Probabilistic Machine Learning

Graphical Models (2)

Alexandre Graell i Amat
alexandre.graell@chalmers.se
https://sites.google.com/site/agraellamat

November 23, 2023

CHALMERS

Markov random fields

In many real world phenomena, can not determine exactly directionality of interaction between random variables $\longrightarrow$ BNs not well suited.

Markov random fields:

- Described by undirected graphs
- Encode a factorization (a set of conditional independence relationships)

Useful for

- Modeling mutual relationships of compatibility among variables
- Imaging and spatial applications

Markov random fields

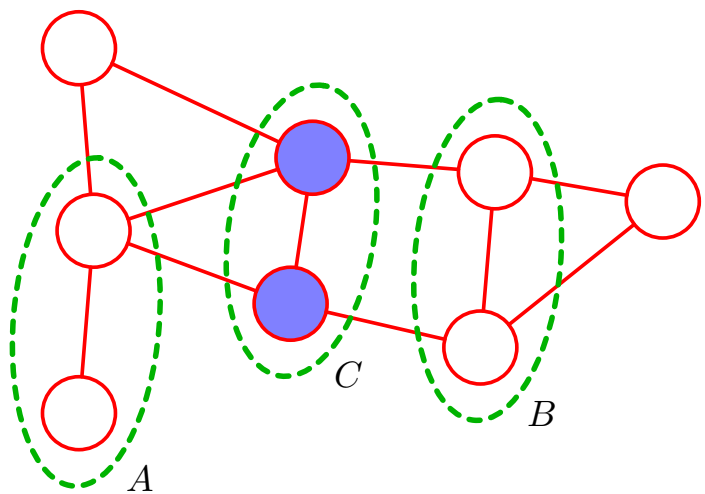
Different types of elements:

- Nodes: represent RVs
  - Empty nodes: unobserved RVs
  - Shaded nodes: observed RVs
- Edges: represent relationships between RVs
- Filled small circles: represent learnable parameters

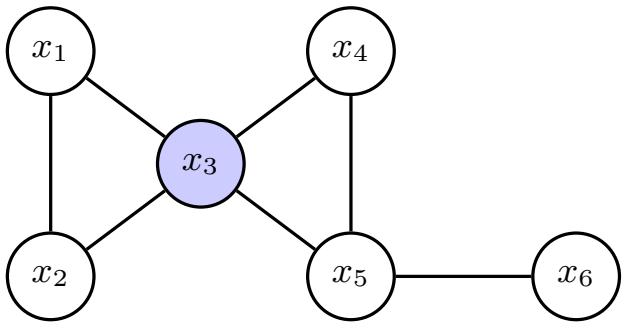
Conditional independence

MRFs allow to graphically assess conditional independence properties by simple separation

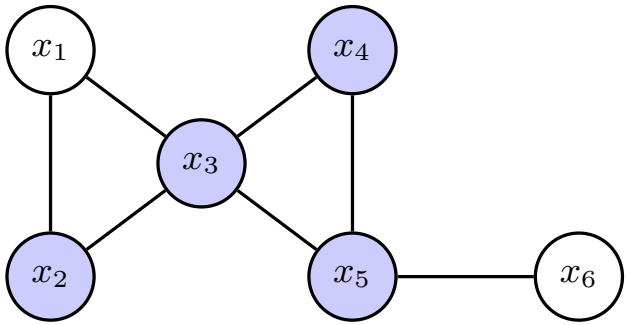
Property: Let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be three disjoint sets. Then $\mathcal{A} \perp \mathcal{B} | \mathcal{C}$ if every path from any node in set $\mathcal{A}$ to any node in set $\mathcal{B}$ passes through at least one node in $\mathcal{C}$ (all paths are blocked).



