Generalized Belief Propagation and Free Energy Minimization

Jonathan Yedidia

Mitsubishi Electric Research Labs (MERL)

Bill Freeman (*MIT*)

Yair Weiss (*Hebrew University*)

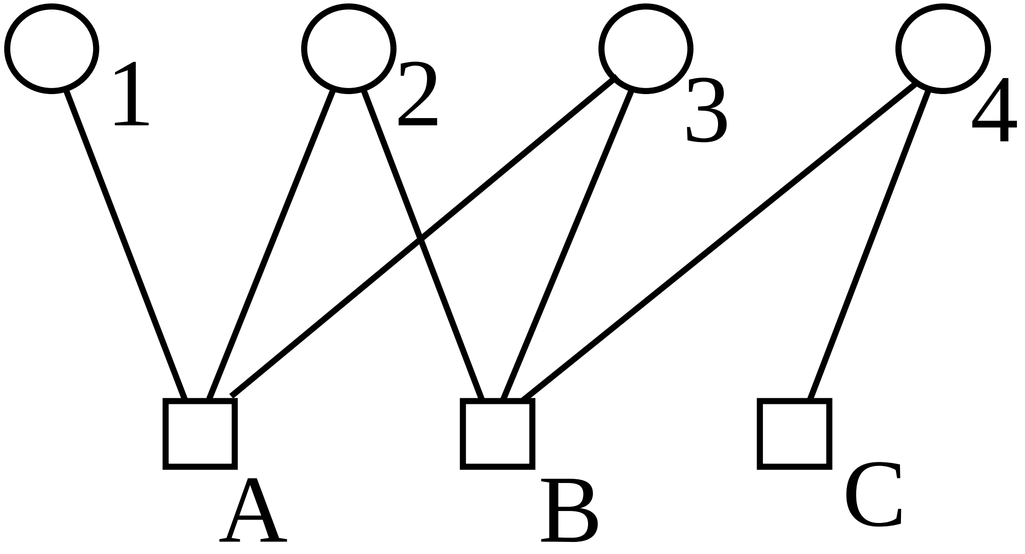
Outline

- Motivation & factor graphs
- Standard belief propagation
- Free energy approximations
- Methods to generate “valid” region-based approximations: (Bethe, junction graphs, cluster variational method, *region graphs*)
- Generalized belief propagation

Factor Graphs

(Kschischang, et.al. 2001)

$$\boxed{p(X) = \frac{1}{Z}\sum_{a=1}^{M} f_a(X_a)}$$



$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} f_A(x_1, x_2, x_3)\, f_B(x_2, x_3, x_4)\, f_C(x_4)$$

Computing Marginal Probabilities

$$\boxed{p_S(X_S) = \sum_{X \setminus X_S} p(X)}$$

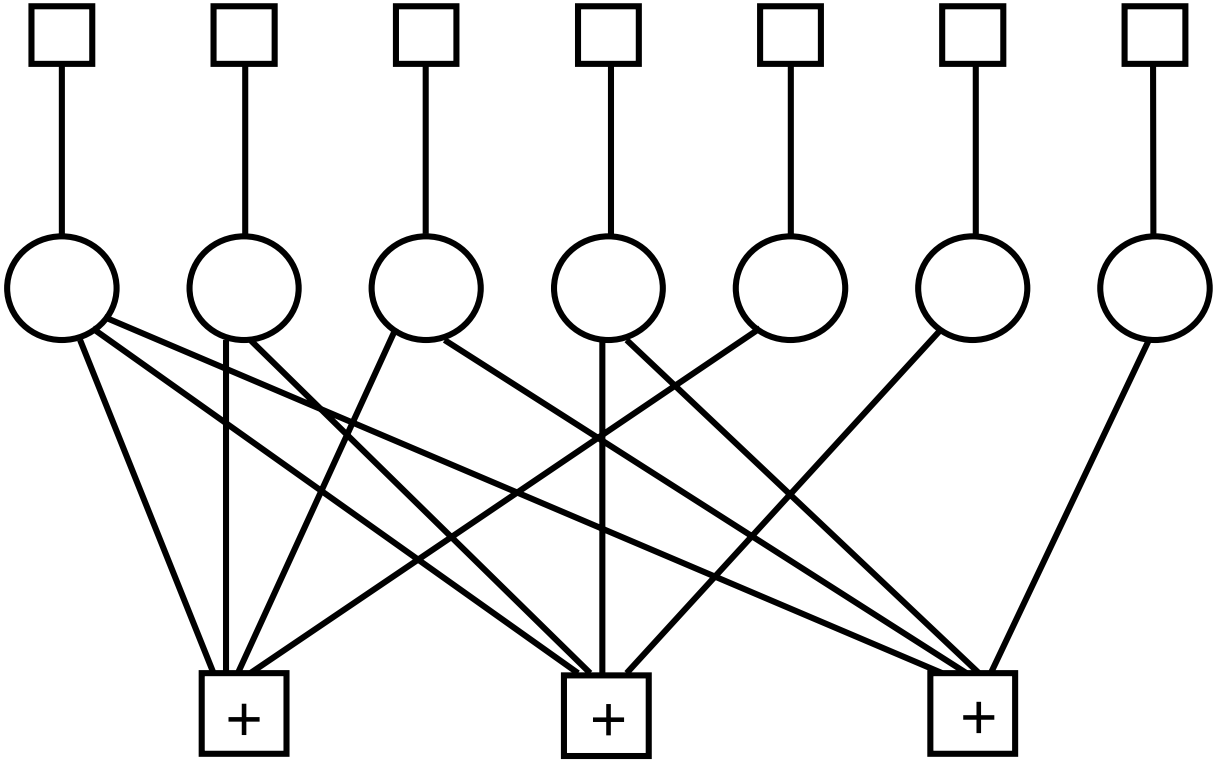
Fundamental for

- Decoding error-correcting codes
- Inference in Bayesian networks
- Computer vision
- Statistical physics of magnets

Non-trivial because of the huge number of terms in the sum.

Error-correcting Codes

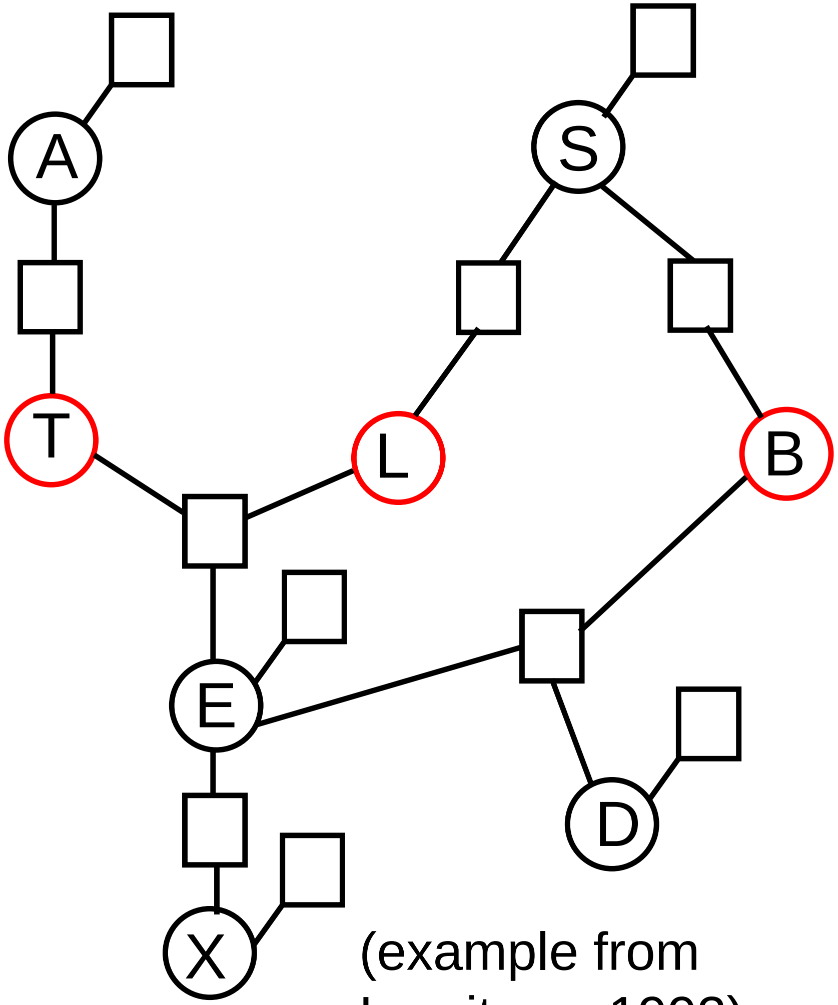
(Tanner, 1981
Gallager, 1963)



Marginal Probabilities = *A posteriori* bit probabilities

Bayesian Networks

(Pearl, 1988)

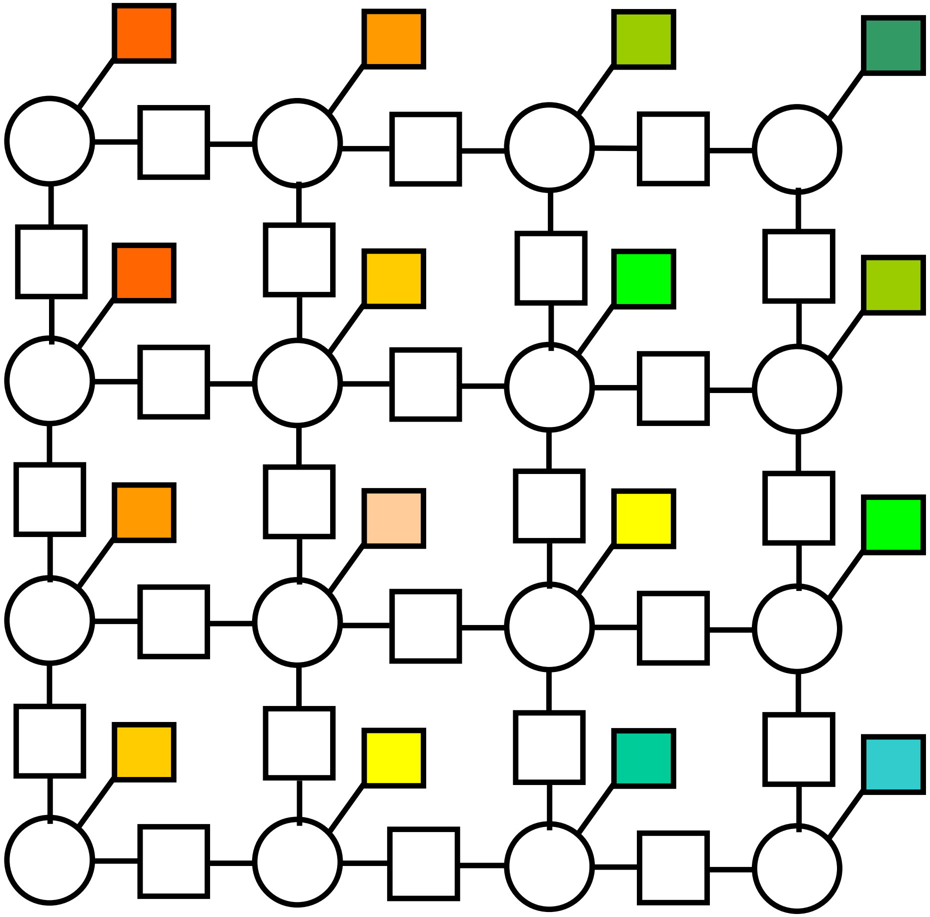


(example from Lauritzen, 1992)

Marginal Probabilities= “beliefs” about possible diagnoses

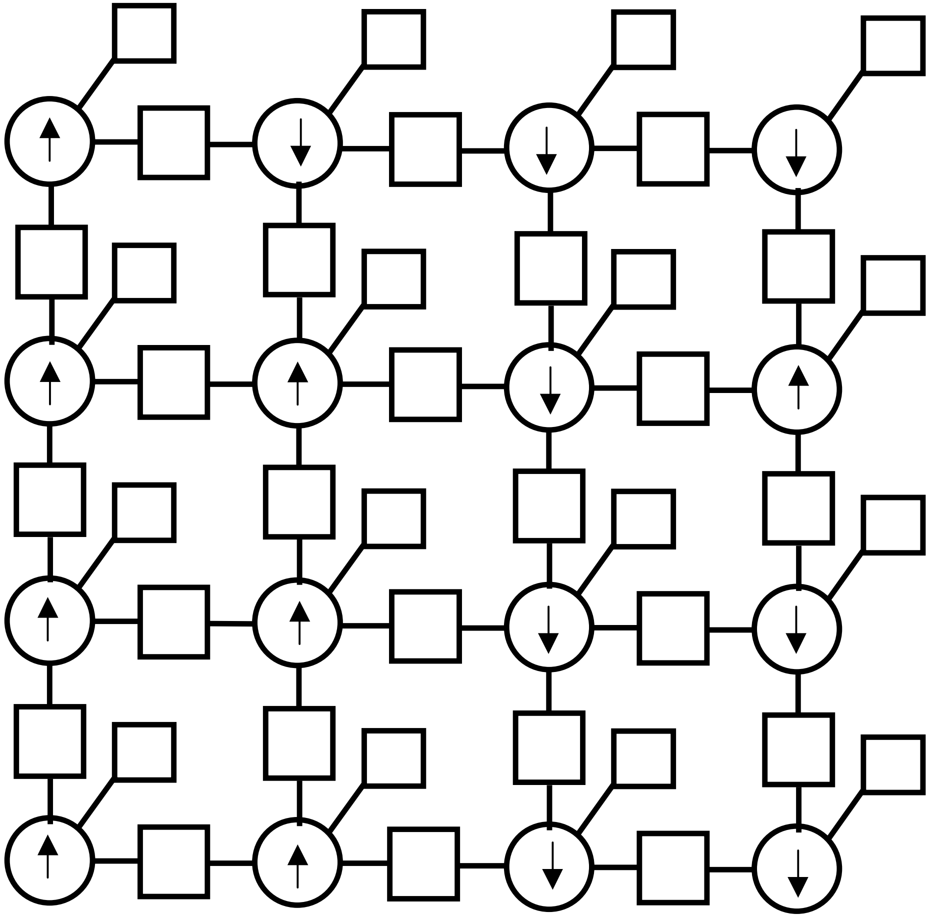
Computer Vision

(Geman & Geman 1984)



