Introduction to Graphical Models

Srikumar Ramalingam

School of Computing

University of Utah

Reference

- Christopher M. Bishop, Pattern Recognition and Machine Learning,
- Jonathan S. Yedidia, William T. Freeman, and Yair Weiss, Understanding Belief Propagation and its Generalizations, 2001.

http://www.merl.com/publications/docs/TR2001-22.pdf

- Jonathan S. Yedidia, Message-passing Algorithms for Inference and Optimization: “Belief Propagation” and “Divide and Concur”

http://people.csail.mit.edu/andyd/CIOG_papers/yedidia_jsp_preprint_princeton.pdf

Inference problems and Belief Propagation

- Inference problems arise in statistical physics, computer vision, error-correcting coding theory, and AI.
- BP is an efficient way to solve inference problems based on passing local messages.

Bayesian networks

- Probably the most popular type of graphical model
- Used in many application domains: medical diagnosis, map learning, language understanding, heuristics search, etc.

Probability (Reminder)

Source: Wikipedia.org

- Sample space is the set of all possible outcomes.

  Example: S = {1,2,3,4,5,6}

- Power set of the sample space is obtained by considering all different collections of outcomes.

  Example Power set = {{},{1},{2},…,{1,2},…,{1,2,3,4,5,6}}

- An event is an element of Power set.

  Example E = {1,2,3}

Probability (Reminder)

- Assigns every event E a number in [0,1] in the following manner:

$$p(A) = \frac{|A|}{|S|}$$

- For example, let A = {2,4,6} denote the event of getting an even number while rolling a dice once:

$$p(A) = \frac{|\{2,4,6\}|}{|\{1,2,3,4,5,6\}|} = \frac{3}{6} = \frac{1}{2}$$

Conditional Probability (Reminder)

- If A is the event of interest and we know that the event B has already occurred then the conditional probability of A given B:

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

- The basic idea is that the outcomes are restricted to only B then this serves as the new sample space.
- Two events A and B are statistically independent if

$$p(A \cap B) = p(A)p(B)$$

- Two events A and B are mutually independent if

$$p(A \cap B) = 0$$

Bayes Theorem (Reminder)

- Let A and B be two events and $p(B) \neq 0$.

$$p(A|B) = \frac{p(A)p(B|A)}{p(B)}$$

Reminder

Summary of probabilities

Event	Probability
A	$P(A) \in [0,1]$
not A	$P(A^{\complement}) = 1 - P(A)$
A or B	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $P(A \cup B) = P(A) + P(B) \qquad \text{if A and B are mutually exclusive}$
A and B	$P(A \cap B) = P(A\|B)P(B) = P(B\|A)P(A)$ $P(A \cap B) = P(A)P(B) \qquad \text{if A and B are independent}$
A given B	$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B\|A)P(A)}{P(B)}$

Source: Wikipedia.org

A murder mystery

A fiendish murder has been committed

Whodunit?

There are two suspects:

– the **Butler**

– the **Cook**

There are three possible murder weapons:

– a butcher’s **Knife**

– a **Pistol**

– a fireplace **Poker**

Slide Source: Christopher M. Bishop

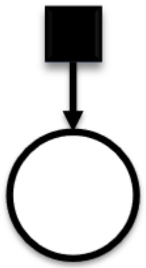
Prior distribution

Butler has served family well for many years
Cook hired recently, rumours of dodgy history

P(Culprit = **Butler**) = 20%

P(Culprit = **Cook**) = 80%

Probabilities add to 100%



P(Culprit)

Culprit = {**Butler**, **Cook**}

This is called a *factor graph*
(we'll see why later)

Slide Source: Christopher M. Bishop

Conditional distribution

Butler is ex-army, keeps a gun in a locked drawer

Cook has access to lots of knives

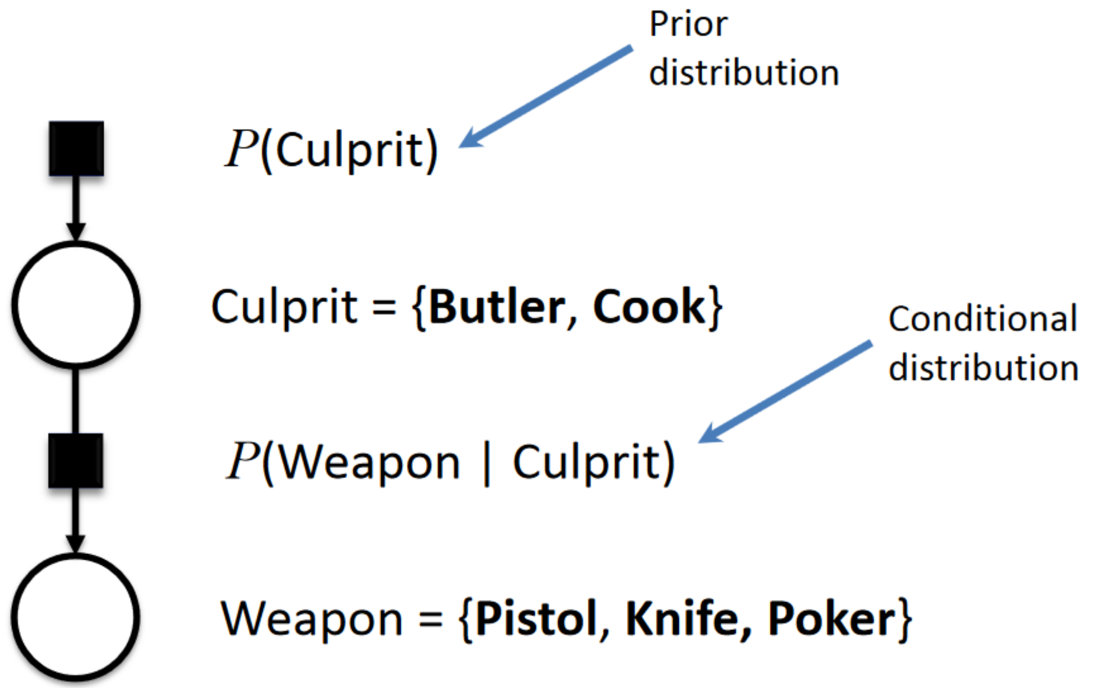
Butler is older and getting frail

	Pistol	**Knife**	**Poker**	
Cook	5%	65%	30%	**= 100%**
Butler	80%	10%	10%	**= 100%**

