Statistical Machine Learning

Christian Walder

Machine Learning Research Group
CSIRO Data61

and

College of Engineering and Computer Science
The Australian National University

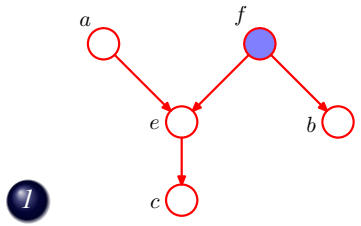
Canberra
Semester One, 2020.

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

©2020
Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

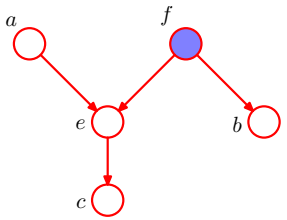
Outlines

Finding $p(a, b \mid f)$ - 'Analytical' method



©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Finding $p(a, b \mid f)$ - 'Analytical' method

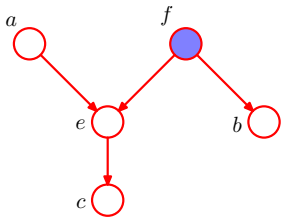


1.
2. Conditional probability with the given observable(s):

$$p(a, b, c, e \mid f)$$

©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Finding $p(a, b \mid f)$ - 'Analytical' method



1.
2. Conditional probability with the given observable(s):

$$p(a, b, c, e \mid f)$$

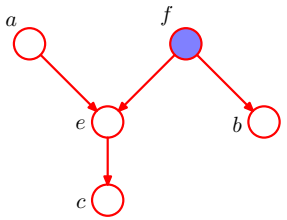
3. Rewrite it as a joint distribution over all variables

$$p(a, b, c, e \mid f) = \frac{p(a, b, c, e, f)}{p(f)}$$

©2020

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Finding $p(a, b \mid f)$ - 'Analytical' method



1. Conditional probability with the given observable(s):

$$p(a, b, c, e \mid f)$$

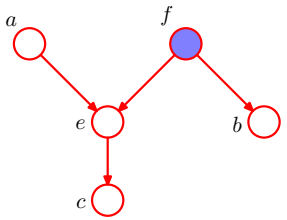
3. Rewrite it as a joint distribution over all variables

$$p(a, b, c, e \mid f) = \frac{p(a, b, c, e, f)}{p(f)}$$

4. Factorise the joint probability according to the graph

$$p(a, b, c, e \mid f) = \frac{p(a, b, c, e, f)}{p(f)} = \frac{p(a)\, p(f)\, p(e \mid a, f)\, p(b \mid f)\, p(c \mid e)}{p(f)}$$

Finding $p(a, b \mid f)$ - 'Analytical' method



1.
2. Conditional probability with the given observable(s):

$$p(a, b, c, e \mid f)$$

3. Rewrite it as a joint distribution over all variables

$$p(a, b, c, e \mid f) = \frac{p(a, b, c, e, f)}{p(f)}$$

4. Factorise the joint probability according to the graph

$$p(a, b, c, e \mid f) = \frac{p(a, b, c, e, f)}{p(f)} = \frac{p(a)\, p(f)\, p(e \mid a, f)\, p(b \mid f)\, p(c \mid e)}{p(f)}$$

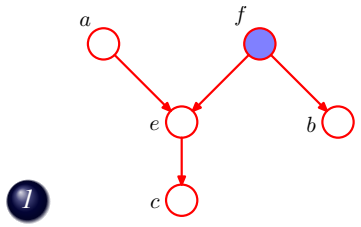
5. Marginalise over all variables we don't care about

$$p(a, b \mid f) = \sum_{c,e} \frac{p(a)\, p(f)\, p(e \mid a, f)\, p(b \mid f)\, p(c \mid e)}{p(f)} = p(a)\, p(b \mid f)$$

©2020

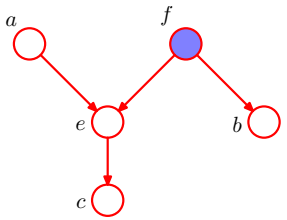
Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Finding $p(a, b \mid f)$ - 'Graphical' method



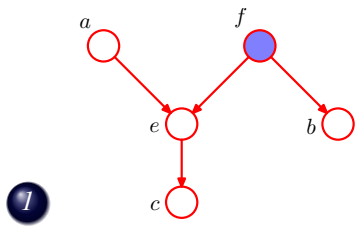
©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Finding $p(a, b \mid f)$ - 'Graphical' method



1.
2. Check whether $a \perp\!\!\!\perp b \mid f$ holds or not.
   Result : $a \perp\!\!\!\perp b \mid f$ holds.
   - Reason : The path from a to b is blocked by f because f is a TT-node and observed. Therefore, $a \perp\!\!\!\perp b \mid f$.

