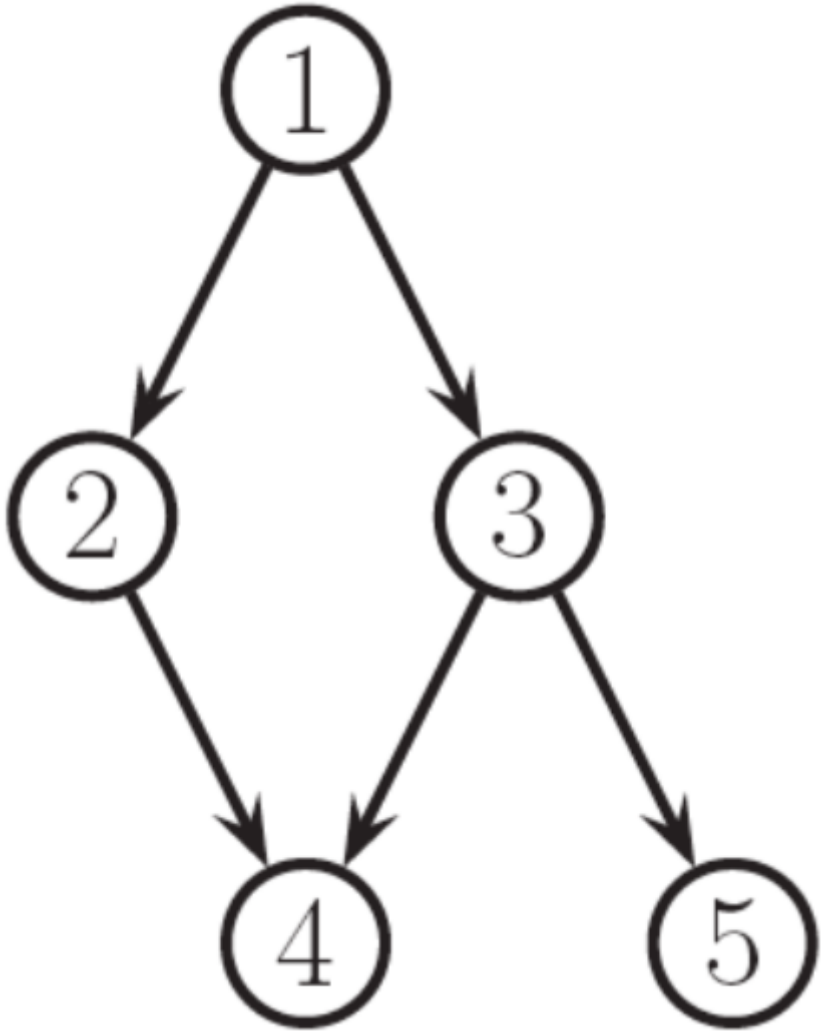
Week 3-1: Markov Random Fields

Assigned Reading

- MLPP: Chapters 19-19.5

Directed Graphical Models (a Review)

So far, we have seen directed acyclic graphical models (DAGMs). These models represent large *joint* distributions using *local* relationships specified by the graph, where each random variable is a **node** and the **edges** specify the *conditional dependence* between random variables (and therefore missing edges imply conditional independence). Graphically, these models looked like