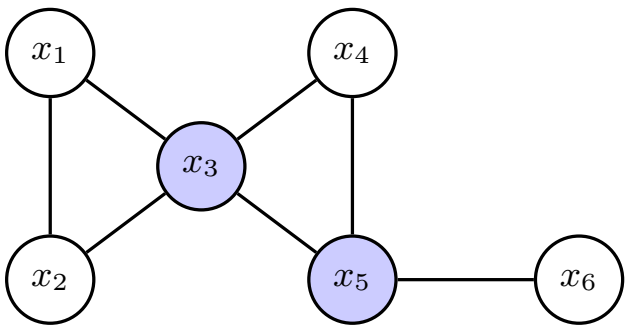
Conditional independence: Examples



$\{x_1, x_2\} \perp \{x_4, x_5, x_6\} | x_3$



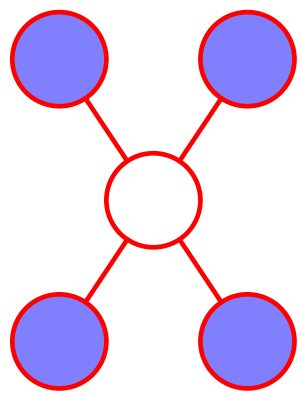
$x_1 \perp x_6 | \{x_2, x_3, x_4, x_5\}$



$x_4 \perp \{x_1, x_2, x_6\} | \{x_3, x_5\}$

Markov blanket

Markov blanket: A node is conditionally independent of all other nodes given all its neighbors.

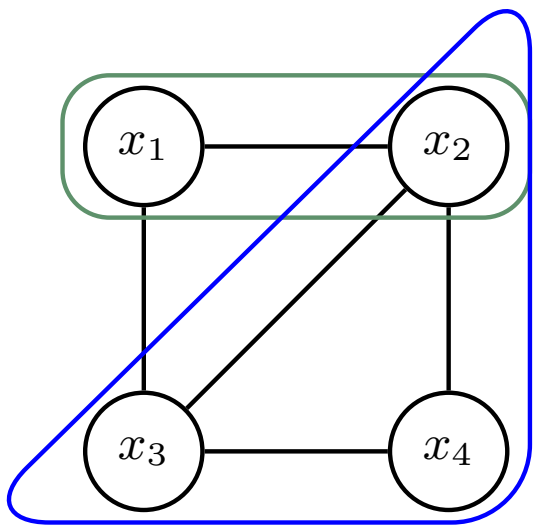


Factorization properties

Goal: Express $p(\boldsymbol{x})$ as a product of functions defined over local sets of variables

Clique: A fully connected subgraph.

Maximal clique: A clique for which it is not possible to add any additional node without it ceasing to be a clique.



Factorization properties

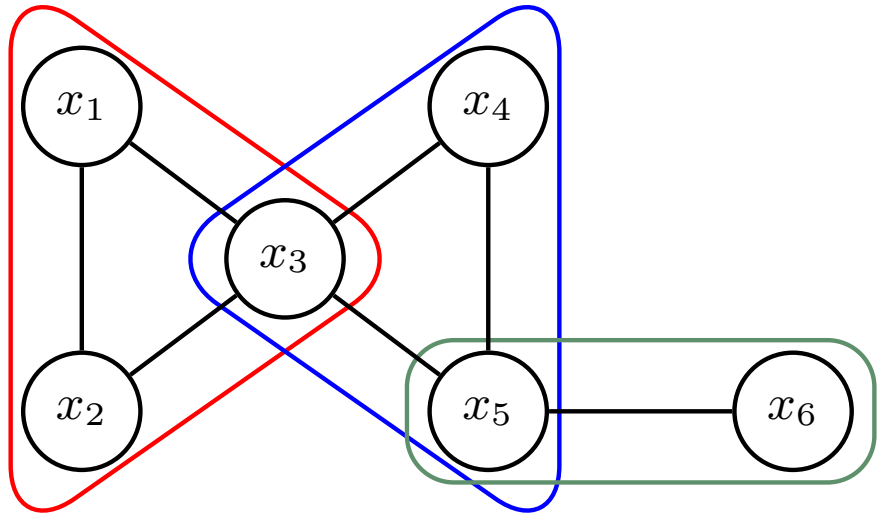
MRF: An undirected graph whose vertices represent random variables associated to a joint probability distribution that factorizes as product of potential functions corresponding to the maximal cliques of the graph.