Marginal Probabilities=
“beliefs” about possible
underlying scenes

Statistical Physics



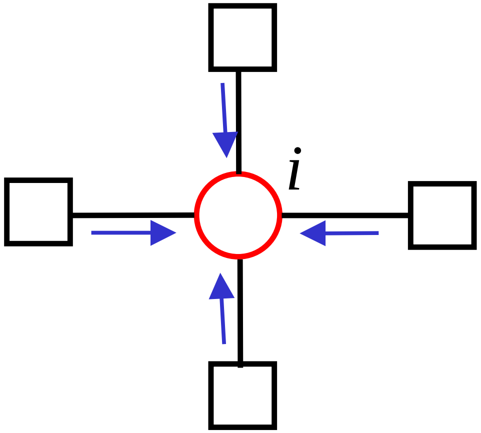
(Ising 1925, Edwards & Anderson 1975)

Marginal Probabilities= local magnetization

Simplifications We Will *Not* Exploit

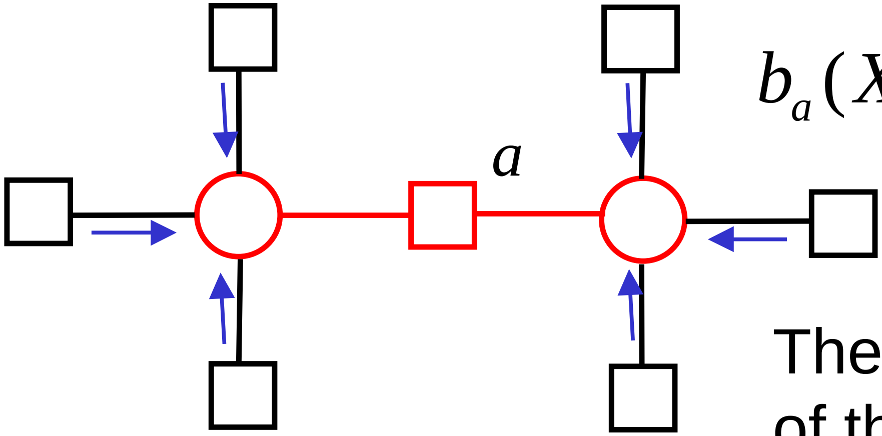
- Functions may be parity checks (codes)
- Functions may be conditional probabilities (Bayes nets)
- Functions may be only pair-wise (computer vision)
- Functions may have translational symmetry (statistical physics)

Standard Belief Propagation



$$b_i(x_i) \propto \prod_{a \in N(i)} m_{a \to i}(x_i)$$

“beliefs” “messages”



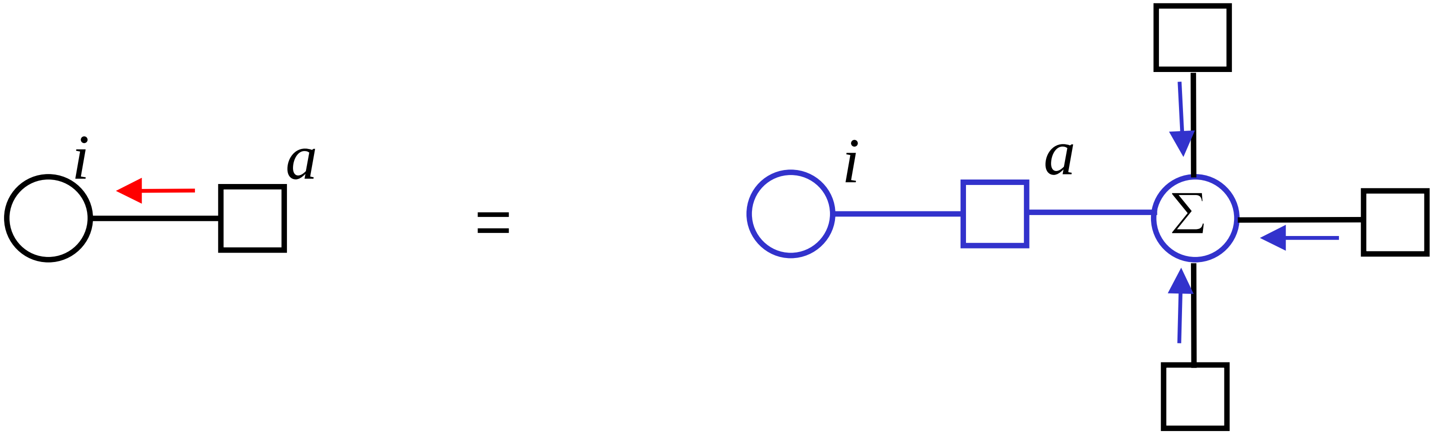
$$b_a(X_a) \propto f_a(X_a) \prod_{i \in N(a)} \prod_{b \in N(i) \backslash a} m_{b \to i}(x_i)$$

The “belief” is the BP approximation of the marginal probability.

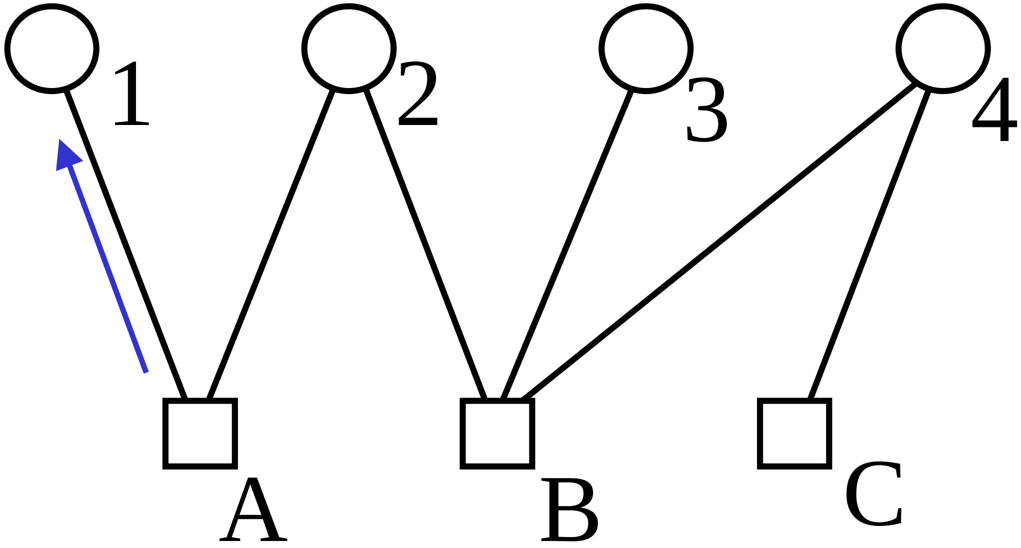
BP Message-update Rules

Using $b_i(x_i) = \sum_{X_a \backslash x_i} b_a(X_a),$ we get

$$\boxed{m_{a\to i}(x_i) = \sum_{X_a \backslash x_i} f_a(X_a) \prod_{j\in N(a)\backslash i} \prod_{b\in N(j)\backslash a} m_{b\to j}(x_j)}$$

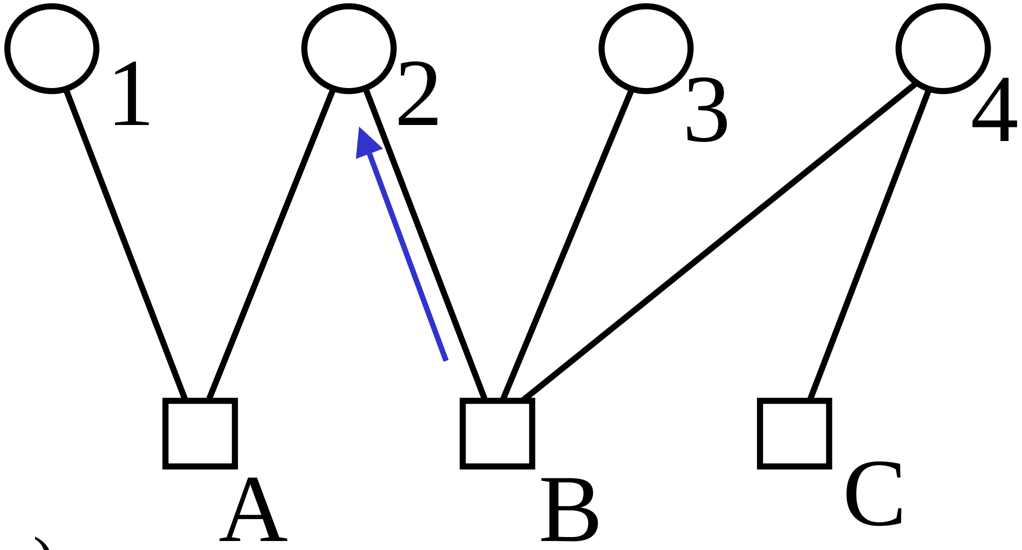


BP Is Exact for Trees



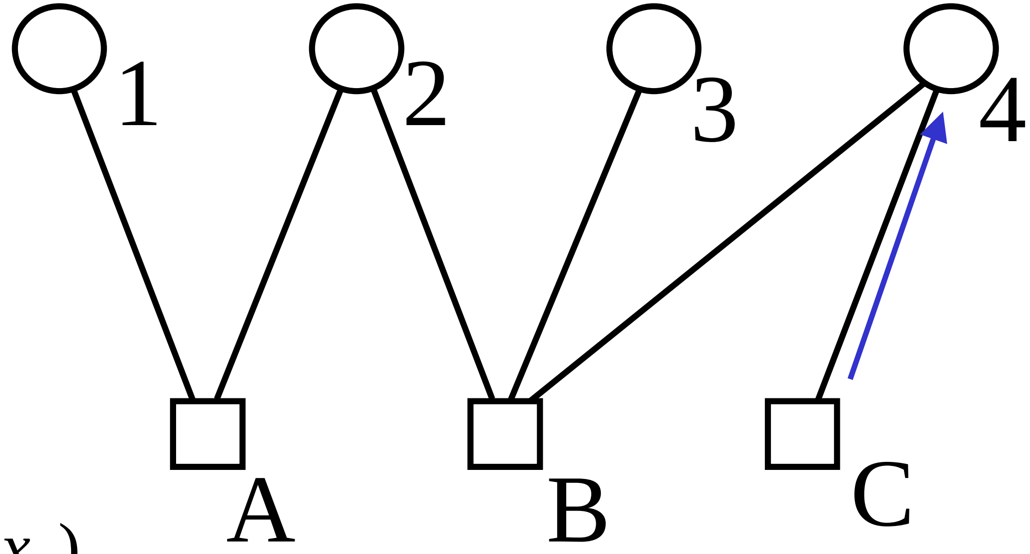
$$b_1(x_1) \propto m_{A\rightarrow 1}(x_1)$$

BP Is Exact for Trees



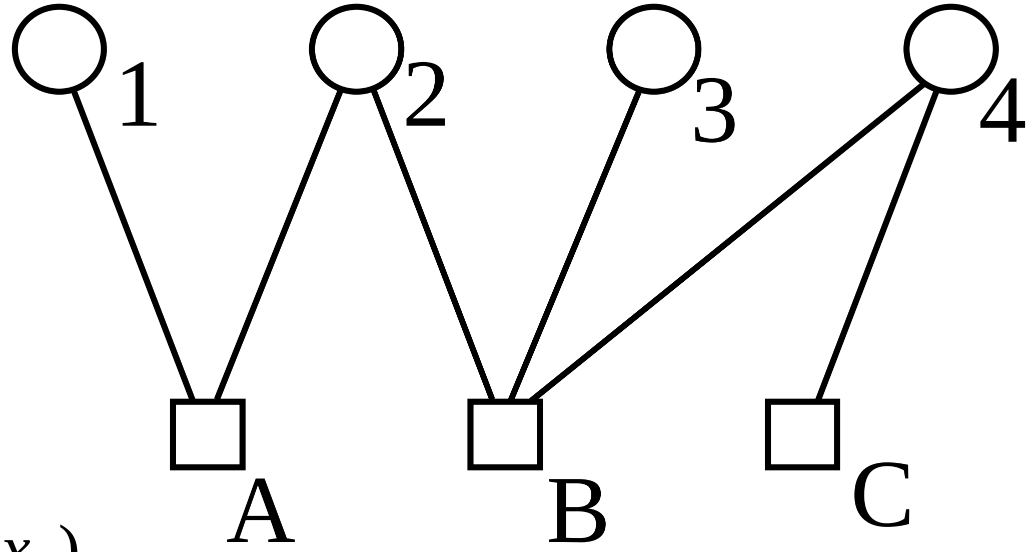
$$b_1(x_1) \propto m_{A\to 1}(x_1)$$
$$\propto \sum_{x_2} f_A(x_1, x_2) m_{B\to 2}(x_2)$$