The graph factorizes according to the local conditional probabilities

$$p(x_{1,\ldots,N}) = \prod_{i}^{N} p(x_i|x_{\pi_i})$$

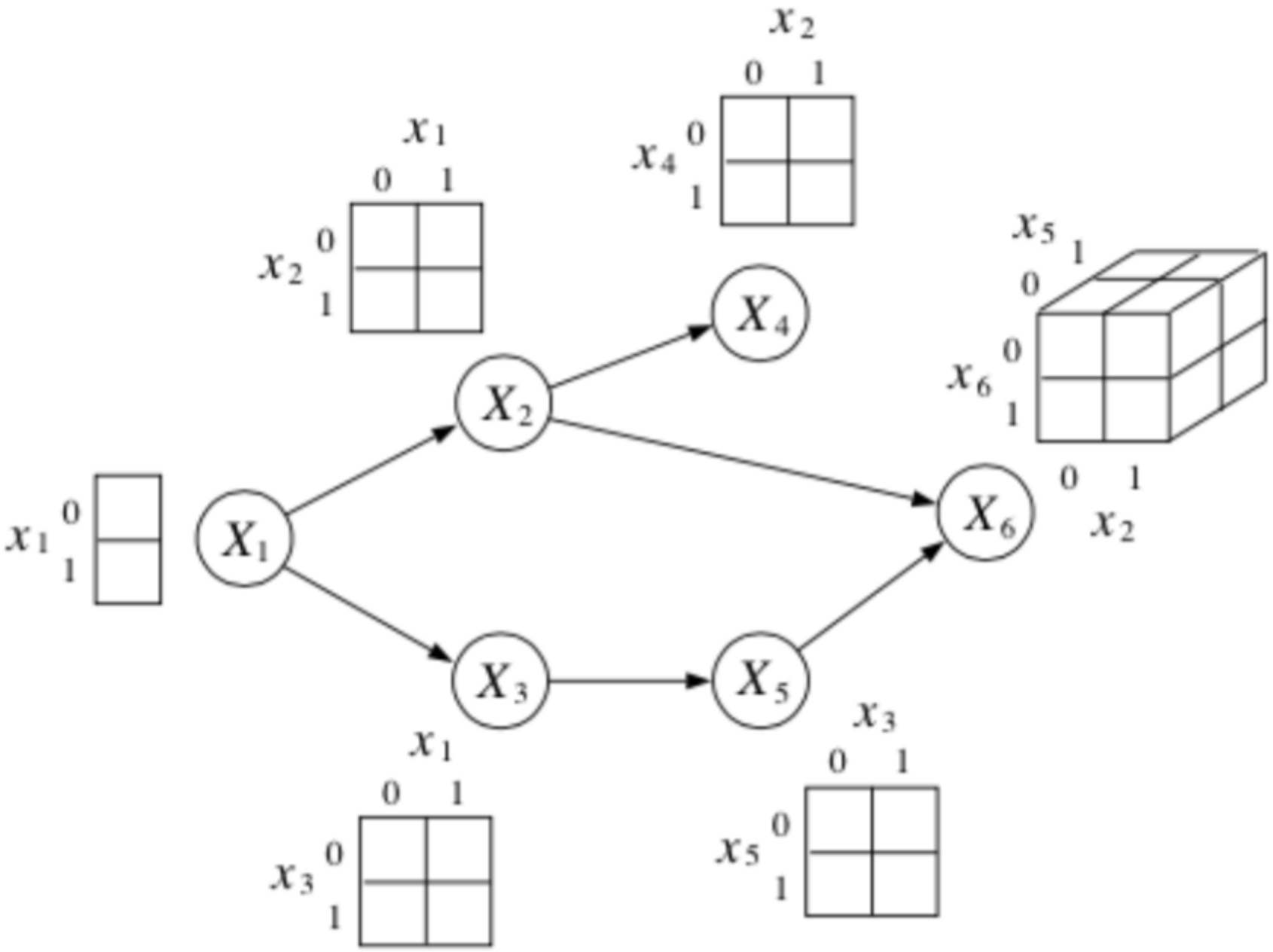
where x_{π_i} are the parents of node x_i.

Each node is conditionally independent of its non-descendents given its parents

$$\{x_i \perp x_{\tilde{\pi}_i} \mid x_{\pi_i}\} \quad \forall i$$

> Recall that this is simply a topological ordering of the graph (i.e. parents have lower numbers than their children)

For discrete variables, *each node* stores a conditional probability table (CPT) of size k^n, where k is the number of discrete states and n the number of parent nodes.



In the DAGM above, X_6 has $k^n = 2^2$ possible configurations.

Are DAGMs always useful?

For some problems, it is not always clear how to choose the direction for the edges in our DAGMs. Take the example of modeling dependencies in an image

