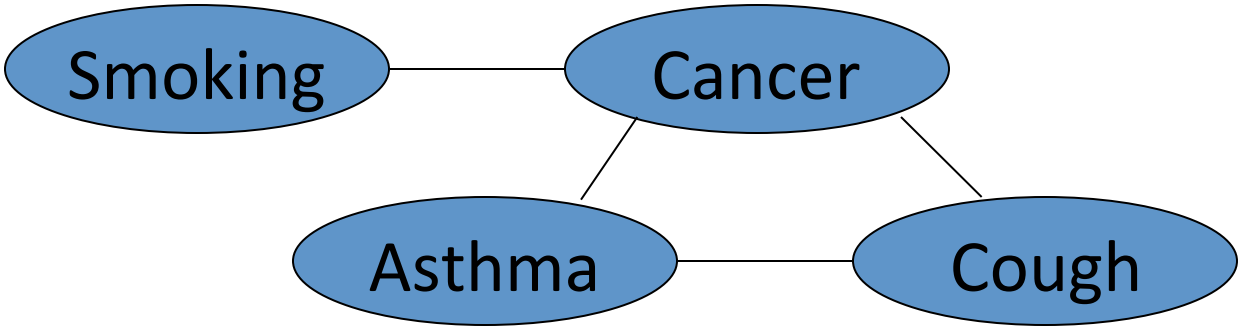
Markov Networks

Alan Ritter

Markov Networks

- **Undirected** graphical models



- Potential functions defined over cliques

$$P(x) = \frac{1}{Z} \prod_c \Phi_c(x_c)$$

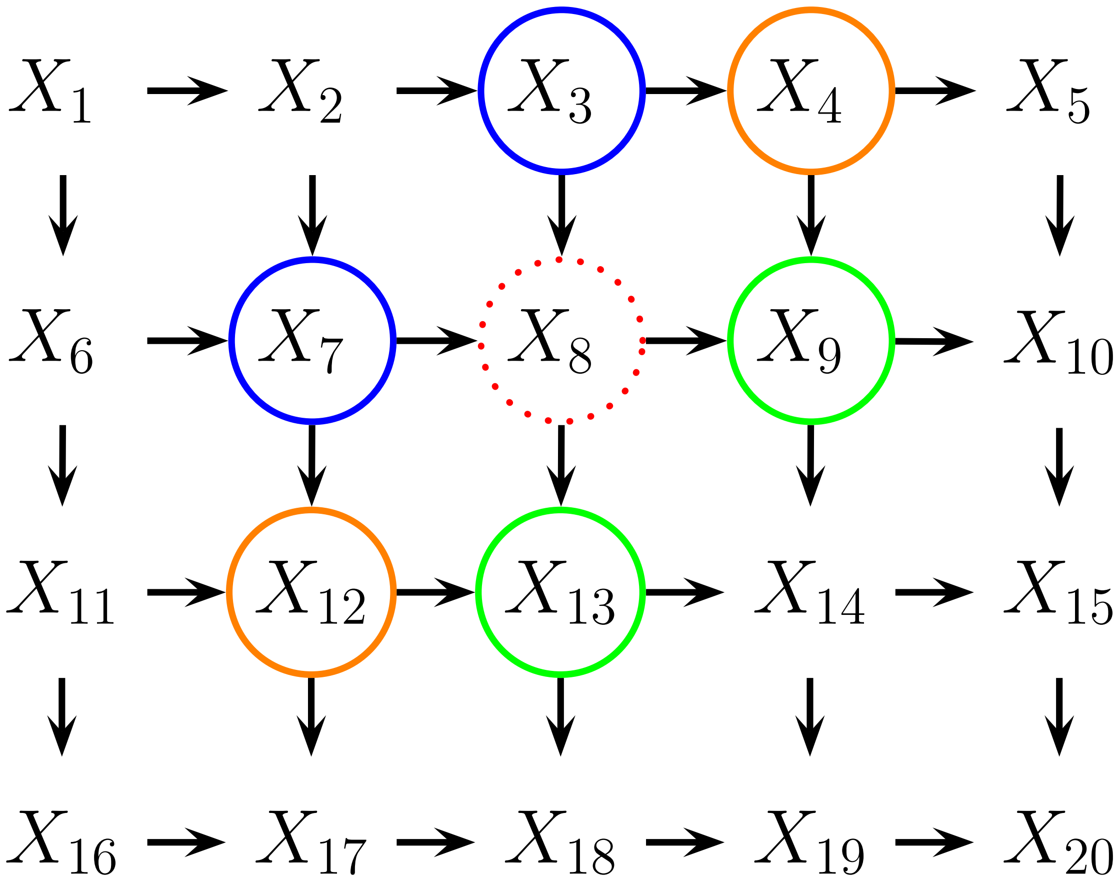
$$Z = \sum_x \prod_c \Phi_c(x_c)$$

Smoking	Cancer	Φ(S,C)
False	False	4.5
False	True	4.5
True	False	2.7
True	True	4.5

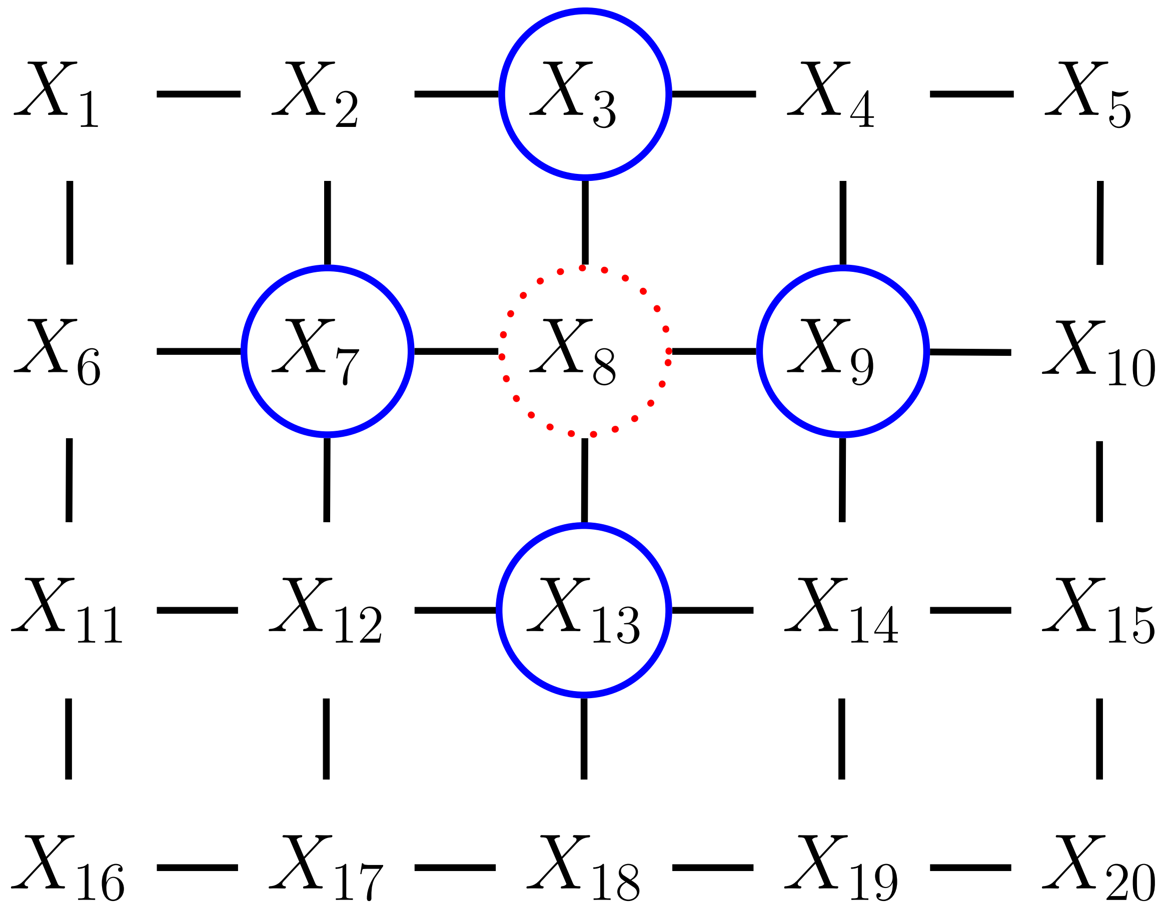
Undirected Graphical Models: Motivation

- Terminology:
  - Directed graphical models = Bayesian Networks
  - Undirected graphical models = Markov Networks
- We just learned about DGMs (Bayes Nets)
- For some domains being forced to choose a direction of edges is awkward.
- Example: consider modeling an image
  - Assumption: neighboring pixels are correlated
  - We could create a DAG model w/ 2D topology

