# Statistical Physics for Discrete Optimization Theory

Parsa Rangriz

Department of Physics,
Sharif University of Technology,
Tehran, Iran

October 3, 2021

# Outline

1. Motivation

2. Graphical Model

3. Cavity Method

4. Free Energy (log-Partition Function)

5. Free Energy Approximations
    - Mean Field Approach
    - Region-Based Free Energy Approximation
    - The Bethe Method

6. Application: Graph Partitioning

# Motivation

Suppose a distribution such as $p(x_1, x_2, \cdots, x_n)$ in which we are interested in computing the marginal distribution or the partition function $Z$.

## Definition: Marginal Distribution

$$p_i(x_i) = \sum_{\{x_s | s \neq i\}} p(x_1, x_2, \cdots, x_n) \tag{1}$$

$$p_I(x_I) = \sum_{\{x_s | s \notin I\}} p(x_1, x_2, \cdots, x_n) \tag{2}$$

The set of random variables $\{x_i\}_{i=1}^{n}$ has some relations in its structure. For example:

$$p(x_1, x_2, \cdots, x_6) = p_1(x_1) p_{23}(x_2, x_3) p_{456}(x_4, x_5, x_6) \tag{3}$$

# Graphical Model: Introduction

In order to define a graphical model, we associate with each vertex $s \in \mathcal{V}$ a random variable $X_s$ taking values in some space $\mathcal{X}_s$. Depending on the application, this state space $\mathcal{X}_s$ may either be continuous, (e.g., $\mathcal{X}_s = \mathbb{R}$) or discrete (e.g., $\mathcal{X}_s = \mathbb{Z}_r$).

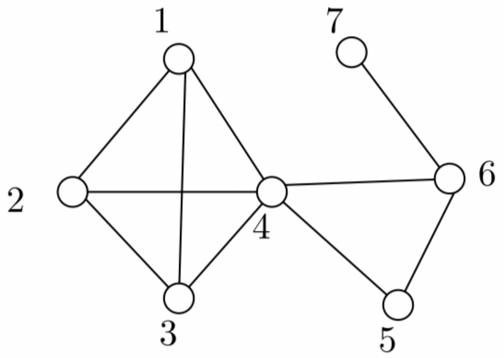


Figure 1: A simple example of graph

$$p(x_1, \cdots, x_6) = \frac{1}{Z} p_{1234}(x_1, x_2, x_3, x_4) p_{456}(x_4, x_5, x_6) p_{67}(x_6, x_7) \quad (4)$$

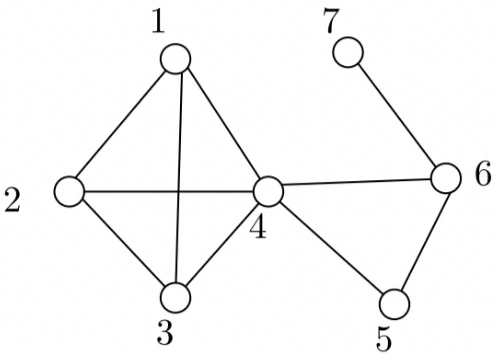
# Graphical Model: Clique

## Definition: Clique

A clique $C$ is a fully connected subset of the vertex set $\mathcal{V}$, meaning that $(s,t) \in \mathcal{E}$ for all $s, t \in C$. Let us associate with each clique $C$ a compatibility function $\psi_C : (\otimes_{s \in C} \mathcal{X}_s) \to \mathbb{R}_+$.