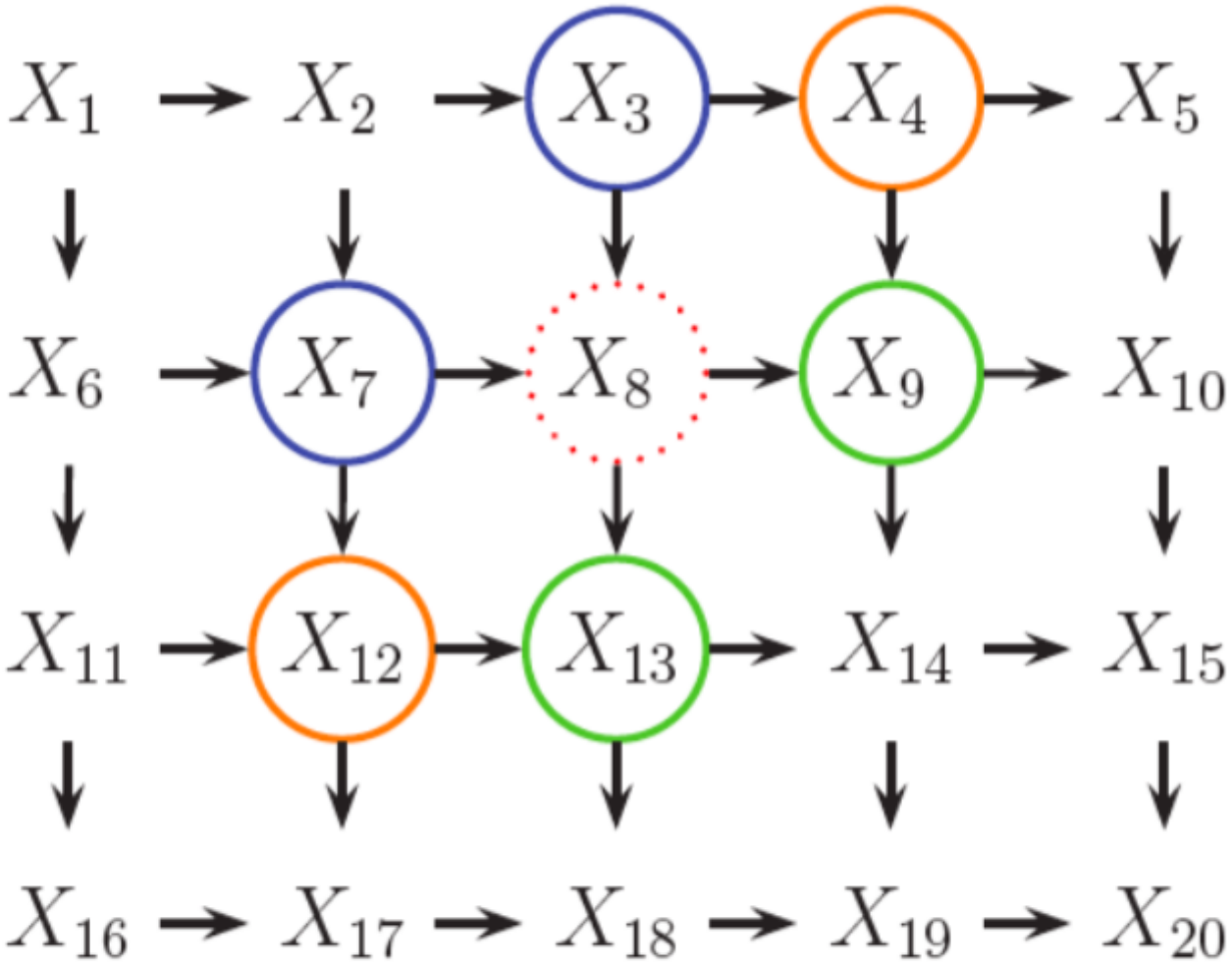


Figure : Causal MRF or a Markov mesh

Our assumptions lead to unatural conditional independence between random variables. Take for example, the Markov blanket of node X_8

$$mb(8) = \{3, 7\} \cup \{9, 13\} \cup \{12, 4\}$$

> The Markov blanket contains the parents, children and co-parents of a node. More generally, it is the set of all variables that shield the node from the rest of the network. I think the point of this example is that one would expect X_2 and X_{14} to be in the Markov blanket $mb(8)$, especially given that X_4 and X_{12} are.

An alternative to DAGMs, is undirected graphical models (UGMs).

Undirected Graphical Models

Undirected graphical models, also called Markov random fields (MRFs) or Markov networks, are a set of random variables described by an undirected graph. As in DAGMs, the **nodes** in the graph represent *random variables*. However, in contrast to DAGMs, **edges** represent *probabilistic interactions* between neighboring variables (as opposed to conditional dependence).