BP Is Exact for Trees



$$
\begin{aligned}
b_1(x_1) &\propto m_{A\to 1}(x_1) \\
&\propto \sum_{x_2} f_A(x_1,x_2) m_{B\to 2}(x_2) \\
&\propto \sum_{x_2,x_3,x_4} f_A(x_1,x_2) f_B(x_2,x_3,x_4) m_{C\to 4}(x_4)
\end{aligned}
$$

BP Is Exact for Trees



$$
\begin{aligned}
b_1(x_1) &\propto m_{A\to 1}(x_1) \\
&\propto \sum_{x_2} f_A(x_1,x_2)\, m_{B\to 2}(x_2) \\
&\propto \sum_{x_{2,}x_3,x_4} f_A(x_1,x_2)\, f_B(x_2,x_3,x_4)\, m_{C\to 4}(x_4) \\
&\propto \sum_{x_{2,}x_3,x_4} f_A(x_1,x_2)\, f_B(x_2,x_3,x_4)\, f_C(x_4)
\end{aligned}
$$

Variational (Gibbs) Free Energy

Kullback-Liebler Distance:

$$D(b \| p) \equiv \sum_X b(X) \ln \frac{b(X)}{p(X)}$$

“Boltzmann’s Law” (definition of “energy”):

$$p(X) = \frac{1}{Z} \exp\left[-E(X)\right]$$

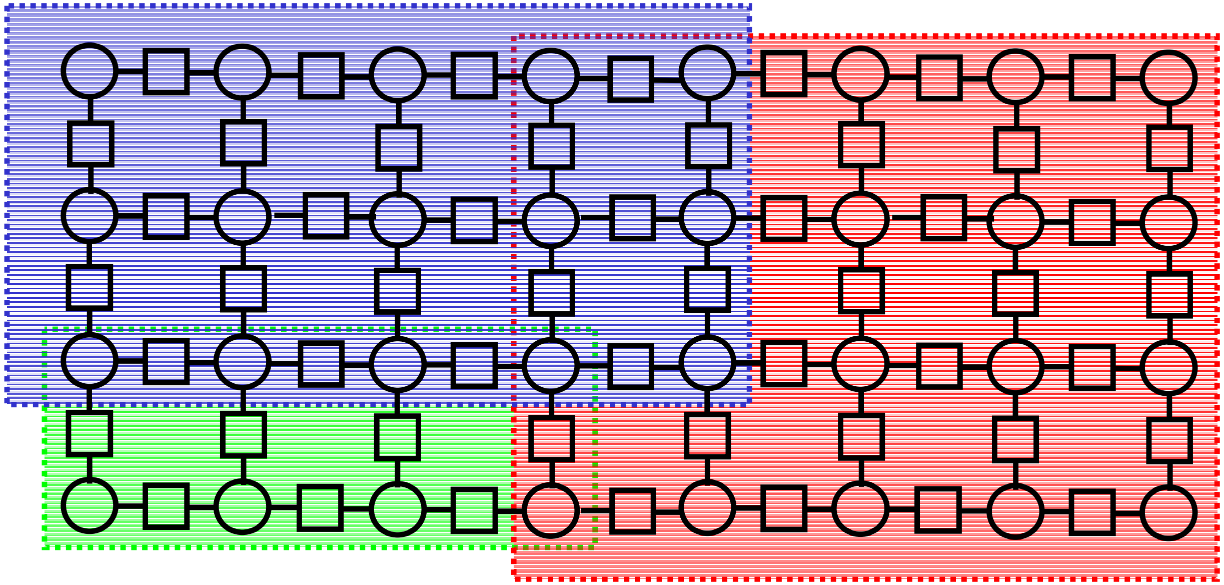
$$D(b \| p) \equiv \underbrace{\sum_X b(X) \overset{\overset{U(b)}{\downarrow}}{} E(X) + \sum_X b(X) \overset{\overset{-H(b)}{\downarrow}}{} \ln b(X)} + \ln Z$$

Gibbs Free Energy $G(b)$;
minimized when $b(X) = p(X)$

Region-based Approximations to the Gibbs Free Energy (Kikuchi, 1951)

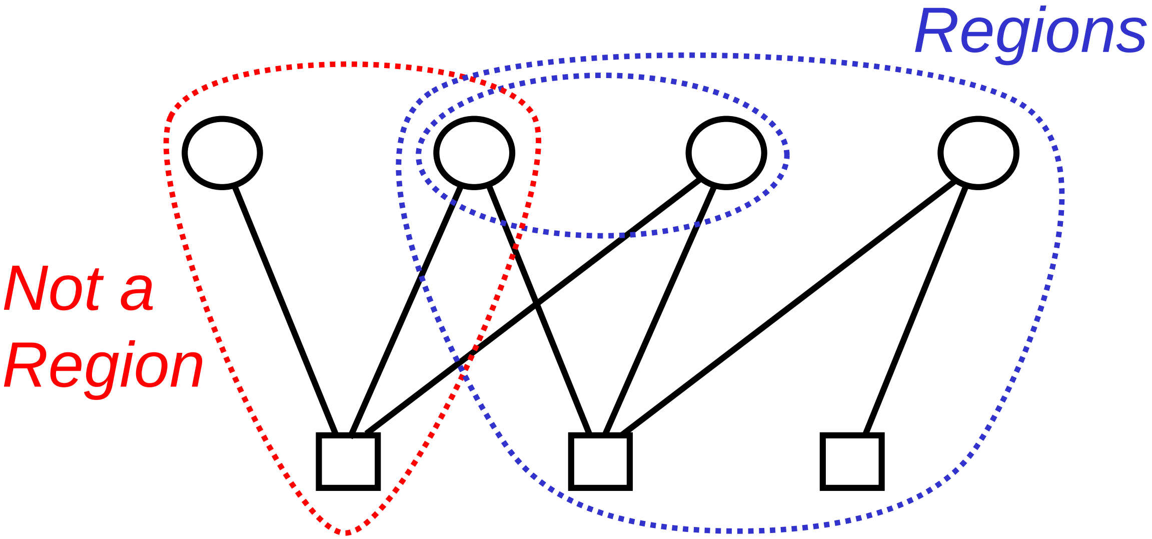
Exact: $G[b(X)]$ *(intractable)*

Regions: $G[\{b_r(X_r)\}]$



Defining a "Region"

A *region r* is a set of variable nodes V_r and factor nodes F_r such that if a factor node *a* belongs to F_r , all variable nodes neighboring *a* must belong to V_r.



Region Definitions

Region *states*: X_r

Region *beliefs*: $b_r\left(X_r\right)$

Region *energy*: $E_r\left(X_r\right) = -\sum_{a \in F_r} \ln f_a\left(X_a\right)$

Region *average energy*: $U_r\left(b_r\right) = \sum_{X_r} b_r\left(X_r\right) E_r\left(X_r\right)$

Region *entropy*: $H_r\left(b_r\right) = -\sum_{X_r} b_r\left(X_r\right) \ln b_r\left(X_r\right)$

Region *free energy*: $G_r\left(b_r\right) = U_r\left(b_r\right) - H_r\left(b_r\right)$

Important Technical Point

For our approximations, we *will* require that all $b_r(X_r)$ are locally consistent with each other, but we do *not* seek, nor even require the existence of, a global $b(X)$ such that $b_r(X_r) = \sum_{X \setminus X_r} b(X)$.

This contrasts with *mean field theory*, where one seeks a tractable global function $b(X)$ that minimizes the Gibbs free energy.

“Valid” Approximations

Introduce a set of regions *R,* and a *counting number* c_r for each region *r* in *R*, such that $c_r=1$ for the largest regions, and for every factor node *a* and variable node *i,*

$$\boxed{\sum_{r\in R} c_r I\left(a\in F_r\right)=\sum c_r\, I\left(i\in V_r\right)=1}$$

Indicator functions

Count every node once!

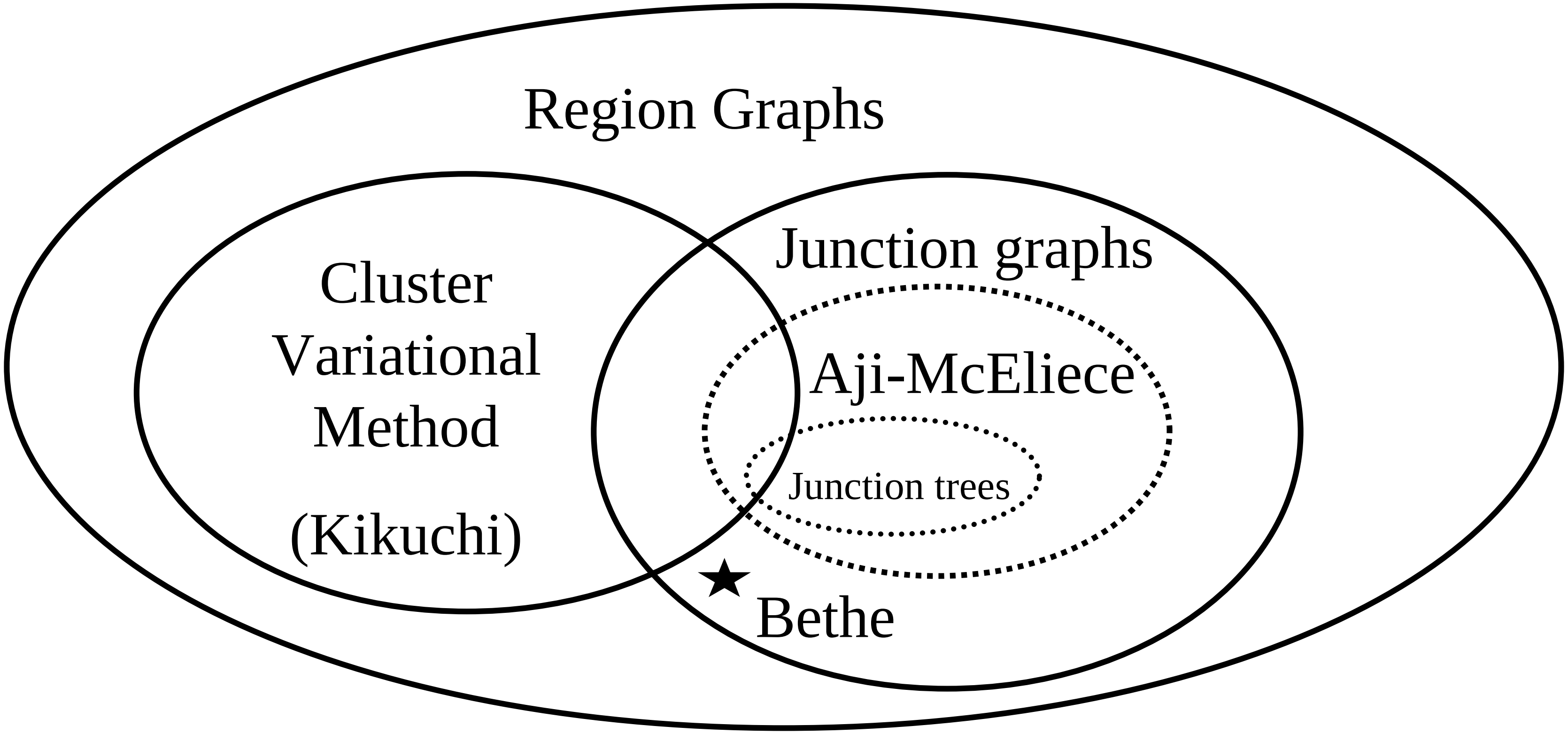
$$\boxed{G\left(\{b_r\}\right)=\sum_{r\in R} c_r G_r\left(b_r\right)}$$

Entropy and Energy

- $\sum c_r I(a \in F_r) = 1$: Counting each factor node once makes the approximate energy *exact* (if the beliefs are).

