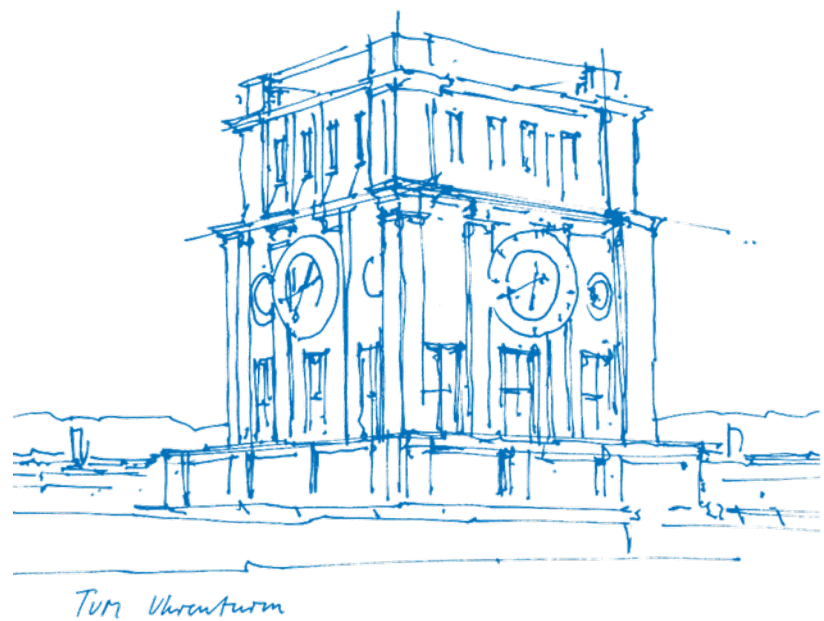
II : Graphical Model Representation

Tao Wu, Yuesong Shen, Zhenzhang Ye

Computer Vision & Artificial Intelligence
Technical University of Munich



Outline of the Chapter

- Bayesian network (directed graphical model).
- Markov random field (undirected graphical model).
- Independence assumption, representation power, parameterization, etc.

Bayesian Network

Bayesian Network (BN)

A **Bayesian network** (BN) is a *directed acyclic graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ together with:

- Random variables $X = (X_i)_{i \in \mathcal{V}}$ over $\mathcal{V}$;
- A (joint probability) distribution P *factorized* as a product of conditional probability distributions (CPDs):

$$p(x) = \prod_{i \in \mathcal{V}} p\big(x_i \big| (x_j)_{j \in \mathrm{Pa}_{\mathcal{G}}(i)}\big),$$

where $\mathrm{Pa}_{\mathcal{G}}(i) = \{j \in \mathcal{V} : (j, i) \in \mathcal{E}\}$ consists of parents of i in $\mathcal{G}$.

Example "Student"

$$P(D, I, G, S, L) = P(D)P(I)P(G|D, I)P(S|I)P(L|G).$$

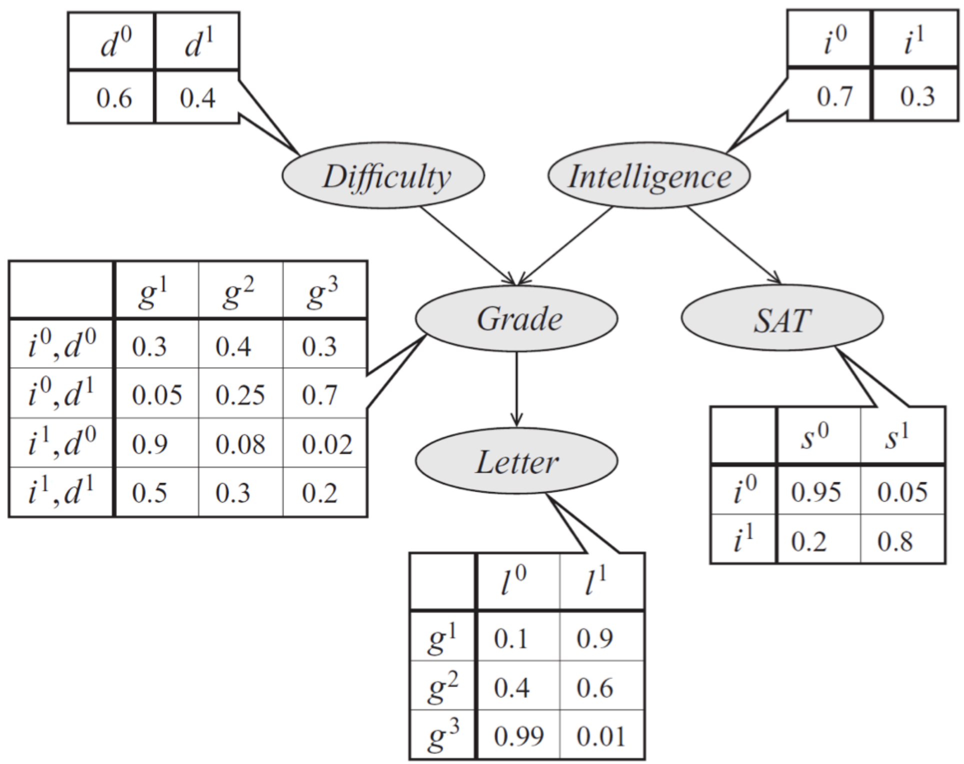


Figure: Bayesian network represented by probability tables.