Dependencies in UGMs

In DAGMs, we used conditional probabilities to represent the distribution of nodes given their parents. In UGMs, we use a more *symmetric* parameterization that captures the affinities between related variables.

The following three properties are used to determine if nodes are conditionally independent:

def. **Global Markov Property** (G): $X_A \perp X_B | X_C$ iff X_C separates X_A from X_B

!!! note
That is, there is no path in the graph between A and B that doesn't go through X_C.

def. **Local Markov Property (Markov Blanket)** (L): The set of nodes that renders a node t conditionally independent of all the other nodes in the graph

$$t \perp \mathcal{V} \setminus cl(t) \mid mb(t)$$

where $cl(t) = mb(t) \cup t$ is the closure of node t.

def. **Pairwise (Markov) Property** (P): The set of nodes that renders two nodes, s and t, conditionally independent of each other.

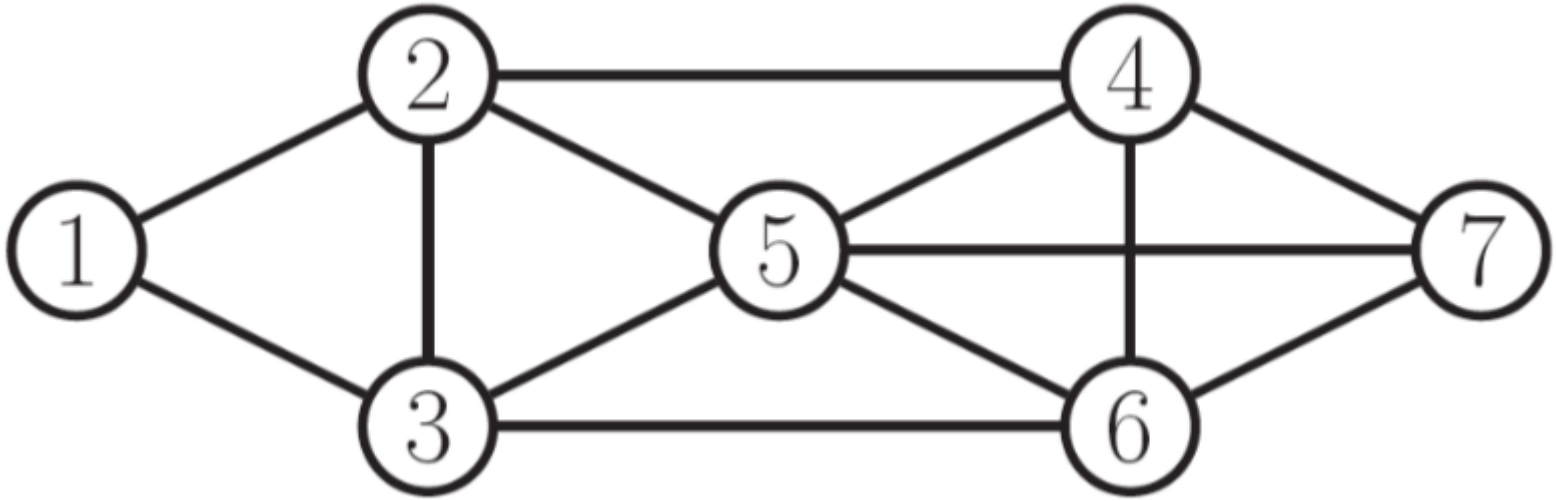
$$s \perp t | \mathcal{V} \setminus \{s, t\} \Leftrightarrow G_{st} = 0$$