With this notation, an undirected graphical model is a collection of distributions that factorizes as

$$p(x_1, \cdots, x_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \tag{5}$$



# Graphical Model: Factor Graph

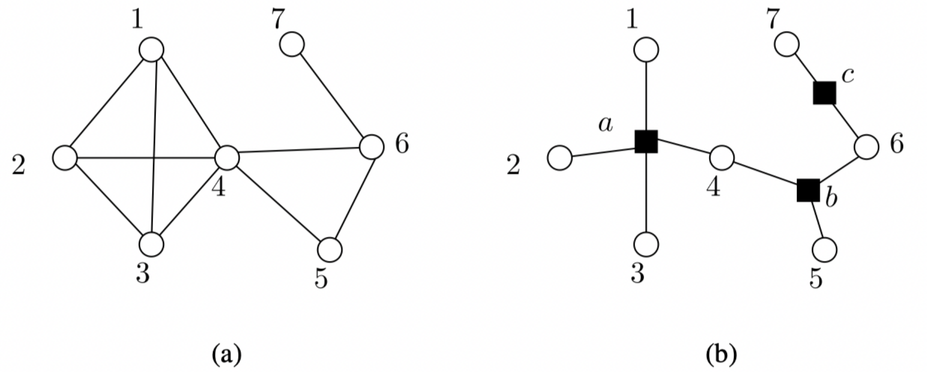


Figure 2: Illustration of Graphs and Factor Graphs

## Definition: Factor Graph

consider a bipartite graph $G = (\mathcal{V}, \mathcal{F}, \mathcal{E}')$ where $\mathcal{V}$ is the original set of vertices (variable nodes), $\mathcal{F}$ is a new set of vertices (factor nodes), and $\mathcal{E}' \subseteq \mathcal{V} \times \mathcal{F}$ is a new edge set.

# Graphical Model: Factor Graph

Now, the probability distribution is formed as follows:

$$p(x_1, \cdots, x_n) = \frac{1}{Z} \prod_{f \in \mathcal{F}} \psi_f(x_{\partial f}) \tag{6}$$

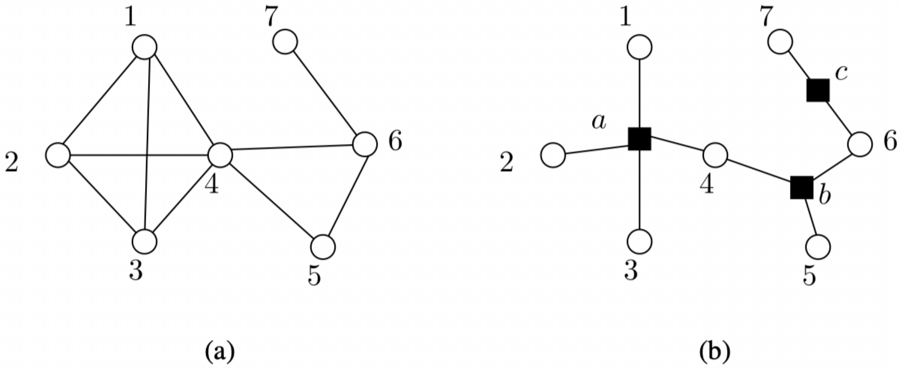


Figure 3: Illustration of Graphs and Factor Graphs

In this case,

$$p(x_1, \cdots, x_6) = \frac{1}{Z} \psi_a(x_1, x_2, x_3, x_4) \psi_b(x_4, x_5, x_6) \psi_c(x_6, x_7) \tag{7}$$

# Cavity Method (Message Passing)

Message passing (belief propagation in artificial intelligence, cavity method in statistical physics and sum-product in computer science) is an iterative algorithm.

$$m_{j\to a}^{(t+1)}(x_j) = \prod_{b\in\partial j\backslash a} m_{b\to j}^{(t)}(x_j) \tag{8}$$

$$m_{a\to j}^{(t)}(x_j) = \sum_{x_{\partial a\backslash j}} F_a(x_{\partial a}) \prod_{k\in\partial a\backslash j} m_{k\to a}^{(t)}(x_k) \tag{9}$$