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{c} \psi_c(\mathbf{x}_c)$$

c : index of a maximal clique
$\mathbf{x}_c$: set of RVs associated with the vertices of c
Z : normalization constant (partition function)

$$Z = \sum_{\mathbf{x}} \prod_{c} \psi_c(\mathbf{x}_c),$$

Factorization properties: An example



$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_c \psi_c(\mathbf{x}_c)$$

$$= \frac{1}{Z} \psi_1(x_1, x_2, x_3) \psi_2(x_3, x_4, x_5) \psi_3(x_5, x_6)$$

Factorization properties

$p(\boldsymbol{x})$ factorizes as

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_c \psi_c(\mathbf{x}_c)$$

A few notes:

- Considering $\psi_c(\mathbf{x}_c) \geq 0$ ensures $p(\boldsymbol{x}) \geq 0$
- In general, $\psi_c(\mathbf{x}_c)$ are not probability distributions

Motivation for $\psi_c(\cdot)$: Which configurations of local variables are preferred to others.

- $\psi_c(\mathbf{x}_c)$ encodes the compatibility of the values $\mathbf{x}_c$ in each clique (larger $\psi_c(\mathbf{x}_c) \Longrightarrow$ configurations $\mathbf{x}_c$ more likely to occur)
- All variables in clique $\mathbf{x}_c$ can play same role in defining the value of $\psi_c(\mathbf{x}_c)$

Factorization and conditional independence

- MRFs allow to easily assess conditional independence (graph separation!)
- A factorization of the joint distribution

How do we connect conditional independence and factorization?

For the graph to represent conditional independence, $\psi_c(\mathsf{x}_c) > 0$.

- $\mathcal{U}_\mathsf{I}$: Set of distributions $p(\boldsymbol{x})$ consistent with conditional independence defined by graph
- $\mathcal{U}_\mathsf{F}$: Set of distributions $p(\boldsymbol{x})$ that factorize as

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_c \psi_c(\mathbf{x}_c)$$

Hammersley-Clifford theorem: $\mathcal{U}_\mathsf{I}$ and $\mathcal{U}_\mathsf{F}$ are identical.

Factorization and conditional independence

For the graph to represent conditional independence, $\psi_c(\mathsf{x}_c) > 0$.

Idea: Parametrize $\psi_c(\mathsf{x}_c)$ using energy-based form

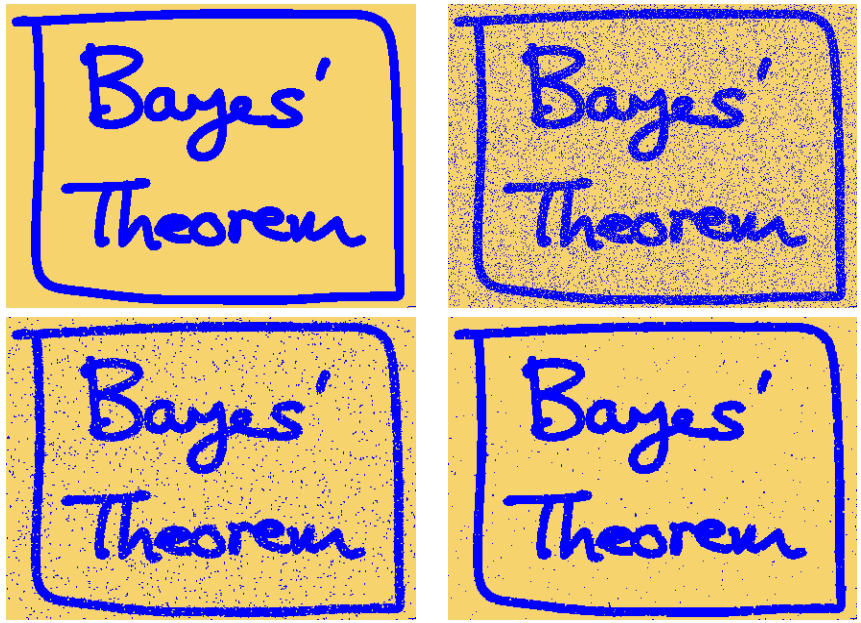
$$\psi_c(\mathsf{x}_c) = e^{-E(\mathsf{x}_c)}$$

Hence:

$$p(\boldsymbol{x}) = \frac{1}{Z}\prod_c \psi_c(\mathbf{x}_c) = \frac{1}{Z}e^{-\sum_c E(\mathsf{x}_c)}$$

Total energy obtained by adding energies of each maximal clique.