Finding $p(a, b \mid f)$ - 'Graphical' method



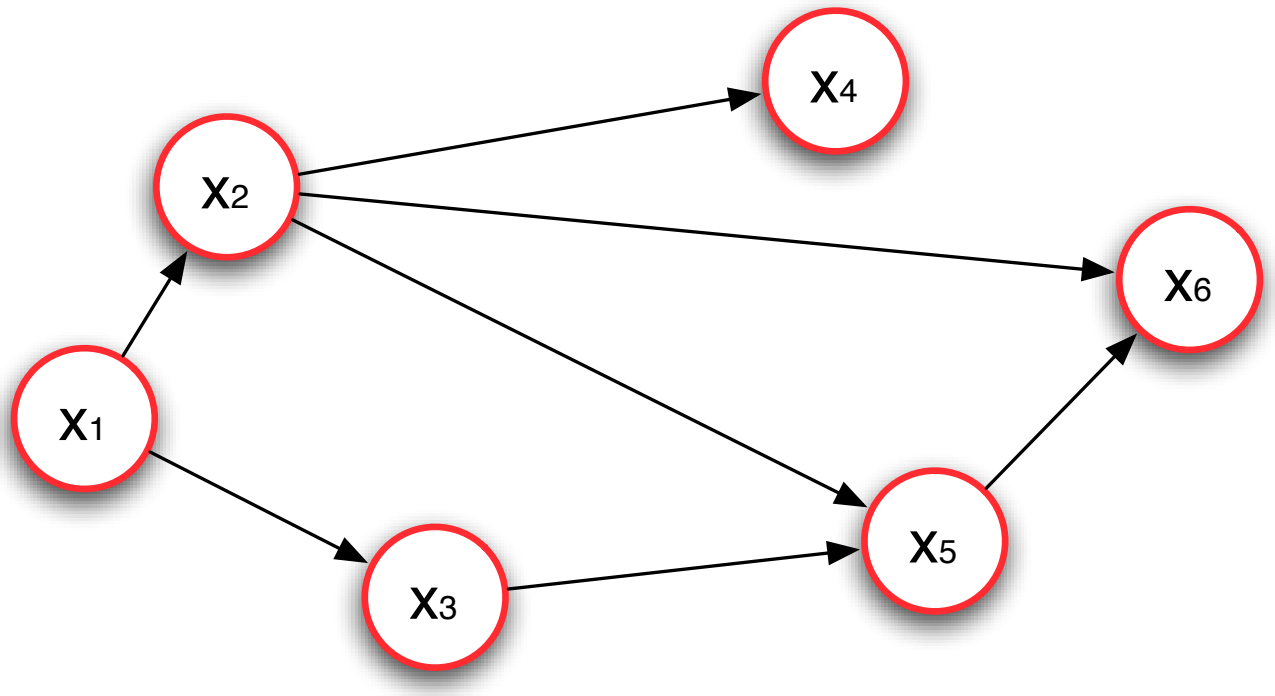
1.
2. Check whether $a \perp\!\!\!\perp b \mid f$ holds or not.
   Result : $a \perp\!\!\!\perp b \mid f$ holds.
   - Reason : The path from a to b is blocked by f because f is a TT-node and observed. Therefore, $a \perp\!\!\!\perp b \mid f$.
3. Write down the factorisation

$$p(a, b \mid f) = p(a \mid f)\, p(b \mid f) = p(a)\, p(b \mid f)$$

Bayesian Network - D-separation

A third example

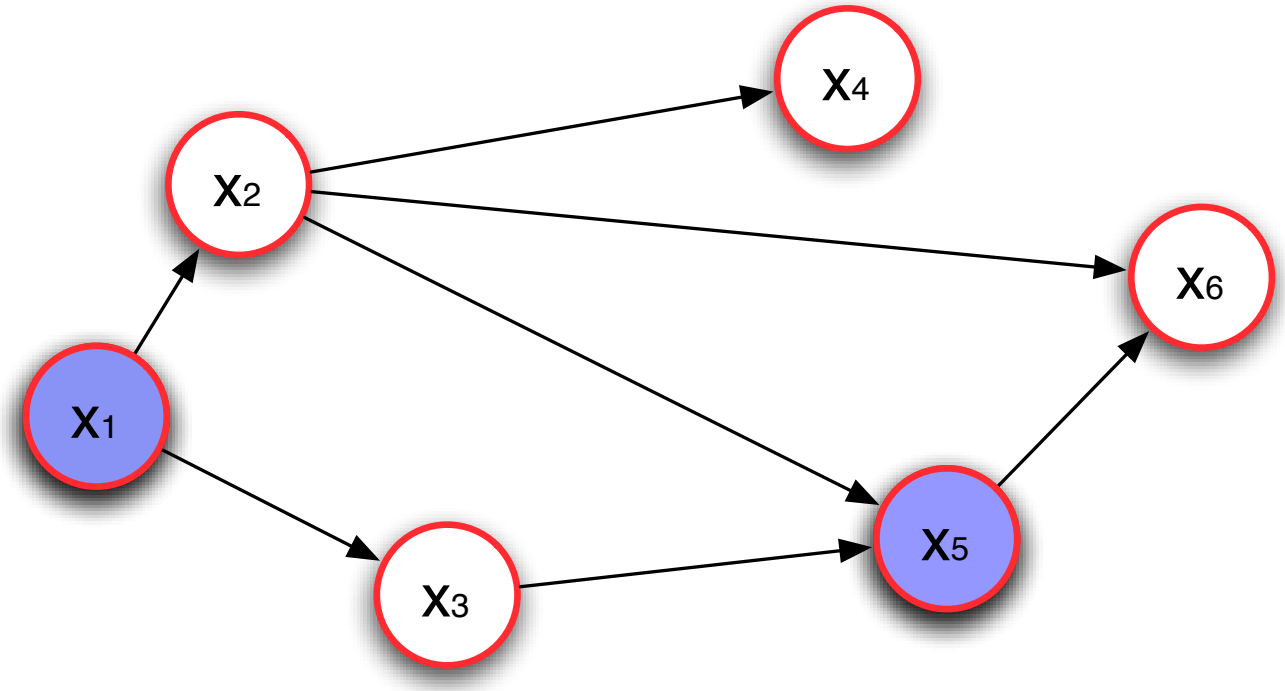
- Is x_3 d-separated from x_6 given x_1 and x_5 ?
- Mark x_1 and x_5 as observed.



