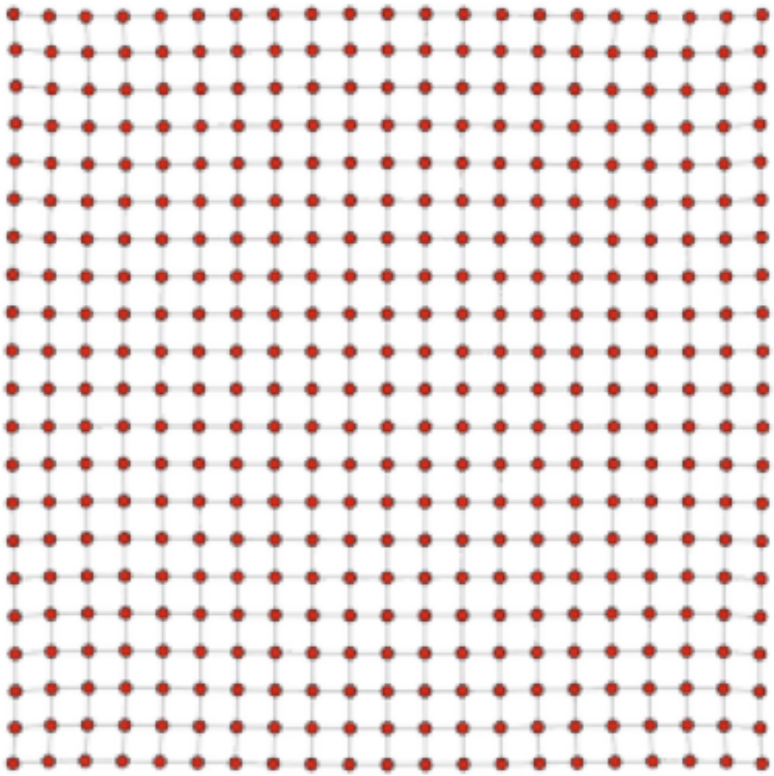
UNDIRECTED MODELS

COMP 4680/8650
Justin Domke
comp8650@anu.edu.au

Reading: Murphy, §19.1-19.5.1

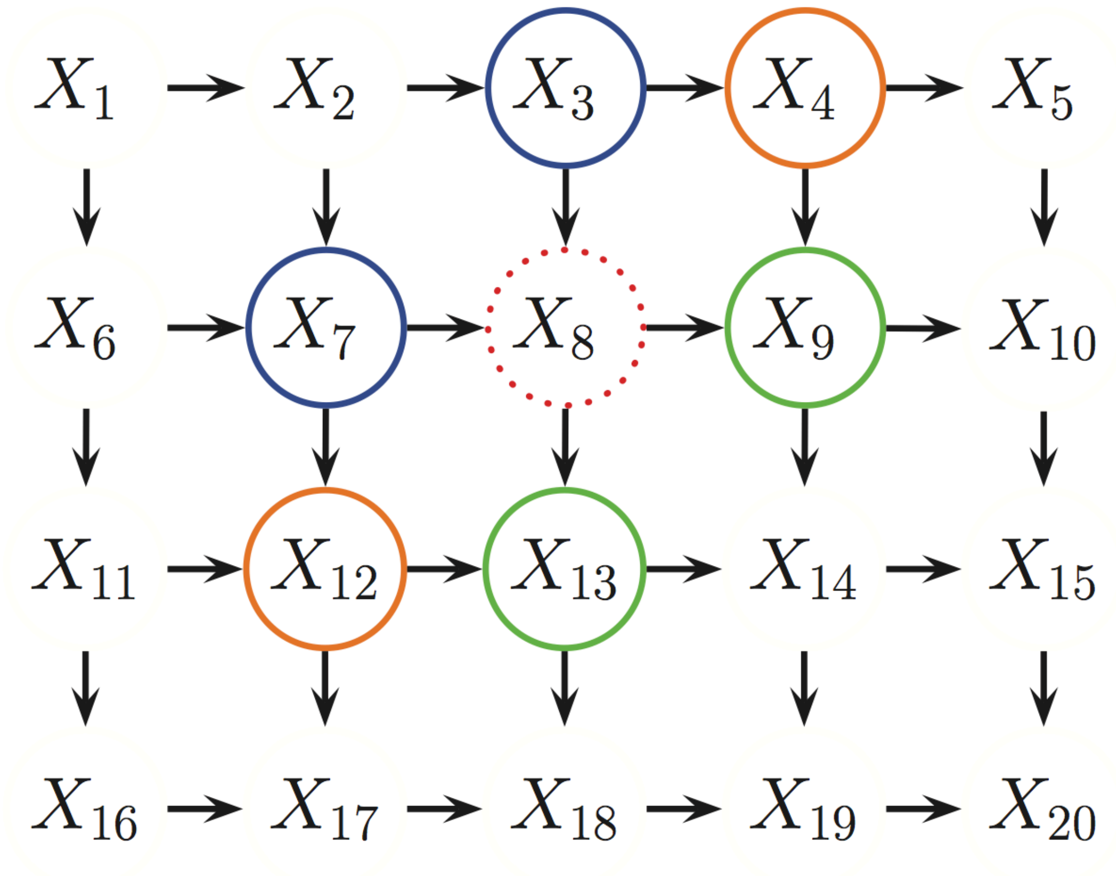
How to model an image?