P(Weapon | Culprit)

Slide Source: Christopher M. Bishop

Factor graph



Slide Source: Christopher M. Bishop

Joint distribution

What is the probability that the **Cook** committed the murder using the **Pistol**?

P(Culprit = **Cook**) = 80%

P(Weapon = **Pistol** | Culprit = **Cook**) = 5%

P(Weapon = **Pistol** , Culprit = **Cook**) = 80% x 5% = 4%

Likewise for the other five combinations of Culprit and Weapon

Slide Source: Christopher M. Bishop

Joint distribution

	Pistol	Knife	Poker
Cook	4%	52%	24%
Butler	16%	2%	2%

= 100%

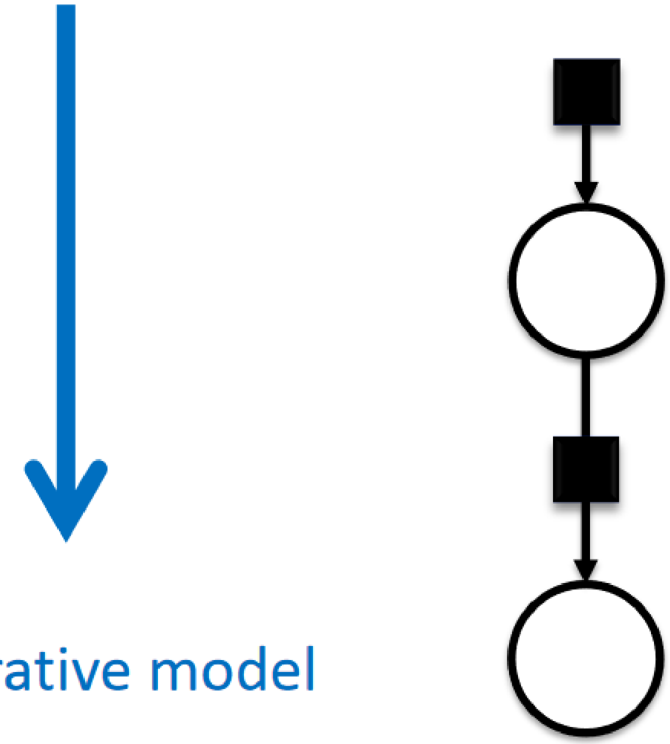
P(Weapon, Culprit) = P(Weapon | Culprit) P(Culprit)

$$P(x, y) = P(y|x)P(x)$$

Product rule

Slide Source: Christopher M. Bishop

Factor graphs



$$P(\text{Weapon, Culprit}) = P(\text{Weapon} \mid \text{Culprit})\, P(\text{Culprit})$$

Slide Source: Christopher M. Bishop

Generative viewpoint

Murderer	Weapon
Cook	Knife
Butler	Knife
Cook	Pistol
Cook	Poker
Cook	Knife
Butler	Pistol
Cook	Poker
Cook	Knife
Butler	Pistol
Cook	Knife
...	...

Marginal distribution of Culprit

	Pistol	Knife	Poker	
Cook	4%	52%	24%	**= 80%**
Butler	16%	2%	2%	**= 20%**

$$P(x) = \sum_{y} P(x, y)$$

Sum rule

Marginal distribution of Weapon

	Pistol	Knife	Poker
Cook	4%	52%	24%
Butler	16%	2%	2%
	= 20%	**= 54%**	**= 26%**

$$P(x) = \sum_{y} P(x, y)$$

Sum rule

Posterior distribution

We discover a **Pistol** at the scene of the crime

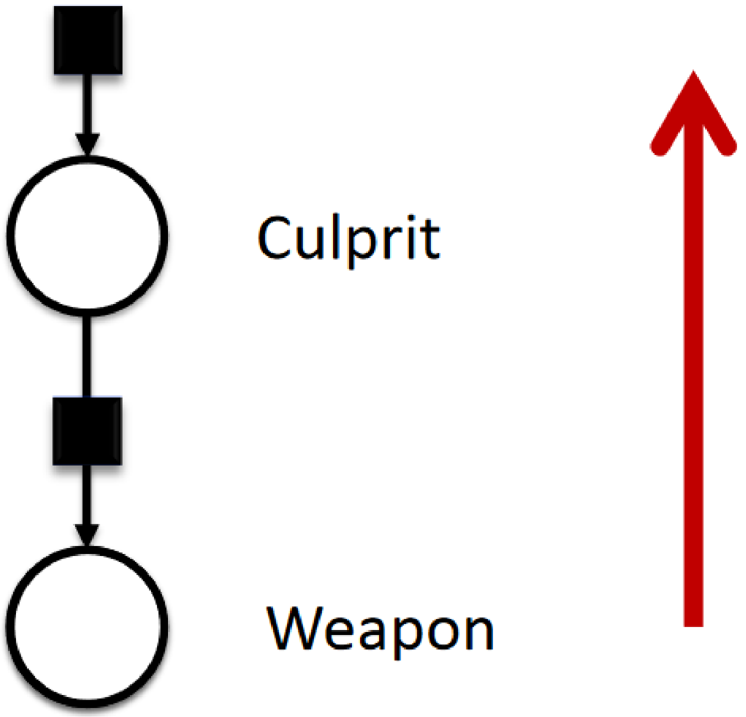
	Pistol	Knife	Poker	
Cook	4%	~~52%~~	~~24%~~	= 20%
Butler	16%	~~2%~~	~~2%~~	= 80%

This looks bad for the Butler!