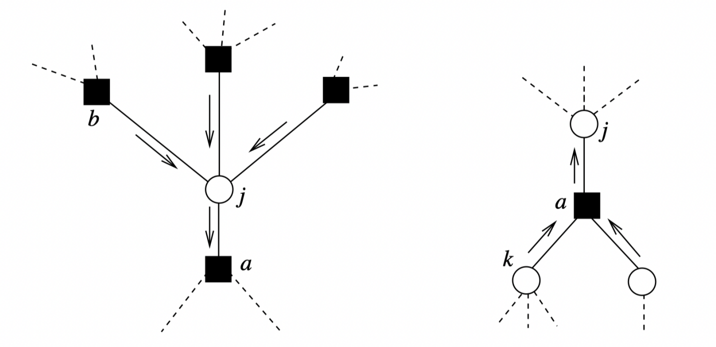


Figure 4: The Messages of a Factor Graph

# Cavity Method (Messages Passing)

$$p_i(x_i) = \sum_{\{x_s | s \neq i\}} p(x_1, x_2, \cdots, x_n) \tag{10}$$

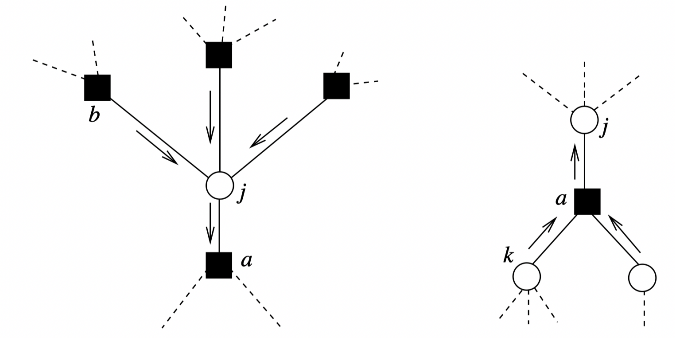


Figure 5: The Messages of a Factor Graph

The estimated marginal is

$$p_i(x_i) \approx b_i(x_i) = \prod_{a \in \partial i} m_{a \to i}(x_i) \tag{11}$$

# Free Energy (log-Partition Function)

A fundamental result of statistical mechanics is, in thermal equilibrium, the probability of a configuration $(x_1, \cdots, x_m)$, the probability of a state will be given by Boltzmann distribution,

## Boltzmann Distribution

$$p(x_1, \cdots, x_n) = \frac{1}{Z} e^{-E(x_1, \cdots, x_n)} \tag{12}$$

where $Z$ is the partition function. Then,

$$Z = \sum_{\{x_i\}_{i=1}^{n}} e^{-E(x_1, \cdots, x_n)} \tag{13}$$

# Free Energy (log-Partition Function)

For the case of a factor graph probability distribution function

$$p(x_1, \cdots, x_m) = \frac{1}{Z} \prod_{a \in \mathcal{F}} \psi_a(x_{\partial a}) \tag{14}$$

we therefore define the energy $E(x_1, \cdots, x_m)$ of a state $(x_1, \cdots, x_m)$ to be

## Definition: Energy of Factor Graph

$$E(x_1, \cdots, x_m) = -\sum_{a \in \mathcal{F}} \ln \psi_a(x_{\partial a}) \tag{15}$$

in order to be consistent with the Boltzmann distribution

$$p(x_1, \cdots, x_m) = \frac{1}{Z} e^{-E(x_1, \cdots, x_m)} \tag{16}$$

# Free Energy (log-Partition Function)

First we define the Helmholtz free Energy as follows:

**Definition: Helmholtz Free Energy**

$$F_H = -\ln Z = U(p) - H(p) \tag{17}$$

One important technique is based on a variational approach. A corresponding variational free energy (Gibbs free energy) defined by

**Definition: Variational (Gibbs) Free Energy**

$$F(b) = U(b) - H(b) \tag{18}$$

# Free Energy (log-Partition Function)

where $U(b)$ is the variational average energy

$$U(b) = \sum_{\{x_i\}_{i=1}^{m} \in S} b(x_1, \cdots, x_m) E(x_1, \cdots, x_m) \tag{19}$$

and $H(b)$ is the variational entropy

$$H(b) = - \sum_{\{x_i\}_{i=1}^{m} \in S} b(x_1, \cdots, x_m) \ln b(x_1, \cdots, x_m) \tag{20}$$