Model Complexity

Consider BN representation for RVs $(X_i)_{i=1}^n$.

- If each RV X_i takes at most d outcomes and has at most k parents, then representation of
$$p(x_i | (x_j)_{j \in \mathrm{Pa}_{\mathcal{G}}(i)})$$
requires $O(d^{k+1})$ free parameters.
- Since the joint distribution for $(X_i)_{i=1}^n$ is a product of n CPDs, the overall model complexity for BN is $O(nd^{k+1})$.
- Compared to a naive representation for the joint distribution which requires $O(d^n)$ parameters (typically $n \gg k$).

The reduction of complexity is due to the underlying independence assumptions.

Independencies in BNs

- For a distribution P for RVs (X_i), we denote by $\mathcal{I}(P)$ the set of all **independence assumptions (assertions)** that hold in P:

$$\mathcal{I}(P) = \big\{(X_i \perp X_j \,|\, X_k)\big\}.$$

Recall conditional independence: $X_i \perp X_j \,|\, X_k$ iff

$$p(x_i, x_j | x_k) = p(x_i | x_k) p(x_j | x_k).$$

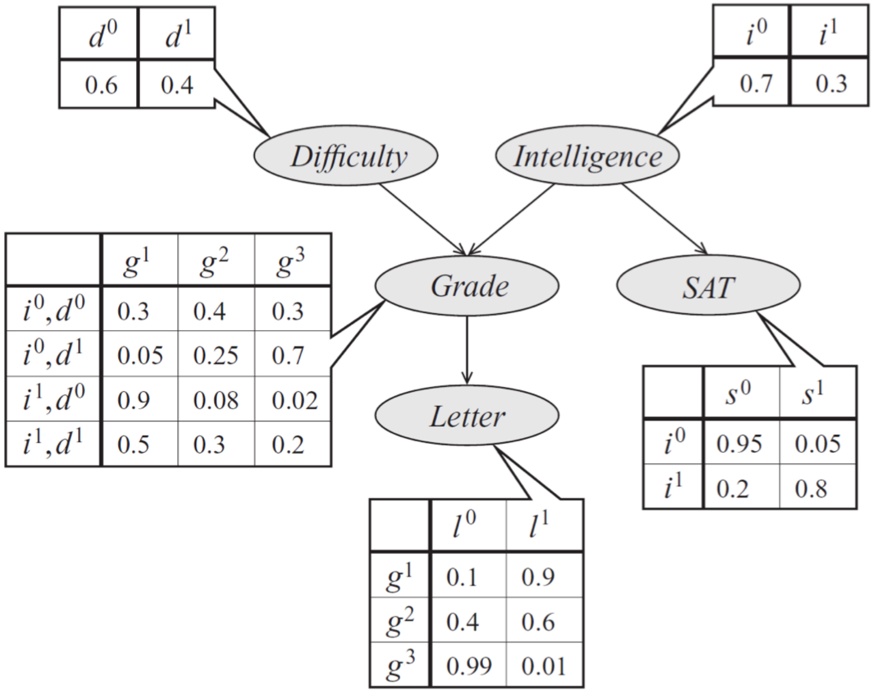
- BN $\mathcal{G}$ implies **local independencies**:

$$\mathcal{I}_\ell(\mathcal{G}) = \Big\{\Big(X_i \perp (X_j)_{j \in \mathrm{NonDes}_\mathcal{G}(i) \setminus \{i\} \setminus \mathrm{Pa}_\mathcal{G}(i)} \,|\, (X_k)_{k \in \mathrm{Pa}_\mathcal{G}(i)}\Big)\Big\},$$

where $\mathrm{NonDes}_\mathcal{G}(i)$ contains the non-descendants of i in $\mathcal{G}$.

Example "Student"

$$\mathcal{I}_\ell(\mathcal{G}) = \left\{ \left(X_i \perp (X_j)_{j \in \mathrm{NonDes}_\mathcal{G}(i) \setminus \{i\} \setminus \mathrm{Pa}_\mathcal{G}(i)} \,\middle|\, (X_k)_{k \in \mathrm{Pa}_\mathcal{G}(i)} \right) \right\}.$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

In this example we have, e.g., $(L \perp \{I, D, S\} \mid G),\ (G \perp S \mid \{I, D\}) \in \mathcal{I}_\ell(\mathcal{G})$.

Beyond Local Independence

- Does $\mathcal{G}$ encode other independence assertions besides $\mathcal{I}_\ell(\mathcal{G})$? (Yes.)
- How to identify a specific independence assertion in $\mathcal{G}$? (D-separation.)