Generative viewpoint

Murderer	Weapon
~~Cook~~	~~Knife~~
~~Butler~~	~~Knife~~
Cook	Pistol
~~Cook~~	~~Poker~~
~~Cook~~	~~Knife~~
Butler	Pistol
~~Cook~~	~~Poker~~
~~Cook~~	~~Knife~~
Butler	Pistol
~~Cook~~	~~Knife~~
…	…

Reasoning backwards



Bayes' theorem

$$P(x, y) = P(y|x)P(x)$$

likelihood

prior

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

posterior

Prior – belief before making a particular obs.

Posterior – belief after making the obs.

Posterior is the prior for the next observation

– Intrinsically incremental

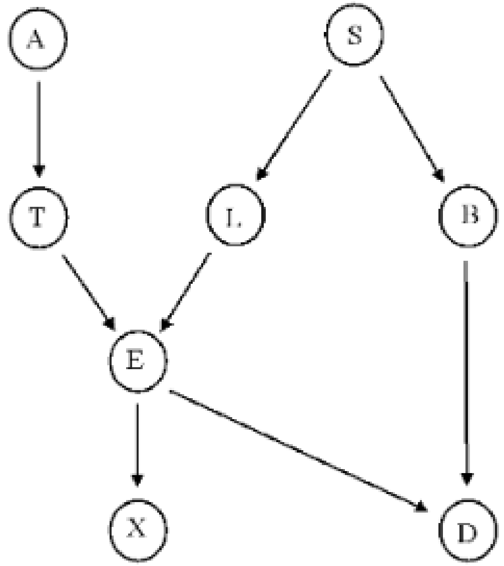
Medical diagnosis problem

- We will have (possibly incomplete) information such as symptoms and test results.
- We would like the probability that a given disease or a set of diseases is causing the symptoms.

Fictional Asia example (Lauritzen and Spiegelhalter 1988)

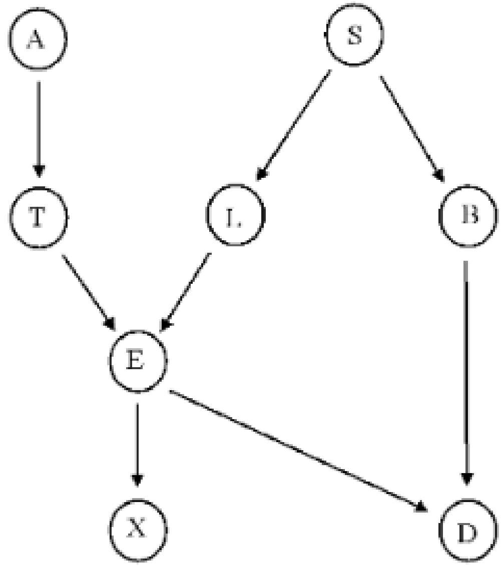
- A recent trip to Asia (A) increases the chance of Tuberculosis (T).
- Smoking is a risk factor for both lung cancer (L) and Bronchitis (B).
- The presence of either (E) tuberculosis or lung cancer can be treated by an X-ray result (X), but the X-ray alone cannot distinguish between them.
- Dyspnea (D) (shortness of breath) may be caused by bronchitis (B), or either (E) tuberculosis or lung cancer.

Bayesian networks



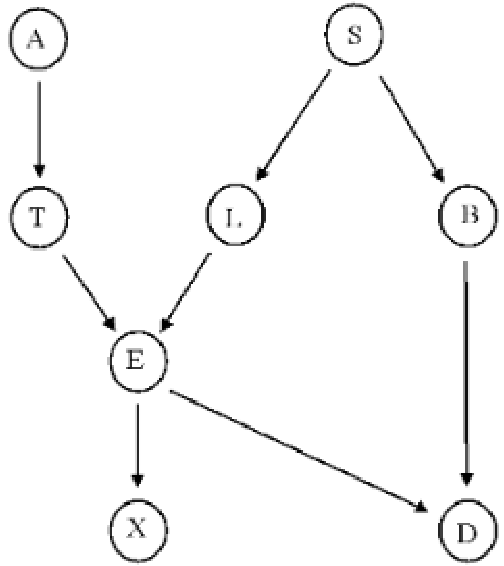
- Let x_i denote the different possible states of the node i.
- Associated with each arrow, there is a conditional probability.
- $p(x_L|x_S)$ denote the conditional probability that a patient has lung cancer given he does or does not smoke.

Bayesian networks



- $p(x_L|x_S)$ denote the conditional probability that a patient has lung cancer given he does or does not smoke.
- Here we say that “S” node is the parent of the “L” node.