It follows directly from our definitions that

$$F(b) = F_H + D(b \| p) \tag{21}$$

where $D$ is the relative entropy (Kullback-Leibler convergence)

## Definition: Relative Entropy

$$D(b \| p) = \sum_{\{x_i\}_{i=1}^{m} \in S} b(x_1, \cdots, x_m) \ln \frac{b(x_1, \cdots, x_m)}{p(x_1, \cdots, x_m)} \tag{22}$$

# Approximate Inferences

- The Mean Field Approach
- Region-Based Free Energy Approximation
- The Bethe Method
- The Region Graph Method
- Maxent-Normal Region Graph Approximation
- Generalized Message Passing Algorithm
- The Junction Trees Approximation
- The Junction Graph Method
- Cluster Variational Method (Kikuchi)
- ...

# Mean Field Approach

From previous slides,

$$F(b) = F_H + D(b \| p) \tag{23}$$

Minimizing the variational free energy $F(b)$ with respect to trail probability function $b$ is therefore an exact procedure for computing $F_H$ and recovering $p$.

A more practical possibility is to upper-bound $F_H$ by minimizing $F(b)$ over a restricted class of probability distributions. This is the basic idea underlying the mean field approach.

One very popular mean-field form for $b$ is the factorized form

$$b_{\mathrm{MF}}(x_1, \cdots, x_m) = \prod_{i=1}^{m} b_i(x_i) \tag{24}$$

# Region-Based Free Energy Approximation

Kikuchi and the other physicists introduced a class of approximations to the variational free energy $F(b)$. The idea behind these approximations is similar, but slightly different from the mean-field free energy $F_{\mathrm{MF}}(b)$. in a Kikuchi approximation the approximate free energy $F_K$ will be a function of larger set of nodes $b_I(x_I)$.

## Definition: Region

We define a region $R$ of a factor graph to be a set $\mathcal{V}_R$ of variable nodes and a set of $\mathcal{F}_R$ of factor nodes, such that if a factor node $a$ belongs to $\mathcal{F}_R$, all the variable nodes neighboring $a$ are in $\mathcal{V}_R$.

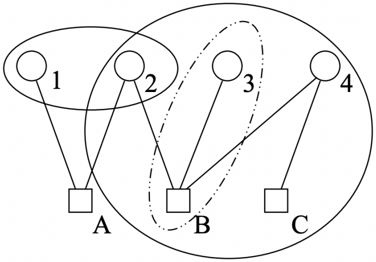


Figure 6: A Simple Region of a Factor Graph

# Region-Based Free Energy Approximation

We define the state $x_R$ of a region to be the collective set of variable node states $\{x_i | i \in \mathcal{V}_R\}$.

## Definition: Region Functions

We define the region energy $E_R(x_R)$ to be

$$E_R(x_R) = -\sum_{a \in \mathcal{F}_R} \ln F_a(x_{\partial a}) \tag{25}$$

For any region $R$, we define the region average energy $U_R(b_R)$, and the region entropy $H_R(b_R)$ by

$$U_R(b_R) = \sum_{x_R} b_R(x_R) E_R(x_R) \tag{26}$$

$$H_R(b_R) = -\sum_{x_R} b_R(x_R) \ln b_R(x_R) \tag{27}$$

# Region-Based Free Energy Approximation

## Region Free Energy

The region free energy is defined by

$$F_R(b_R) = U_R(b_R) - H_R(b_R) \tag{28}$$

The intuitive idea behind a region-based free energy approximation is that we will try to break up the factor graph into a set of large regions that include every factor and variable node, and say that the overall free energy is the sum of the free energies of all the regions.