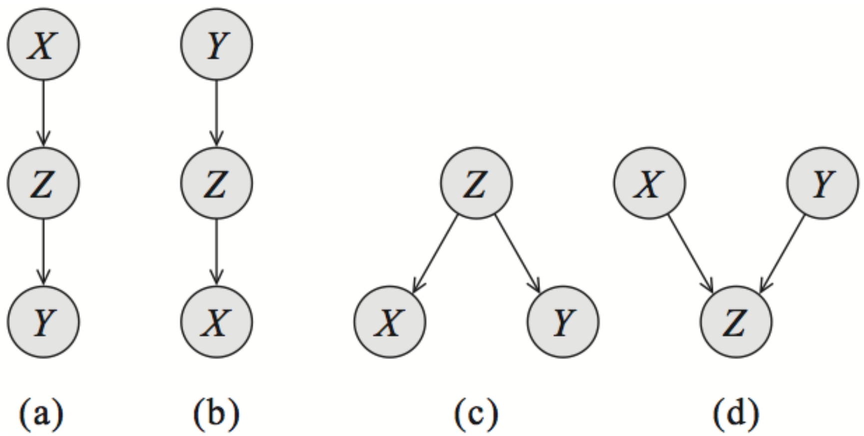
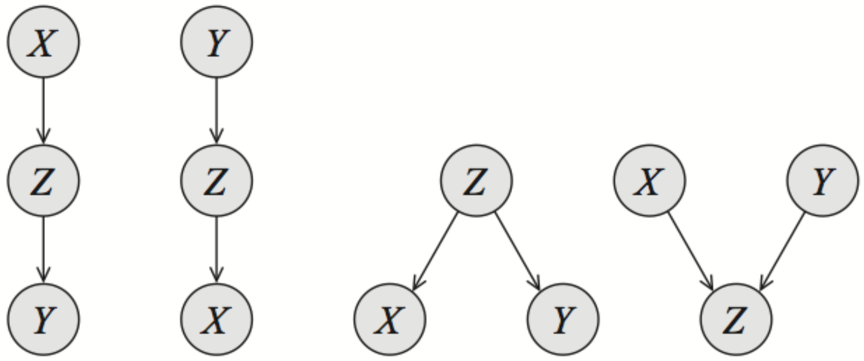


Figure: Two-edge trails from X to Y via Z. (d) is called the **v-structure**.

In the above figure, information/dependence flows from X to Y if the **trail** $X \leftrightarrow Z \leftrightarrow Y$ is **active**. This is the case if:

- In (a)–(c), Z is unobserved. (In contrast, $X \perp Y \mid Z$.)
- In (d), Z or one of its descendants is observed. (In contrast, $X \perp Y$ o.w.)

Active Trail



(a) (b) (c) (d)

Let $X_1 \leftrightarrow X_2 \leftrightarrow \ldots \leftrightarrow X_n$ be a trail in a BN $\mathcal{G}$, and Z be a set of observed nodes (RVs). The trail is **active** given Z if

- Whenever there is a v-structure (case (d)) in the trail $X_{i-1} \leftrightarrow X_i \leftrightarrow X_{i+1}$, then X_i or one of its descendants are in Z.
- No other node along the trail belongs to Z.

Intuitively, information/dependence flows from X_1 to X_n (and vice versa) through the active trail $X_1 \leftrightarrow X_2 \leftrightarrow \ldots \leftrightarrow X_n$.

D-separation, Global Independence

Let X, Y, Z be three sets of nodes in a BN $\mathcal{G}$. If there is no active trail between any node in X and Y given Z, we say X and Y are **d-separated** given Z.