Or crop yields, or the height of the earth, or…



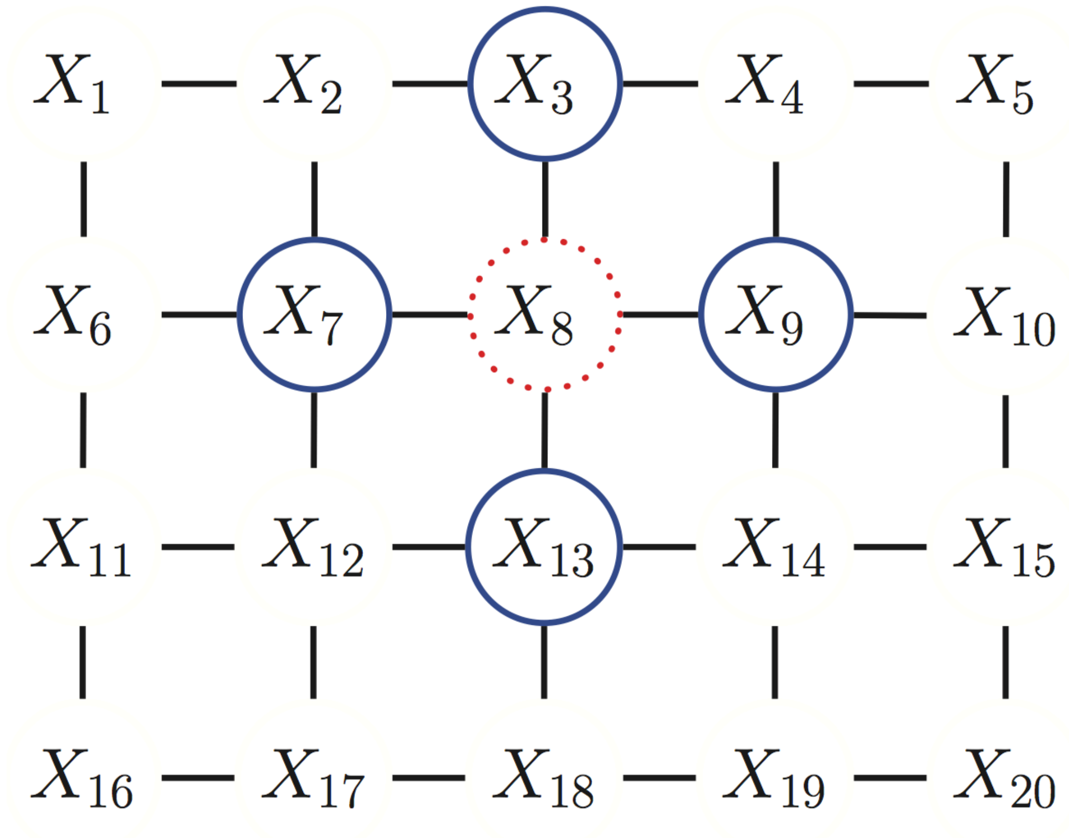
How to model an image?

- Awkward with a directed model



How to model an image?

- Sometimes, prefer the Markov blanket is all neighbors.



Definition

- A probability distribution over x is a Markov random field (or undirected model) with respect to a graph G if

$$x_A \perp x_B | x_C$$ if and only if C separates A from B in G.

for all A, B, C.

Which is true?