2D Bayesian Network



Markov Random Field (Markov Network)



UGMs (Bayes Nets) vs DGMs (Markov Nets)

- **Advantages**
  1. Symmetric
     - More natural for certain domains (e.g. spatial or relational data)
  2. Discriminative UGMs (A.K.A Conditional Random Fields) work better than discriminative UGMs
- **Disadvantages**
  1. Parameters are less interpretable and modular
  2. Parameter estimation is computationally more expensive

Conditional Independence Properties

- Much Simpler than Bayesian Networks
  - No d-seperation, v-structures, etc…
- UGMs define CI via simple graph separation

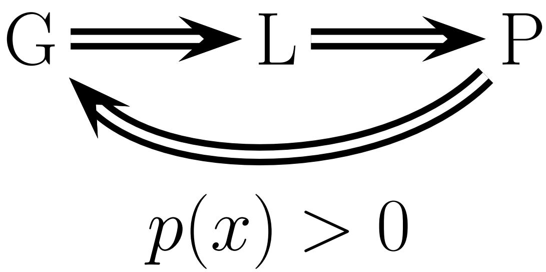
$$X_A \perp_G X_B | X_E \iff E \text{ separates } A \text{ from } B \text{ in } G$$

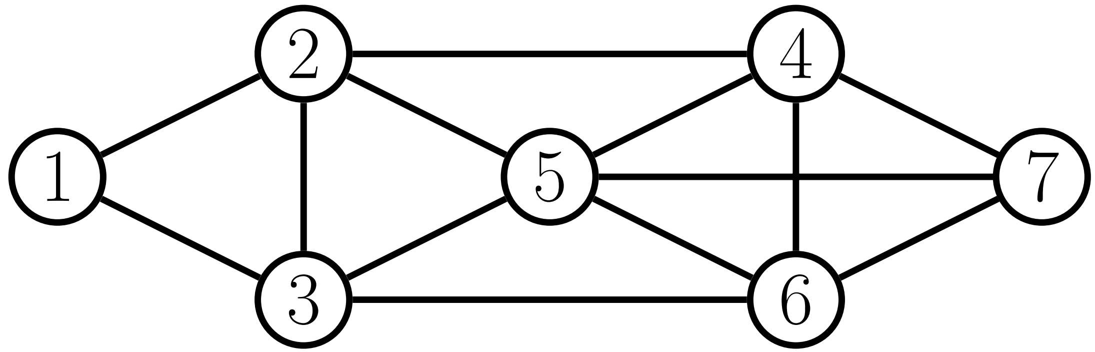
- E.g. if we remove all the evidence nodes from the graph, are there any paths connecting A and B?

Markov Blanket

- Also Simple
  - Markov blanket of a node is just the set of it’s immediate neighbors
  - Don’t need to worry about co-parents