Application of MRFs: Image denoising



Goal: Given a noisy image ($\boldsymbol{y}$), recover original noise-free image ($\boldsymbol{x}$).

- Encode images using an array representing numerical values of pixels
- Pixels take on values $+1, -1$, $x_i \in \{+1, -1\}$
- $\{y_i\}$: distorted version of $\{x_i\}$

Image denoising: Graphical model

Assumptions:

- Neighboring (noiseless) pixels are strongly correlated
- Noise acts independently on each pixel
- If noise level low: strong correlation between x_i and y_i

A Markov random field!

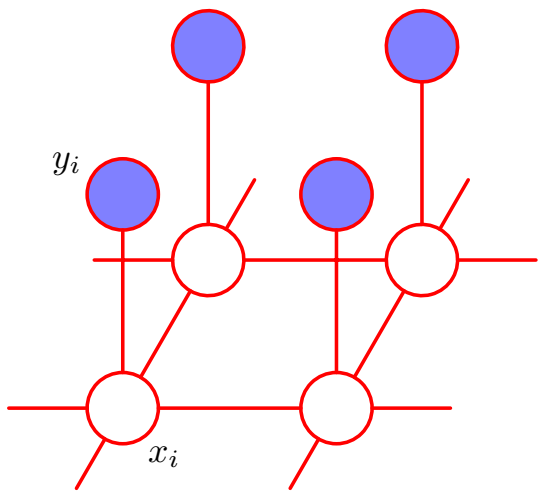
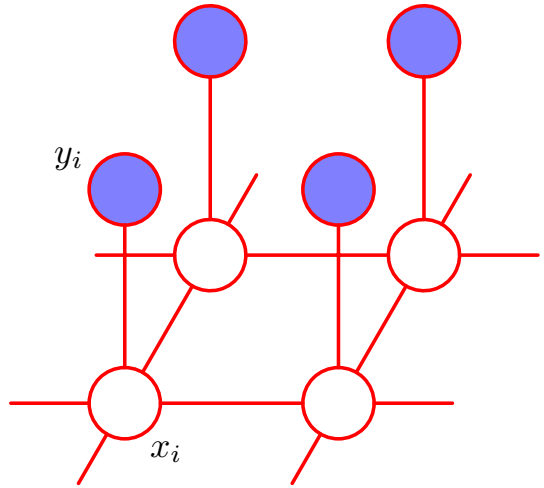


Image denoising: Graphical model

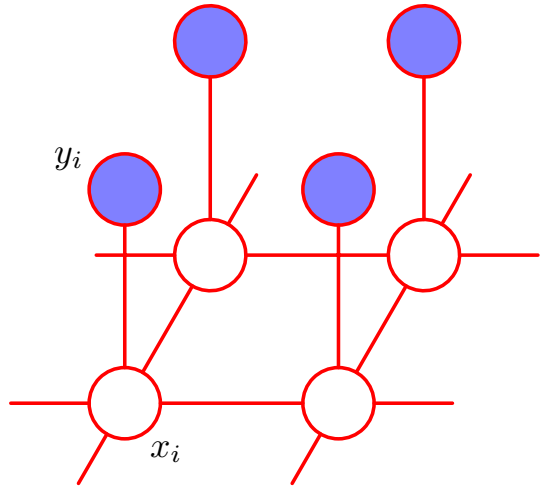


- Two types of cliques: $\{x_i, x_j\}$ and $\{x_i, y_i\}$ connected by edges
- Factorization:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \prod_{i,j} \psi_{i,j}(x_i, x_j) \prod_i \psi_i(x_i, y_i)$$

- We will assume $\mathsf{x}_i, \mathsf{y}_i \in \{+1, -1\}$ (Ising model)