!!! note
G_{st} is a function that counts the number of edges between nodes s, t

where

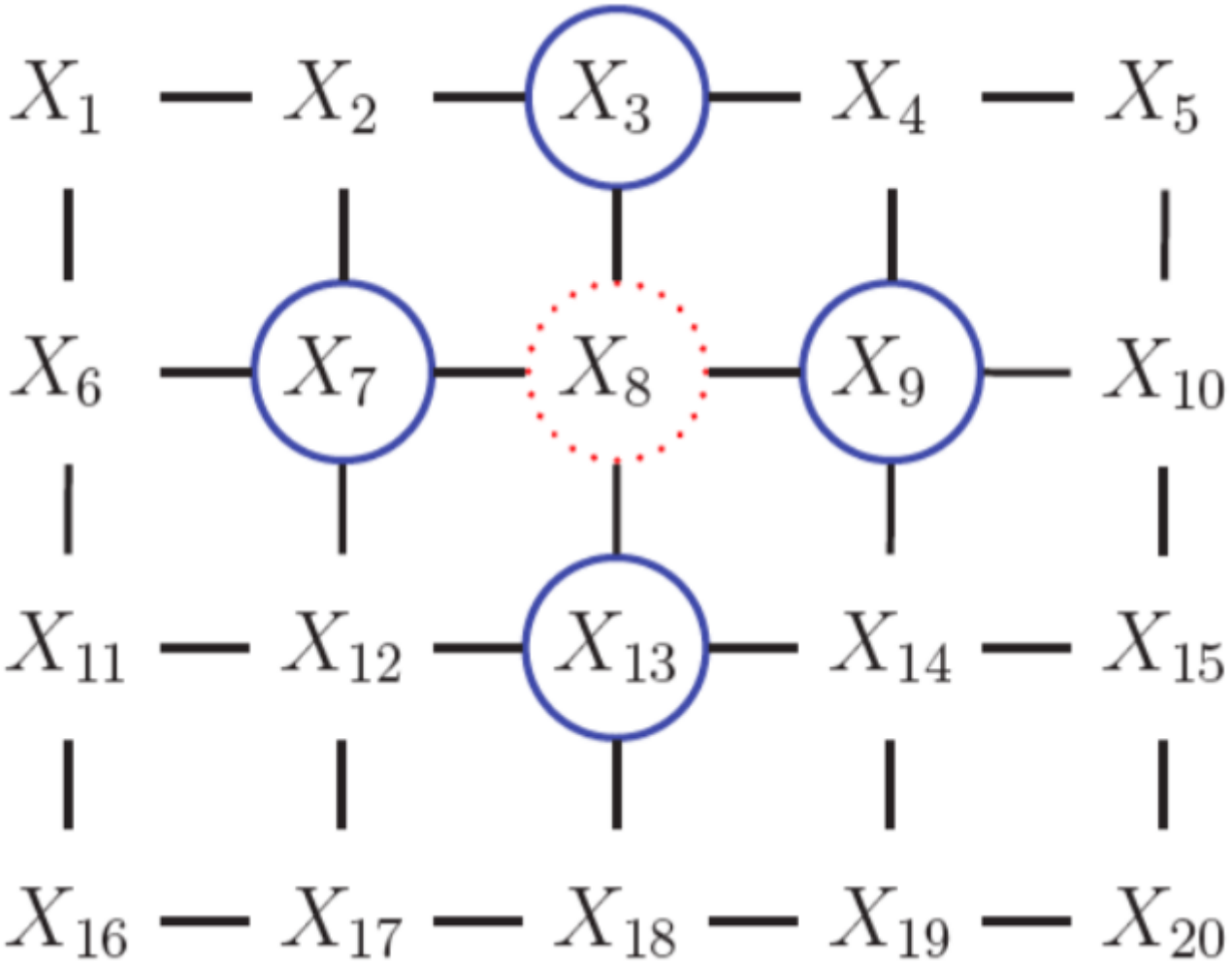
$$G \Rightarrow L \Rightarrow P \Rightarrow G \quad p(x) > 0$$

Simple example



- Global: $\{1, 2\} \perp \{6, 7\} | \{3, 4, 5\}$
- Local: $1 \perp \text{rest} | \{2, 3\}$
- Pairwise: $1 \perp 7 | \text{rest}$