$A)\ x_1 \perp x_3 | x_2$

$B)\ x_1 \perp x_3 | x_{2,4}$

$C)\ x_1 \perp x_3 | x_{2,5}$

$D)\ x_6 \perp x_1 | x_{2,3,4,5}$

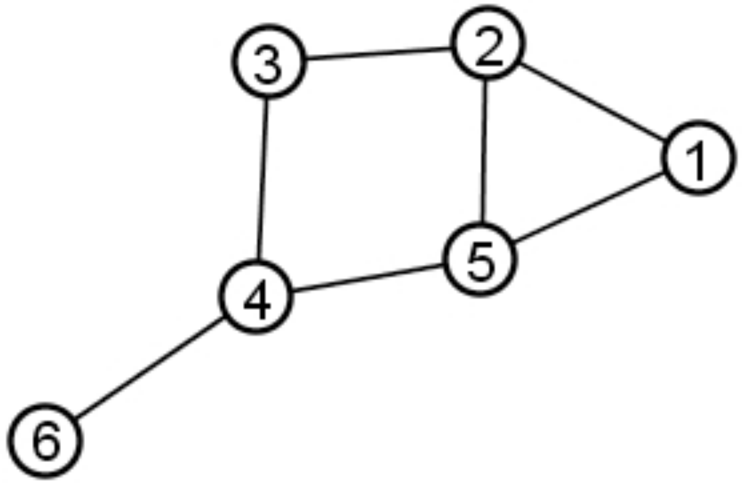
$E)\ x_6 \perp x_1 | x_{2,4}$

$F)\ x_6 \perp x_1 | x_4$

$G)\ x_6 \perp x_1 | x_2$

$H)\ x_{6,1} \perp x_{3,5} | x_{2,4}$

$I)\ x_{6,1} \perp x_{3,5} | x_4$



Which is true?

$\cancel{A)\ x_1 \perp x_3 | x_2}$

true $B)\ x_1 \perp x_3 | x_{2,4}$

true $C)\ x_1 \perp x_3 | x_{2,5}$

true $D)\ x_6 \perp x_1 | x_{2,3,4,5}$

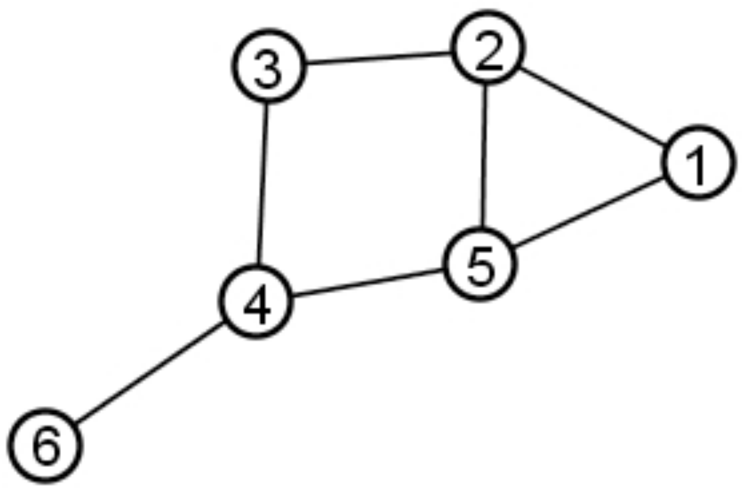
true $E)\ x_6 \perp x_1 | x_{2,4}$

$\cancel{F)\ x_6 \perp x_1 | x_4}$

true $G)\ x_6 \perp x_1 | x_2$

true $H)\ x_{6,1} \perp x_{3,5} | x_{2,4}$

$\cancel{I)\ x_{6,1} \perp x_{3,5} | x_4}$



Definition

- A probability distribution over x is a Markov random field (or undirected model) with respect to a graph G if

$$x_A \perp x_B | x_C \quad \text{if and only if C separates A from B in G.}$$

for all A, B, C.

… But how to we actually write down such a $p(x)$?

How do we write down <u>all</u> such $p(x)$??

Hammersley-Clifford Theorem

- Varying esteem for this theorem among your lecturers.

- History:
  - Proved by Hammersley and Clifford in 1971 using a “blackening algebra”
  - H&C didn’t like the positivity condition. Delayed publishing in the hopes of getting rid of it.
  - Besag proved/published theorem in 1974. (5073 citations to date!)
    - Also, Grimmett, Preston and Sherman, same year.
  - Moussouris in 1974 gave an example with 4 nodes that needs positivity.

Hammersley-Clifford Theorem

- A positive distribution $p(x) > 0$ is an MRF with respect to a graph G if and only if $p(x)$ can be represented as

$$p(x|\theta) = \frac{1}{Z(\theta)} \prod_{c \in \mathcal{C}} \psi_c(x_c|\theta_c)$$

where $\mathcal{C}$ is the set of all cliques, and

$$Z(\theta) = \sum_x \prod_{c \in \mathcal{C}} \psi_c(x_c|\theta_c)$$

is the partition function.