©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Bayesian Network - D-separation

- Is x_3 d-separated from x_6 given x_1 and x_5 ?

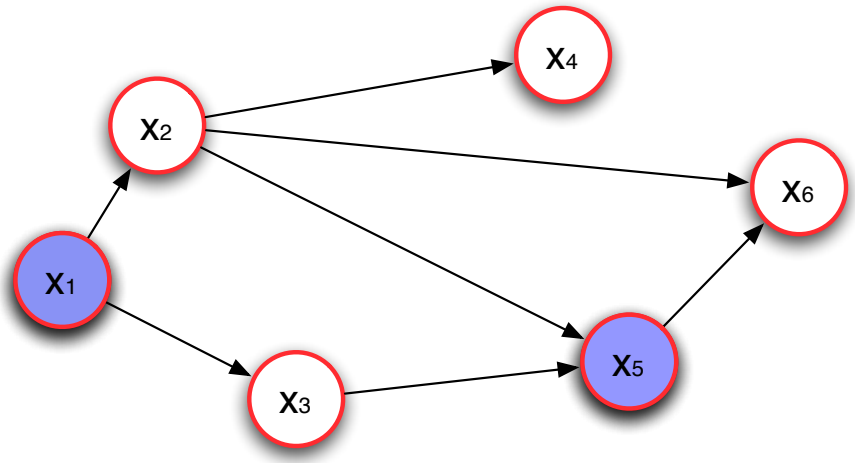


©2020 Ong & Walder & Webers

Data61 | CSIRO
The Australian National University

Bayesian Network - D-separation

- Is x_3 d-separated from x_6 given x_1 and x_5 ?
- 4 junctions of interest between x_3 and x_6:
  - $\{x_3, x_1, x_2\}$ is blocked because x_1 is TT-node and observed.
  - $\{x_3, x_5, x_6\}$ is blocked because x_5 is a HT-node and observed.
  - $\{x_3, x_5, x_2\}$ is not blocked because x_5 is a HH-node and observed.
  - $\{x_5, x_2, x_6\}$ is not blocked because x_2 is a TT-node and unobserved.
- The path $\{x_3, x_5, x_2, x_6\}$ prevents D-separation.
- Therefore, x_3 is not d-separated from x_6 given x_1 and x_5 as not all paths between x_3 and x_6 are blocked.



©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Part XVIII

Probabilistic Graphical Models 2

Markov Random Fields (MRFs)

- Markov Random Fields (Markov network, undirected graphical model) are defined over a graph with undirected edges.
- MRFs allow for different conditional independence statements than Bayesian networks.
- In a Bayesian network, the definition of a blocked path was subtle for a HH-node because it did include that all descendants were unobservable.
- Is there an alternative graphical semantics for probability distributions such that conditional independence is determined by simple graph separation?
- Yes, removing the direction from the edges removes the asymmetry between parent and child nodes and subsequently the subtleties associated with the HH-node.

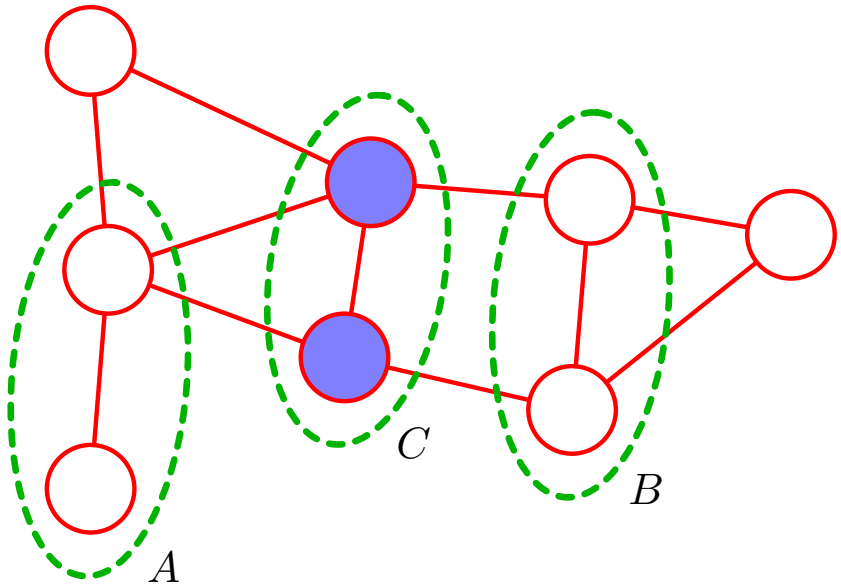
©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Graph Separation

©2020
Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Definition (Graph Separation)

In an undirected graph G, having A, B and C disjoint subsets of nodes, if every path from A to B includes at least one node from C, then C is said to separate A from B in G.



