Probabilistic Graphical Models

David Sontag

New York University

Lecture 2, February 2, 2012

Bayesian networks

Reminder of last lecture

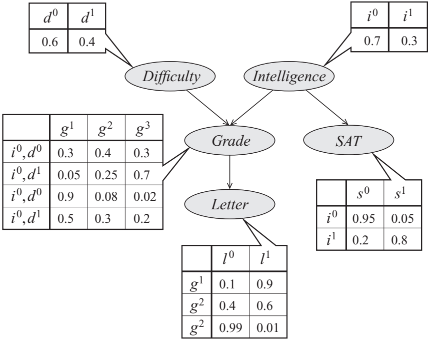
- A **Bayesian network** is specified by a directed *acyclic* graph $G = (V, E)$ with:
  1. One node $i \in V$ for each random variable X_i
  2. One conditional probability distribution (CPD) per node, $p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values
- Corresponds 1-1 with a particular factorization of the joint distribution:

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

- Powerful framework for designing *algorithms* to perform probability computations

Example

- Consider the following Bayesian network:



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^0,d^0	0.9	0.08	0.02
i^0,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

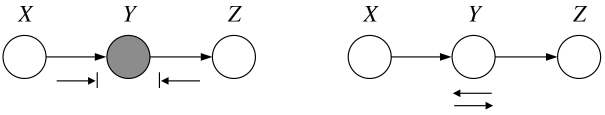
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

- What is its joint distribution?

$$
\begin{aligned}
p(x_1, \ldots x_n) &= \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)}) \\
p(d, i, g, s, l) &= p(d)p(i)p(g \mid i, d)p(s \mid i)p(l \mid g)
\end{aligned}
$$

D-separation ("directed separated") in Bayesian networks

- Algorithm to calculate whether $X \perp Z \mid \mathbf{Y}$ by looking at graph separation
- Look to see if there is **active path** between X and Y when variables $\mathbf{Y}$ are observed:



(a) (b)

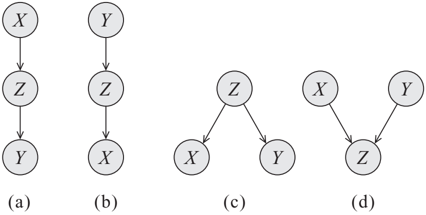
- If no such path, then X and Z are **d-separated** with respect to $\mathbf{Y}$
- d-separation reduces statistical independencies (hard) to connectivity in graphs (easy)
- Important because it allows us to quickly prune the Bayesian network, finding just the relevant variables for answering a query

Independence maps

- Let $I(G)$ be the set of all conditional independencies implied by the directed aylic graph (DAG) G
- Let $I(p)$ denote the set of all conditional independencies that hold for the joint distribution p.
- A DAG G is an **I-map** (independence map) of a distribution p if $I(G) \subseteq I(p)$
  - A fully connected DAG G is an I-map for *any* distribution, since $I(G) = \emptyset \subseteq I(p)$ for all p
- G is a **minimal I-map** for p if the removal of even a single edge makes it not an I-map
  - A distribution may have several minimal I-maps
  - Each corresponds to a specific node-ordering
- G is a **perfect map** (P-map) for distribution p if $I(G) = I(p)$

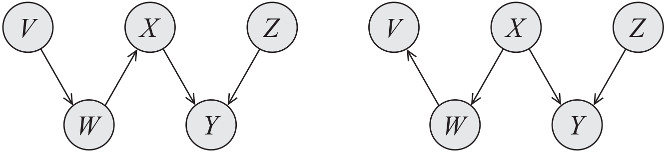
Equivalent structures

- Different Bayesian network structures can be **equivalent** in that they encode precisely the same conditional independence assertions (and thus the same distributions)
- Which of these are equivalent?



Equivalent structures

- Different Bayesian network structures can be **equivalent** in that they encode precisely the same conditional independence assertions (and thus the same distributions)
- Which of these are equivalent?