Bayesian networks

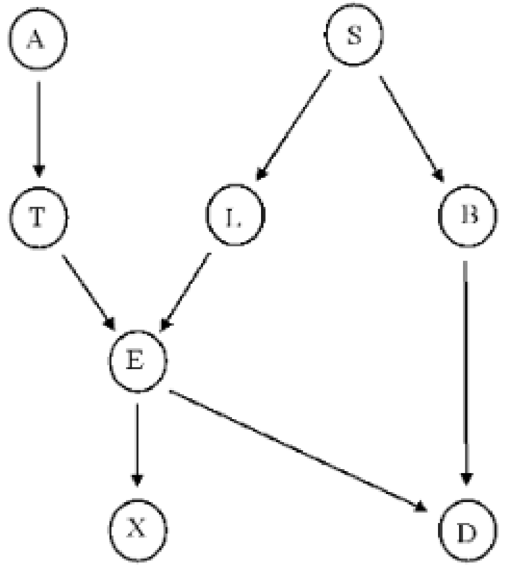


- Some nodes like D might have more than one parent.
- We can write the conditional probability as follows

$$p(x_D|x_E, x_B)$$

- Bayesian networks and other graphical models are most useful if the graph structure is sparse.

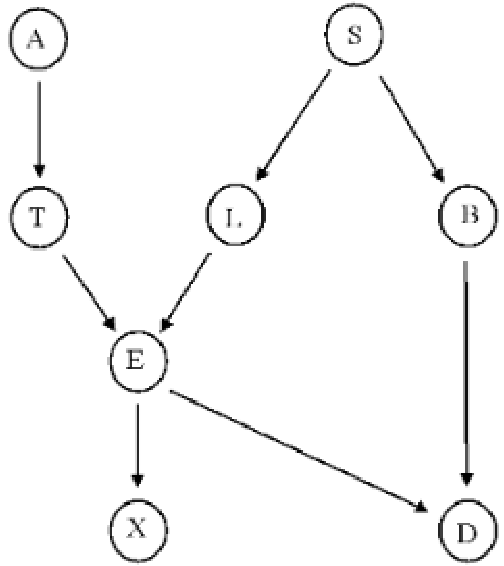
Joint probability in Bayesian networks



- The joint probability that the patient has some combination of the symptoms, test results, and diseases is just the product of the probabilities of the parents and the conditional ones:

$$p(\{\boldsymbol{x}\}) = p(\{x_A, x_S, x_T, x_L, x_B, x_E, x_X, x_D\})$$

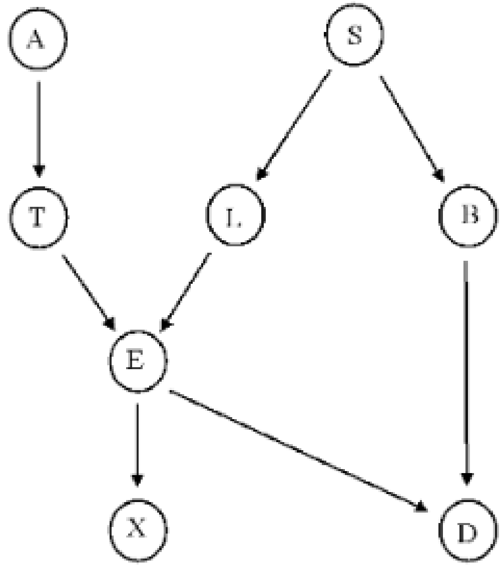
Joint probability in Bayesian networks



$$p(\{\boldsymbol{x}\}) = p(\{x_A, x_S, x_T, x_L, x_B, x_E, x_X, x_D\})$$

$$= p(x_A)p(x_S)p(x_T|x_A)p(x_L|x_S)p(x_B|x_S)p(x_E|x_T, x_L)p(x_X|x_E)p(x_D|x_E, x_B)$$

Joint probability in Bayesian networks



In general, Bayesian network is an acyclic directed graph with N random variables x_i that defines a joint probability function:

$$p(x_1, x_2, x_3, \ldots, x_N) = \Pi_{\{i=1\}}^{\{N\}} p(x_i | Par(x_i))$$

Marginal Probabilities

- Probability that a patient has a certain disease:

$$p(x_N) = \sum_{x_1} \sum_{x_2} \dots \sum_{x_{\{N-1\}}} p(x_1, x_2, \dots, x_N)$$

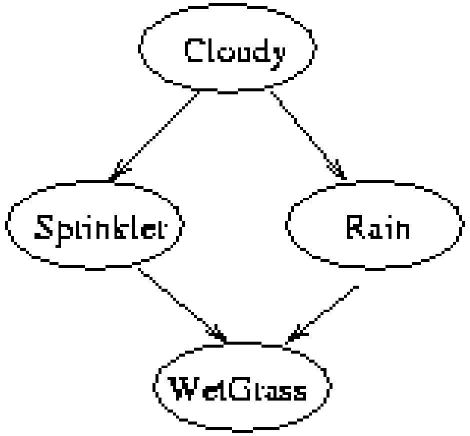
- Marginal probabilities are defined in terms of sums of all possible states of all other nodes.
- We refer to approximate marginal probabilities computed at a node x_i as beliefs and denote it as follows:

$$b(x_i)$$

- The virtue of BP is that it can compute the beliefs (at least approximately) in graphs that can have a large number of nodes efficiently.