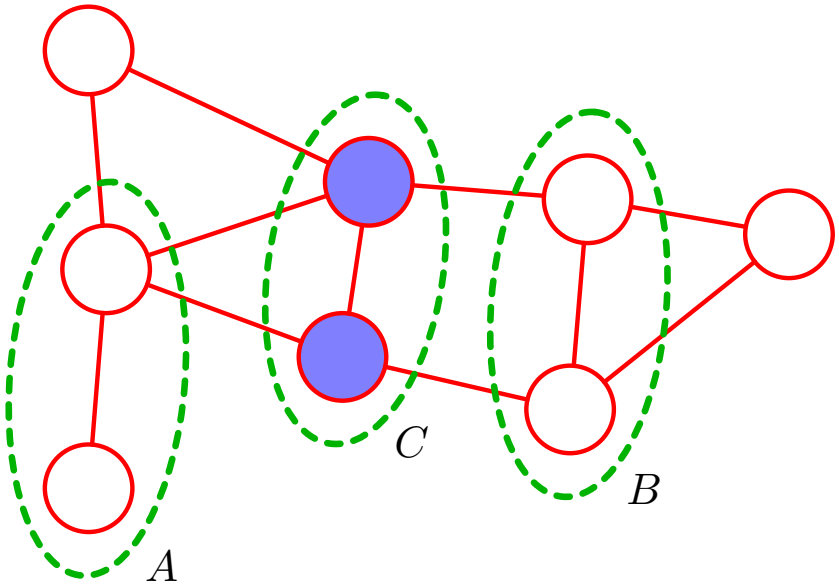
Conditional Independence in MRFs

Definition (Conditional Independence in Markov Random Field)

In an undirected graph G, having A, B and C disjoint subsets of nodes, A is conditionally independent of B given C

$$A \perp\!\!\!\perp B \,|\, C$$

iff C separates A from B in G.



©2020

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Markov Random Field

©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Definition (Markov Random Field)

A Markov Random Field is a set of probability distributions $\{ p(\mathbf{x}) \mid p(\mathbf{x}) > 0, \forall \mathbf{x} \}$ such that there exists an undirected graph G with disjoint subsets of nodes A, B and C, in which whenever C separates A from B in G,

$$A \perp\!\!\!\perp B \,|\, C.$$

- Although we sometimes say "the MRF is such an undirected graph", we mean "the MRF represents the set of all probability distributions whose conditional independency statements are precisely those given by graph separation in the graph".

Factorisation in MRFs

- Assume two nodes x_i and x_j that are not connected by an edge.
- Given all other nodes in the graph, x_i and x_j must be conditionally independent as all paths between x_i and x_j are blocked by observed nodes.

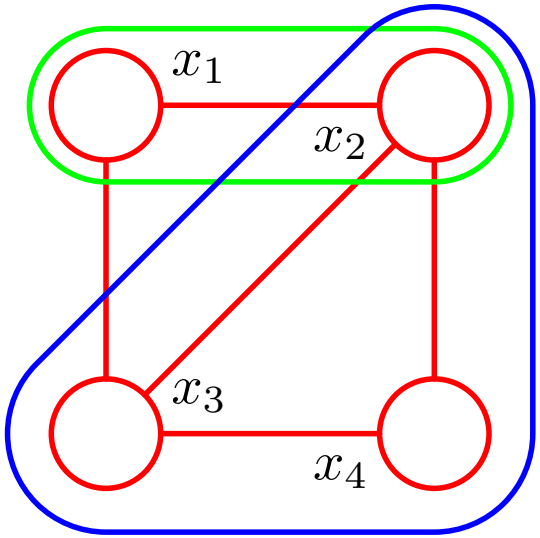
$$p(x_i, x_j \mid \mathbf{x}_{\setminus\{i,j\}}) = p(x_i \mid \mathbf{x}_{\setminus\{i,j\}})\, p(x_j \mid \mathbf{x}_{\setminus\{i,j\}})$$

  where $\mathbf{x}_{\setminus\{i,j\}}$ denotes the set of all variables $\mathbf{x}$ with x_i and x_j removed.
- This is suggestive of the importance of considering sets of connected nodes (cliques). Can we use this for the factorisation of the graph?

©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Cliques in a graph

- A clique is a subset of nodes in a graph such that there exists an edge between all pairs of nodes in the subset. (The nodes in a clique are fully connected.)
- A maximal clique of a graph is a clique which is not a proper subset of another clique. (No other nodes of the graph can be added to a maximal clique without destroying the property of full connectedness.)

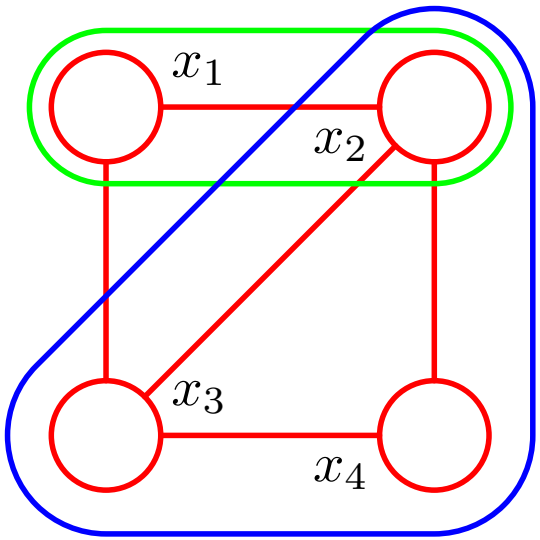


©2020

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Factorisation using Cliques

- We can express the factorisation with the help of the cliques.
- In fact, we only need the maximal cliques, as any function over the variables of a subset of a maximal clique can be expressed as a function of the members of the maximal clique.
- Denote by $\mathcal{C}$ the set of maximal cliques of a graph.
- For a maximal clique $C \in \mathcal{C}$, denote by $\mathbf{x}_C$ the subset of the variables $\mathbf{x}$ which belong to C.