Image denoising: Graphical model



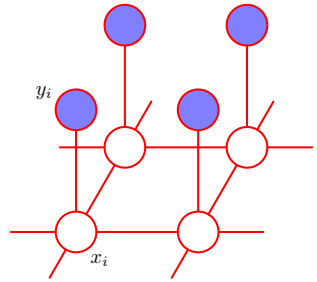
- Cliques $\{x_i, y_i\}$: $E(x_i, y_i)$ expresses correlation between x_i and y_i,

$$E(x_i, y_i) = -\eta x_i y_i, \quad \eta > 0 \quad \Longrightarrow \quad \psi_i(x_i, y_i) = e^{\eta x_i y_i}$$

- Cliques $\{x_i, x_j\}$:

$$E(x_i, x_j) = -\beta x_i x_j, \quad \beta > 0 \quad \Longrightarrow \quad \psi_{i,j}(x_i, x_j) = e^{\beta x_i x_j}$$

Image denoising: Graphical model



$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \prod_{i,j} \psi_{i,j}(x_i, x_j) \prod_i \psi_i(x_i, y_i)$$

- An energy term hx_i is often added for each pixel i of $\boldsymbol{x}$ to biasing the model toward pixel values of a particular sign,

$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \prod_i \psi_i(x_i) \prod_{i,j} \psi_{i,j}(x_i, x_j) \prod_i \psi_i(x_i, y_i)$$

  with $\psi_i(x_i) = e^{-E(x_i)} = e^{-hx_i}$

- Equivalently,

$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \exp\left(-h \sum_i x_i + \beta \sum_{i,j} x_i x_j + \eta \sum_i x_i y_i \right)$$

Image denoising

Goal: Given a noisy image $\boldsymbol{y}$, recover the original noise-free image $\boldsymbol{x}$.

Corresponds to maximizing

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left(-h\sum_i x_i + \beta\sum_{i,j} x_i x_j + \eta\sum_i x_i y_i\right) = \exp(-E(\boldsymbol{x},\boldsymbol{y}))$$

Some notes:

- $h = 0$: prior probabilities of the two states of x_i are equal
- $\beta = 0$: removes links between neighboring pixels $\longrightarrow$ most probable solution is $x_i = y_i\ \forall i$ (observed noisy image)
- Can be solved iteratively (iterated conditional modes)