Bayesian Networks

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

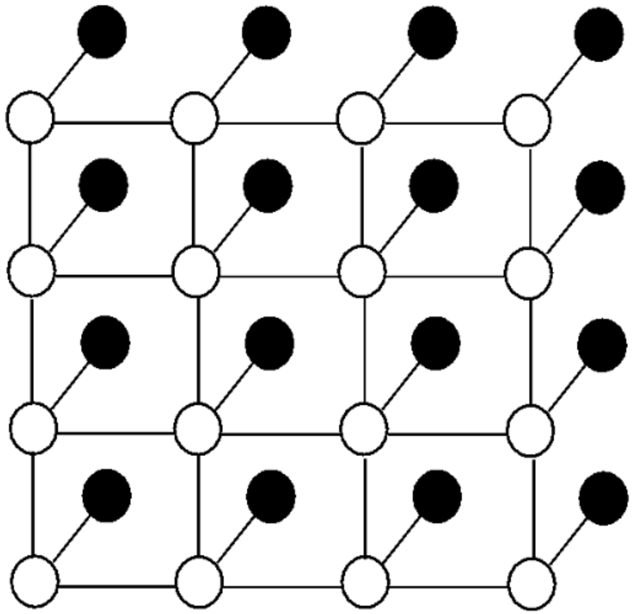
S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Courtesy: Keven Murphy

Pairwise Markov Random Fields

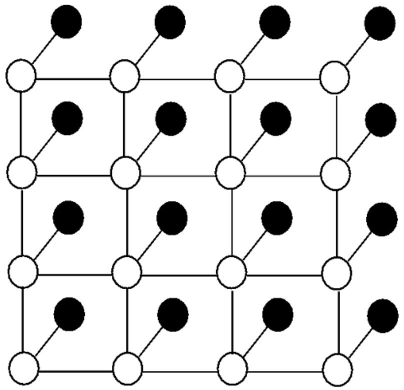
- Attractive theoretical model for many computer vision tasks (Geman 1984).
- Many computer vision problems such as segmentation, recognition, stereo reconstruction are solved.

Pairwise Markov Random Fields



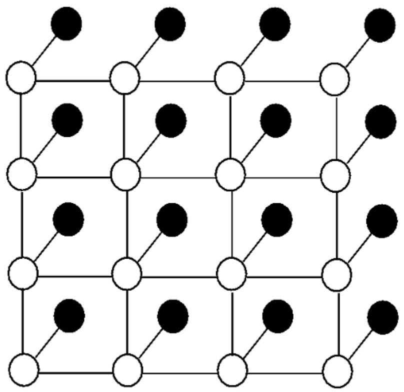
- In a simple depth estimation problem on an image of size 1000 x 1000, every node can have states from 1 to D denoting different distances from the camera center.

Pairwise Markov Random Fields



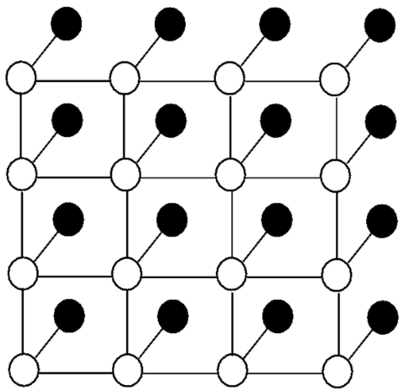
- Let us observe certain quantities about the image y_i and we are interested in computing other entities about the underlying scene x_i.
- The indices i denote certain pixel locations.
- Assume that there is some statistical dependency between x_i and y_i and let us denote it by some compatibility function $\phi_i(x_i, y_i)$, also referred to as the evidence.

Pairwise Markov Random Fields



- To be able to infer anything about the scene, there should be some kind of structure on x_i.
- In a 2D grid, x_i should be compatible with nearby scene elements x_j.
- Let us consider a compatibility function $\psi_{ij}(x_i, y_j)$ where the function connects only nearby pixel elements.

Pairwise Markov Random Fields



$$p(\{\boldsymbol{x}\},\{\boldsymbol{y}\}) = \frac{1}{\mathrm{Z}}\Pi_{\{ij\}}\psi_{ij}\left(x_i, x_j\right)\Pi_i\phi_i(x_i, y_i)$$

- Here Z is the normalization constant.
- The Markov Random fields is pairwise because the compatibility function depends only on pairs of adjacent pixels.
- There is no parent-child relationship in MRFs and we don’t have directional dependencies.

Potts Model

- The interaction $J_{ij}(x_i, x_j)$ between two neighboring nodes is given by

$$J_{ij}(x_i, x_j) = \ln \psi_{ij}(x_i, x_j)$$

- The field $h_i(x_i)$ at each node is given by

$$h_i(x_i) = \ln \phi_i(x_i, y_i)$$

Potts Model

- The Potts model energy is defined as below:

$$E(\{x_i\}) = -\sum_{ij} J_{ij}(x_i, x_j) - \sum_{i} h(x_i)$$

Boltzmann's law from statistical mechanics

- The pairwise MRF exactly corresponds to the Potts model energy at temperature T = 1.

$$p(\{x_i\}) = \frac{1}{Z} e^{-\frac{E(\{x_i\})}{T}}$$

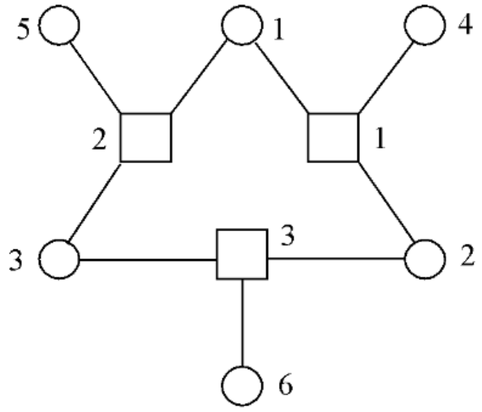
- The normalization constant Z is called the partition function.