Factorisation using Cliques

- A probability distribution $p(\mathbf{x})$ factorises with respect to a given undirected graph G if it can be written as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C)$$

where $\mathcal{C}$ is the set of maximal cliques of G, and potential functions $\psi_C(\mathbf{x}_C) \geq 0$. The constant $Z = \sum_{\mathbf{x}} p(\mathbf{x})$ ensures the correct normalisation of $p(\mathbf{x})$.

©2020
Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Conditional Independence $\Leftrightarrow$ Factorisation

Theorem (Factorisation $\Rightarrow$ Conditional Independence)

If a probability distribution factorises according to an undirected graph, and if A, B and C are disjoint subsets of nodes such that C separates A from B in the graph, then the distribution satisfies $A \perp\!\!\!\perp B \mid C$.

Theorem (Conditional Independence $\Rightarrow$ Factorisation (Hammersley-Clifford Theorem))

If a strictly positive probability distribution $p(\mathbf{x}) > 0, \forall\, \mathbf{x}$, satisfies the conditional independence statements implied by graph separation over a particular undirected graph, then it also factorises according to the graph.

©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Factorisation with strictly positive potential functions

- As the potential functions $\psi_C(\mathbf{x}_C)$ are strictly positive, one can express them as exponential

$$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\}$$

  of an energy function $E(\mathbf{x}_C)$.
- The exponential distribution is called the Boltzmann distribution.
- The joint distribution is defined as the product of the potentials, and so the total energy is obtained by adding the energies of each of the maximal cliques.

$$p(\mathbf{x}) = \frac{1}{Z}\prod_{C\in\mathcal{C}}\psi_C(\mathbf{x}_C) = \frac{1}{Z}\exp\left\{-\sum_{C\in\mathcal{C}}E(\mathbf{x}_C)\right\}$$

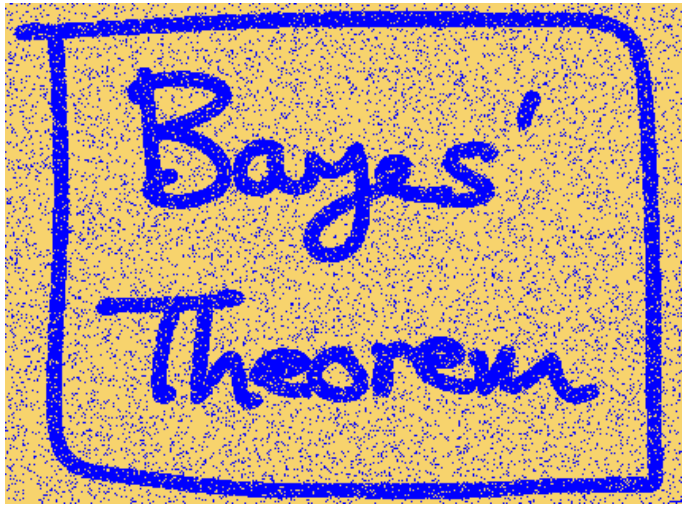
©2020

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Example of Image Denoising

- Given an unknown and noise-free image described by binary pixels $x_i \in \{-1, +1\}$ for $i = 1, \ldots, D$.
- Randomly flip the sign of some pixels with some small probability and denote the pixels of the noisy image by y_i for $i = 1, \ldots, D$.
- Goal : Recover the original noise-free image.

Original image.

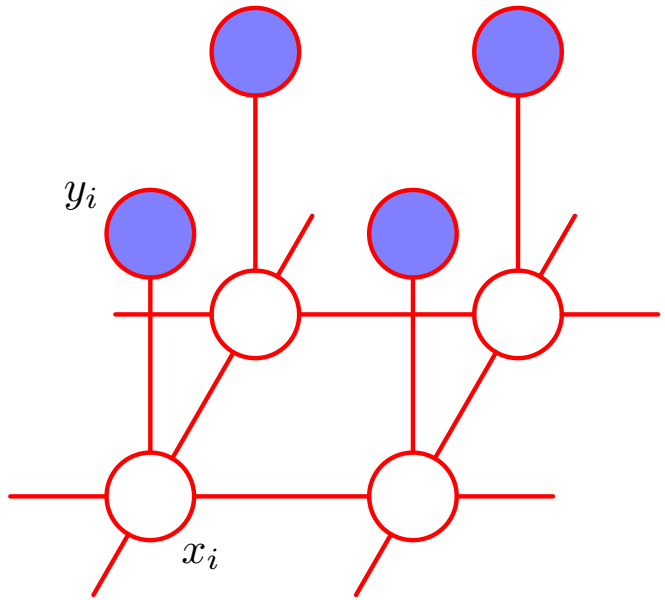


After randomly changing 10% of the pixels.