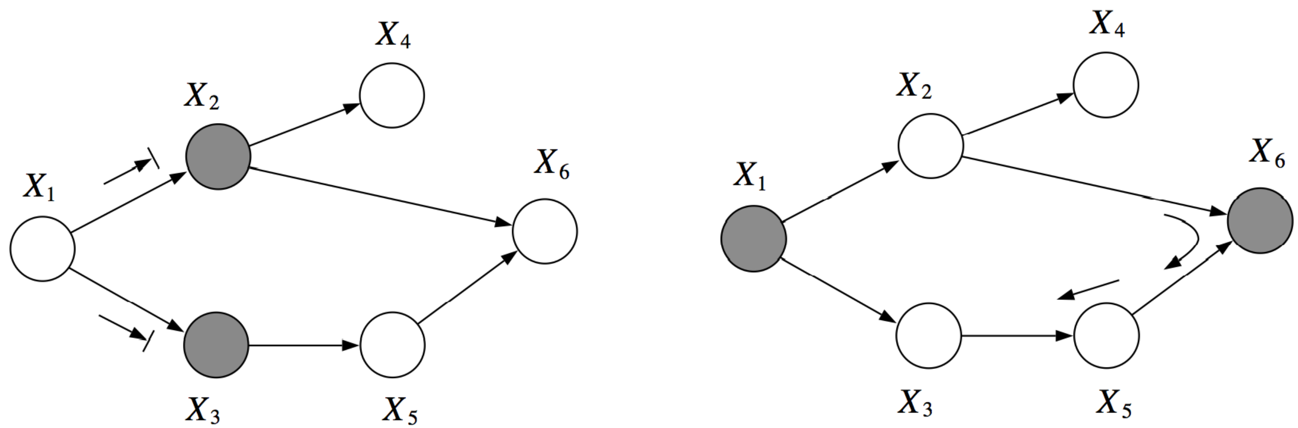


Figure: (left) X_1 and X_6 are d-sep. given $\{X_2, X_3\}$; (right) X_2 and X_3 are not d-sep. given $\{X_1, X_6\}$.

We denote by $\mathcal{I}(\mathcal{G})$ the set of **global Markov independencies**:

$$\mathcal{I}(\mathcal{G}) = \{(X \perp Y \mid Z) : X \text{ and } Y \text{ are d-separated given } Z\}.$$

Facts about D-separation

F1. (Soundness) If a distribution P factorizes according to $\mathcal{G}$, then $\mathcal{I}(\mathcal{G}) \subset \mathcal{I}(P)$. The converse is also true. In this case, we call $\mathcal{G}$ an **I-map** for P.

F2. (Sharpness) If nodes X and Y are not d-separated given Z in $\mathcal{G}$, then X and Y are dependent given Z in some distribution P that factorizes over $\mathcal{G}$.

F3. (Completeness) When a distribution P factorizes according to $\mathcal{G}$, $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$ does not necessarily holds. Obviously, one can add superfluous edges to $\mathcal{G}$ s.t. $\mathcal{I}(\mathcal{G}) \subsetneq \mathcal{I}(P)$.

$p(b \mid a)$	b_0	b_1
a_0	0.4	0.6
a_1	0.4	0.6

Figure: Here $A \perp B$. Note that $A \rightarrow B$ is an I-map for P, but $\emptyset = \mathcal{I}(\mathcal{G}) \subsetneq \mathcal{I}(P)$.

Remark: For almost all P (except for a set of measure zero in the space of CPD parameterizations) for which $\mathcal{G}$ is an I-map, we have $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$.

I-equivalence

We can compare two BNs using their independence assertions.

- Two BNs $\mathcal{G}_1$ and $\mathcal{G}_2$ are said to be **I-equivalent** if $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$.
- The **skeleton** of a BN $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an *undirected* graph $(\mathcal{V}, \mathcal{E}')$ such that $\{X, Y\} \in \mathcal{E}'$ whenever $(X, Y) \in \mathcal{E}$.
- Fact: If two BNs have the same skeleton and the same set of v-structures, then they are I-equivalent.

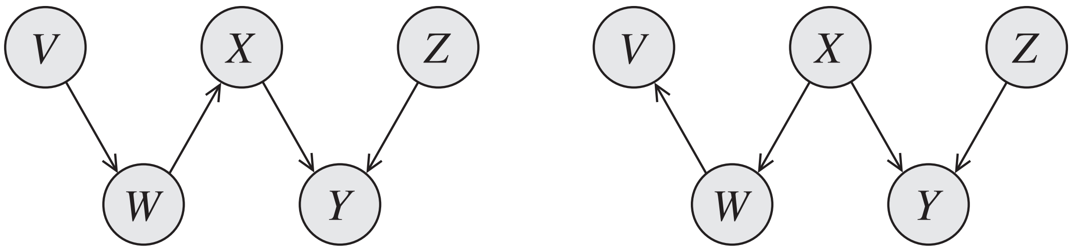


Figure: Example of two I-equivalent BNs.

Perfect Map and Counterexamples

- We say a BN $\mathcal{G}$ is a **perfect map** for a distribution P if $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$.
- Certain independencies cannot be expressed perfectly by BN.