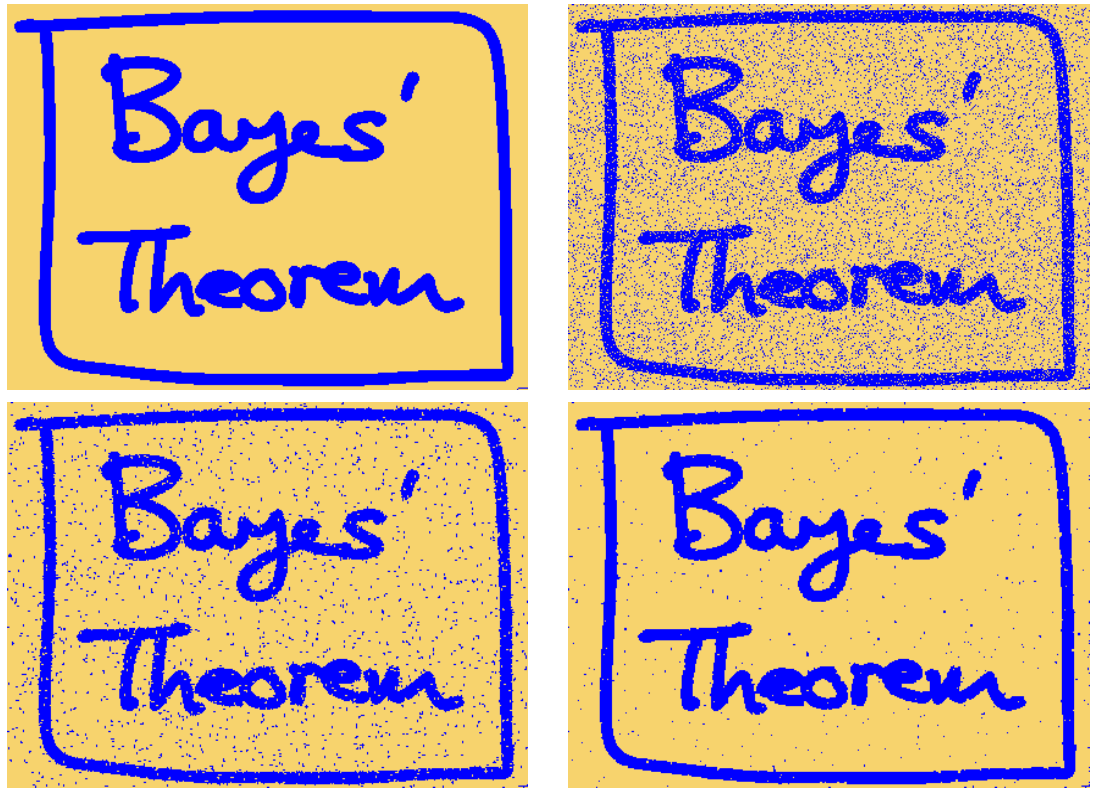
Image denoising

Iterated conditional modes:

Initialization: $x_i = y_i$ for all i

1. For $j = 1, \ldots, N$
   - Evaluate total energy for states $x_j = +1$ and $x_j = -1$ keeping all other variables fixed
   - Set x_j to state with lower energy
2. Repeat 1 until convergence or stopping criterion

Application of MRFs: Image denoising



$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \exp\left(-h \sum_i x_i + \beta \sum_{i,j} x_i x_j + \eta \sum_i x_i y_i\right)$$

- $\beta = 1.0$, $\eta = 2.1$ and $h = 0$

Markov random fields

Useful for

- Modeling mutual relationships of compatibility among variables
- Imaging and spatial applications

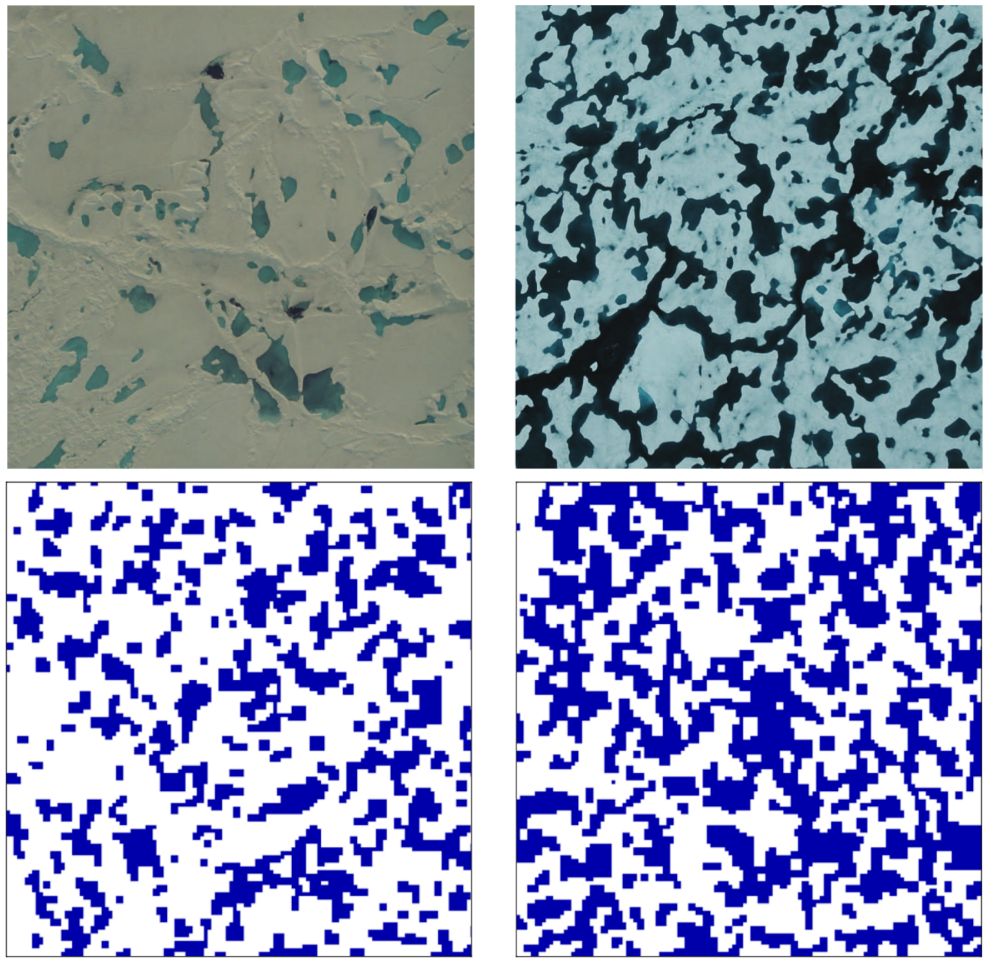
Pitfalls:

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_c \psi_c(\mathsf{x}_c)$$

- Difficult to evaluate joint probability distribution and sample from it
- Computing Z intractable for large alphabets
- Ancestral sampling not possible

Markov random fields: Example

Modeling of Arctic ice melting



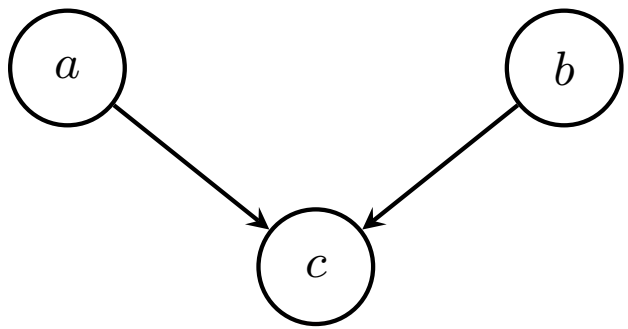
Y.P. M, I. Sudakov, C. Strong, and K. M. Golden, "Ising model for melt ponds on Arctic sea ice," New Journal of Pyhisics, 2019

From BNs to MRFs

BNs and MRFs encode different types of statistical dependencies:

- BNs: capture causality
- MRFs: capture mutual compatibility

Independencies captured by a BN cannot always be captured by a MRF:



(a $\perp$ b$|\emptyset$ and a $\not\perp$ b|c)

From BNs to MRFs

BN with factorization

$$p(\boldsymbol{x}) = \prod_{k=1}^{K} p(x_k | x_{\mathcal{P}(x_k)})$$

Defining

$$\psi_k(x_k, x_{\mathcal{P}(x_k)}) = p(x_k | x_{\mathcal{P}(x_k)})$$

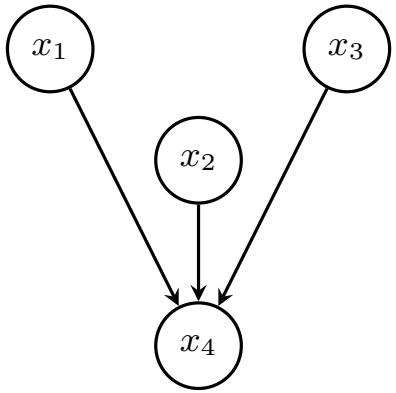
we obtain

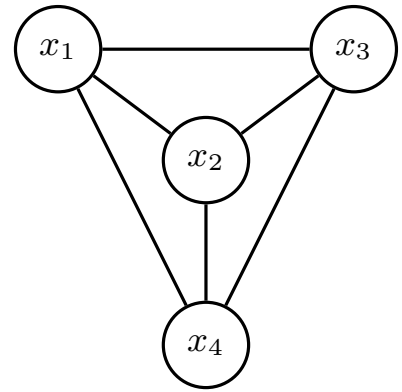
$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{k=1}^{K} \psi_k(x_k, x_{\mathcal{P}(x_k)})$$

BN $\longrightarrow$ MRF

1. Connect all pairs of parents by an undirected edge
2. Make all edges undirected

From BNs to MRFs





$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3) \qquad \psi_{1,2,3,4} = p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)$$

Resulting MRF may not account for all independencies encoded by original BN.

Reading

"Pattern recognition and machine learning,"

Chapter 8 (8.3)