- $\sum c_r I(i \in V_r) = 1$: Counting each variable node once makes the approximate entropy *reasonable.*

Methods to Generate Valid Region-based Approximations



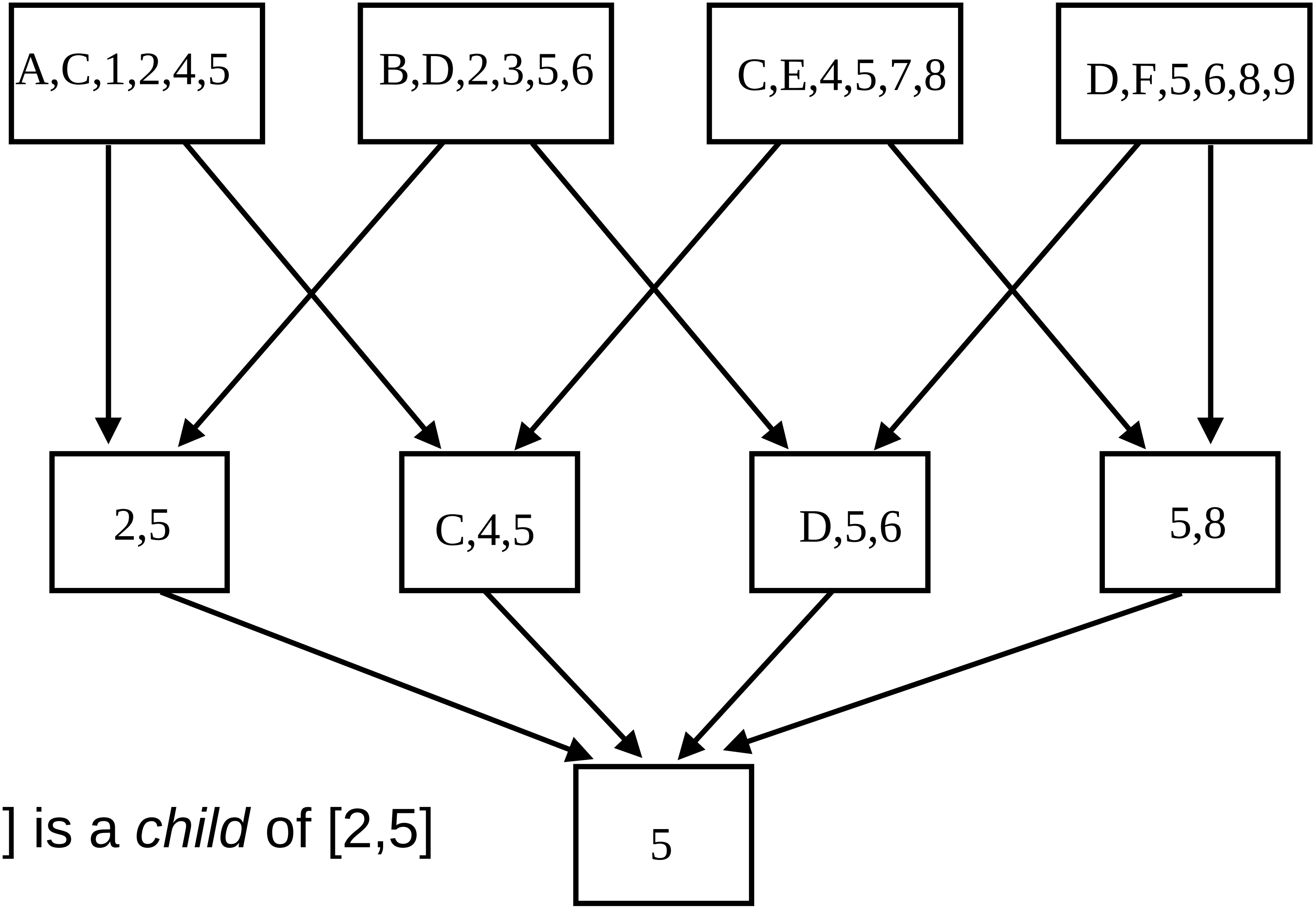
(Bethe is example of Kikuchi for Factor graphs with no 4-cycles; Bethe is example of Aji-McEliece for “normal” factor graphs.)

Definition of a Region Graph

- Labeled, directed graph of regions.
- Arc may exist from region *A* to region *B* if *B* is a subset of *A.*
- Sub-graphs formed from regions containing a given node are connected.
- $c_r = 1 - \sum_{s \in A(r)} c_s$ where $A(r)$ is the set of *ancestors* of region *r.*
- We insist that

$$\forall a, \forall i : \sum_{r \in R} c_r I\left(a \in F_r\right) = \sum c_r I\left(i \in V_r\right) = 1.$$

Example of a Region Graph



[5] is a *child* of [2,5]

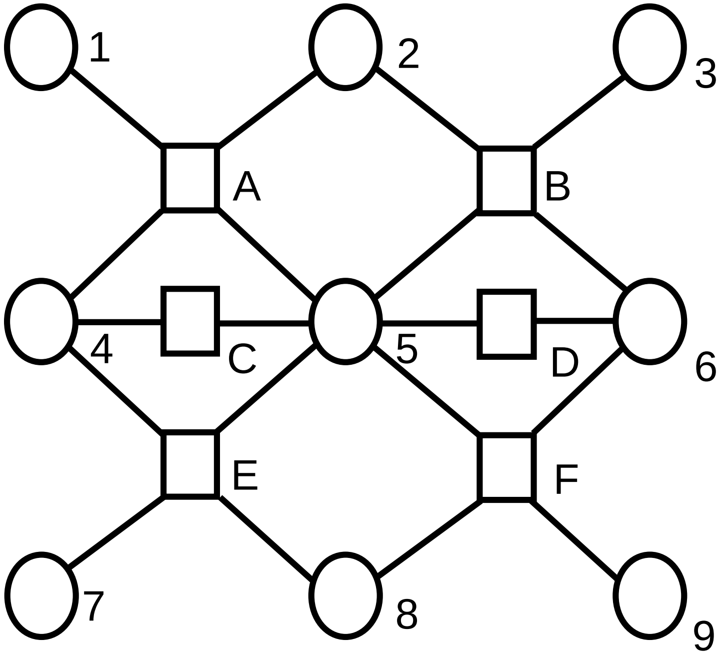
Bethe Method

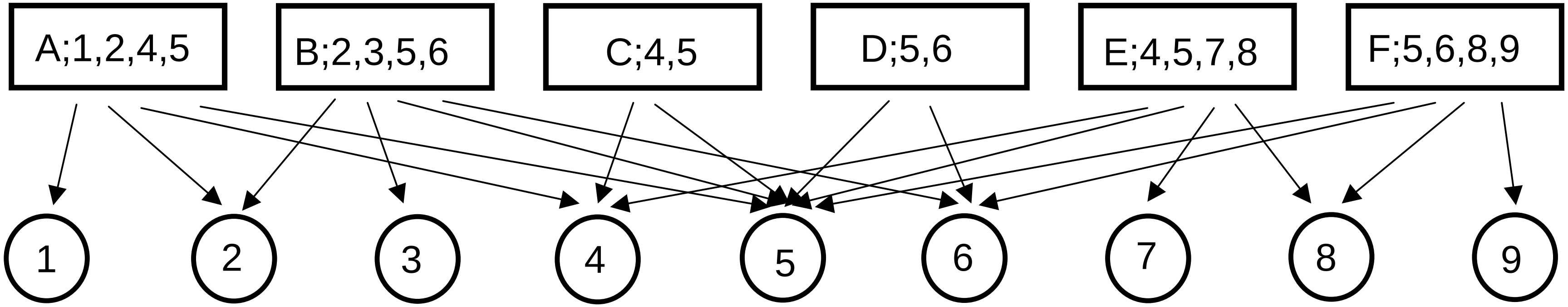
(after Bethe, 1935)

Two sets of regions:

Large regions containing a single factor node *a* and all attached variable nodes. $\boxed{c_r = 1}$

Small regions containing a single variable node *i*. $\boxed{c_r = 1 - d_i}$




A;1,2,4,5
B;2,3,5,6
C;4,5
D;5,6
E;4,5,7,8
F;5,6,8,9

Bethe Approximation to Gibbs Free Energy

$$G_{Bethe} = \sum_{a} \sum_{X_a} b_a(X_a) \ln\left(\frac{b_a(X_a)}{f_a(X_a)} \right) + \sum_{i} (1 - d_i) \sum_{x_i} b_i(x_i) \ln b_i(x_i)$$

Equal to the exact Gibbs free energy when the factor graph is a tree because in that case,

$$b(X) = \prod_{a} b_a(X_a) \prod_{i} b_i(x_i)^{1-d_i}$$

Minimizing the Bethe Free Energy

$$L = G_{Bethe} + \sum_i \gamma_i \{ \sum_{x_i} b_i(x_i) - 1 \}$$

$$+ \sum_a \sum_{i \in N(a)} \sum_{x_i} \lambda_{ai}(x_i) \left\{ \sum_{X_a \setminus x_i} b_a(X_a) - b_i(x_i) \right\}$$

$$\frac{\partial L}{\partial b_i(x_i)} = 0 \quad \Longrightarrow \quad b_i(x_i) \propto \exp\left(\frac{1}{d_i - 1} \sum_{a \in N(i)} \lambda_{ai}(x_i) \right)$$

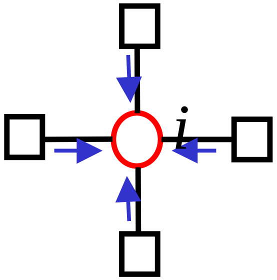
$$\frac{\partial L}{\partial b_a(X_a)} = 0 \quad \Longrightarrow \quad b_a(X_a) \propto \exp\left(-E_a(X_a) + \sum_{i \in N(a)} \lambda_{ai}(x_i) \right)$$

Bethe = BP

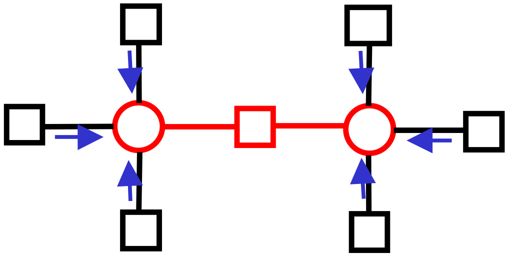
Identify

$$\boxed{\lambda_{ai}(x_i) = \ln \prod_{b\in N(i)\neq a} m_{b\to i}(x_i)}$$

to obtain BP equations:



$$b_i(x_i) \propto \prod_{a\in N(i)} m_{a\to i}(x_i)$$



$$b_a(X_a) \propto f_a(X_a) \prod_{i\in N(a)} \prod_{b\in N(i)\backslash a} m_{b\to i}(x_i)$$

Junction Graphs

A labeled directed graph with two types of regions:

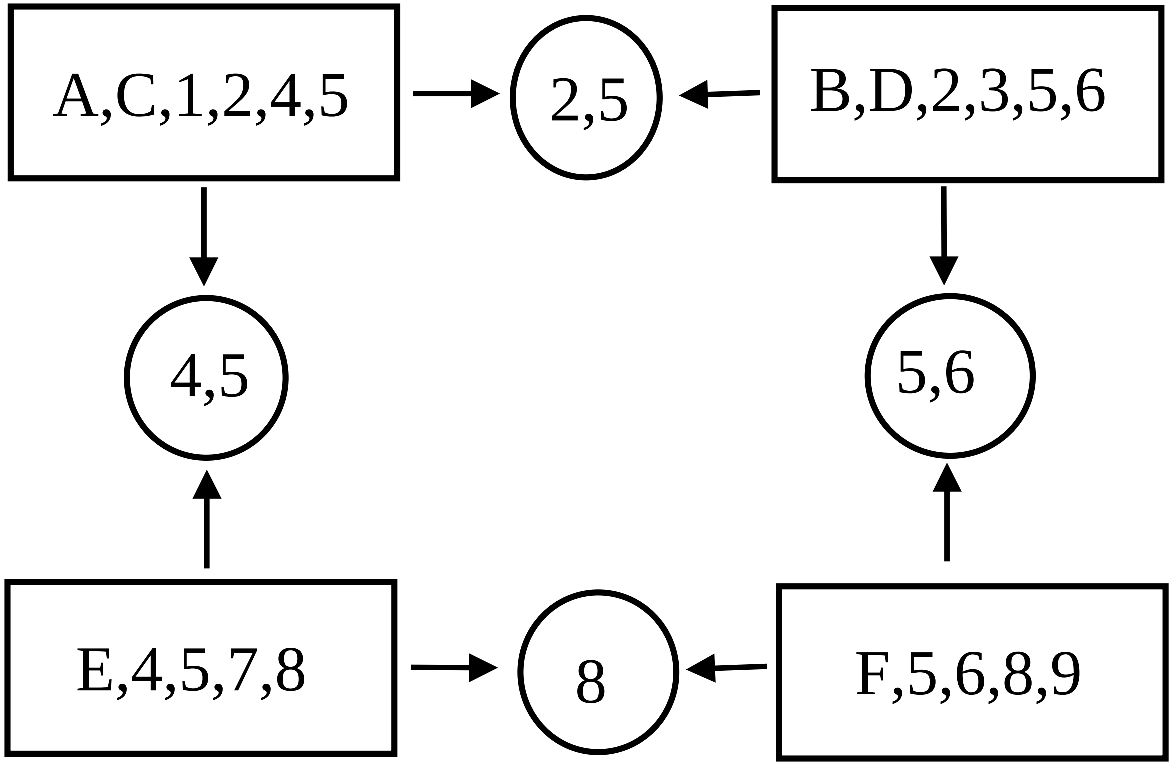
Large regions that are not sub-regions of any other region. $\boxed{c_r = 1}$

Small regions that are sub-regions of every region they are connected to. $\boxed{c_r = 1 - d_r}$

Must obey the *junction graph condition*:

Every sub-graph obtained by selecting only those regions containing any particular node will be a tree.

Example of (Aji-McEliece) Junction Graph



(But small regions can contain factor nodes or be connected to more than two large regions in general.)

Theorems About Junction Graphs

- Stationary points of junction graph free energy are fixed points of “generalized distributive law” message passing algorithm.

- When the junction graph is a tree, the message-passing algorithm is exact (*junction tree algorithm*).