# Region-Based Free Energy Approximation

## Region-Based Approximation

We define a region-based approximate entropy $H_{\mathcal{R}}$ by

$$H_{\mathcal{R}}(\{b_R\}) = \sum_{R \in \mathcal{R}} c_R H_R(b_R) \tag{29}$$

and the region-based average energy $U_{\mathcal{R}}$ by

$$U_{\mathcal{R}}(\{b_R\}) = \sum_{R \in \mathcal{R}} c_R U_R(b_R) \tag{30}$$

where the chosen set of regions $\mathcal{R}$, and the associated set of counting numbers $c_R$ instantiate the approximation. We define the region-based free energy by

$$F_{\mathcal{R}}(\{b_R\}) = U_{\mathcal{R}}(\{b_R\}) - H_{\mathcal{R}}(\{b_R\}) \tag{31}$$

# The Bethe Method

The origins of the Bethe method date back to 1935 and Bethe's famous approximation method for magnets.

## Definition: Sub and Super Region

First, we make a small preliminary definition: if $R_1$ and $R_2$ are two regions, we say that $R_1$ is a sub-region of $R_2$ and $R_2$ is a super-region of $R_1$ if the set of variable and factor nodes in $R_1$ are a subset of those in $R_2$.

The counting numbers $c_R$ for each region $R \in \mathcal{R}$ are given by

$$c_R = 1 - \sum_{S \in \mathcal{S}(R)} c_S \tag{32}$$

where $\mathcal{S}(R)$ is the set of regions that are super-regions of $R$.

# The Bethe Method

## The Bethe Method

In the region-based approximation generated by the Bethe method, we take the set of regions included in $\mathcal{R}$ to be of two types. First, we have a set of large regions $\mathcal{R}_L$ such that $|\mathcal{F}|$ regions in $\mathcal{R}_L$ each contain exactly one factor node and all the variable node neighboring that factor node. Second, we have a set of small regions $\mathcal{R}_S$, such that the $|\mathcal{V}|$ regions in $\mathcal{R}_S$ each contain a single variable node.

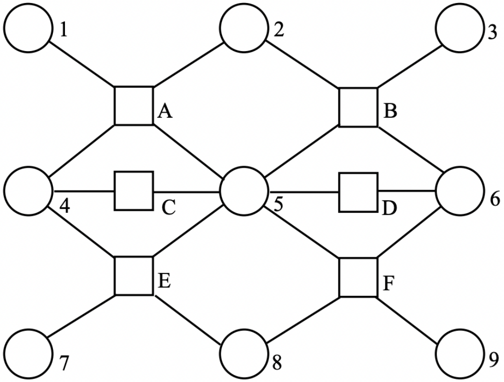


Figure 7: A Factor Graph

# The Bethe Method

Using the definition, we see that for every region $R \in \mathcal{R}_L$, $c_R = 1$, while for every region $R \in \mathcal{R}_S$, $c_R = 1 - d_i$. where $d_i$ is the degree (number of neighboring factor nodes) of the variable node $i$.

## The Bethe Functions

The Bethe free energy is $F_B = U_B - H_B$, where the Bethe average energy is

$$U_B = -\sum_{a \in \mathcal{F}} \sum_{x_{\partial a}} b_a(x_{\partial a}) \ln F_a(x_{\partial a}) \tag{33}$$

and the Bethe entropy is

$$H_B = -\sum_{a \in \mathcal{F}} \sum_{x_{\partial a}} b_a(x_{\partial a}) \ln b_a(x_{\partial a}) + \sum_{i \in \mathcal{V}} (d_i - 1) \sum_{x_i} b_i(x_i) \ln b_i(x_i) \tag{34}$$

# Graph Partitioning

Suppose a graph $G = (\mathcal{V}, \mathcal{E})$ with $N$ vertices which is to be cut into two subgraphs with specific value of vertex difference and subject to minimize the number of connections between vertices from different groups. This is called min-cut problem.