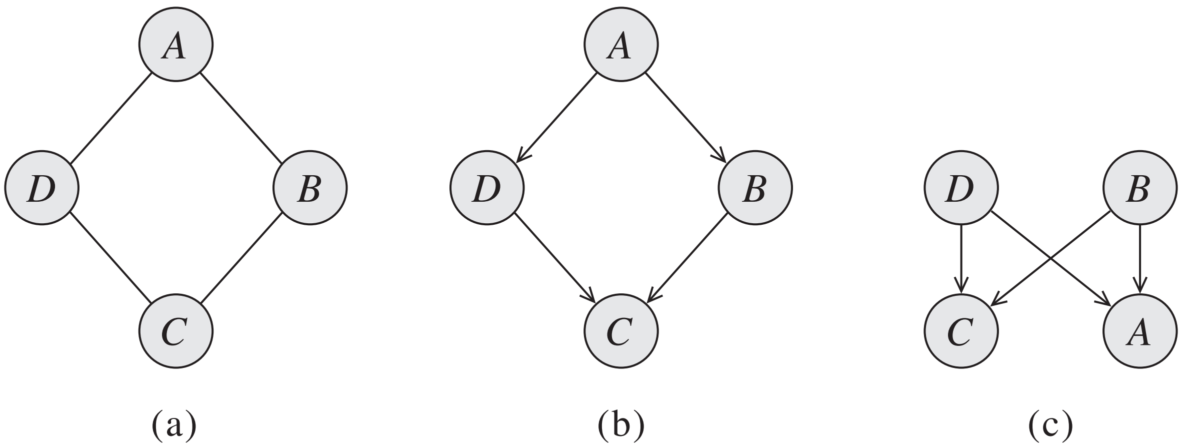


Figure: A counterexample where a perfect map does not exist.

(a) Desired independence assertions: $A \perp C \mid \{B, D\}$, $B \perp D \mid \{A, C\}$.

(b) In this BN: $(A \perp C \mid \{B, D\}) \in \mathcal{I}(\mathcal{G})$, but $(B \perp D \mid \{A, C\}) \notin \mathcal{I}(\mathcal{G})$.

(c) Again, $(A \perp C \mid \{B, D\}) \in \mathcal{I}(\mathcal{G})$, but $(B \perp D \mid \{A, C\}) \notin \mathcal{I}(\mathcal{G})$.

Topics which are not covered here ...

- Algorithm for detecting d-separation in a BN $\mathcal{G}$.
- Algorithm for finding minimal I-map $\mathcal{G}$ for a given distribution P.
- Algorithm for finding perfect map $\mathcal{G}$ (if exists) for a given distribution P.
- Further reading: Koller & Friedman, Chapter 3.

Markov Random Field

Markov Random Field (MRF)

A Markov Random Field (MRF) is an *undirected graph* $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, together with a (joint probability) distribution P for RVs $X = (X_i)_{i \in \mathcal{V}}$ s.t.

$$p(x) = \frac{1}{Z} \prod_{C \in \mathrm{Clique}(\mathcal{H})} \phi_C(x_C), \qquad (\dagger)$$

- $\mathrm{Clique}(\mathcal{H})$ is the set of **cliques** (i.e. *fully connected subgraphs*) of $\mathcal{H}$.
- Each ϕ_C is a (nonnegative) **factor** on the clique C, and $x_C = (x_i)_{i \in \mathcal{V}_C}$.
- Z is the **partition function** ("Z" from German word "Zustandssumme"):

$$Z = \sum_x \prod_{C \in \mathrm{Clique}(\mathcal{H})} \phi_C(x_C),$$

  which is a normalization constant ensuring $\sum_x p(x) = 1$.

Distributions that can be factorized in form of $(\dagger)$ are called **Gibbs distributions**.

Illustration of MRF

$$p(a, b, c, d) = \frac{1}{Z}\phi_{\{A,B\}}(a, b)\phi_{\{B,C\}}(b, c)\phi_{\{C,D\}}(c, d)\phi_{\{D,A\}}(d, a),$$

$$Z = \sum_{a,b,c,d} \phi_{\{A,B\}}(a, b)\phi_{\{B,C\}}(b, c)\phi_{\{C,D\}}(c, d)\phi_{\{D,A\}}(d, a).$$

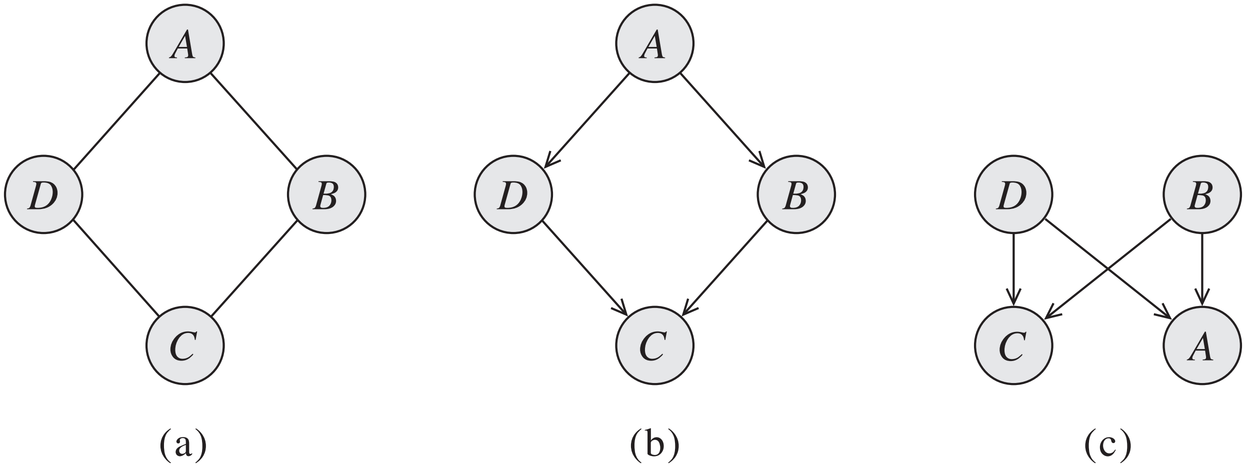


Figure: MRF in (a) cannot be perfectly represented by BN in (b) or (c).