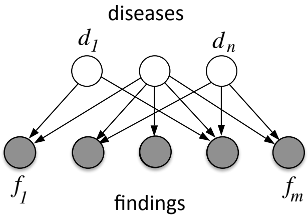


- A **causal network** is a Bayesian network with an explicit requirement that the relationships be causal
  - Bayesian networks are not the same as **causal networks**

What are some frequently used graphical models?

Quick Medical Reference (decision theoretic)

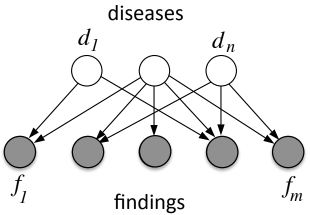
(Miller et al. '86, Shwe et al. '91)



- Joint distribution factors as $p(\mathbf{f}, \mathbf{d}) = \prod_j p(d_j) \prod_i p(f_i \mid \mathbf{d})$
  $p(d_j = 1)$ is the prior probability of having disease j
- Model assumes the following independencies: $d_i \perp d_j, \quad f_i \perp f_j \mid \mathbf{d}$
- Common findings can be caused by hundreds of diseases – too many parameters required to specify the CPD $p(f_i \mid \mathbf{d})$ as a table

Quick Medical Reference (decision theoretic)

(Miller et al. '86, Shwe et al. '91)

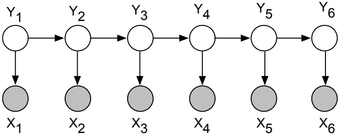


- Instead, we use a **noisy-or parameterization**:

$$p(f_i = 0 \mid \mathbf{d}) = (1 - q_{i0}) \prod_{j \in \mathrm{Pa}(i)} (1 - q_{ij})^{d_j}$$

- $q_{ij} = p(f_i = 1 \mid d_j = 1)$ is the probability that the disease j, if present, could alone cause the finding to have a positive outcome
- $q_{i0} = p(f_i = 1 \mid L)$ is the "leak" probability – the probability that the finding is caused by something other than the diseases in the model

Hidden Markov models

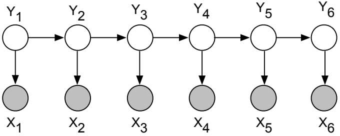


- Frequently used for speech recognition and part-of-speech tagging
- Joint distribution factors as:

$$p(\mathbf{y}, \mathbf{x}) = p(y_1)p(x_1 \mid y_1) \prod_{t=2}^{T} p(y_t \mid y_{t-1})p(x_t \mid y_t)$$

  - $p(y_1)$ is the distribution for the starting state
  - $p(y_t \mid y_{t-1})$ is the *transition* probability between any two states
  - $p(x_t \mid y_t)$ is the *emission* probability
- What are the conditional independencies here? For example, $Y_1 \perp \{Y_3, \ldots, Y_6\} \mid Y_2$

Hidden Markov models



- Joint distribution factors as:

$$p(\mathbf{y}, \mathbf{x}) = p(y_1)p(x_1 \mid y_1) \prod_{t=2}^{T} p(y_t \mid y_{t-1})p(x_t \mid y_t)$$

- A **homogeneous** HMM uses the same parameters (β and α below) for each transition and emission distribution (**parameter sharing**):

$$p(\mathbf{y}, \mathbf{x}) = p(y_1)\alpha_{x_1,y_1} \prod_{t=2}^{T} \beta_{y_t,y_{t-1}}\alpha_{x_t,y_t}$$

How many parameters need to be learned?