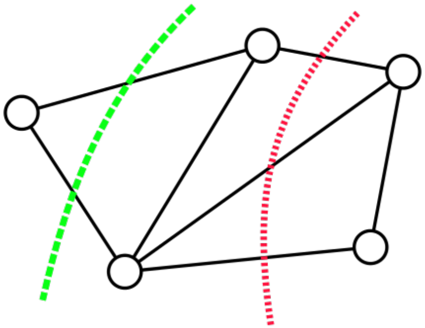


Figure 8: A graph and two of its cuts. The dotted line in red represents a cut with three crossing edges. The dashed line in green represents one of the minimum cuts of this graph, crossing only two edges

# Ising Model

There are some solution and ways to come up with this problem. One of these is a statistical mechanics approach which corresponds a weight to any vertex and edge. For this purpose, the spin intuition becomes important. This would be possible if we set a spin $S_i$ to each node $i \in \mathcal{V}$. If $i$ in in the first group, $S_i = 1$ and otherwise, $S_i = -1$. Then the Hamiltonian of the system is given by

$$H = -\sum_{(ij)\in\mathcal{E}} S_i S_j \tag{35}$$

and the difference between the two subgraphs is

$$m = \frac{1}{N}\sum_{i\in\mathcal{V}} S_i \tag{36}$$

# The Legendre Transformation

As we explained in the introduction, the graph partitioning is equivalent to the Ising model at fixed magnetization $m$. The magnetization will be fixed via an external magnetic field $h$ which appears in the Hamiltonian as

$$H(\{S_i\}_i) = -\sum_{(ij)\in\mathcal{E}} S_i S_j - h\sum_{i\in\mathcal{V}} S_i \tag{37}$$

This Hamiltonian is a Legendre transformation function of the previous Hamiltonian with changing $m$ with $h$.

The main aim of this course project is to find out the value of bisection width, $b$, which it means the number of edges between different groups divided by the total number of vertices.

$$b = \frac{1}{2}\left(\frac{E}{N} + hm + \frac{\alpha}{2}\right) \tag{38}$$

where $\alpha$ is the mean degree of the graph.

# Motivation I

As we discussed above, the problem of graph partitioning is equal to the Ising model with the following Hamiltonian

$$H(\{S_i\}_i) = -\sum_{(ij)\in\mathcal{E}} S_i S_j - h\sum_{i\in\mathcal{V}} S_i \tag{39}$$

By using the statistical mechanical ensemble theory, we have the Boltzmann distribution

## The Boltzmann Distribution

Suppose a state $\{S_i\}_i$ with the Hamiltonian, $H(\{S_i\}_i)$, and the temperature, $T = \frac{1}{\beta}$, then the probability of this state is equal to

$$p(\{S_i\}_i) = \frac{1}{Z} e^{-\beta H(\{S_i\}_i)} \tag{40}$$

# Motivation II

For our problem, the probability distribution is given by

$$p(\{S_i\}_i) = \frac{1}{Z} \prod_{(ij)\in\mathcal{E}} e^{\beta S_i S_j} \prod_{i\in\mathcal{V}} e^{\beta h S_i} \tag{41}$$

Then the partition function is

$$Z = \sum_{\{S_i\}_i} \prod_{(ij)\in\mathcal{E}} e^{\beta S_i S_j} \prod_{i\in\mathcal{V}} e^{\beta h S_i} \tag{42}$$

As we know in statistical physics, if we could calculate the partition function, it would be possible to obtain any thermodynamic variables such as the magnetization, free energy, and so on.