ISING model

- If the number of states is just 2 then the model is called an ising model.
- The problem of computing beliefs can be seen as computing local magnetizations in Ising model.
- The spin glass energy function is written below using two-state spin variables $s_i = \{+1, -1\}$:

$$E(\{s_i\}) = -\sum_{ij} J_{ij}(s_i, s_j) - \sum_i h(s_i)$$

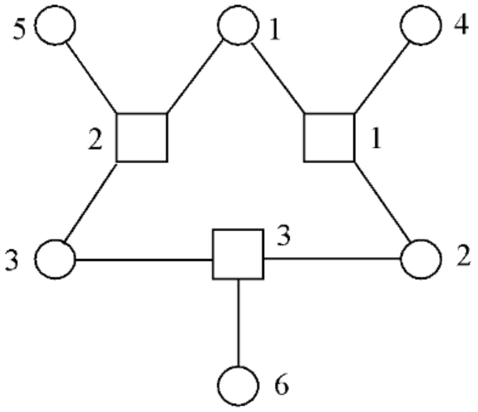
Tanner Graphs and Factor Graphs



We have transmitted N = 6 bits with k = 3 parity check constraints.

- Error-correcting codes: We try to decode the information transmitted through noisy channel.
- The first parity check code forces the sum of bits from #1, #2, and #5 to be even.

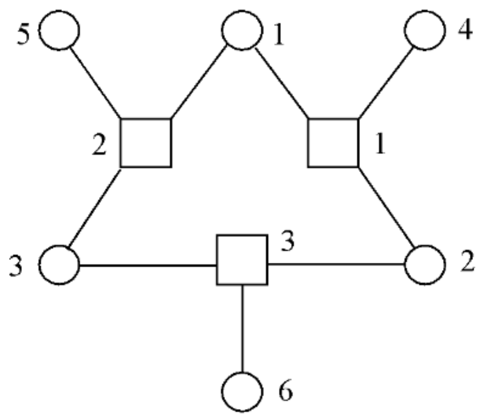
Tanner Graphs and Factor Graphs



We have transmitted N = 6 bits with k = 3 parity check constraints.

- Let y_i be the received bit and the transmitted bit be given by x_i.
- Joint probability can be written as follows:
- $p(\{x, y\}) = \frac{1}{Z}\psi_{124}(x_1, x_2, x_4)\psi_{135}(x_1, x_3, x_5)\psi_{236}(x_2, x_3, x_6)\Pi_i p(y_i|x_i)$

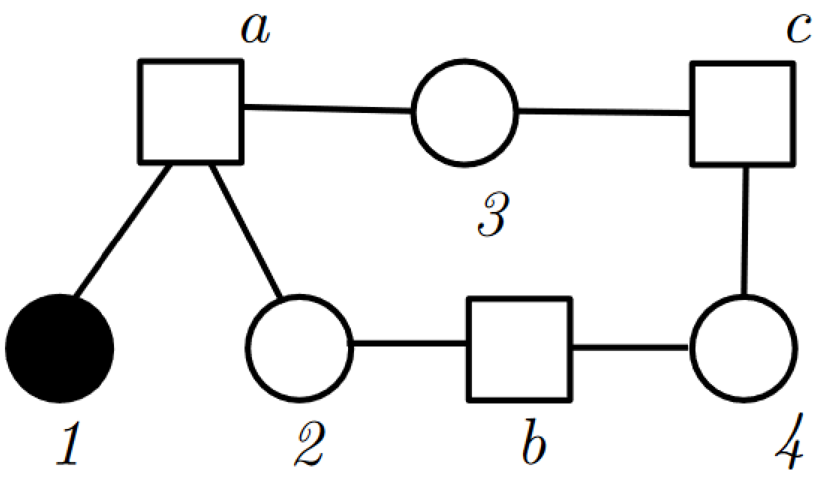
Tanner Graphs and Factor Graphs



We have transmitted N = 6 bits with k = 3 parity check constraints.

- The parity check functions have values 1 when the bits satisfy the constraint and 0 if they don't.
- A decoding algorithm typically tries to minimize the number of bits that are decoded incorrectly.

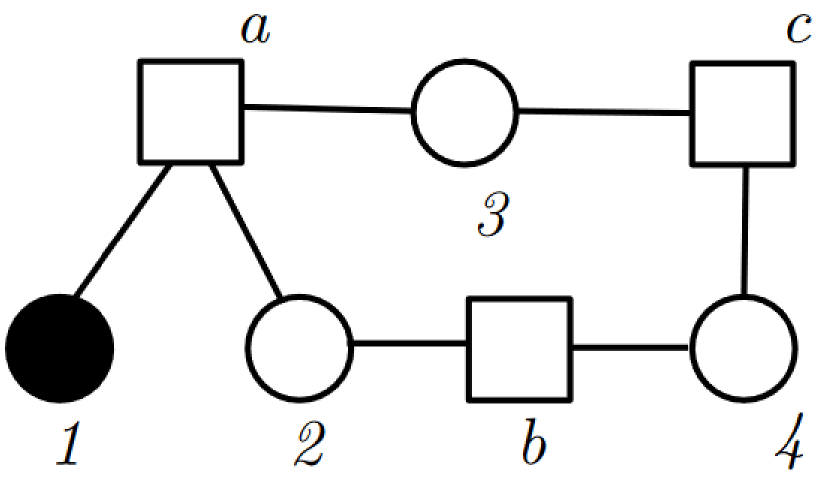
Factor Graphs (Using Energy or Cost functions)



Toy factor graph with one observed variable, 3 hidden variables, and 3 factor nodes

- Factor graphs are bipartite graphs containing two types of nodes: variable nodes (circles) and factor nodes (squares).