Mixture of Gaussians

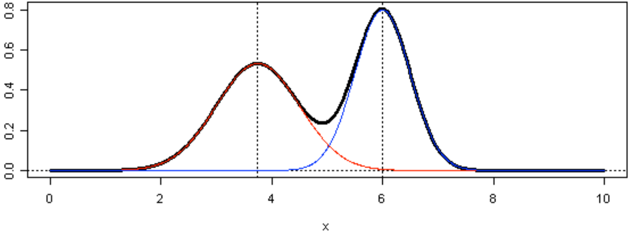
- The N-dim. multivariate normal distribution, $\mathcal{N}(\mu, \Sigma)$, has density:

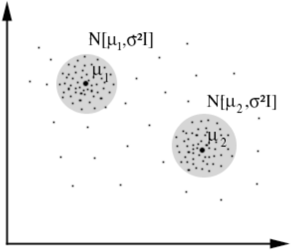
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\Big)$$

- Suppose we have k Gaussians given by μ_k and Σ_k, and a distribution θ over the numbers $1, \ldots, k$
- Mixture of Gaussians distribution $p(y, \mathbf{x})$ given by
  1. Sample $y \sim \theta$ (specifies which Gaussian to use)
  2. Sample $x \sim \mathcal{N}(\mu_y, \Sigma_y)$

Mixture of Gaussians

- The marginal distribution over **x** looks like:





Latent Dirichlet allocation (LDA)

- **Topic models** are powerful tools for exploring large data sets and for making inferences about the content of documents

Documents

Topics

politics	religion	sports
president	hindu	baseball
obama	judiasm	soccer
washington	ethics	basketball
religion	buddhism	football
...	...	...

- Many applications in information retrieval, document summarization, and classification

New document

Words $w_1, \ldots, w_N$

What is this document about?

weather	.50
finance	.49
sports	.01

Distribution of topics θ

- LDA is one of the simplest and most widely used topic models

Generative model for a document in LDA

1. Sample the document's **topic distribution** θ (aka topic vector)

$$\theta \sim \mathrm{Dirichlet}(\alpha_{1:T})$$

   where the $\{\alpha_t\}_{t=1}^T$ are fixed hyperparameters. Thus θ is a distribution over T topics with mean $\theta_t = \alpha_t / \sum_{t'} \alpha_{t'}$

2. For $i = 1$ to N, sample the **topic** z_i of the i'th word

$$z_i | \theta \sim \theta$$

3. ... and then sample the actual **word** w_i from the z_i'th topic

$$w_i | z_i, \ldots \sim \beta_{z_i}$$

   where $\{\beta_t\}_{t=1}^T$ are the *topics* (a fixed collection of distributions on words)

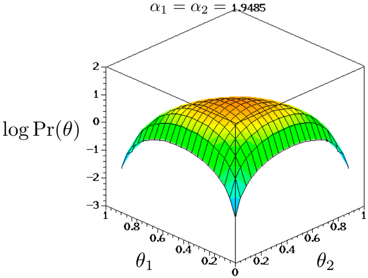
Generative model for a document in LDA

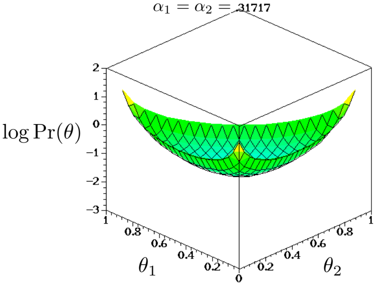
1. Sample the document's **topic distribution** θ (aka topic vector)

$$\theta \sim \mathrm{Dirichlet}(\alpha_{1:T})$$

where the $\{\alpha_t\}_{t=1}^{T}$ are hyperparameters. The Dirichlet density is:

$$p(\theta_1, \ldots, \theta_T) \propto \prod_{t=1}^{T} \theta_t^{\alpha_t - 1}$$





Generative model for a document in LDA

3. … and then sample the actual **word** w_i from the z_i'th topic

$$w_i | z_i, \ldots \sim \beta_{z_i}$$

where $\{\beta_t\}_{t=1}^{T}$ are the *topics* (a fixed collection of distributions on words)