©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Example of Image Denoising

- Prior knowledge 1 : We know the noise level is small. Therefore: Strong correlation between original pixels x_i and noisy pixels y_i.
- Prior knowledge 2 : Neighbouring pixels x_i and x_j in an image are strongly correlated (given a decent resolution).
- Prior knowledge can be captured in a Markov Random Field.



©2020 Ong & Walder & Webers

Example of Image Denoising

- Cliques $\{x_i, y_i\}$: Choose an energy function of the form $-\eta x_i y_i$ for $\eta > 0$ resulting in a potential function $\exp\{\eta x_i y_i\}$. This favours equal signs for x_i and y_i.
- Cliques $\{x_i, x_j\}$ where i and j are indices of neighbouring pixels : Choose an energy function of the form $-\beta x_i x_j$ for $\beta > 0$ again favouring equal signs for x_i and x_j.
- We need to accomodate for the fact that one kind of pixels might exist more often than the other kind. This can be done by adding a term hx_i for each pixel in the noise-free image. (Potential function is an arbitrary, nonnegative function over maximal cliques, so we are allowed to multiply it by any nonnegative function of subsets of the clique.)

©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

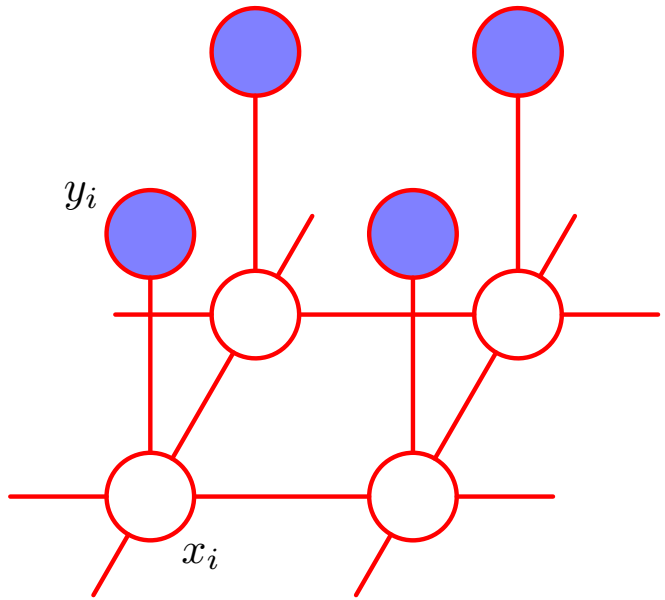
Example of Image Denoising

- Energy function

$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{i,j} x_i x_j - \eta \sum_i x_i y_i$$

- which defines a joint distribution over $\mathbf{x}$ and $\mathbf{y}$

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}.$$

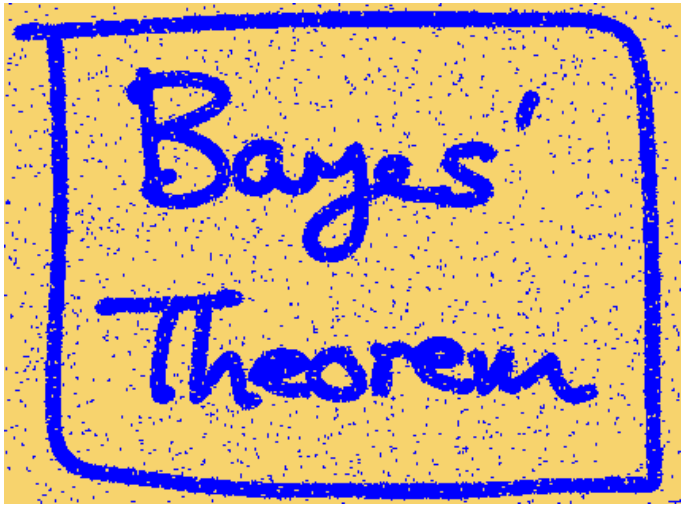


©2020 Ong & Walder & Webers

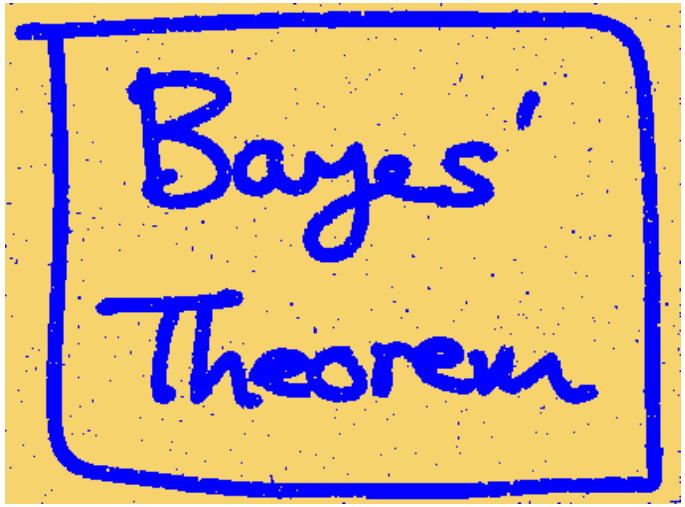
Data61 | CSIRO
The Australian National University

Example - Iterated Conditional Modes (ICM)

1. Fix the elements for $\mathbf{y}$ as we have them observed. (Implicitily defines $p(\mathbf{x} \mid \mathbf{y})$).
2. Initialise $x_i = y_i$ for $i = 1, \ldots, D$.
3. Take one node x_j and evaluate the total energy for both possible states of $x_j = \{-1, +1\}$ keeping all other variables fixed. Set x_j to the state having the lower energy.
4. Repeat for another node until stopping criterion is satisfied.



