Conditional Independence: Representation and Learning

Generative and Deep Learning (GDL)

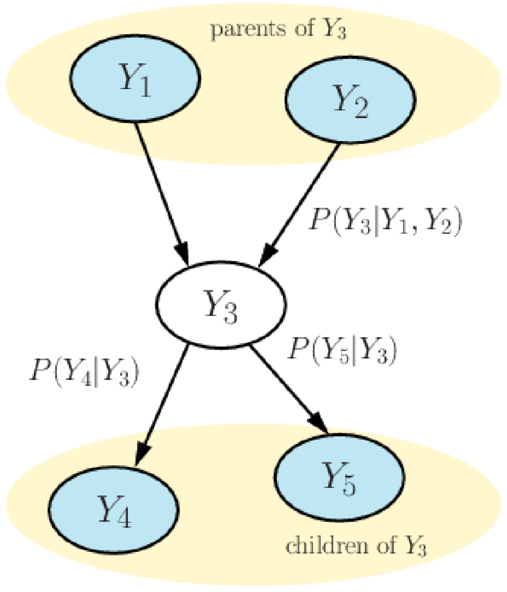
Riccardo Massidda (riccardo.massidda@unipi.it)
Davide Bacciu (davide.bacciu@unipi.it)

Università di Pisa

Graphical Models: Probability and Causality

- Bayesian Networks (Tuesday 24th)
- Bayesian Networks (Wednesday 25th, **today**!)
  - d-separation
  - Markov Property and Faithfulness
  - Markov Blanket
  - Introduction to Markov Random Fields
- Graphical Causal Models (Thursday 26th)
- Structure Learning and Causal Discovery (Tuesday 3rd)

Bayesian Network



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

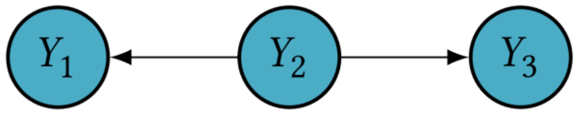
In a Bayesian Network, the **joint probability** is decomposed as

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

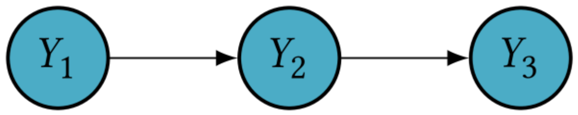
Fundamental BN structures

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

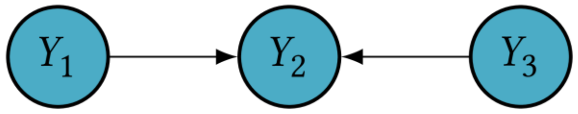
- **Tail-to-Tail** (Fork, "Common Cause")



- **Head-to-Tail** (Chain, "Causal Effect")



- **Head-to-Head** (Collider, "Common Effect")



Blocked Path

Let $r = (Y_1 \leftrightarrow \cdots \leftrightarrow Y_2)$ 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) $Y_i \leftarrow Y_c \rightarrow Y_j$ such that $Y_c \in Z$, or
- r contains a **chain** (head-to-tail) $Y_i \rightarrow Y_c \rightarrow Y_j$ such that $Y_c \in Z$, or
- r contains a **collider** (head-to-head) $Y_i \rightarrow Y_c \leftarrow Y_j$ such that **neither Y_c nor its descendants are in Z**.

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.