Independence Properties





Pairwise: $1 \perp 7 | \mathrm{rest}$

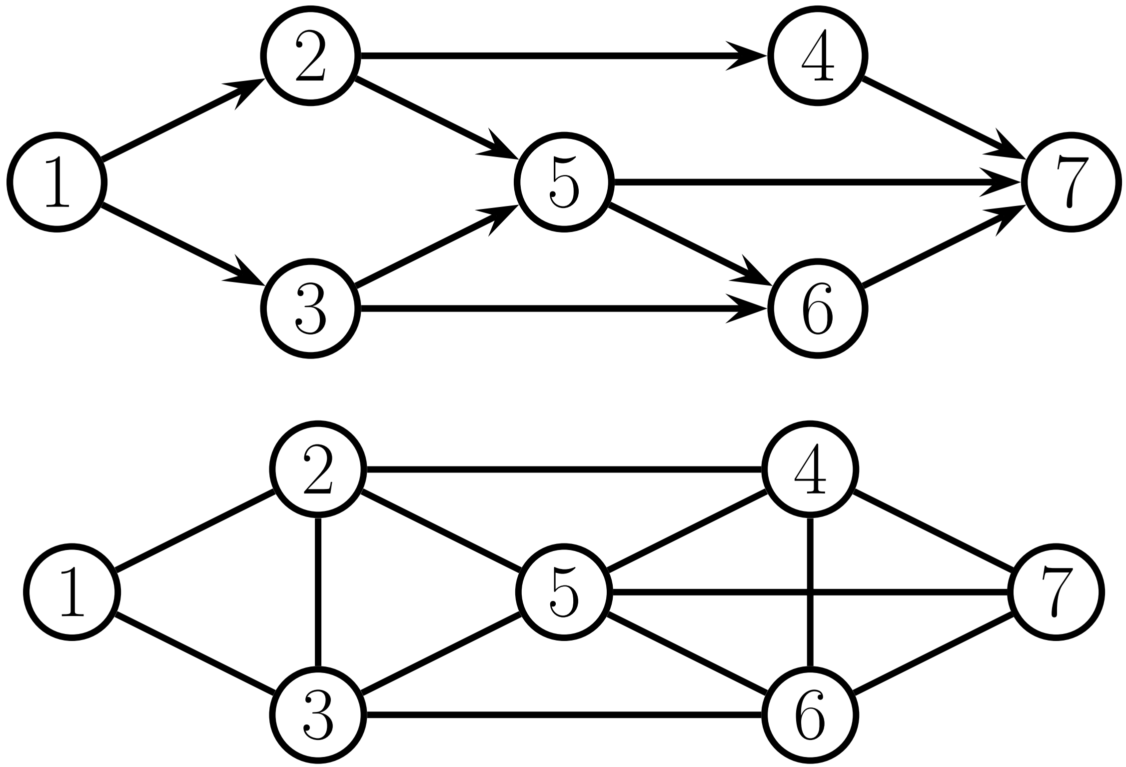
Local: $1 \perp \mathrm{rest} | 2, 3$

Global: $1, 2 \perp 6, 7 | 3, 4, 5$

Converting a Bayesian Network to a Markov Network

- Tempting:
  - Just drop directionality of the edges
  - But this is clearly incorrect (v-structure)
  - Introduces incorrect CI statements
- Solution:
  - Add edges between “unmarried” parents
  - This process is called **moralization**