Cluster Variational Method

(Kikuchi, 1951)

Form a region graph with an arbitrary number of different sized regions. Start with largest regions. $\boxed{c_r = 1}$

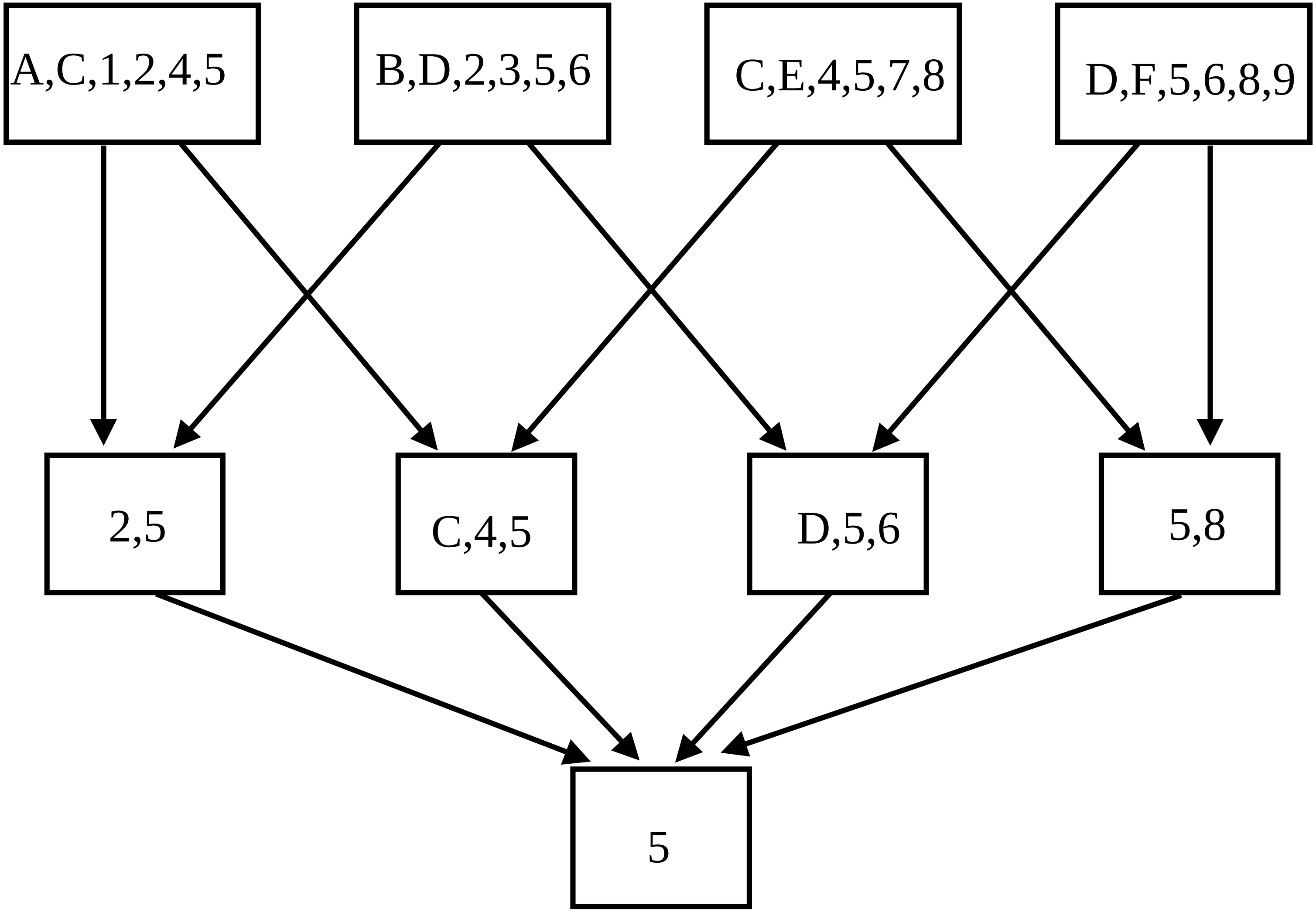
Then find intersection regions of the largest regions, discarding any regions that are sub-regions of other intersection regions.

Continue finding intersections of those intersection regions, etc.

All intersection regions obey $\boxed{c_r = 1 - \sum_{s \in S(r)} c_s}$, where

S(r) is the set of super-regions of region *r.*

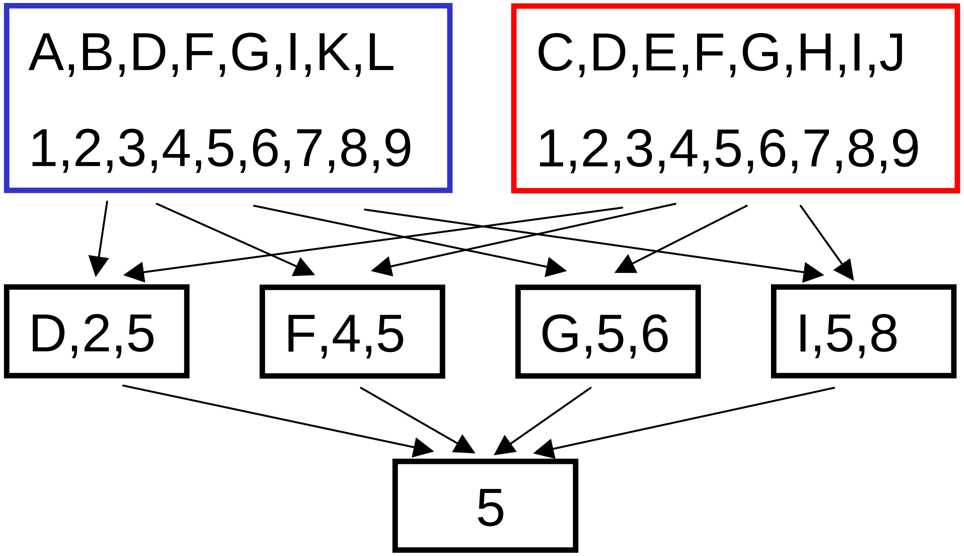
Region Graph Created Using CVM



Just count every node once!

Bethe Method, Junction Graph Method, and Cluster Variational Method are essentially ways to guarantee that every variable and factor node is counted once.

Example of region graph that cannot be created using other methods:



A,B,D,F,G,I,K,L

1,2,3,4,5,6,7,8,9

C,D,E,F,G,H,I,J

1,2,3,4,5,6,7,8,9

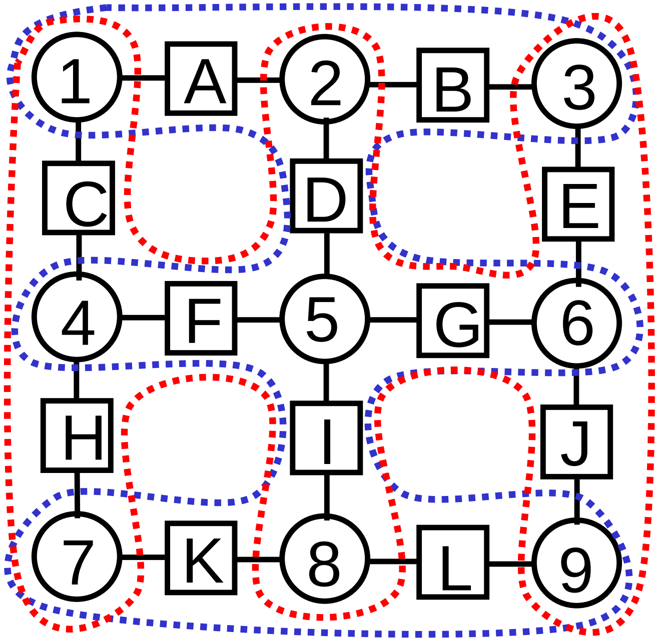
D,2,5

F,4,5

G,5,6

I,5,8

5



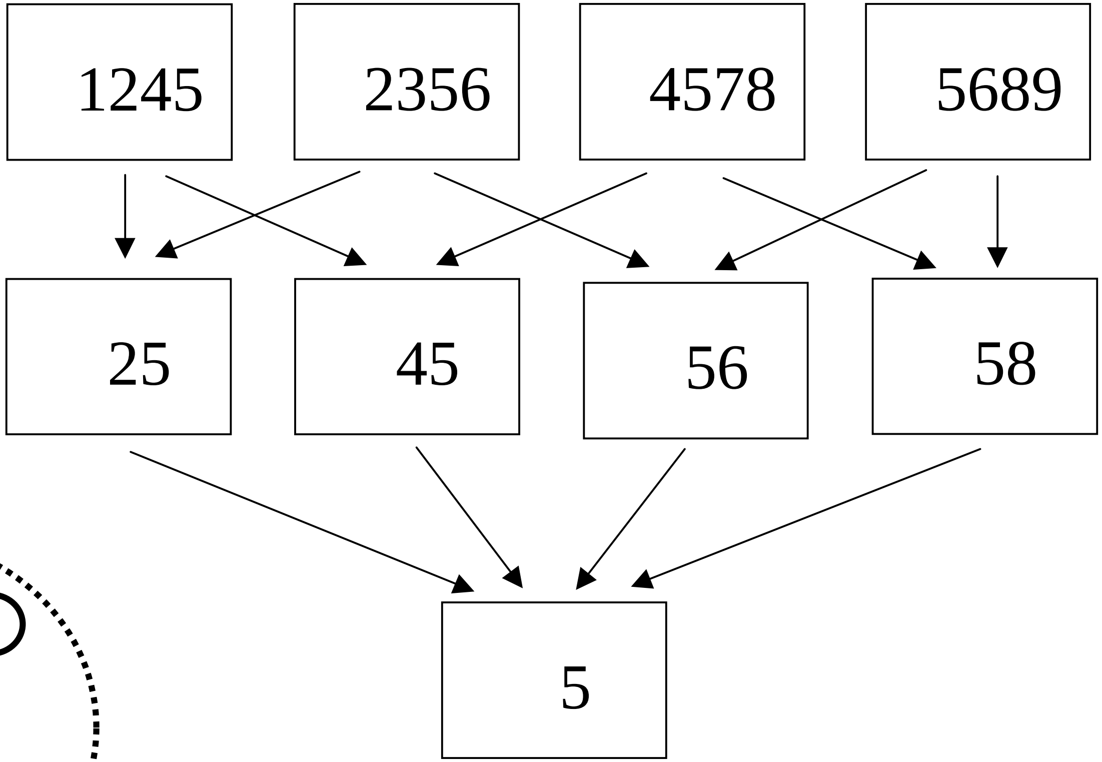
Minimizing a Region Graph Free Energy

- Minimization is possible, but it may be awkward because of all the constraints that must be satisfied. (Yuille 2001, Welling & Teh 2001)
- We introduce *generalized* belief propagation algorithms whose fixed points are provably identical to the stationary points of the region graph free energy.

Generalized Belief Propagation

- Belief in a region is the product of:
  - Local information (factors in region)
  - Messages from parent regions
  - Messages into descendant regions from parents who are not descendants.
- Message-update rules obtained by enforcing marginalization constraints.