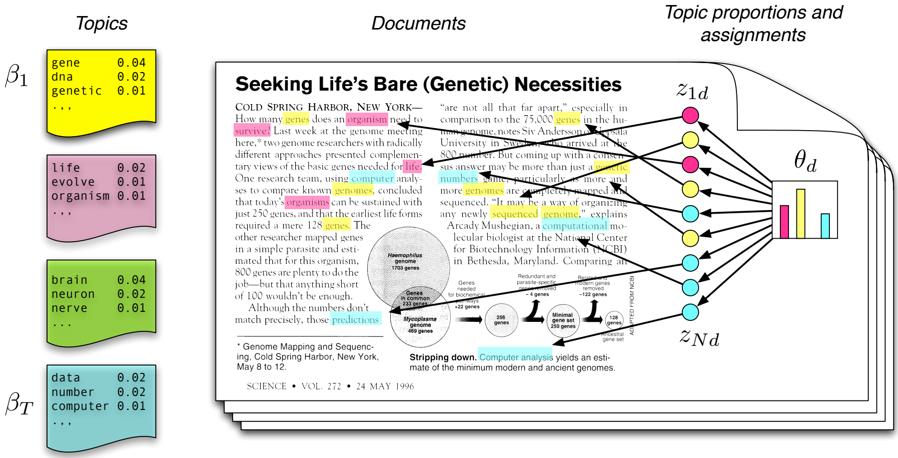
Documents

Topics

politics .0100	religion .0500	sports .0105
president .0095	hindu .0092	baseball .0100
obama .0090	judiasm .0080	soccer .0055
washington .0085	ethics .0075	basketball .0050
religion .0060	buddhism .0016	football .0045
...	...	...

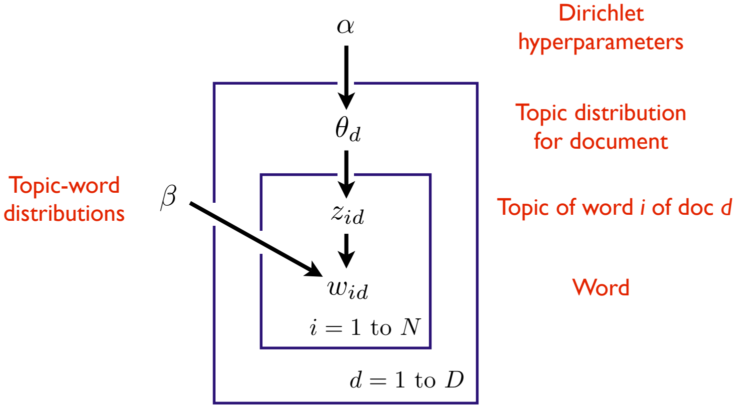
$$\beta_t = \{\, p(w \mid z = t) \,\}$$

Example of using LDA



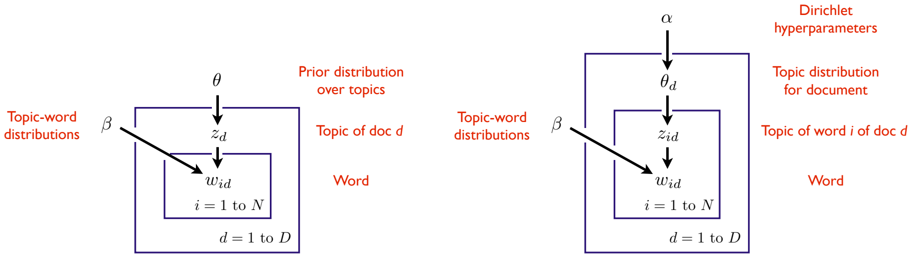
(Blei, *Introduction to Probabilistic Topic Models*, 2011)

"Plate" notation for LDA model



Variables within a plate are replicated in a conditionally independent manner

Comparison of mixture and admixture models



- Model on left is a **mixture model**
  - Called *multinomial* naive Bayes (a word can appear multiple times)
  - Document is generated from a single topic
- Model on right (LDA) is an **admixture model**
  - Document is generated from a distribution over topics