Notes:

- This isn't obvious!
- No direct probabilistic interpretation for ψ.

Hammersley-Clifford Theorem

Proof sketch:

- Assumes $x_i \in 0, 1, 2, ..., K_i$.
- It's easy to show that $p(x|\theta) = \frac{1}{Z(\theta)} \prod_{c \in \mathcal{C}} \psi_c(x_c|\theta_c)$ obeys this conditional independence assumptions of a graph.
- Instead, we start with a arbitrary distribution that obeys the conditional independence assumptions, and show that it can indeed be written like this.

Define $x^* = (0, 0, ..., 0)$ and $Q(x) := \ln(p(x)/p(x^*))$.

Step 1: Can write Q **uniquely** as:

$$Q(x) = \sum_i x_i G_i(x_i) + \sum_{i<j} x_i x_j G_{ij}(x_i, x_j) + \sum_{i<j<k} x_i x_j x_k G_{ijk}(x_i, x_j, x_k)$$
$$+ ... + x_1 x_2 ... x_n G_{12...n}(x_1, x_2, ..., x_n)$$

Step 2: Define $x^i = (x_1, ..., x_{i-1}, 0, x_{i+1}, ..., x_n)$. Then,

$$\exp(Q(x) - Q(x^i)) = \frac{p(x)}{p(x^i)} = \frac{p(x_i | x_{-i})}{p(0 | x_{-i})}$$

Step 3: Pick node 1 w.o.l.o.g. Then, write

$$Q(x) - Q(x^1) = x_1 (G_1(x_1) + \sum_{1<j} x_j G_{1j}(x_1, x_j) + \sum_{1<j<k} x_j x_k G_{1jk}(x_i, x_j, x_k)$$
$$+ ... + x_2 ... x_n G_{12...n}(x_1, x_2, ..., x_n))$$

Step 4: Suppose t is not a neighbor of 1. All terms involving x_t must be zero.

Why? A) $Q(x) - Q(x^1)$ is independent of x_t since $p(x_i | x_{-i})$ is.

B) If we set $x_i = 0, i \notin \{1, t\}$ then G_{1t} must be zero.

C) Similarly for third/fourth/n-th order terms

Example: Ising Model

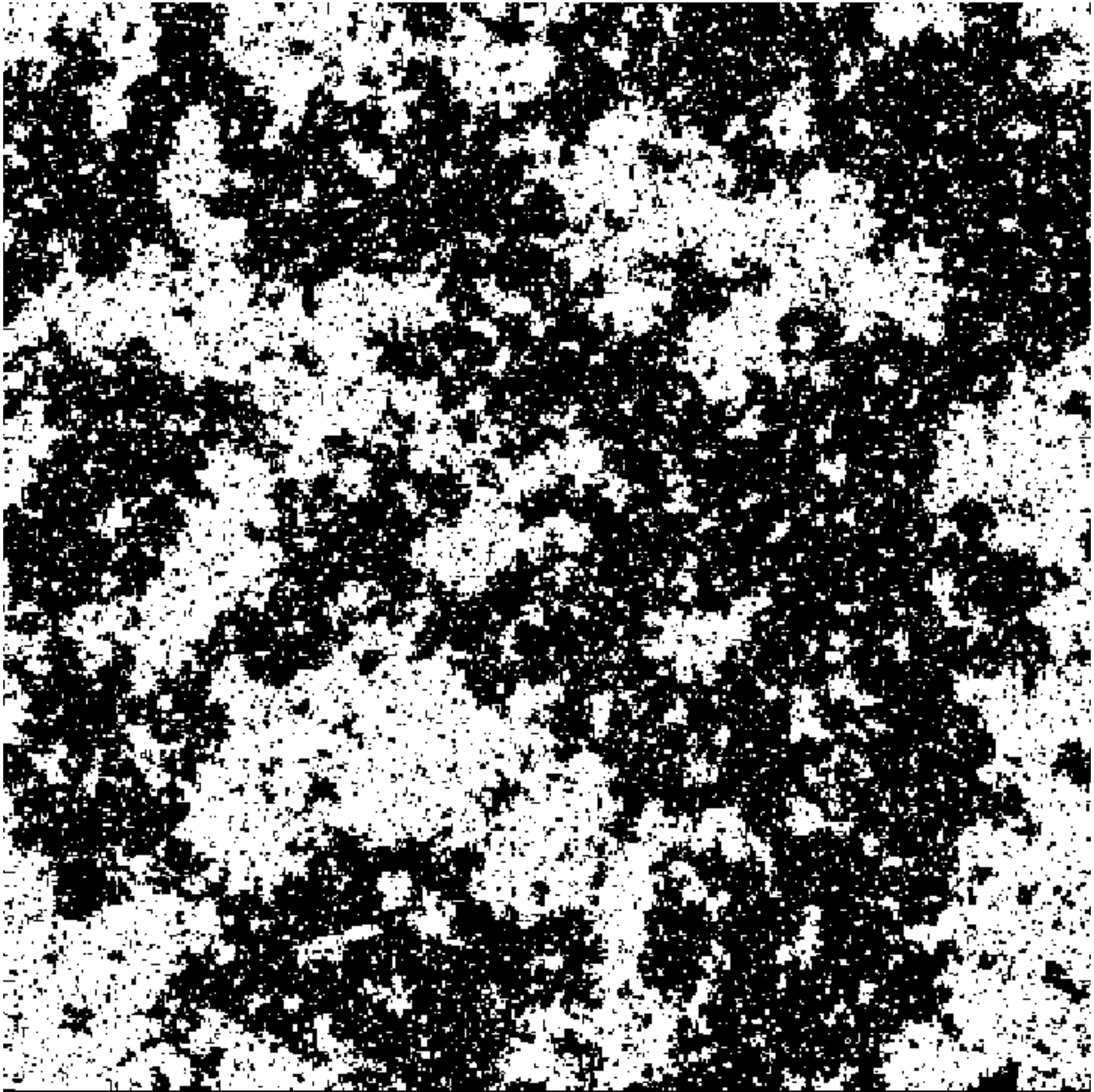
- Invented by Lenz in 1920, given to Ising.
  - Used by physicists to understand phase transitions and magnetism.