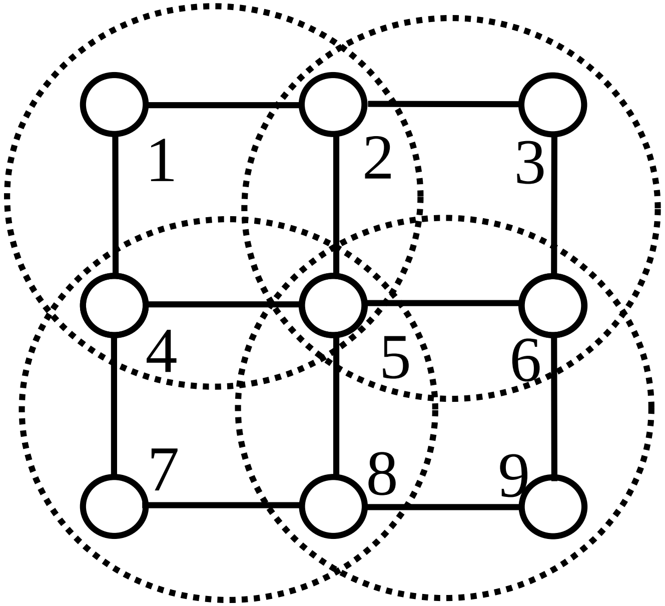
Generalized Belief Propagation



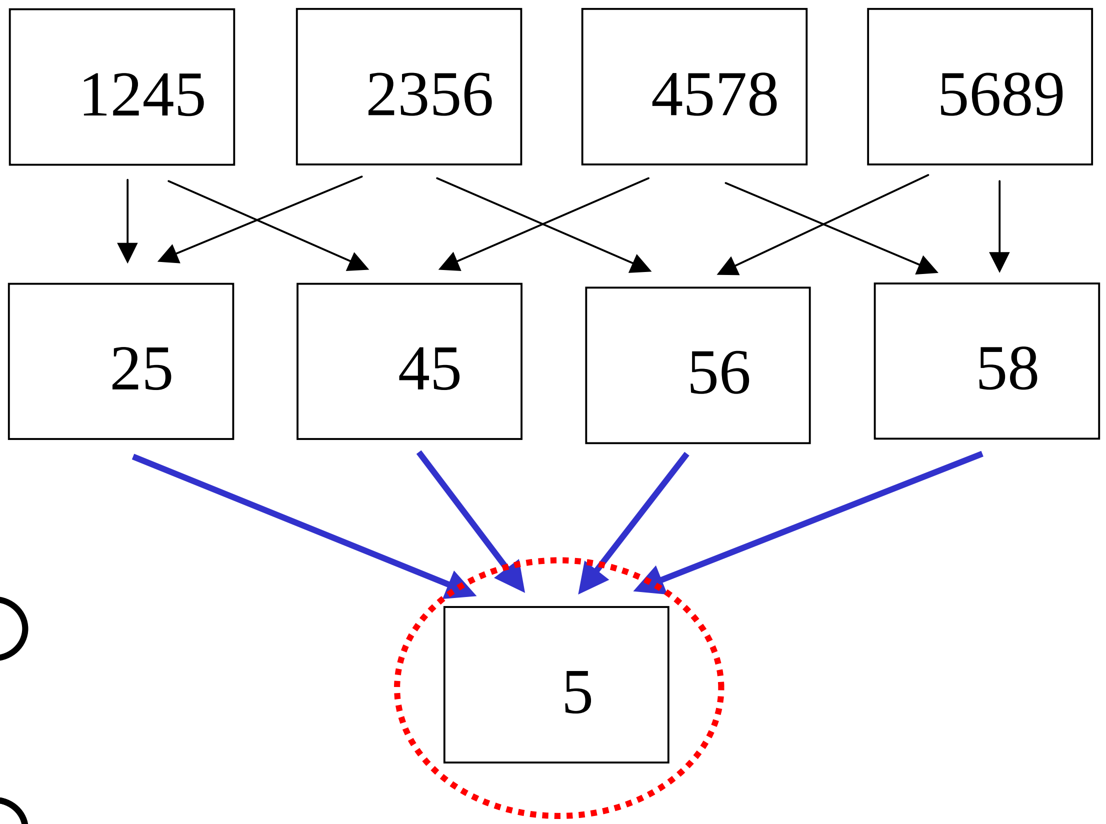
1245 2356 4578 5689

25 45 56 58

5



Generalized Belief Propagation



1245

2356

4578

5689

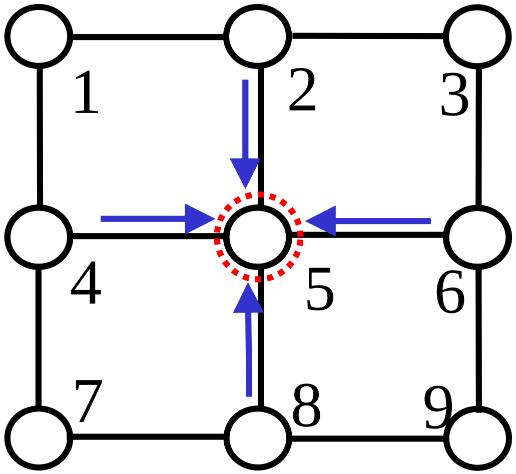
25

45

56

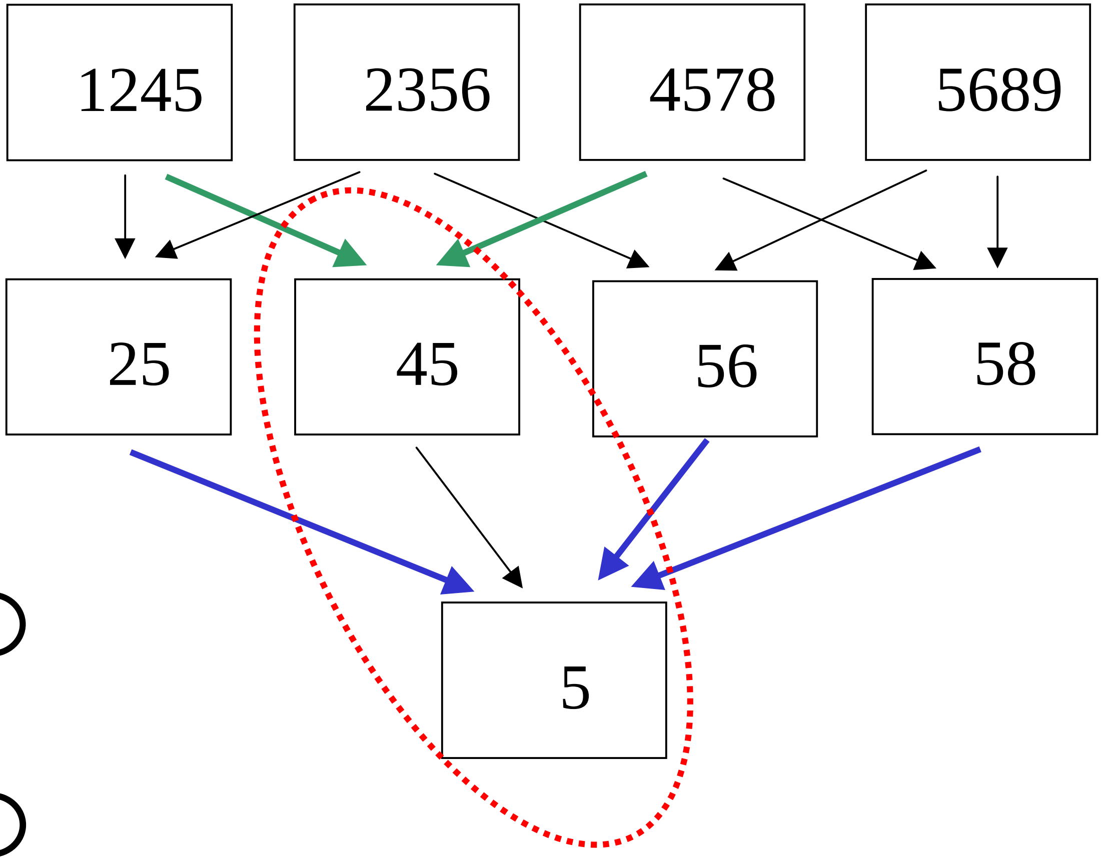
58

5



$$b_5 \propto m_{2\to5} m_{4\to5} m_{6\to5} m_{8\to5}$$

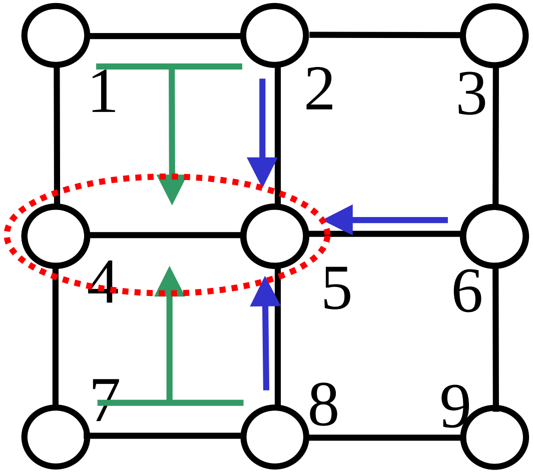
Generalized Belief Propagation



1245 2356 4578 5689

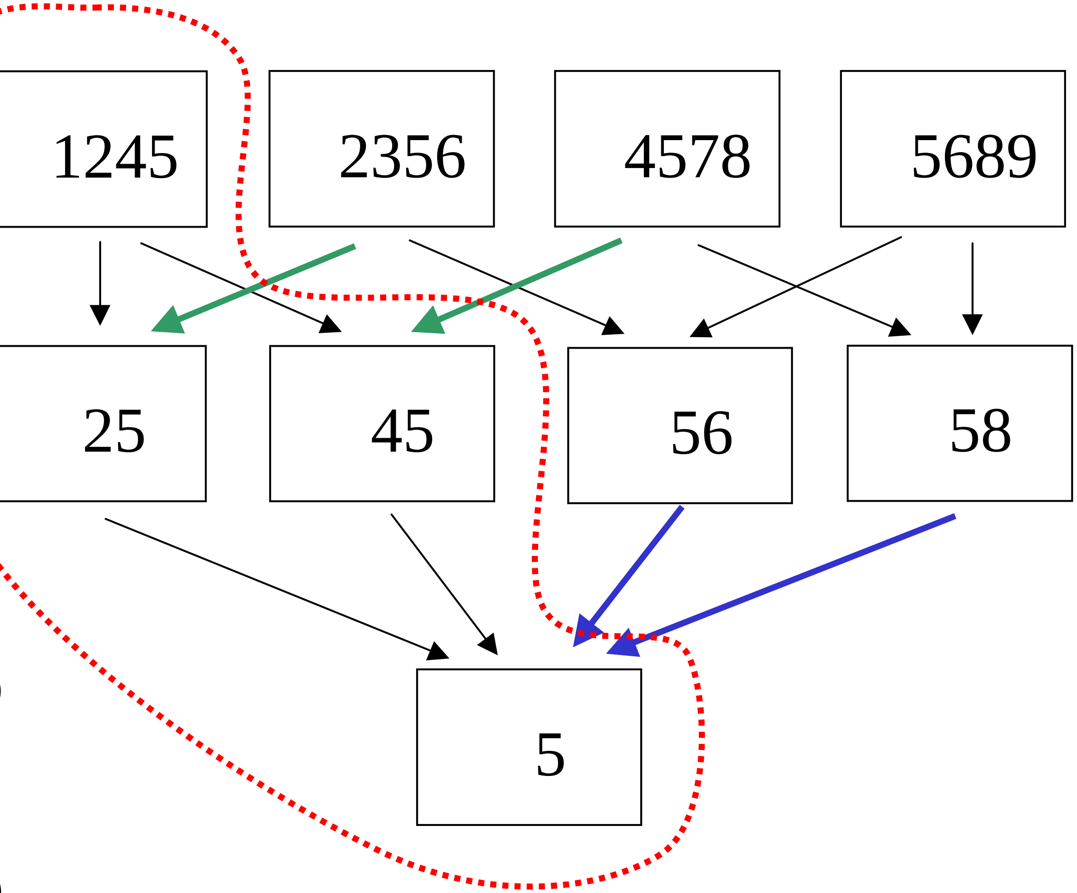
25 45 56 58

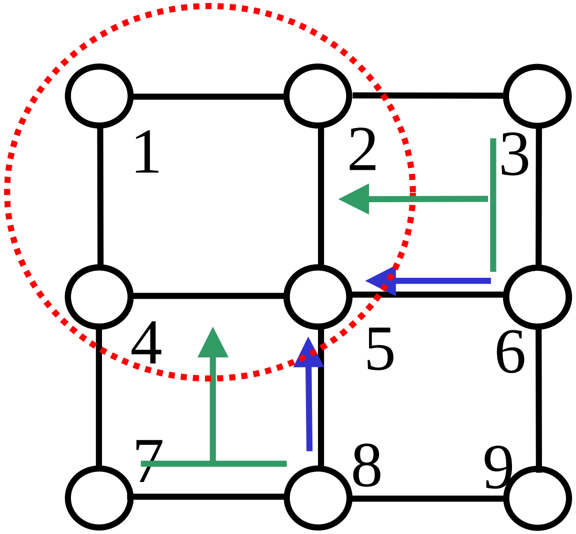
5



$$b_{45} \propto [\, f_{45} \,][m_{12\to 45} m_{78\to 45} m_{2\to 5} m_{6\to 5} m_{8\to 5}]$$

Generalized Belief Propagation

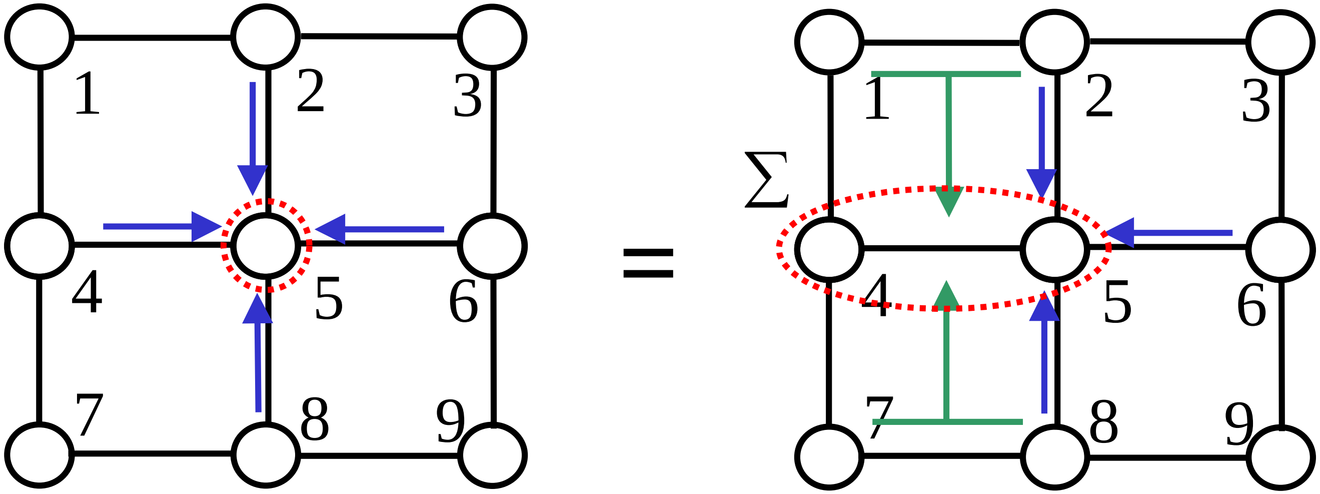




$$b_{1245} \propto [\, f_{12}\, f_{14}\, f_{25}\, f_{45}\,][m_{36\to25} m_{78\to45} m_{6\to5} m_{8\to5}\,]$$