Summary

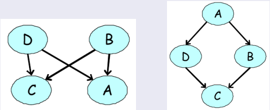
- **Bayesian networks** given by (G, P) where P is specified as a set of local **conditional probability distributions** associated with G's nodes
- One interpretation of a BN is as a **generative model**, where variables are sampled in topological order
- Local and global independence properties identifiable via **d-separation** criteria
- Computing the probability of any assignment is obtained by multiplying CPDs
  - **Bayes' rule** is used to compute conditional probabilities
  - Marginalization or **inference** is often computationally difficult
- Examples (will show up again): **naive Bayes**, **hidden Markov models**, **latent Dirichlet allocation**

Bayesian networks have limitations

- Recall that G is a **perfect map** for distribution p if $I(G) = I(p)$
- **Theorem:** Not every distribution has a perfect map as a DAG

Proof.

(By counterexample.) There is a distribution on 4 variables where the only independencies are $A \perp C \mid \{B, D\}$ and $B \perp D \mid \{A, C\}$. This cannot be represented by any Bayesian network.



(a) (b)

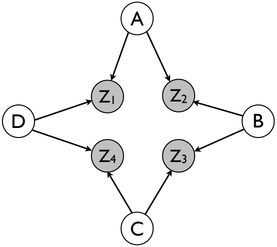
Both (a) and (b) encode $(A \perp C | B, D)$, but in both cases $(B \not\perp D | A, C)$. $\square$

Example

- Let's come up with an example of a distribution p satisfying $A \perp C \mid \{B, D\}$ and $B \perp D \mid \{A, C\}$
- A=Alex's hair color (red, green, blue)
  B=Bob's hair color
  C=Catherine's hair color
  D=David's hair color
- Alex and Bob are friends, Bob and Catherine are friends, Catherine and David are friends, David and Alex are friends
- Friends *never* have the same hair color!

Bayesian networks have limitations

- Although we could represent any distribution as a fully connected BN, this obscures its structure
- Alternatively, we can introduce "dummy" binary variables $\mathbf{Z}$ and work with a **conditional** distribution:



- This satisfies $A \perp C \mid \{B, D, \mathbf{Z}\}$ and $B \perp D \mid \{A, C, \mathbf{Z}\}$
- Returning to the previous example, we would set:

$$p(Z_1 = 1 \mid a, d) = 1 \text{ if } a \neq d, \text{ and } 0 \text{ if } a = d$$

Z_1 is the observation that Alice and David have different hair colors

Undirected graphical models

- An alternative representation for joint distributions is as an **undirected graphical model**
- As in BNs, we have one node for each random variable
- Rather than CPDs, we specify (non-negative) **potential functions** over sets of variables associated with cliques C of the graph,

$$p(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \phi_c(\mathbf{x}_c)$$

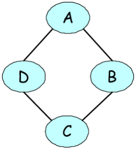
Z is the **partition function** and normalizes the distribution:

$$Z = \sum_{\hat{x}_1, \dots, \hat{x}_n} \prod_{c \in C} \phi_c(\hat{\mathbf{x}}_c)$$

- Like CPD's, $\phi_c(\mathbf{x}_c)$ can be represented as a table, but it is *not normalized*
- Also known as **Markov random fields** (MRFs) or Markov networks
Potential functions are also called **factors**

Hair color example as a MRF

- We now have an **undirected** graph:



- The joint probability distribution is parameterized as

$$p(a,b,c,d) = \frac{1}{Z}\phi_{AB}(a,b)\phi_{BC}(b,c)\phi_{CD}(c,d)\phi_{AD}(a,d)\ \phi_A(a)\phi_B(b)\phi_C(c)\phi_D(d)$$

- **Pairwise potentials** enforce that no friend has the same hair color:

$$\phi_{AB}(a,b) = 0 \text{ if } a = b, \quad \text{and 1 otherwise}$$

- **Single-node potentials** specify an affinity for a particular hair color, e.g.