- Assume Binary States

$$x_i \in \{-1, +1\}$$

- Univariate (“field”) and pairwise terms

$$p(x) = \frac{1}{Z} \prod_s \exp(b_s x_s) \prod_{(s,t)} \exp(w_{st} x_s x_t)$$



Learning Markov Random Fields

- For now, we are interested in <u>parameter learning</u>.

- Maximum-Likelihood Learning. Given $x^1, x^2, ..., x^D$, we want to pick θ to maximize

$$\prod_{d=1}^{D} p(x^d|\theta)$$

- Convenient to re-formulate this as

$$\arg\max_\theta \frac{1}{D}\sum_{d=1}^{D} \log p(x^d|\theta)$$

- First problem: Given θ, how high does it score?

Learning Markov Random Fields

$$\arg\max_\theta \frac{1}{D}\sum_{d=1}^{D} \log p(x^d|\theta)$$

$$= \arg\max_\theta \frac{1}{D}\sum_{d=1}^{D} \log\left(\frac{1}{Z(\theta)}\prod_{c\in\mathcal{C}} \psi_c(x_c^d|\theta_c)\right)$$

$$= \arg\max_\theta \underline{\frac{1}{D}\sum_{d=1}^{D}\sum_{c\in\mathcal{C}} \log\psi_c(x_c^d|\theta_c)} - \underline{\log(Z(\theta))}$$

This is easy to compute.

This is hard, since

$$Z(\theta) = \sum_x \prod_{c\in\mathcal{C}} \psi_c(x_c|\theta_c)$$

Computing Z

Does this remind you of belief propagation?

- Suppose we have a simple chain:

$$p(x) = \frac{1}{Z}\psi(x_{1,2})\psi(x_{2,3})\psi(x_{3,4})\ldots\psi(x_{n-1,n})$$

$$Z = \sum_x \psi(x_{1,2})\psi(x_{2,3})\psi(x_{3,4})\ldots\psi(x_{n-1,n})$$

How to compute Z? Well, we could define

$$T_i(x_i) = \sum_{x_1,\ldots x_{i-1}} \psi(x_{1,2})\ldots\psi(x_{i-1},x_i)$$

Then we have the simple recurrence that

$$T_1(x_1) = 1 \qquad T_{i+1}(x_{i+1}) = \sum_{x_1,\ldots x_i} T_i(x_i)\psi(x_i,x_{i+1})$$

And, finally,

$$Z = \sum_{x_n} T_n(x_n)$$

What to take home

- Definition of MRFs in terms of conditional independence and separation.
- The Hammersley-Clifford theorem, why it is very surprising and wondeful, and how to prove it.
- How learning an MRF requires computing the (log) partition function, and how the difficulty of computing the log-partition function is connected to the difficulty of computing marginals.