the conditional independence relationships in a Bayesian Network.

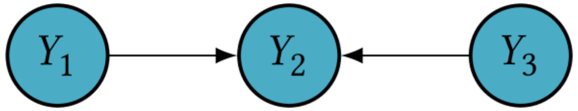
- **Tail-to-Tail** (Fork, “Common Cause”)



- **Head-to-Tail** (Chain, “Causal Effect”)



- **Head-to-Head** (Collider, “Common Effect”)



Blocked Path

Let $\mathrm{r} = (\mathrm{Y}_1 \leftrightarrow \cdots \leftrightarrow \mathrm{Y}_2)$ be an **undirected path** between Y_1 and Y_2.

The path r is **blocked** by a set Z if one of the following holds:

- r contains a **fork** (tail-to-tail) $\mathrm{Y}_\mathrm{i} \leftarrow \mathrm{Y}_\mathrm{c} \rightarrow \mathrm{Y}_\mathrm{j}$ such that $\mathrm{Y}_\mathrm{c} \in \mathrm{Z}$, or
- r contains a **chain** (head-to-tail) $\mathrm{Y}_\mathrm{i} \rightarrow \mathrm{Y}_\mathrm{c} \rightarrow \mathrm{Y}_\mathrm{j}$ such that $\mathrm{Y}_\mathrm{c} \in \mathrm{Z}$, or
- r contains a **collider** (head-to-head) $\mathrm{Y}_\mathrm{i} \rightarrow \mathrm{Y}_\mathrm{c} \leftarrow \mathrm{Y}_\mathrm{j}$ such that **neither Y_c nor its descendants are in Z**.

UNIVERSITÀ DI PISA

d-Separation

Definition (d-separated path)

Let $r = Y_1 \leftrightarrow \cdots \leftrightarrow Y_2$ be an undirected path between Y_1 and Y_2, then r is d-separated by Z if there exist at least one node $Y_c \in Z$ for which path r is blocked.

UNIVERSITÀ DI PISA

d-Separation

Definition (d-separation)

Two nodes Y_i and Y_j in a BN $\mathcal{G}$ are said to be d-separated by $Z \subset \mathcal{V}$ (denoted by $Dsep_{\mathcal{G}}(Y_i\ , Y_j|Z)$ if and only if all undirected paths between Y_i and Y_j are d-separated by Z

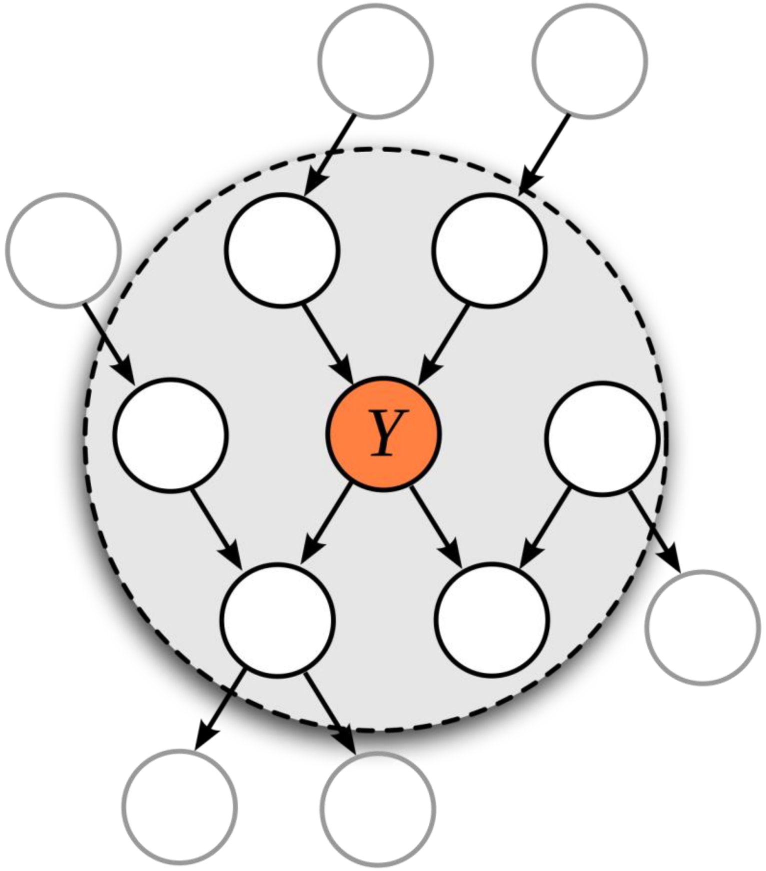
$$Y_1 \perp_{\mathcal{G}} Y_2 \mid Z$$

Global Markov Property

$$Y_1 \perp_{\mathcal{G}} Y_2 \mid Z \implies Y_1 \perp Y_2 \mid Z$$

- A Bayesian Network respects the **Global Markov** condition whenever **d-separations** in the graph imply **conditional independence** relations.
- **Global** and **local** Markov properties are **equivalent**.