## Thermodynamic Properties

$$F = -\frac{1}{\beta} \ln Z, \quad E = -\frac{\partial}{\partial \beta} \ln Z, \quad m = -\frac{1}{N} \frac{\partial}{\partial \beta h} \ln Z \tag{43}$$

# The BP Algorithm

As far we know, it takes too long to do so naively by summing each term in the partition function. For this reason, there are some algorithms such as the sum-product that help us to use some methods for calculation of each marginal distribution. Thus,

$$p_i(S_i) = e^{\beta h S_i} \prod_{j \in \partial i} m_{j \to i}(S_i) \tag{44}$$

$$p_{ij}(S_i, S_j) = e^{\beta h S_i} e^{\beta h S_j} e^{\beta S_i S_j} \prod_{k \in \partial i \backslash j} m_{k \to i}(S_i) \prod_{l \in \partial j \backslash i} m_{l \to j}(S_k) \tag{45}$$

The messages are obtained as follows,

$$m_{i \to j}(S_i) = \frac{1}{Z_{i \to j}} e^{\beta h S_i} \prod_{k \in \partial i \backslash j} \left( \sum_{S_k} e^{\beta S_i S_k} m_{k \to i}(S_k) \right) \tag{46}$$

# The Bethe Method

For calculating the partition function, it is equal to compute the free energy.

$$F = E - \frac{1}{\beta} S \tag{47}$$

The Bethe free energy is given by

$$-\beta F = \ln Z = \sum_{i \in \mathcal{V}} \ln Z_i - \sum_{(ij) \in \mathcal{E}} \ln Z_{ij} \tag{48}$$

where,

$$Z_i = \sum_{S_i} e^{\beta h S_i} \prod_{k \in \partial i} \left( \sum_{S_k} e^{\beta S_i S_k} m_{k \to i}(S_k) \right) \tag{49}$$

$$Z_{ij} = \sum_{S_i, S_j} e^{\beta S_i S_j} m_{i \to j}(S_i) m_{j \to i}(S_j) \tag{50}$$

# Zero Temperature Limit

Minimum value of the Hamiltonian is obtained when the zero temperature $\beta \to \infty$, limit is get used. Now, we define a new parameter, namely cavity field $h_{i\to j}$,

$$e^{2\beta h_{i\to j}} = \frac{m_{i\to j}(+1)}{m_{i\to j}(-1)} \tag{51}$$

By some calculations, the self-consistent equations for messages become

$$h_{i\to j} = h + \sum_{k\in\partial i\setminus j} \max(J + h_{k\to i}, 0) - \max(h_{k\to i}, J) \tag{52}$$

The magnetization is given by

$$m = \frac{1}{N}\sum_{i\in\mathcal{V}} m_i \tag{53}$$

where,

$$m_i = \begin{cases} +1 & h + \sum_{k\in\partial i}(\max(J + h_{k\to i}, 0) - \max(h_{k\to i}, 1) > 0 \\ -1 & \text{otherwise} \end{cases} \tag{54}$$

# Erdős–Rényi Model

Now it is time to use the BP equations in Erdős–Rényi model. The following plot is the bisection width $b$ of the model with $N = 100$ respect to the average degree $\alpha$. The data compared with the exact average bisection for $N = 2000$

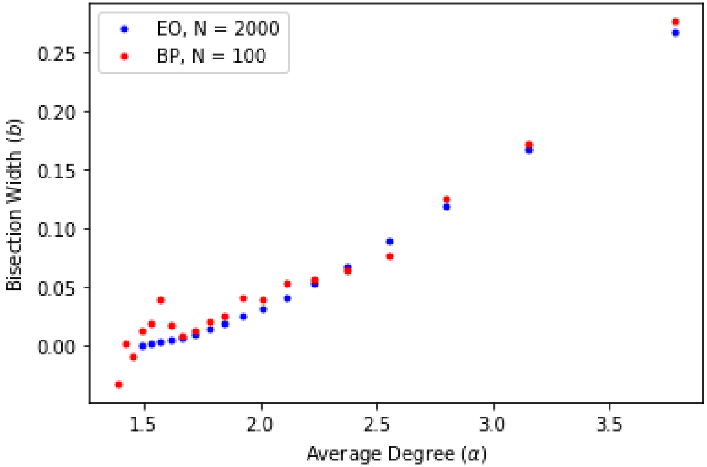


Figure 9: Bisection width, $b$, of Erdős–Rényi model as a function of the mean connectivity with $N = 100$