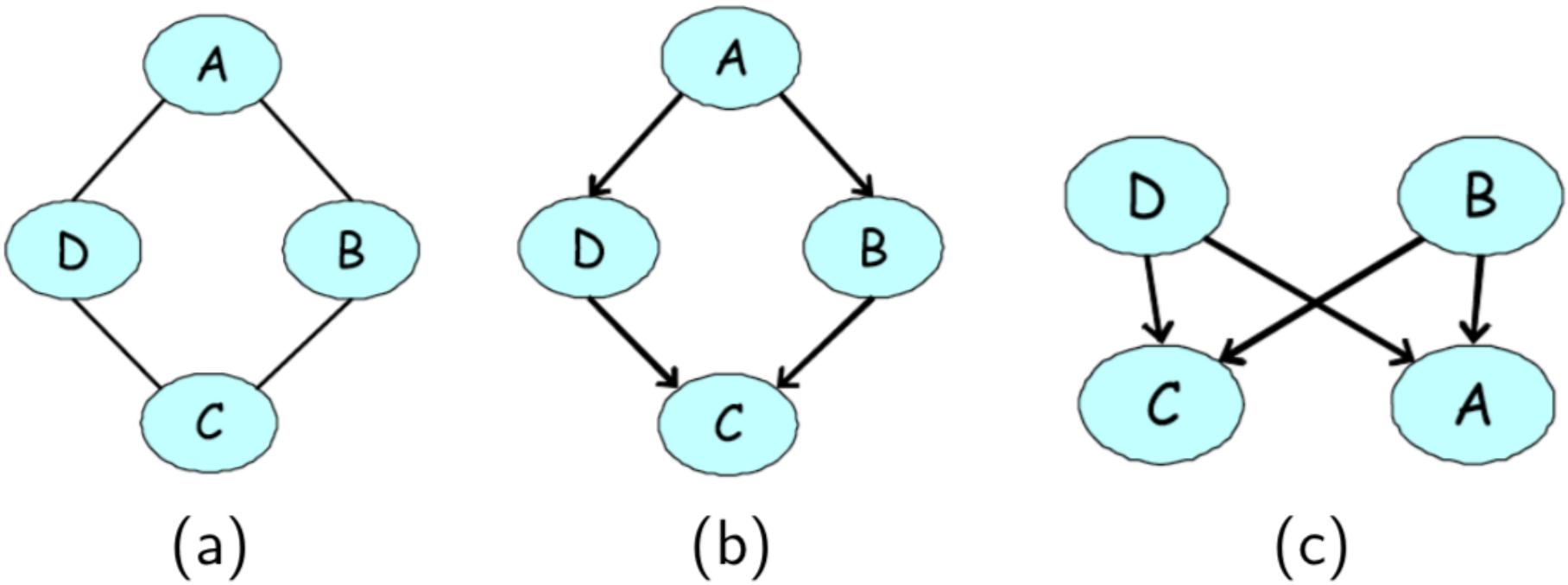
Image example



- Global: $\{X_1, X_2\} \perp \{X_{15}, X_{20}\} | \{X_3, X_6, X_7\}$
- Local: $1 \perp \text{rest} | \{X_2, X_6\}$
- Pairwise: $1 \perp 20 | \text{rest}$

Not all UGMs can be represented as DGMs

Take the following UGM for example (a) and our attempts at encoding this as a DGM (b, c).



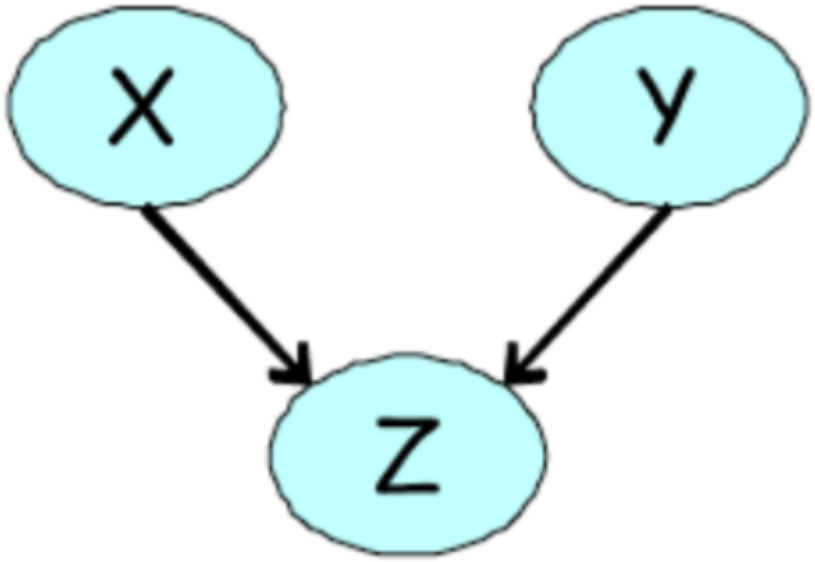
First, note the two conditional independencies of our UGM in (a):

1. $A \perp C | D, B$
2. $B \perp D | A, C$

In (b), we are able to encode the first independence, but not the second (i.e., our DGM implies that B is dependent on D given A and C). In (c), we are again able to encode the first independence, but our model also implies that B and D are marginally independent.