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

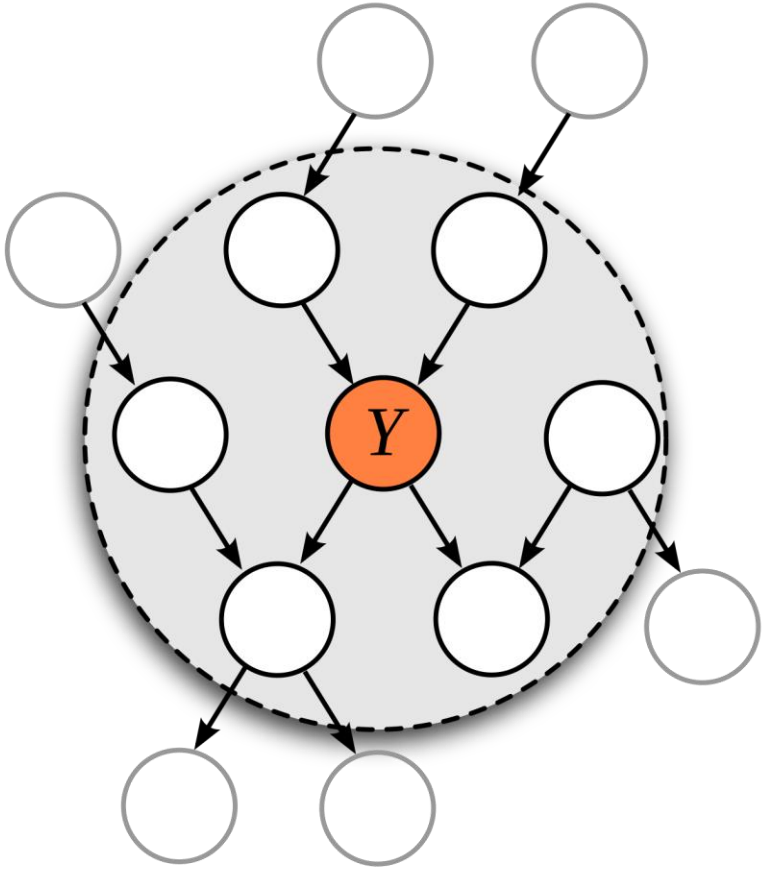
$$Y_1 \perp_{\mathcal{G}} Y_2 \mid Z$$

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** $\mathrm{Mb}(Y)$ 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

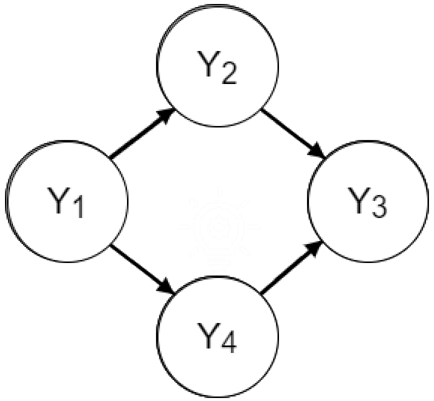
$$Y_1 \perp Y_2 \mid Z \implies Y_1 \perp_{\mathcal{G}} Y_2 \mid Z$$

- **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

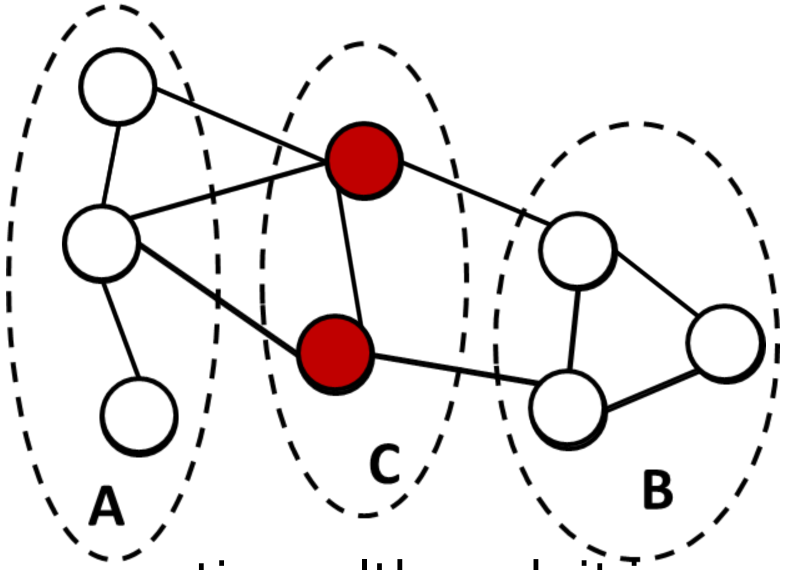


What if we want to represent $Y_1 \perp Y_3 | Y_2, Y_4$?
What if we also want $Y_2 \perp Y_4 | Y_1, Y_3$?