Independencies in MRFs

Recall that global independencies in BNs are characterized by "active trail" and "d-separation". We do the equivalent for MRFs.

- Let X_1 — ... — X_n be a path in MRF $\mathcal{H}$, and O the set of observed nodes. The path X_1 — ... — X_n is **active** given O if none of $(X_i)_{i=1}^n$ belongs to O.
- Let X, Y, O be three sets of nodes in MRF $\mathcal{H}$. If there is no active path between any node in X and Y given O, then we say X and Y are **separated** given O.
- We define the **global independencies** given by $\mathcal{H}$ as:

$$\mathcal{I}(\mathcal{H}) = \{(X \perp Y \mid O) : X \text{ and } Y \text{ are separated given } O\}.$$

Facts about Separation in MRF

F1. (Soundness) If a distribution P factorizes according to MRF $\mathcal{H}$, then $\mathcal{H}$ is an I-map for P, i.e. $\mathcal{I}(\mathcal{H}) \subset \mathcal{I}(P)$.

F2. (Hammersley-Clifford theorem) Converse to (F1), if $\mathcal{H}$ is an I-map for a *positive* distribution P, then P factorizes according to $\mathcal{H}$. (A **positive distribution** has strictly positive probability for any (non-empty) event.)

F3. (Sharpness) If nodes X and Y are not separated given O in $\mathcal{H}$, then X and Y are dependent given O in some distribution P that factorizes over $\mathcal{H}$.

F4. (Completeness) When a distribution P factorizes according to $\mathcal{H}$, $\mathcal{I}(\mathcal{H}) = \mathcal{I}(P)$ does not necessarily hold.

Markov Blanket

- Let RVs $X = (X_i)_{i\in\mathcal{V}}$ and a distribution P for X be given. the **Markov blanket** of nodes $Y \subset X$ ("$\subset$" meaning $Y = (X_i)_{i\in\mathcal{V}'}$ with $\mathcal{V}' \subset \mathcal{V}$) under P is the minimal set of nodes $U \subset X\backslash Y$ s.t.

$$\big(Y \perp X\backslash Y\backslash U \,|\, U\big) \in \mathcal{I}(P).$$

- Fact: If a distribution P factorizes according to MRF $\mathcal{H}$, then the **Markov blanket** of any node is given by its neighbors in $\mathcal{H}$.

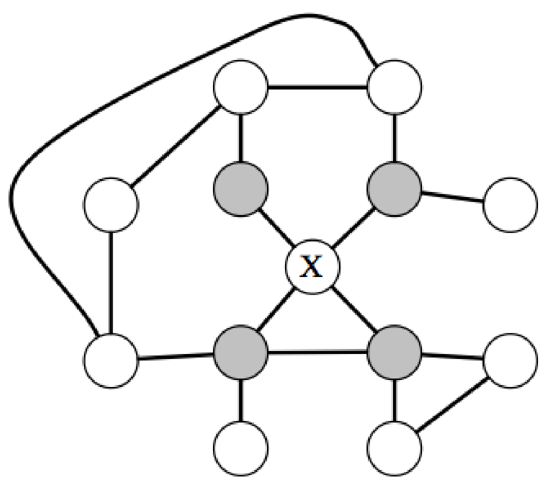


Figure: Markov blanket for node X.

Applying Markov Blanket

- An I-map $\mathcal{H}$ for P is **minimal** if removing any edge from $\mathcal{H}$ renders it no longer an I-map for P. Note that a minimal I-map is not necessarily perfect.

- One can use Markov blanket (MB) to construct "minimal I-map":
  $\forall i \in \mathcal{V}$: identify MB of $i \rightsquigarrow$ forge edge(s) from i to its MB.

- To construct a minimal I-map $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, set

$$\mathcal{E} = \Big\{ \{i, j\} \in \mathcal{V} \times \mathcal{V} : X_j \text{ belongs to the Markov blanket of } X_i \text{ under } P \Big\}.$$

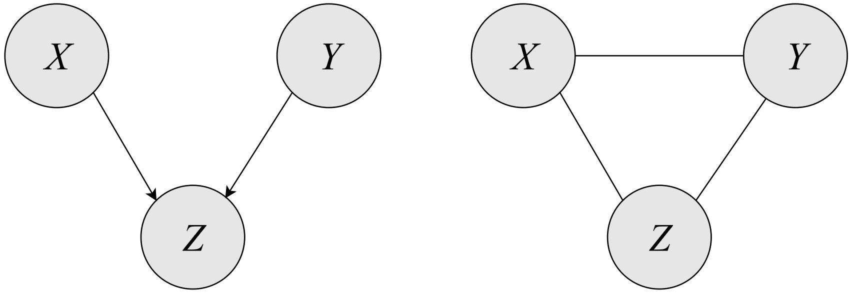


Figure: (left) P factorized according to BN (v-structure) indicates dependence of X and Y given Z observed. (right) Hence, an I-map for P by MRF must have the edge $\{X, Y\}$.