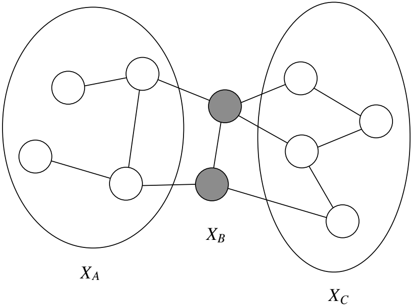
$$\phi_D(\text{“red”}) = 0.6, \quad \phi_D(\text{“blue”}) = 0.3, \quad \phi_D(\text{“green”}) = 0.1$$

The normalization Z makes the potentials **scale invariant**! Equivalent to

$$\phi_D(\text{“red”}) = 6, \quad \phi_D(\text{“blue”}) = 3, \quad \phi_D(\text{“green”}) = 1$$

Markov network structure implies conditional independencies

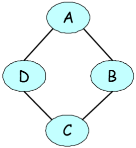
- Let G be the undirected graph where we have one edge for every pair of variables that appear together in a potential
- Conditional independence is given by **graph separation**!



- $X_{\mathbf{A}} \perp X_{\mathbf{C}} \mid X_{\mathbf{B}}$ if there is no path from $a \in \mathbf{A}$ to $c \in \mathbf{C}$ after removing all variables in $\mathbf{B}$

Example

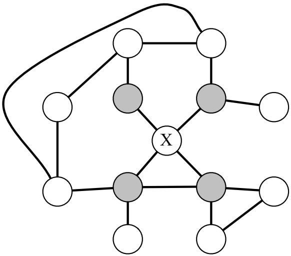
- Returning to hair color example, its undirected graphical model is:



- Since removing A and C leaves no path from D to B, we have $D \perp B \mid \{A, C\}$
- Similarly, since removing D and B leaves no path from A to C, we have $A \perp C \mid \{D, B\}$
- No other independencies implied by the graph

Markov blanket

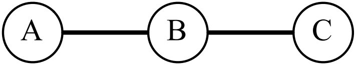
- A set $\mathbf{U}$ is a **Markov blanket** of X if $X \notin \mathbf{U}$ and if $\mathbf{U}$ is a minimal set of nodes such that $X \perp (\mathcal{X} - \{X\} - \mathbf{U}) \mid \mathbf{U}$
- In undirected graphical models, the Markov blanket of a variable is precisely its **neighbors** in the graph:



- In other words, X is independent of the rest of the nodes in the graph given its immediate neighbors

Proof of independence through separation

- We will show that $A \perp C \mid B$ for the following distribution:



$$p(a, b, c) = \frac{1}{Z}\phi_{AB}(a, b)\phi_{BC}(b, c)$$

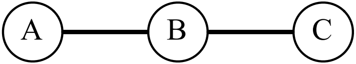
- First, we show that $p(a \mid b)$ can be computed using only $\phi_{AB}(a, b)$:

$$\begin{aligned}
p(a \mid b) &= \frac{p(a, b)}{p(b)} \\
&= \frac{\frac{1}{Z}\sum_{\hat{c}} \phi_{AB}(a, b)\phi_{BC}(b, \hat{c})}{\frac{1}{Z}\sum_{\hat{a},\hat{c}} \phi_{AB}(\hat{a}, b)\phi_{BC}(b, \hat{c})} \\
&= \frac{\phi_{AB}(a, b)\sum_{\hat{c}} \phi_{BC}(b, \hat{c})}{\sum_{\hat{a}} \phi_{AB}(\hat{a}, b)\sum_{\hat{c}} \phi_{BC}(b, \hat{c})} = \frac{\phi_{AB}(a, b)}{\sum_{\hat{a}} \phi_{AB}(\hat{a}, b)}.
\end{aligned}$$

- More generally, the probability of a variable conditioned on its Markov blanket depends *only* on potentials involving that node

Proof of independence through separation

- We will show that $A \perp C \mid B$ for the following distribution:



$$p(a,b,c) = \frac{1}{Z}\phi_{AB}(a,b)\phi_{BC}(b,c)$$

Proof.