Markov Blanket



- The **Markov Blanket** $\mathrm{Mb}(\mathrm{Y})$ of a node Y is the minimal set of vertices that **shield the node** from the rest of the Bayesian Network.
- In a DAG, the Markov Blanket of Y contains
  - Its parents Pa(Y)
  - Its children Ch(Y)
  - Its children's parents Pa(Ch(Y))
- The behavior of a node can be **completely determined and predicted** from the knowledge of its Markov Blanket.

$$P(Y \mid \mathrm{Mb}(Y), Z) = P(Y \mid \mathrm{Mb}(Y)) \; \forall Z \notin \mathrm{Mb}(Y)$$

Faithfulness Property

$$Y_1 \perp Y_2 \mid Z \implies Y_1 \perp_{\mathcal{G}} Y_2 \mid Z$$

- A Bayesian Network is faithful whenever **conditional independence** relations imply **d-separations.**
- While the **global Markov Condition** requires the graph to represent **only** conditional independences, the **Faithfulness** condition requires to represent **all** conditional independences.

Faithfulness Property

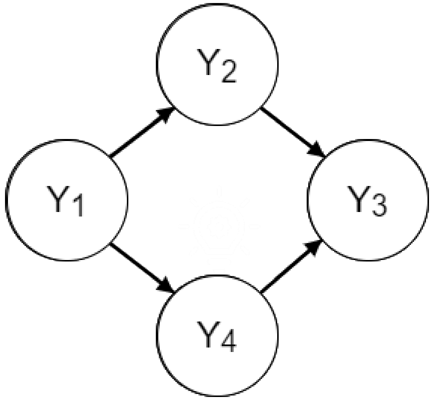
$$Y_1 \perp Y_2 \mid Z \implies Y_1 \perp_{\mathcal{G}} Y_2 \mid Z$$

- **Faithfulness** is fundamental to **concisely represent** joint distributions.
- Intuitively, the **more conditional independences** we represent, the **less parameters** we need to store in the model.

Are Directed Models Enough?

- Bayesian Networks are used to model **asymmetric dependencies**
- What if we want to model **symmetric dependencies**?
  - Bidirectional effects, e.g. spatial dependencies
  - Need **undirected** approaches

Directed models cannot represent some (bidirectional) dependencies in the distributions

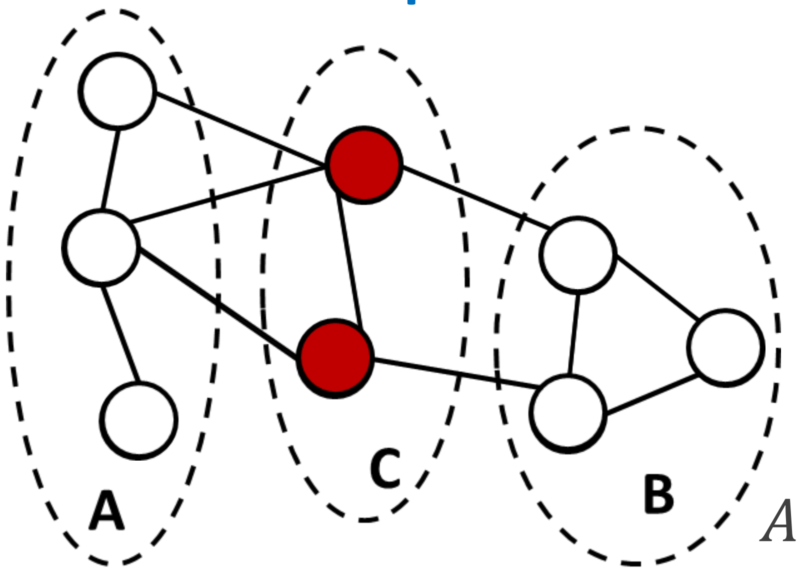


What if we want to represent $Y_1 \perp Y_3 | Y_2, Y_4$?
What if we also want $Y_2 \perp Y_4 | Y_1, Y_3$?

Cannot be done in BN! Need undirected model

Markov Random Fields

What is the **undirected equivalent** of **d-separation** in directed models?



$A \perp B | C$

Again it is based on node separation, although it is way simpler!

- Node subsets $A, B \subset \mathcal{V}$ are conditionally independent given $C \subset \mathcal{V} \backslash \{A, B\}$ if all paths between nodes in A and B pass through at least one of the nodes in C
- The Markov Blanket of a node includes all and only its neighbors

Joint Probability Factorization

What is the undirected equivalent of conditional probability factorization in directed models?

- We seek a product of functions defined over a set of nodes associated with some local property of the graph
- Markov blanket tells that nodes that are not neighbors are conditionally independent given the remainder of the nodes

$$P\left(X_v, X_i \middle| X_{\mathcal{V}\setminus\{v,i\}}\right) = P\left(X_v \middle| X_{\mathcal{V}\setminus\{v,i\}}\right) P\left(X_i \middle| X_{\mathcal{V}\setminus\{v,i\}}\right)$$

- Factorization should be chosen in such a way that nodes X_v and X_i are not in the same factor

What is a **well-known graph structure** that **includes only nodes that are pairwise connected**?