# Random $d$-Regular Graphs

Random $d$-regular graphs are one of the most important graph models which each vertex has $d$ edges and the edges connect the two corresponded vertices in a random way.
Now, as same as the previous section, we want to use the BP and also BP-guided decimation in random $d$-regular graphs. The following plots show the solutions

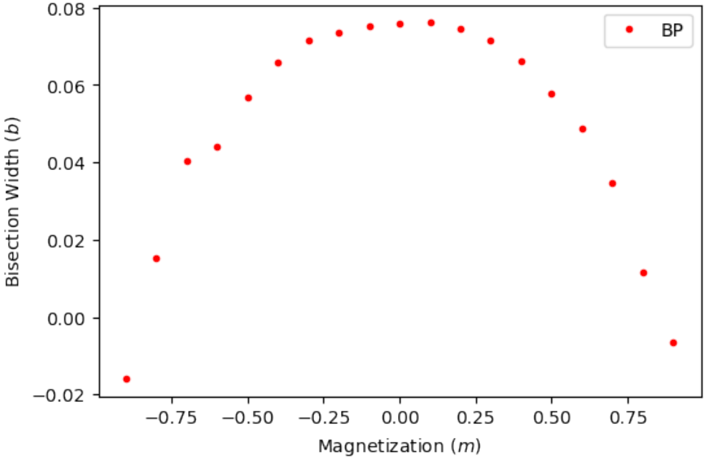


Figure 10: Bisection width, $b$, of random 3-regular graph as a function of the magnetization $m$ with $N = 1000$

# Random *d*-Regular Graphs

Also, we can show that the bisection width is converged to the specific value as we increase the size of system. This illustrates that as we have the larger graph the BP will have a better and accurate solution.

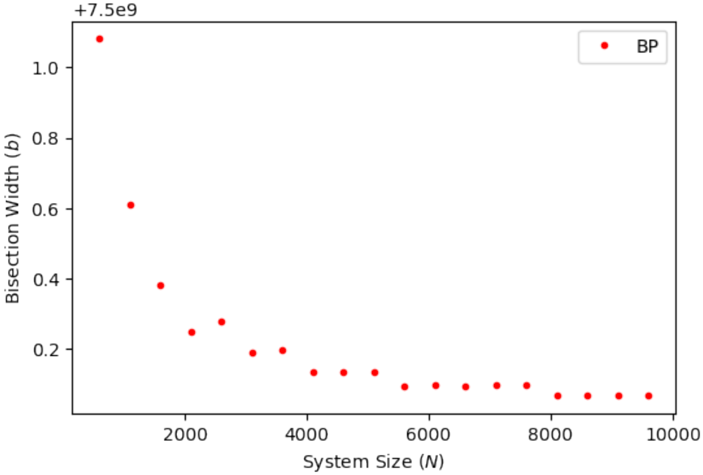


Figure 11: Bisection width $b$ of random 3-regular graph model as a function of the system with $m = 0$

# Discussion

Discussion and Q/A

# References

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss (2005)
Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms
*IEEE Transactions on Information Theory* Vol. 51, July 2005, 2282 – 2313.

Marc Mézard, Andrea Montanari (2009)
Information, Physics, and Computation
*Oxford University Press*

Martin J. Wainwright, Micheal I. Jordan (2008)
Graphical Models, Exponential Families, and Variational Inference
*Foundations and Trends in Machine Learning*, Vol. 1, Nos. 1-2, 1 – 305

Petr Sulc, Lenka Zdeborová (2010)
Belief propagation for graph partitioning
*P*hys. A: Math. Theor. 43 - 285003

Allon G. Percus, et. al. (2008)
The peculiar phase structure of random graph bisection
Journal of Mathematical Physics 49, 125219