- Converting BN (left fig.) to MRF (right fig.) is called **moralization**.

Factor Graph

In an MRF, the joint distribution is factorized into a product of factors. It is possible to make factor-node interaction explicit in a "factor graph".

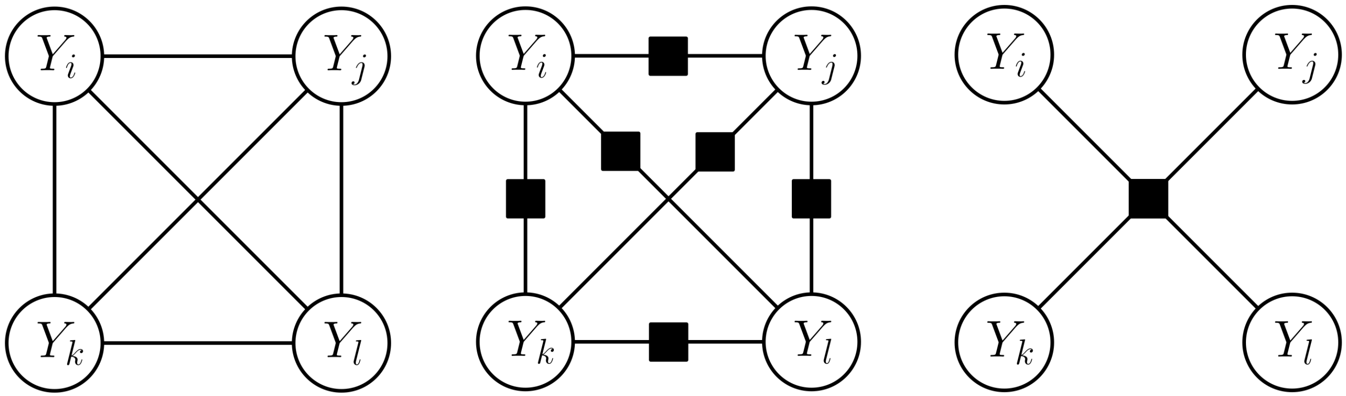
- A **factor graph** is a tuple $\mathcal{G} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$ consisting of a set $\mathcal{V}$ of **variable nodes**, a set $\mathcal{F} \subset 2^{\mathcal{V}}$ of **factor nodes**, and a set $\mathcal{E} \subset \mathcal{V} \times \mathcal{F}$ of **edges**.
- Each edge in $\mathcal{E}$ connects one variable node and a factor node, hence the overall factor graph $\mathcal{G}$ is **bipartite**.
- The factor graph $\mathcal{G}$ defines a family of joint distributions for $X = (X_i)_{i \in \mathcal{V}}$ factorized as

$$p(x) = \frac{1}{Z} \prod_{F \in \mathcal{F}} \phi_F(x_F),$$
$$Z = \sum_x \prod_{F \in \mathcal{F}} \phi_F(x_F),$$

with each ϕ_F being a factor for $X_F = (X_i)_{i \in \mathcal{V} : (i,F) \in \mathcal{E}}$.

Illustration of Factor Graph

Figure: (left) A fully connected MRF with four nodes; (mid) Factor graph with pairwise factors; (right) Factor graph with a single joint factor.



- Factor graphs in (mid) and (right) are both valid for the MRF in (left). Hence, the ambiguity in the factorization of MRF is resolved by factor graph representation.
- A **pairwise MRF** contains only *unary* and *pairwise* (but no higher-order) factors. Note: A pairwise MRF is a tree $\Leftrightarrow$ its factor graph is a tree.

Parameterization of MRFs

- In a factor graph, we often rewrite factor ϕ_F using **energy function** E_F:

$$\phi_F(x_F) =: \exp(-E_F(x_F)) \quad \Rightarrow$$
$$p(x) = \exp\Big(-\sum_{F\in\mathcal{F}} E_F(x_F) - \log Z\Big),$$
$$\log Z = \log \sum_x \exp\Big(-\sum_{F\in\mathcal{F}} E_F(x_F)\Big).$$

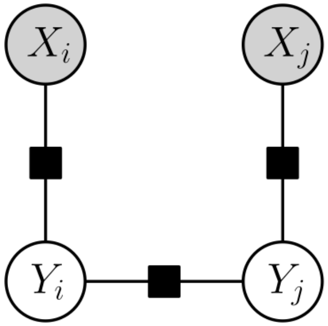
- MRF in **log-linear form** (useful for learning):

$$p(x;\theta) = \exp\Big(-\sum_{C\in\text{Clique}(\mathcal{H})} \theta_C^\top \psi_C(x_C) - \log Z(\theta)\Big),$$
$$\log Z(\theta) = \log \sum_x \exp\Big(-\sum_{C\in\text{Clique}(\mathcal{H})} \theta_C^\top \psi_C(x_C)\Big).$$

Each ψ_C maps x_C to a set of "features"; θ_C are weights which yield a linear function of features.