Local minimum (ICM).



Global minimum (graph-cut).

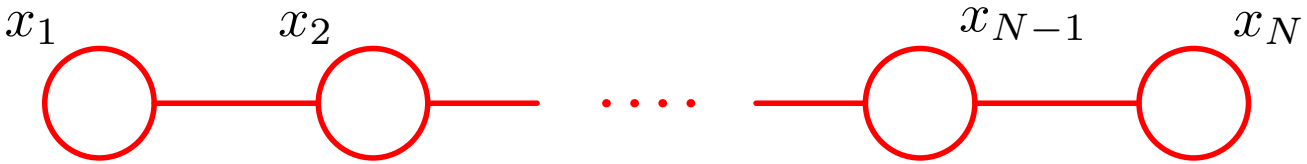
©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Bayesian Networks vs. Markov Random Fields

- Two frameworks for graphical models, can we convert between them?
- Bayesian Network

$$p(\mathbf{x}) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_2) \ldots p(x_N \mid x_{N-1})$$

- Random Markov Field



$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\, \psi_{2,3}(x_2, x_3) \ldots \psi_{N-1,N}(x_{N-1}, x_N)$$

©2020

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Chains are the same

- Bayesian Network

$$p(\mathbf{x}) = p(x_1)\, p(x_2 \,|\, x_1)\, p(x_3 \,|\, x_2) \ldots p(x_N \,|\, x_{N-1})$$

- Random Markov Field

$$p(\mathbf{x}) = \frac{1}{Z} \psi_{1,2}(x_1, x_2)\, \psi_{2,3}(x_2, x_3) \ldots \psi_{N-1,N}(x_{N-1}, x_N)$$

- Find corresponding terms

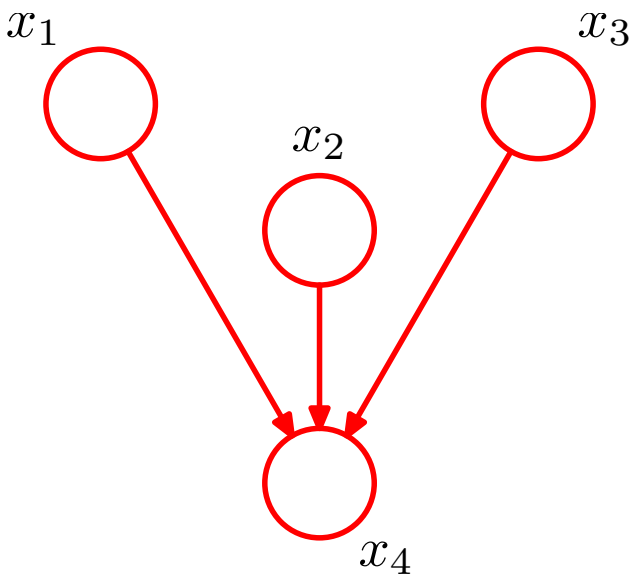
$$\begin{aligned} \psi_{1,2}(x_1, x_2) &= p(x_1)\, p(x_2 \,|\, x_1) \\ \psi_{2,3}(x_2, x_3) &= p(x_3 \,|\, x_2) \\ &\vdots \\ \psi_{N-1,N}(x_{N-1}, x_N) &= p(x_N \,|\, x_{N-1}) \end{aligned}$$

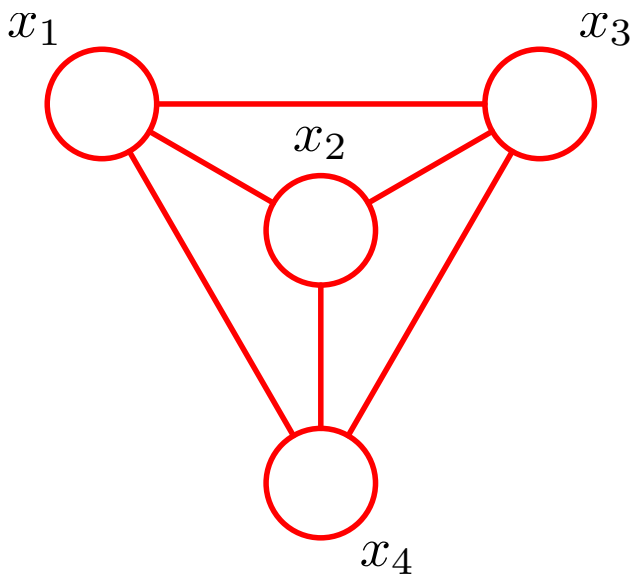
and note that $Z = 1$ in this case.

©2020 Ong & Walder & Webers Data61 | CSIRO The Australian National University

Moralisation

- For other kind of Bayesian Networks (BNs), create the cliques of a MRF by adding undirected edges between all pairs of parents for each node in the graph.
- This process of 'marrying the parents' is called moralisation, and the result is a moral graph.
- BUT the resulting MRF may represent different conditional independence statements than the original BN.
- Example: The MRF is fully connected, and exhibits NO conditional independence properties, in contrast to the original directed graph.





Distribution vs Graph

Definition (D-map)

A graph is a D-map (dependency map) of a distribution if every conditional independence statement satisfied by the distribution is reflected in the graph.