UNIVERSITÀ DI PISA

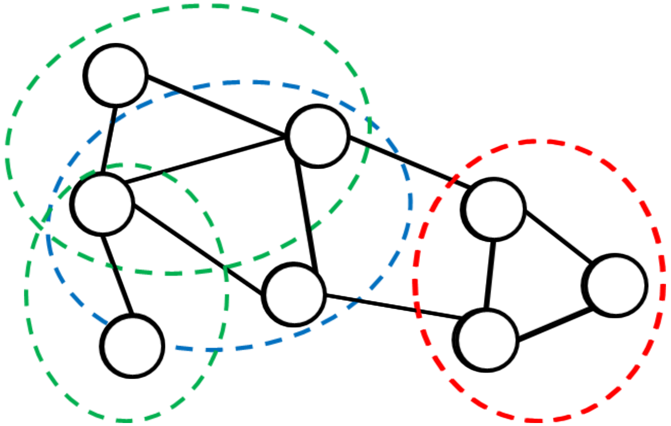
Cliques

Definition (Clique)

A subset of nodes C in graph $\mathcal{G}$ such that $\mathcal{G}$ contains an edge between all pair of nodes in C

Definition (Maximal Clique)

A clique C that cannot include any further node from the graph without ceasing to be a clique



Maximal Clique Factorization

Define $\boldsymbol{X} = X_1, \ldots, X_N$ as the RVs associated to the N nodes in the undirected graph $\mathcal{G}$

$$P(\boldsymbol{X}) = \frac{1}{Z} \prod_{C} \psi(\boldsymbol{X}_C)$$

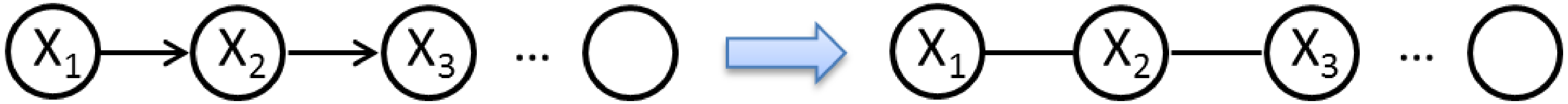
- $\boldsymbol{X}_C \rightarrow$ RV associated with nodes in the maximal clique C
- $\psi(\boldsymbol{X}_C) \rightarrow$ potential function over the maximal cliques C
- $Z \rightarrow$ partition function ensuring normalization

$$Z = \sum_{\boldsymbol{X}} \prod_{C} \psi(\boldsymbol{X}_C)$$

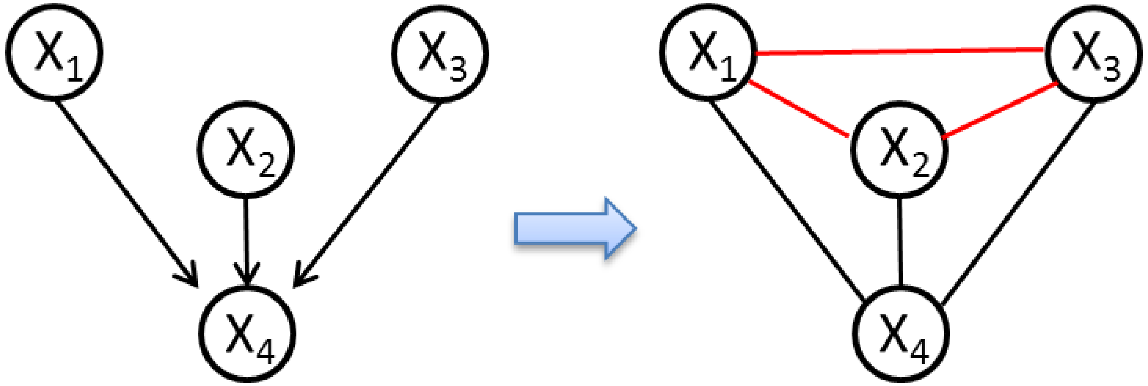
Partition function is the **computational bottleneck** of undirected modes: e.g. $O(K^N)$ for N discrete RV with K distinct values

From Directed To Undirected

Straightforward in some cases



Requires a little bit of thinking for v-structures



Moralization a.k.a. marrying of the parents

Next Lectures: Causal Learning

- Graphical Causal Models (Tuesday 11th)
  - Causation and Correlation
  - Causal Bayesian Networks
  - Structural Causal Models
  - Causal Inference
- Structure Learning and Causal Discovery (Wednesday 12th)
  - Constraint-Based Methods (PC, FCI)
  - Score-Based Methods (GES)
  - Parametric Assumptions (LiNGAM)