Factor Graphs (Using Energy or Cost functions)

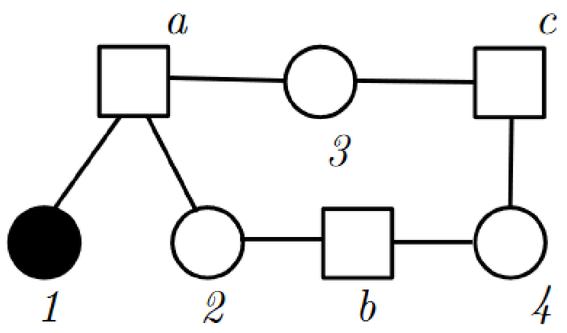


Toy factor graph with one observed variable, 3 hidden variables, and 3 factor nodes

- $C(x_1, x_2, x_3, x_4) = C_a(x_1, x_2, x_3) + C_b(x_2, x_4) + C_c(x_3, x_4)$

Factor Graphs (Using Energy or Cost functions)

x_1	x_2	x_3	C_a
0	0	0	∞
0	0	1	0
0	1	0	0
0	1	1	∞
1	0	0	0
1	0	1	∞
1	1	0	∞
1	1	1	0



x_2	x_4	C_b
0	0	1.2
0	1	1.7
0	2	3.2
1	0	1.9
1	1	0.6
1	2	1.4

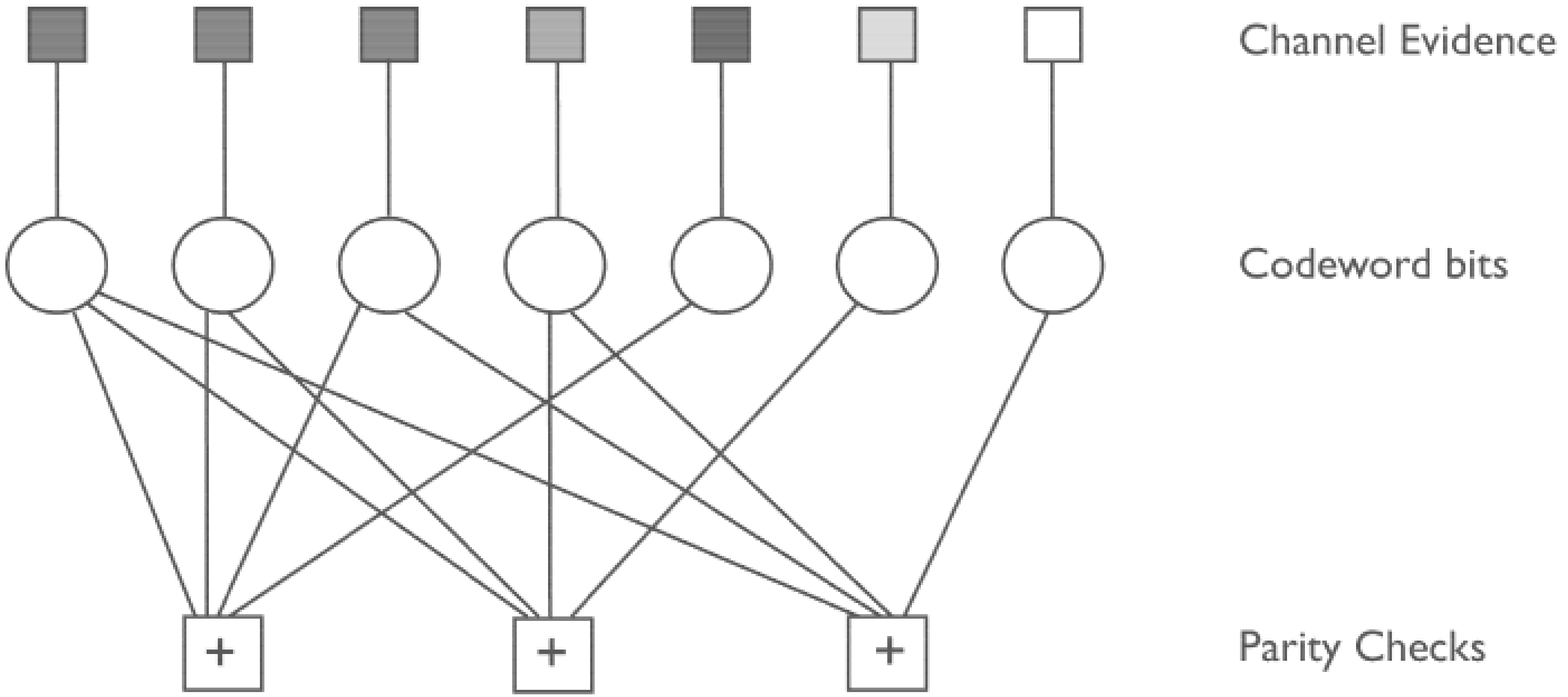
x_3	x_4	C_c
0	0	0.4
0	1	1.9
0	2	0.2
1	0	4.9
1	1	0.3
1	2	2.4

Lowest Energy Configurations

- $C(x_1, x_2, x_3, x_4) = C_a(x_1, x_2, x_3) + C_b(x_2, x_4) + C_c(x_3, x_4)$

- Finding the lowest energy state and computing the corresponding variable assignments is a hard problem
- In most general cases, the problem is NP-hard.