Example: moralization

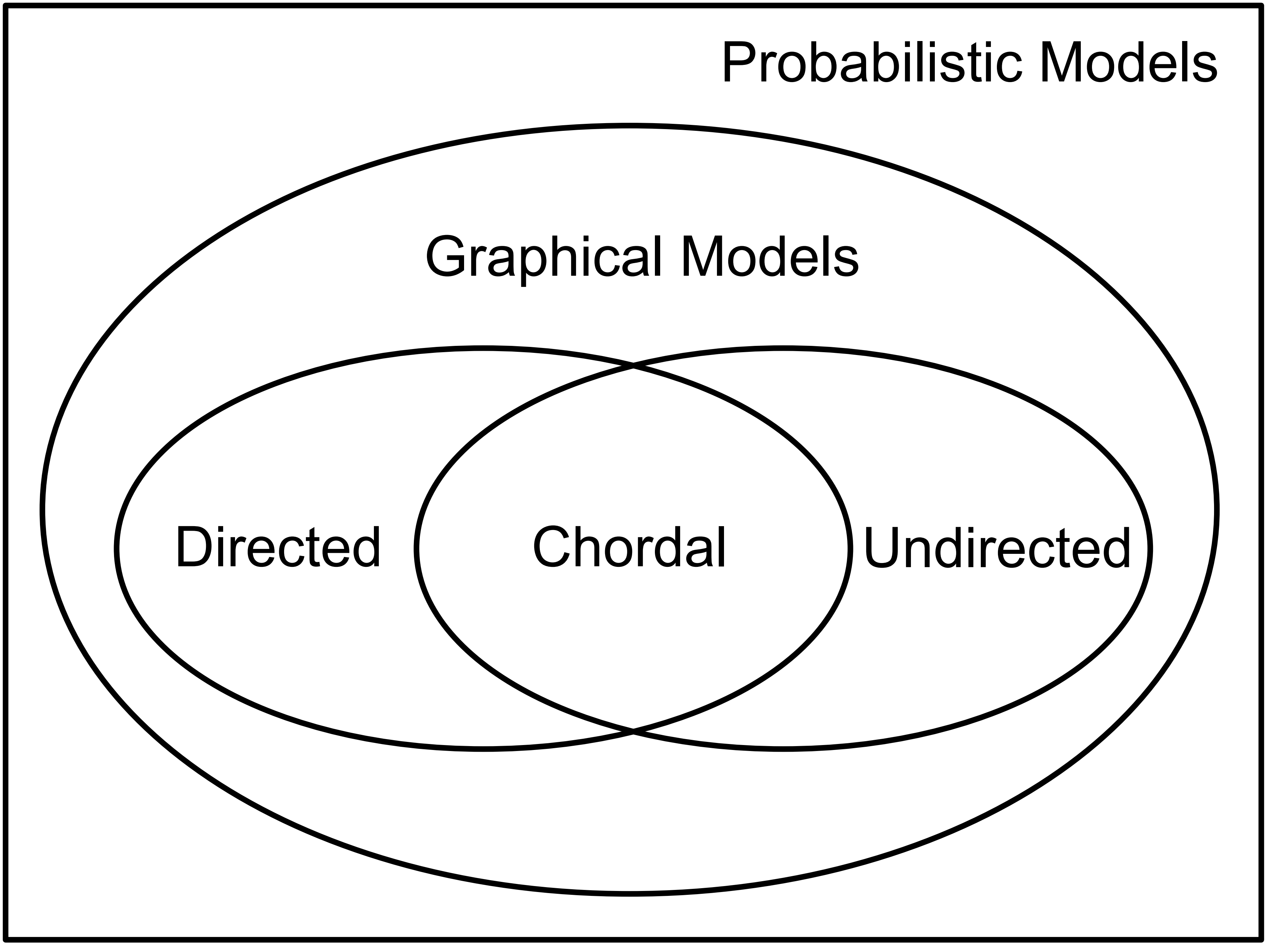


- Unfortunately, this looses some CI information
  - Example: $4 \perp 5 | 2$

Directed vs. Undirected GMs

- Q: which has more "expressive power"?
- Recall:
  - G is an I-map of P if: $I(G) \subseteq I(P)$
- Now define:
  - G is a **perfect I-map** of P if: $I(G) = I(P)$
    - Graph can represent all (and only) CIs in P

Bayesian Networks and Markov Networks are perfect maps for different sets of distributions


Probabilistic Models
Graphical Models
Directed
Chordal
Undirected

Parameterization

- No topological ordering on undirected graph
- Can't use the chain rule of probability to represent P(y)
- Instead we will use **potential functions:**
  - associate potential functions with each maximal clique in the graph $\psi_c(y_c|\theta_c)$
  - A potential can be any non-negative function
- **Joint distribution is defined to be proportional to product of clique potentials**

Parameterization (con't)

- **Joint distribution is defined to be proportional to product of clique potentials**
- **Any positive distribution whose CI properties can be represented by an UGM can be represented this way.**

Hammersly-Clifford Theorem

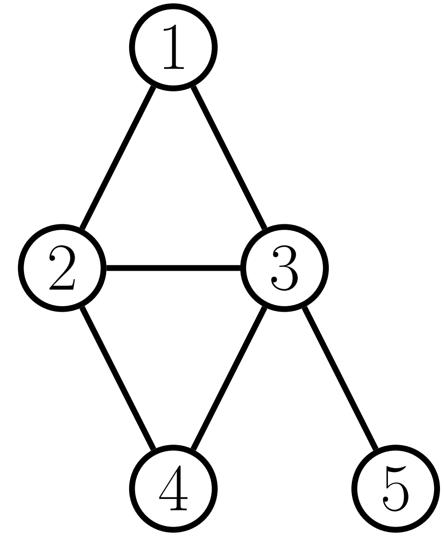
- A positive distribution $P(Y) > 0$ satisfies the CI properties of an undirected graph G iff P can be represented as a product of factors, one per maximal clique