$$\begin{aligned}
p(a,c \mid b) = \frac{p(a,c,b)}{\sum_{\hat{a},\hat{c}} p(\hat{a},b,\hat{c})} &= \frac{\phi_{AB}(a,b)\phi_{BC}(b,c)}{\sum_{\hat{a},\hat{c}} \phi_{AB}(\hat{a},b)\phi_{BC}(b,\hat{c})} \\
&= \frac{\phi_{AB}(a,b)\phi_{BC}(b,c)}{\sum_{\hat{a}} \phi_{AB}(\hat{a},b)\sum_{\hat{c}} \phi_{BC}(b,\hat{c})} \\
&= p(a \mid b)p(c \mid b)
\end{aligned}$$

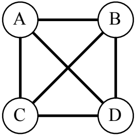
$\square$

Higher-order potentials

- The examples so far have all been **pairwise MRFs**, involving only node potentials $\phi_i(X_i)$ and pairwise potentials $\phi_{i,j}(X_i, X_j)$
- Often we need **higher-order** potentials, e.g.

$$\phi(x, y, z) = x \otimes y \otimes z,$$

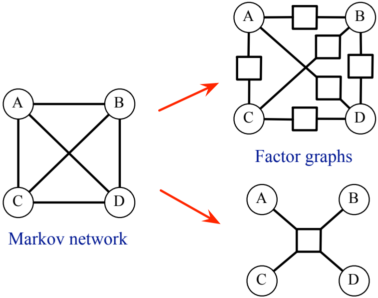
where X, Y, Z are binary and $\otimes$ is the XOR function, enforcing that an odd number of the variables take the value 1
- Although Markov networks are useful for understanding independencies, they hide much of the distribution's structure:



Does this have pairwise potentials, or one potential for all 4 variables?

Factor graphs

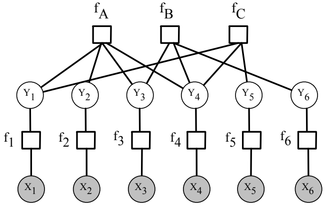
- G does not reveal the structure of the distribution: maximum cliques vs. subsets of them
- A **factor graph** is a bipartite undirected graph with variable nodes and factor nodes. Edges are only between the variable nodes and the factor nodes
- Each factor node is associated with a single potential, whose scope is the set of variables that are neighbors in the factor graph



- The distribution is same as the MRF – this is just a different data structure

Example: Low-density parity-check codes

- Error correcting codes for transmitting a message over a noisy channel (invented by Galleger in the 1960's, then re-discovered in 1996)



- Each of the top row factors enforce that its variables have even parity:

$$f_A(Y_1, Y_2, Y_3, Y_4) = 1 \text{ if } Y_1 \otimes Y_2 \otimes Y_3 \otimes Y_4 = 0, \text{ and } 0 \text{ otherwise}$$

- Thus, the only assignments $\mathbf{Y}$ with non-zero probability are the following (called **codewords**): *3 bits encoded using 6 bits*

$$000000,\ 011001,\ 110010,\ 101011,\ 111100,\ 100101,\ 001110,\ 010111$$

- $f_i(Y_i, X_i) = p(X_i \mid Y_i)$, the likelihood of a bit flip according to noise model

Probabilistic inference

- The *decoding* problem for LDPCs is to find

$$\text{argmax}_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x})$$

  This is called the **maximum a posteriori** (MAP) assignment
- Since Z and $p(\mathbf{x})$ are constants with respect to the choice of $\mathbf{y}$, can equivalently solve (taking the log of $p(\mathbf{y}, \mathbf{x})$):

$$\text{argmax}_{\mathbf{y}} \sum_{c \in C} \theta_c(\mathbf{x}_c),$$

  where $\theta_c(\mathbf{x}_c) = \log \phi_c(\mathbf{x}_c)$
- This is a discrete optimization problem!
  - For general factor graphs, this is NP-hard to solve
  - Next week, you will see a general technique for approximately solving it called **dual decomposition**