Not all DGMs can be represented as UGMs

It is also true that not all DGMs can be represented as UGMs. One such example is the 'V-structure' that we saw in the **explaining away** case in lecture 3.



An undirected model is unable to capture the marginal independence, $X \perp Y$ that holds at the same time as $\neg(X \perp Y | Z)$.

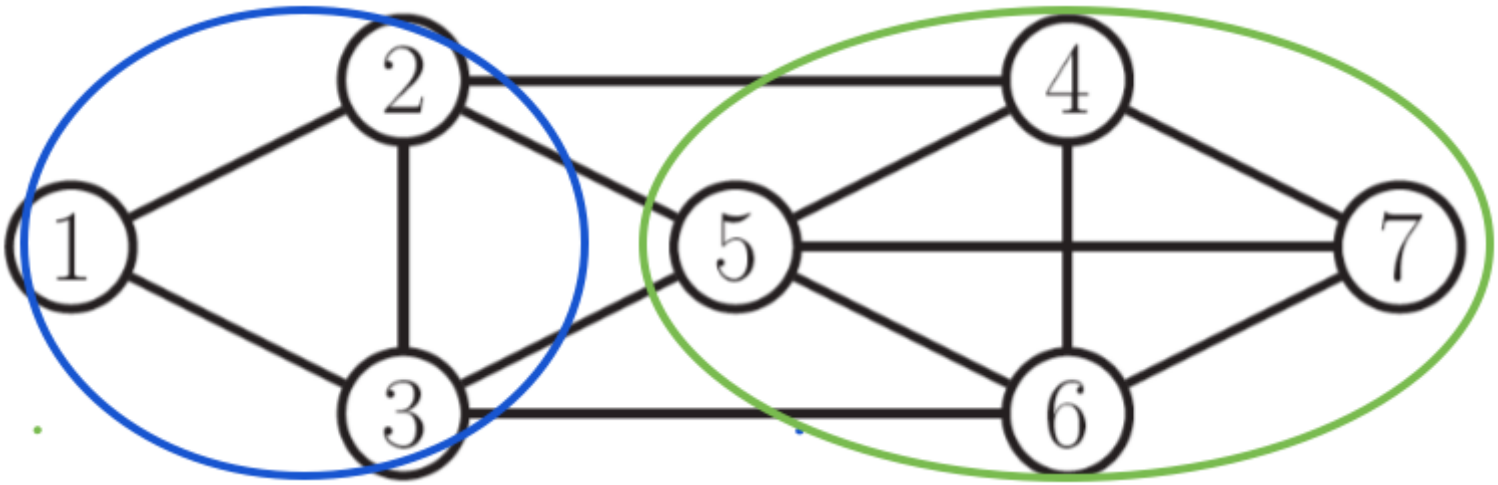
Cliques

A **clique** in an undirected graph is a subset of its vertices such that every two vertices in the subset are connected by an edge (i.e., the subgraph induced by the clique is complete).

def. A **maximal clique** is a clique that cannot be extended by including one more adjacent vertex.

def. A **maximum clique** is a clique of the *largest possible size* in a given graph.

For example, in the following graph a *maximal clique* is show in blue, while a *maximum clique* is shown in green.



Parameterization of an UGM

Let $x = (x_1, \ldots, x_m)$ be the set of all random variables in our graph. Unlike in DGMs, there is no topological ordering associated with an undirected graph, and so we *cannot* use the chain rule to represent the joint distribution $p(x)$.

Instead of associating conditional probabilities with each node, we associate **potential functions** or **factors** with each *maximal clique* in the graph.

For a given clique c, we define the potential function or factor

$$\psi_c(x_c|\theta_c)$$

to be any non-negative function, where x_c is some subset of variables in x involved in a unique, maximal clique.

The joint distribution is *proportional* to the *product of clique potentials*

$$p(x) \propto \prod_{c \in \mathcal{C}} \psi_c(x_c|\theta_c)$$

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

> Any positive distribution whose conditional independencies are represented with an UGM can be represented this way.