Generalized Belief Propagation

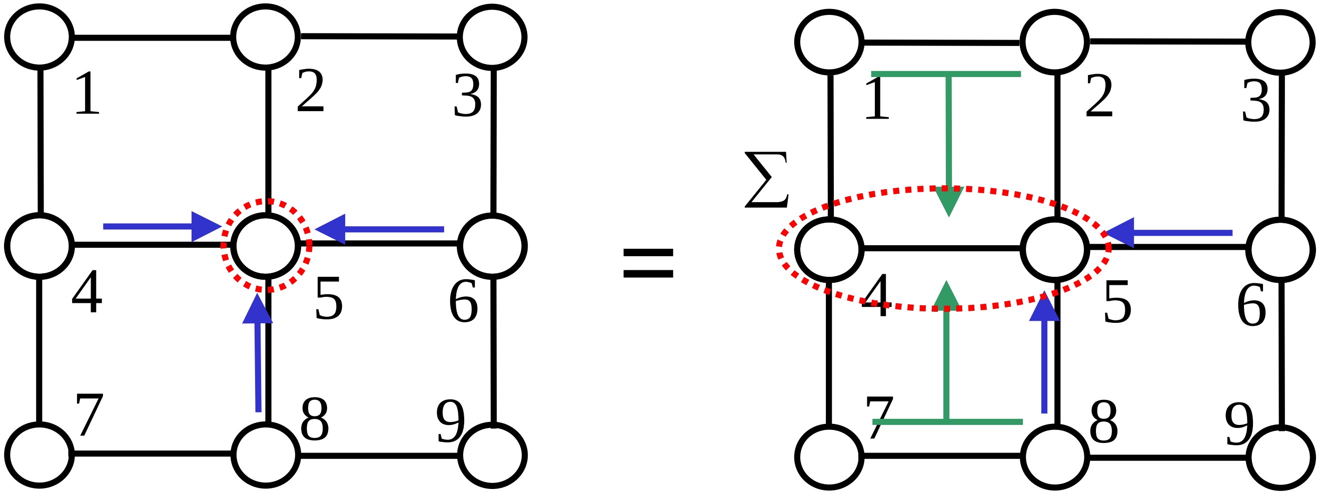
Use Marginalization Constraints to Derive Message-Update Rules



$$b_5(x_5) = \sum_{x_4} b_{45}(x_4, x_5)$$

Generalized Belief Propagation

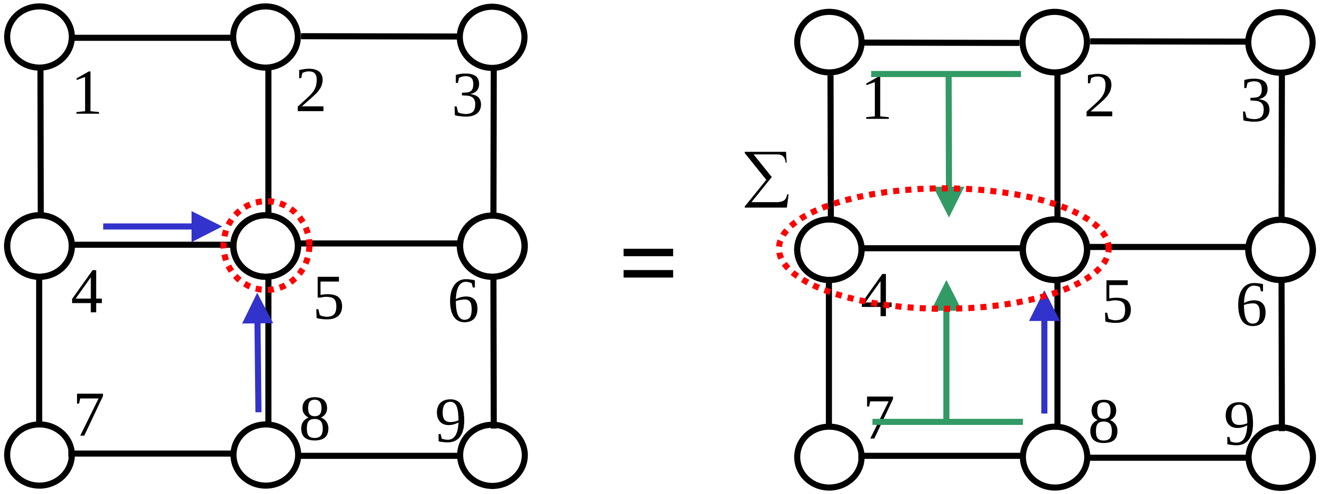
Use Marginalization Constraints to Derive Message-Update Rules



$$b_5(x_5) = \sum_{x_4} b_{45}(x_4, x_5)$$

Generalized Belief Propagation

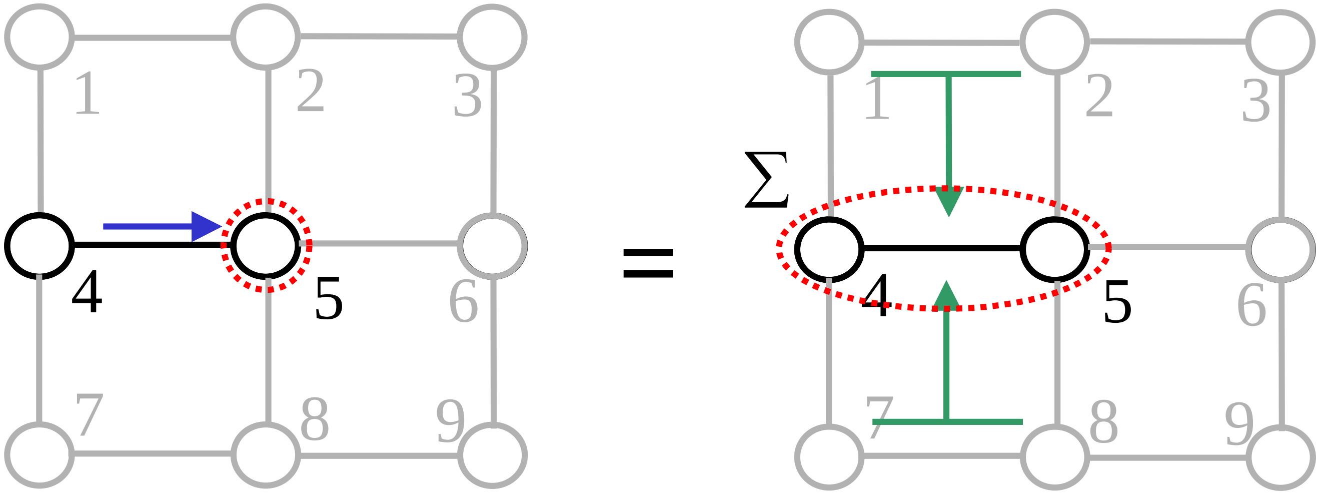
Use Marginalization Constraints to Derive Message-Update Rules



$$b_5(x_5) = \sum_{x_4} b_{45}(x_4, x_5)$$

Generalized Belief Propagation

Use Marginalization Constraints to Derive Message-Update Rules



$$m_{4\rightarrow5}(x_5) \propto \sum_{x_4} f_{45}(x_4, x_5) m_{12\rightarrow45}(x_4, x_5) m_{78\rightarrow45}(x_4, x_5)$$

Generalized Belief Propagation

- Theorems:
  - Fixed points equivalent to stationary points of region graph free energy (messages are complicated combinations of Lagrange multipliers).
  - Exact when region graph is a tree.