$$P(y|\theta) = \frac{1}{Z(\theta)} \prod_{c \in C} \psi_c(y_c|\theta_c)$$

Z is the partition function

$$Z(\theta) = \sum_{y} \prod_{c \in C} \psi_c(y_c|\theta_c)$$

Example



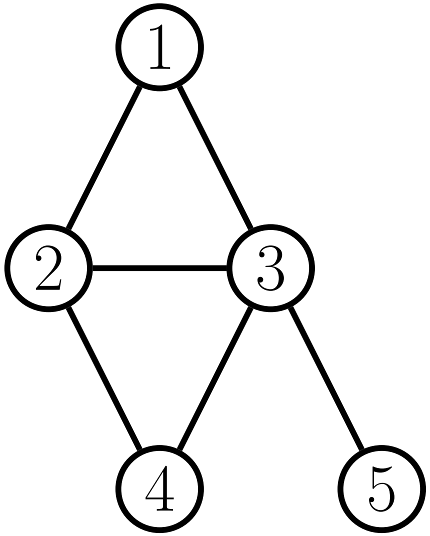
- If P satisfies the conditional independence assumptions of this graph, we can write

$$P(y|\theta) = \frac{1}{Z(\theta)}\psi_{123}(y_1, y_2, y_3)\psi_{234}(y_2, y_3, y_4)\psi_{35}(y_3, y_5)$$

$$Z(\theta) = \sum_{\mathbf{y}} \psi_{123}(y_1, y_2, y_3)\psi_{234}(y_2, y_3, y_4)\psi_{35}(y_3, y_5)$$

Pairwise MRF

- Potentials don't need to correspond to maximal cliques
- We can also restrict parameterization to edges (or any other cliques)
- **Pairwise MRF:**



$$P(y|\theta) = \psi_{12}(y_1, y2)\psi_{13}(y_1, y_3)\psi_{23}(y_2, y_3)\psi_{24}(y_2, y_4)\psi_{34}(y_3, y_4)\psi_{35}(y_3, y_5)$$

Representing Potential Functions

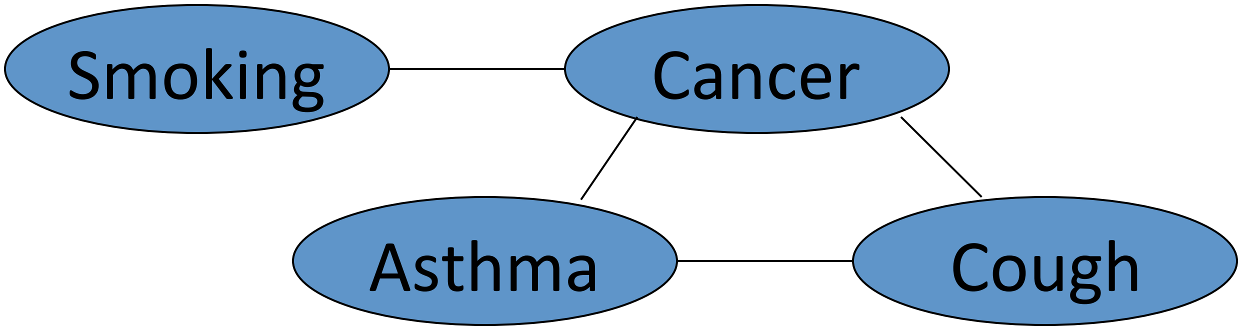
- Can represent as CPTs like we did for Bayesian Networks (DGMs)
  - But, potentials are **not** probabilities
  - Represent relative "compatibility" between various assignments

Representing Potential Functions

- More general approach:
  - Represent the log potentials as a linear function of the parameters
  - **Log-linear (maximum entropy) models**

$$\log P(y|\theta) = \sum_c \psi_c(y_c)^T \theta_c - \log Z(\theta)$$

Log-Linear Models



- Log-linear model:

$$P(x) = \frac{1}{Z} \exp\left(\sum_i w_i . f_i(x) \right)$$

Weight of Feature *i*

Feature *i*

$$f_1(\text{Smoking}, \text{Cancer}) = \begin{cases} 1 & \text{if } \neg \text{Smoking} \vee \text{Cancer} \\ 0 & \text{otherwise} \end{cases}$$

$$w_1 = 0.51$$

Log-Linear models can represent Table CPTs

- Consider pairwise MRF where each edge has an associated potential w/ K^2 features:

$$\phi(y_s, y_t) = [\ldots, \mathbb{I}(y_s = j, y_t = k), \ldots]$$

- Then we can convert into a potential function using the weight for each feature:

$$\psi(y_s, y_t) = exp([\theta_{st}^T \phi_{st}]_{jk}]) = exp(\theta_{st}(j, k))$$

- But, log-linear model is more general
  - Feature vectors can be arbitrarily designed