Definition (I-map)

A graph is an I-map (independence map) of a distribution if every conditional independence statement implied by the graph is satisfied in the distribution.

Definition (P-map)

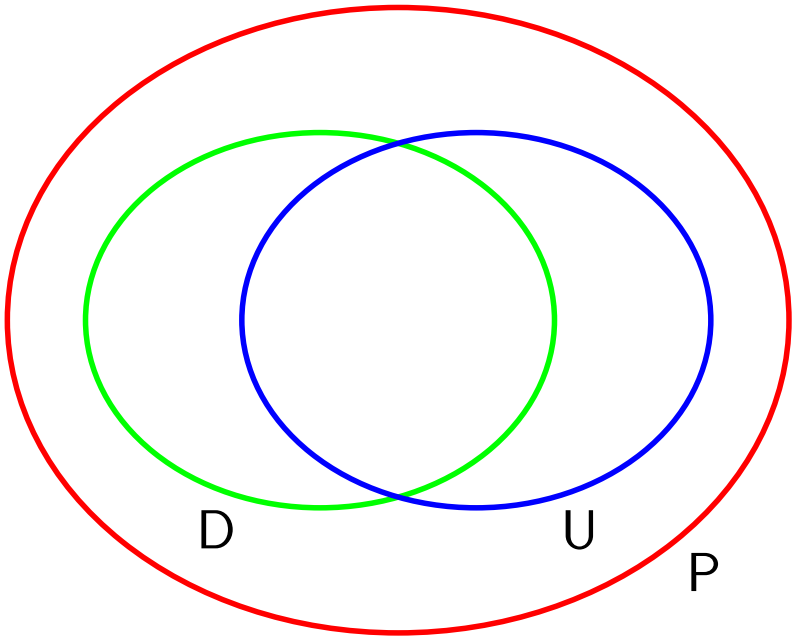
A graph is a P-map (perfect map) of a distribution if it is both a D-map and an I-map for the distribution.

©2020
Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Bayesian Networks vs. Markov Random Fields

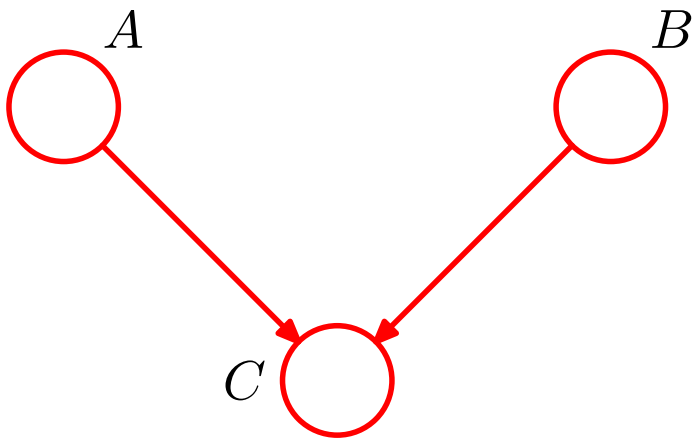
Consider probability distributions depending on N variables

- P set of all probability distributions
- D set of all probability distributions that can be represented as a perfect map by an directed graph
- U set of all probability distributions that can be represented as a perfect map by an undirected graph



©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

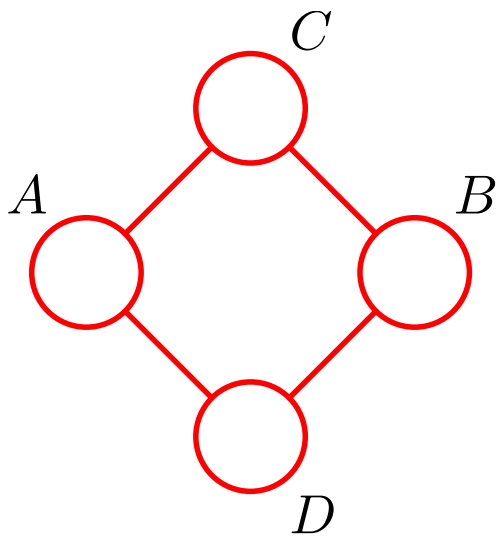
Only as Bayesian Network



A directed graph whose conditional independence properties cannot be expressed using an undirected graph over the same three variables.

©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Only as Markov Random Field



An undirected graph whose conditional independence properties cannot be expressed using a directed graph over the same four variables.

©2020

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

BNs vs. MRFs: Similarities

In both types of Graphical Models

- A relationship between the conditional independence statements satisfied by a distribution and the associated simplified algebraic structure of the distribution is made in term of graphical objects.
- The conditional independence statements are related to concepts of separation between variables in the graph.
- The simplified algebraic structure (factorisation of $p(\mathbf{x})$) is related to 'local pieces' of the graph (child + its parents in BNs, cliques in MRFs).

©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

BNs vs. MRFs: Differences

©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Differences

- The set of probability distributions that can be represented as MRFs is different from the set that can be represented as BNs.
- Although both MRFs and BNs are expressed as a factorisation of local functions on the graph, the MRF has a normalisation constant $Z = \sum_{\mathbf{x}} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C)$ that couples all factors, whereas the BN has not.
- The local 'pieces' of the BN are probability distributions themselves, whereas in MRFs they need only be non-negative functions (i.e. they may not have range $[0, 1]$ as probabilities do).