- Distributions of this form are members of the **exponential family**.

Conditional Random Field (CRF)



In some applications, a subset of nodes of an MRF are always observable. In this case, we can simplify MRF as conditional random field. A **conditional random field** (CRF) is a factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$, with

- $\mathcal{V} = \mathcal{X} \cup \mathcal{Y}$ with observable var. $X = (X_i)_{i \in \mathcal{X}}$ and target var. $Y = (Y_j)_{j \in \mathcal{Y}}$.
- $\mathcal{F}$ does not have any element being a subset of $\mathcal{X}$.
- The conditional distribution $P(Y|X)$ is factorized as

$$p(y|x) = \frac{1}{Z(x)} \prod_{F \in \mathcal{F}} \phi_F(y_{F \cap Y}; x_{F \cap X}),$$

$$Z(x) = \sum_{y} \prod_{F \in \mathcal{F}} \phi_F(y_{F \cap Y}; x_{F \cap X}).$$

MAP Inference on CRF

- CRF parameterized by energies:

$$p(y|x) = \exp\left(-\sum_{F\in\mathcal{F}} E_F(y_F; x_F) - \log Z(x)\right),$$

$$\log Z(x) = \log \sum_y \exp\left(-\sum_{F\in\mathcal{F}} E_F(y_F; x_F)\right).$$

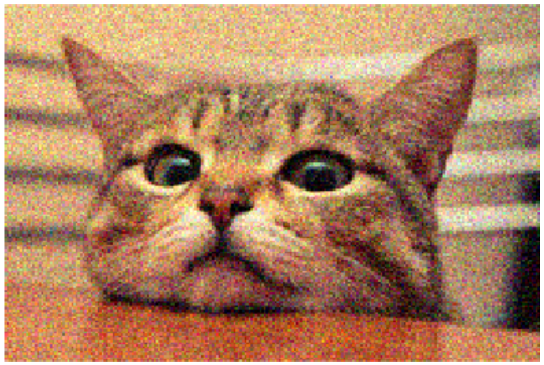
- MAP inference given $x, (\theta_F)_{F\in\mathcal{F}}, (E_F)_{F\in\mathcal{F}}$:

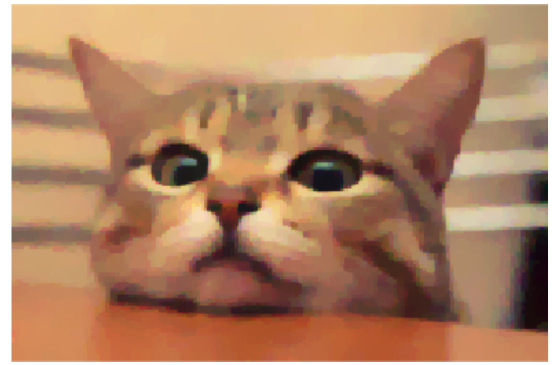
$$\max_y p(y|x) \;\Leftrightarrow\; \max_y \exp\left(-\sum_{F\in\mathcal{F}} E_F(y_F; x_F)\right)$$

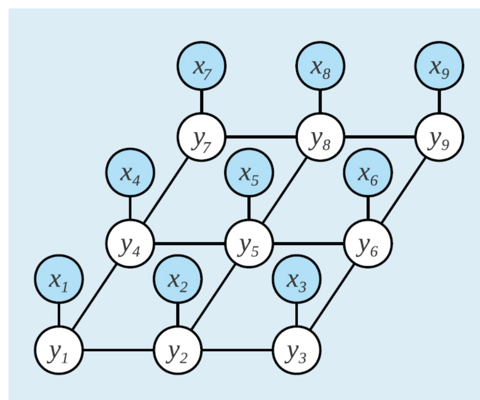
$$\Leftrightarrow\; \min_y \sum_{F\in\mathcal{F}} E_F(y_F; x_F) =: E(y; x).$$

- $\max_y p(y|x)$ is a special case of **structured prediction**: $\max_y g(y, x)$.

Example: Image Denoising by MAP on CRF







$$\min_y E(y;x) := \sum_{i\in\mathcal{V}} |y_i - x_i|^2 + \alpha \sum_{i\in\mathcal{V}} \sum_{j\in \mathrm{nbh}(i)} |y_i - y_j|.$$

unary factors pairwise factors

Summary

- Markov random field: definition, independence assertions.
- Factor graph: explicit representation of factors in MRF.
- Parameterization of MRF: energy, log-linear form.
- Conditional random field: MRF conditioning on observable nodes.
- Further reading: Koller & Friedman, Chapter 4; Murphy, Chapter 19.