Cannot be done in BN!
Need **undirected** models

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

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

- 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**?

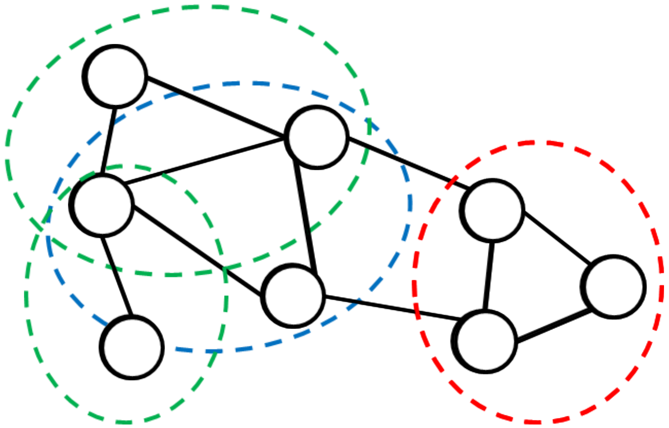
Cliques

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 $\mathcal{G}$

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

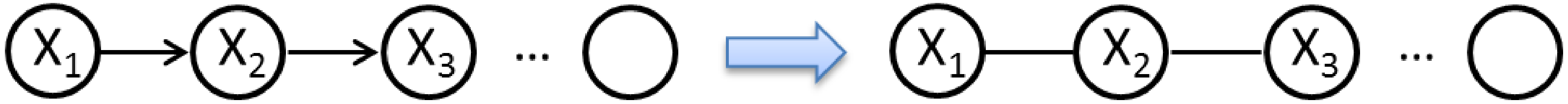
- $\boldsymbol{X}_C \rightarrow$ RV associated with nodes in the maximal clique C
- $\psi(\boldsymbol{X}_C) \rightarrow$ potential function over the maximal cliques C
- $Z \rightarrow$ partition function ensuring normalization

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

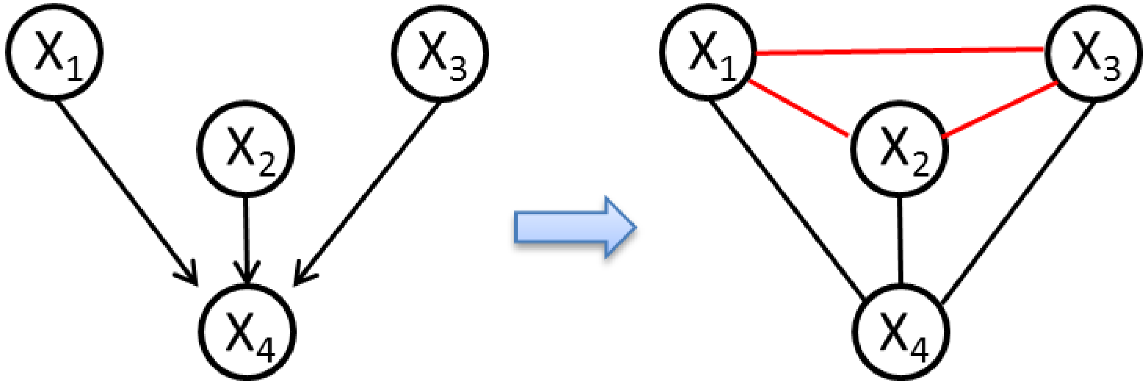
Partition function is the **computational bottleneck** of undirected modes: e.g. $O(K^N)$ for N discrete RV with K distinct values

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

Graphical Models: Probability and Causality

- Graphical Causal Models (Thursday 26th, **next!**)
  - Causation and Correlation
  - Causal Bayesian Networks
  - Structural Causal Models
  - Causal Inference
- Probabilistic and Causal Structure Learning (Thursday 3rd)
  - Constraint-Based Methods (PC, FCI)
  - Score-Based Methods (GES)
  - Parametric Assumptions (LiNGAM)