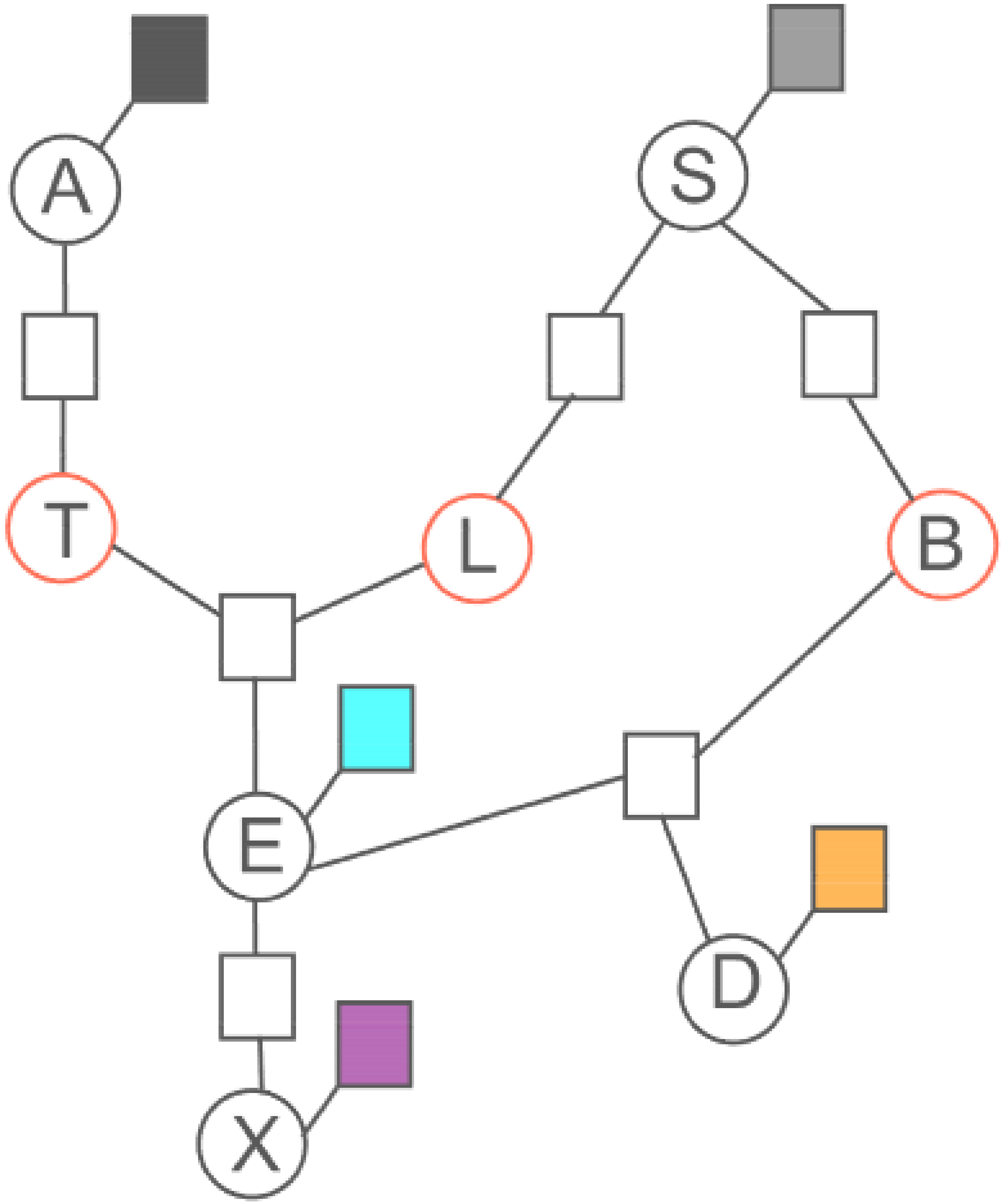
Factor Graphs for Error Correction



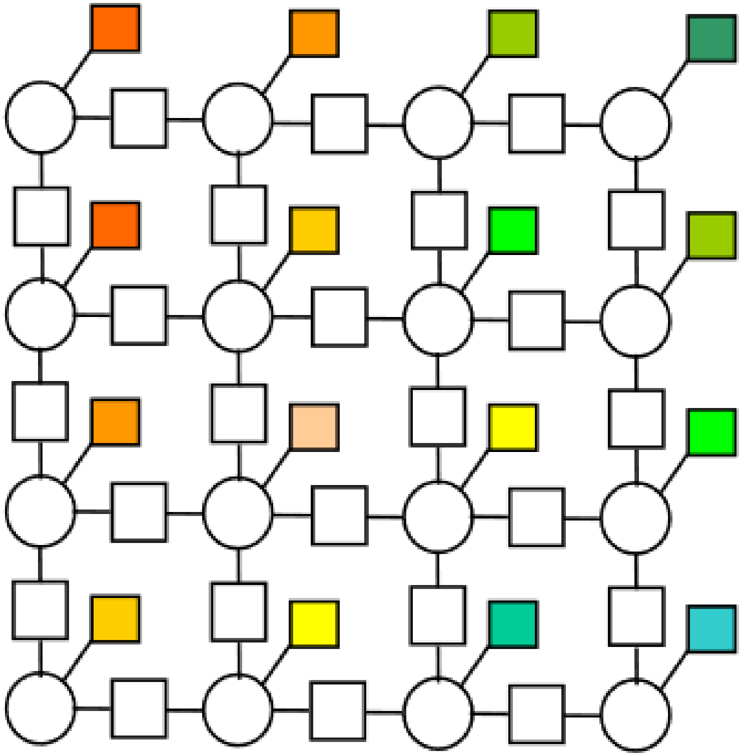
A factor graph for (N=7,k=3) Hamming code, which has 7 codeword bits, of the left-most four are information bits and the last 3 are parity bits.

Factor graph for the medical expert system

- Here the variables are given by Asia (A) , Tuberculosis (T), Lung cancer (L), Smoker (S), Bronchitis (B), Either (E), X-ray (X), and D.



Stereo reconstruction in Computer Vision



Set up the Factor graphs

- Point matching between 2 images given the Fundamental matrix.
- Point correspondences between 2 sets of 3D points.
- The classical problem of line-labeling.