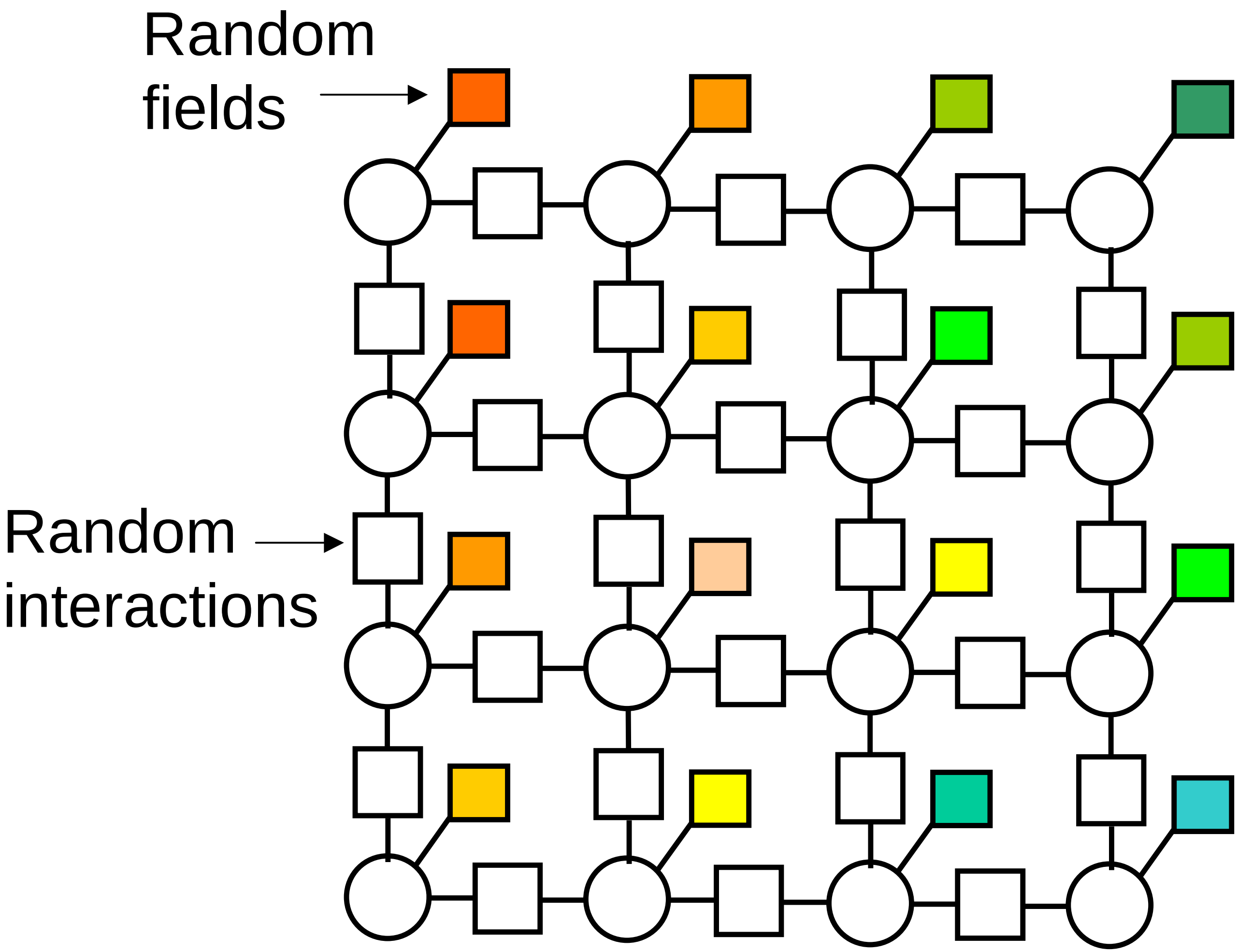
Generalized Belief Propagation

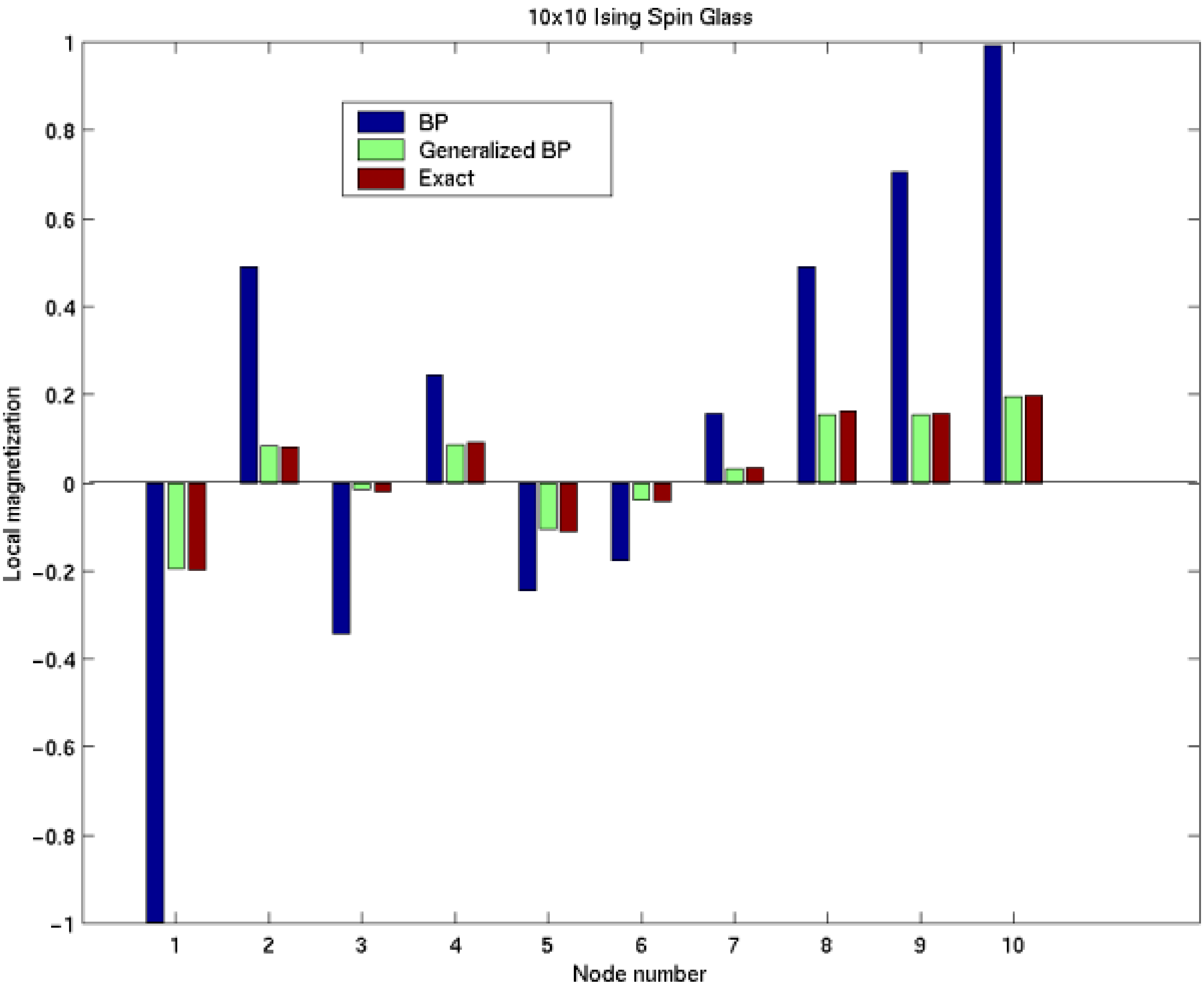
- Theorems:
  - Fixed points equivalent to stationary points of region graph free energy (messages are complicated combinations of Lagrange multipliers).
  - Exact when region graph is a tree.
- Empirically, more likely to converge than ordinary BP, but not guaranteed.

Generalized Belief Propagation

- Theorems:
  - Fixed points equivalent to stationary points of region graph free energy (messages are complicated combinations of Lagrange multipliers).
  - Exact when region graph is a tree.
- Empirically, more likely to converge than ordinary BP, but not guaranteed.
- Can be nearly as fast as ordinary BP, but much more accurate. Complexity depends on details of region graph.

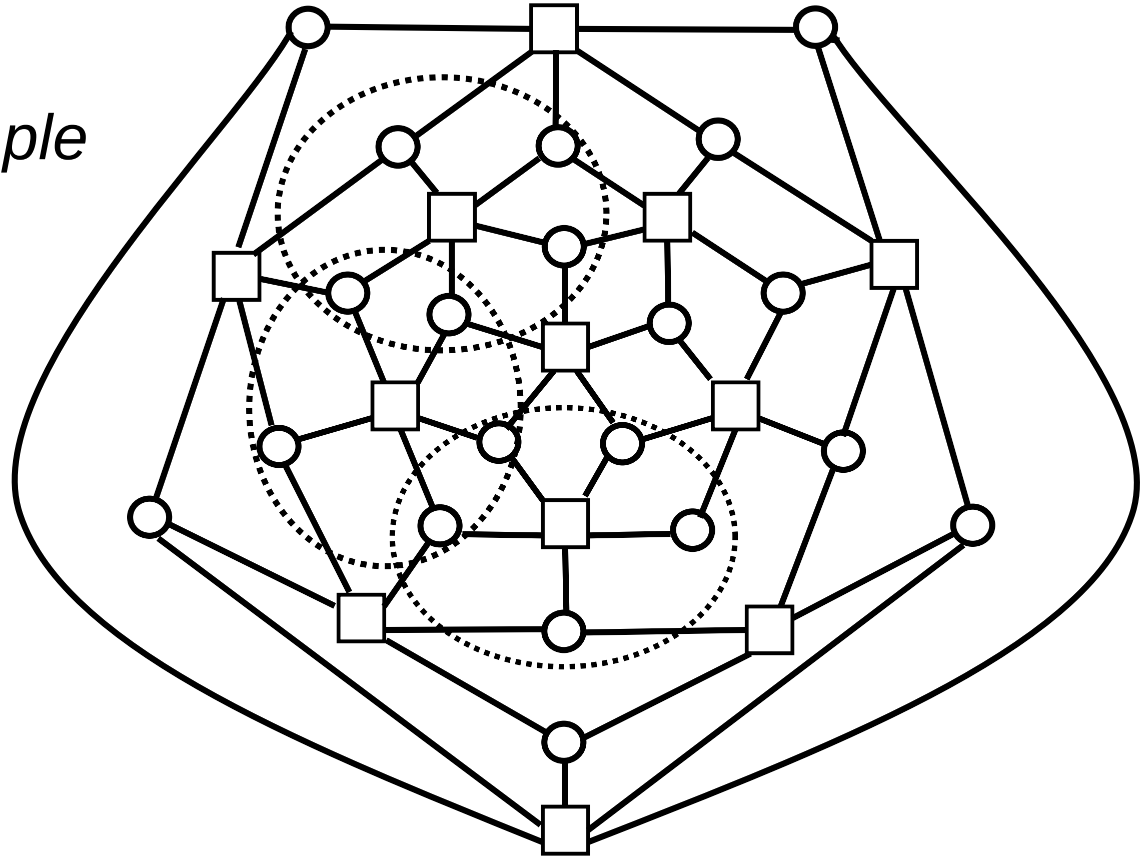
10x10 Ising Spin Glass




10x10 Ising Spin Glass
BP
Generalized BP
Exact
Local magnetization
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
1
2
3
4
5
6
7
8
9
10
Node number

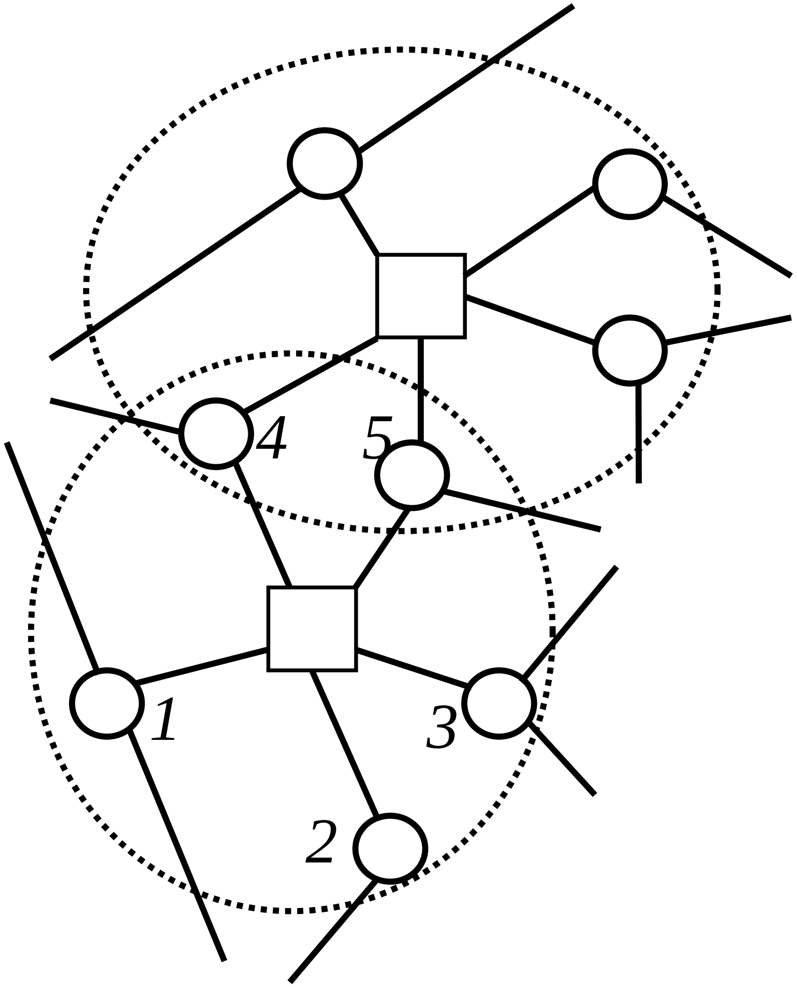
GBP Decoding of Parity Check Codes

Toy example

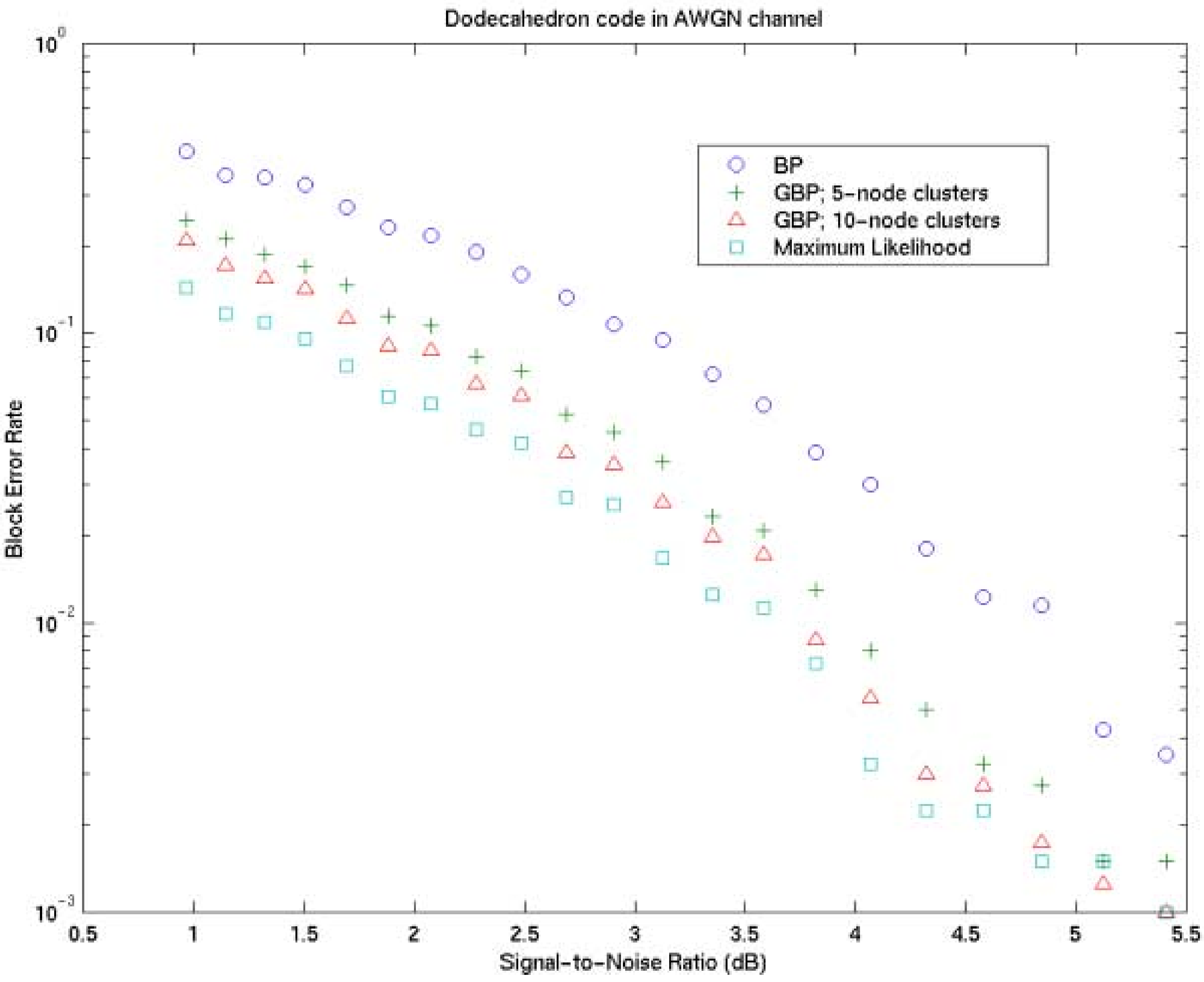


Cluster Variational Method with largest clusters consisting of all the nodes in a parity check.

New messages



$$m_{123\rightarrow 45}(x_4, x_5)$$


Dodecahedron code in AWGN channel
○ BP
+ GBP; 5-node clusters
△ GBP; 10-node clusters
□ Maximum Likelihood
Block Error Rate
10^{0}
10^{-1}
10^{-2}
10^{-3}
Signal-to-Noise Ratio (dB)
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5

Future Directions

- GBP for soft-decoding of BCH codes and other interesting codes.
- GBP for Bayes nets:

  Daphne Koller (Stanford): “GBP made possible an order of magnitude increase in accuracy over standard BP, at very little additional cost, for networks with hundreds of thousands of nodes.”
- GBP in physics: although cluster variational method was known, region graphs and the equivalence with belief propagation is new.

Conclusions

- Standard BP equivalent to minimizing the Bethe free energy.
- Bethe method, junction graph method, and cluster variational method are all sub-classes of the more general region graph method for generating valid free energy approximations.
- GBP is equivalent to minimizing region graph free energy, and is exact when the region graph is a tree.
- GBP is a straightforward and efficient method for obtaining accurate estimates of marginal probabilities.