More formally,

A distribution $p(x) > 0$ satisfies the conditional independence properties of an undirected graph G iff p can be represented as a product of factors, one per maximal clique, i.e.,

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

where $\mathcal{C}$ is the set of all (maximal) cliques of G, and $Z(\theta)$ the **partition function**, defined as

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

The factored structure of the distribution makes it possible to more efficiently do the sums/integrals needed to compute it.

Lets see how to factorize the undirected graph of our running example:

"../img/lecture_4_7A.png" is not created yet. Click to create.

$$p(x) \propto \psi_{1,2,3}(x_1, x_2, x_3)\psi_{2,3,5}(x_2, x_3, x_5)\psi_{2,4,5}(x_2, x_4, x_5)\psi_{3,5,6}(x_3, x_5, x_6)\psi_{4,5,6,7}(x_4, x_5, x_6, x_7)$$

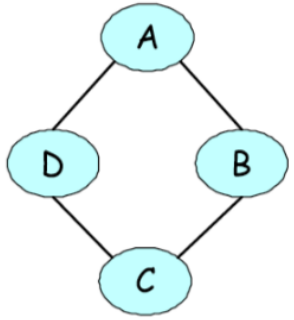
Representing potentials

Recall how we parameterized DAGMs of discrete random variables by using conditional probability tables to represent the possible configurations of each node. In this case, the parameters were a valid probability distribution (i.e. the probabilities of the possible configurations summed to 1).

In UGMs, we do something similar. If the variables are discrete, we can represent the *potential* (or energy) functions as tables of (non-negative) numbers

$$p(A,B,C,D) = \frac{1}{Z}\psi_{A,B}(A,B)\psi_{B,C}(B,C)\psi_{C,D}(C,D)\psi_{A,D}(A,D)$$

where



$\phi_1[A,B]$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

$\phi_2[B,C]$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

$\phi_3[C,D]$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_4[D,A]$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

It is important to note that these potentials are *not* probabilities, but instead encode *relative affinities* between the different assignments. For example, in the above table, a^0, b^0 is taken to be 30X more likely than a^1, a^0.

A more general approach is to define the log potentials as a linear function of the parameters

$$\log \psi_c(y_c) = \phi_c(y_c)^T \theta_c$$

where $\phi_c(y_c)$ is a feature vector.

The resulting distribution is given by

$$p(y|\theta) = \exp\{\sum_c \phi_c(y_c)^T \theta_c - \log Z(\theta)\}$$

which has an exponential familiy form. Thus its moments can be computed using the properties of the exponential families.

Example MRF: Ising model

Below, we will use some terms from statistical physics which may be confusing (it is ok if it is).

The Ising model is an example of an MRF that arose from statistical physics which was originally deveoped for modeling the behavior of magnets. The nodes variables are spins, i.e., we use $y_s \in \{-1, +1\}$ to denote the spin of an atom, which is either down or up. In some magnets, called ferro-magnets, neighboring spins tend to line up in the same direction whereas in others termed as anti-ferro-magnets, the spins tend to be different from their neighbors. We focus on ferro-magnets.

For example

"../img/ising2.png" is not created yet. Click to create.

and define the pairwise clique potentials as

$$\psi_{st}(y_s, y_t) = \begin{pmatrix} e^J & e^{-J} \\ e^{-J} & e^J \end{pmatrix}$$

where J is the coupling strength between nodes s and t.
This means that if the nodes s and t are connected by an edge (denoted with $s \sim t$), then in compact form we have $\psi_{st}(y_s, y_t) = e^{J y_s y_t}$. For instance, if $(y_s, y_t) = (1, -1)$ we have $\psi_{st}(1, -1) = e^{-J}$. Notice that if the nodes s and t have the same spin, this increases the potential, hence the probability of that configuration.

Sometimes there is an external field, which is an energy term added to each spin. This is equivalent to considering a node potential for each variable given as

$$\psi_s(y_s) = e^{b_s y_s}$$

The overall distribution becomes

$$
\begin{aligned}
p(y) \propto & \prod_{s \sim t} \psi_{st}(y_s, y_t) \prod_s \psi_s(y_s) \\
= & \exp\{J \sum_{s \sim t} y_s y_t + \sum_s b_s y_s\}
\end{aligned}
$$

Observe that if $y_s = y_t$, i.e. the neighboring spins are the same, then the likelihood is larger. We will use this Ising model for image denoising in the Assignment 2!