**Jin Gu**
Department of Automation, Tsinghua University
Email: jgu@tsinghua.edu.cn
Phone: (010) 62794294-866

# Chapter 4
# Markov Networks

2021 Fall

Jin Gu (古槿)

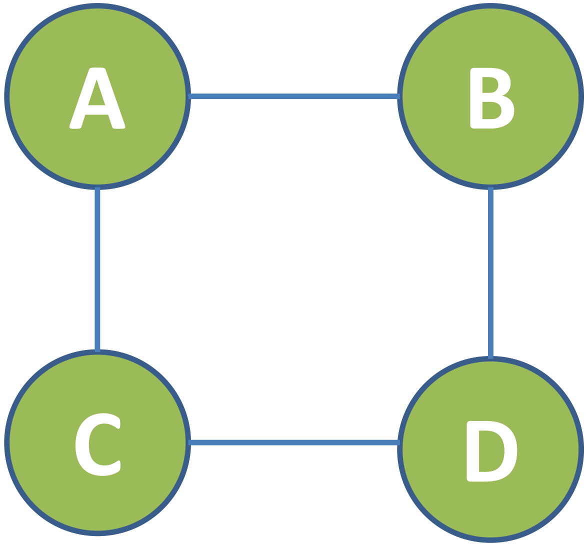
# Outlines

- Why we need undirected models?
- The representation of Markov networks
  - The intuitive factorization of Markov Networks
  - Hammersley-Clifford Theorem
  - From factor product to log-linear models
  - A constructive proof of HC theorem
- Ising models and Boltzmann machines
- “Unique” independences in BNs & MNs
- Extended to continuous models (GMRF)

# Textbook References

- Textbook 1
  - Chapter 4.1 ~ 4.3
  - Chapter 4.5
  - Chapter 7.1, 7.3

- Textbook 2
  - Chapter 19.1 ~ 19.4
  - Chapter 19.6 (CRF)




Four molecules connected as a rectangle (lattice) structure. Each molecule has two different states 0/1. And the states of two interacted molecules determine the local interaction strength.

| A | C | $\pi_1[A,C]$ |
|---|---|---|
| $a^0$ | $c^0$ | 20 |
| $a^0$ | $c^1$ | 3 |
| $a^1$ | $c^0$ | 5 |
| $a^1$ | $c^1$ | 35 |

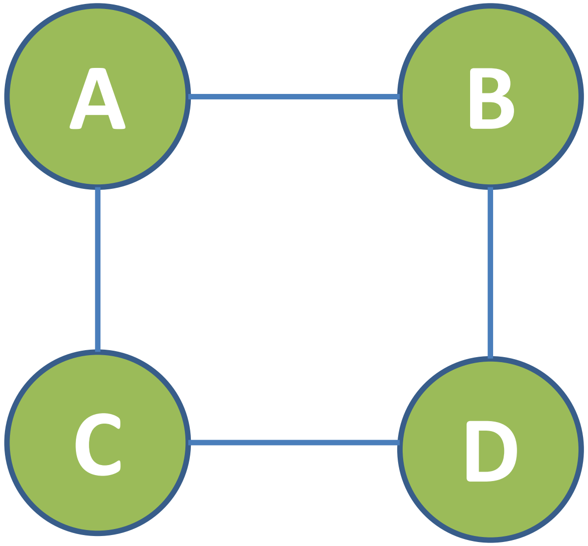
| A | B | $\pi_2[A,B]$ |
|---|---|---|
| $a^0$ | $b^0$ | 15 |
| $a^0$ | $b^1$ | 8 |
| $a^1$ | $b^0$ | 6 |
| $a^1$ | $b^1$ | 10 |

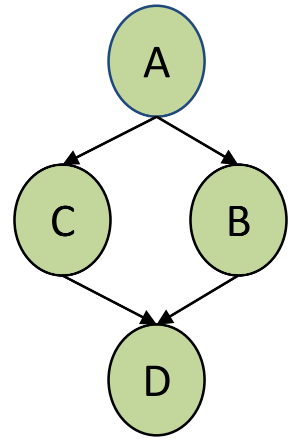
| C | D | $\pi_3[C,D]$ |
|---|---|---|
| $c^0$ | $d^0$ | 50 |
| $c^0$ | $d^1$ | 1 |
| $c^1$ | $d^0$ | 2 |
| $c^1$ | $d^1$ | 20 |

| B | D | $\pi_4[B,D]$ |
|---|---|---|
| $b^0$ | $d^0$ | 100 |
| $b^0$ | $d^1$ | 1 |
| $b^1$ | $d^0$ | 1 |
| $b^1$ | $d^1$ | 1000 |

If we define $P(ABCD)$ as <u>the probability of any state configuration</u> of the four molecules.

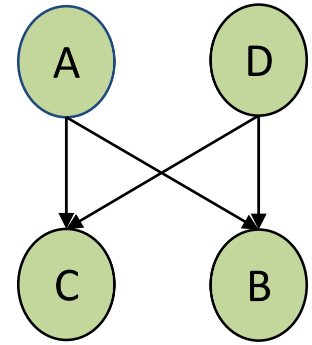
What are the **independences** in this distribution?
Can we find a BN which is a **Perfect Map** of it?




Ind(A;D|B,C) holds

Ind(B;C|A,D) does not




Ind(B;C|A,D) holds

Ind(A;D) also holds

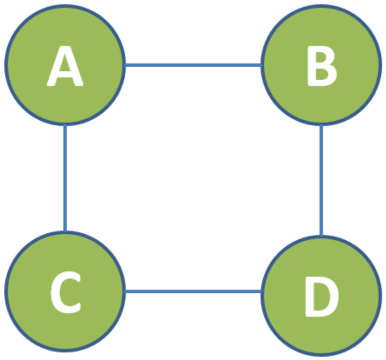
- Definition reminder: *G* is a perfect map (*P*-Map) for *P* if I(*P*)=I(*G*)
- Can you find a P-Map for this problem?
  - **No**: cannot be perfectly represented in a BN
    - Independencies in P: Ind(A;D | B,C), and Ind(B;C | A,D)

?

# What Will You Learn This Week?

- How to represent the joint probability P($ABCD$) over all configurations as Gibbs distribution?

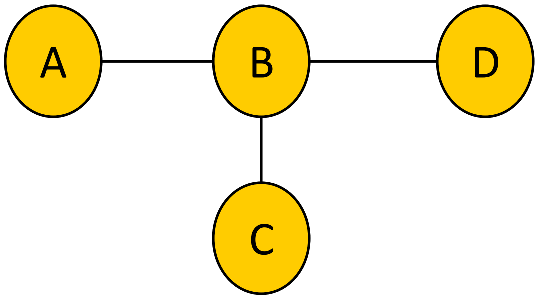
- How to perfectly represent a Gibbs distribution by an undirected graph



# Markov Networks (Intuitive)

- Undirected graph $H$
  - Nodes $X_1, \cdots, X_n$ represent random variables
- $H$ encodes independence assumptions
  - A path $X_i, \cdots, X_j$ is active if none of the $X_k$ variables along the path are observed
  - $\boldsymbol{X}$ and $\boldsymbol{Y}$ are separated in $H$ given $\boldsymbol{Z}$ if there is no active path between any node $x \in \boldsymbol{X}$ and any node $y \in \boldsymbol{Y}$ given $\boldsymbol{Z}$
    - Denoted $\mathrm{sep}_H(\boldsymbol{X}; \boldsymbol{Y} | \boldsymbol{Z})$

$D \perp \{A, C\} \mid B$ ⬅




Global Markov assumptions: $I(H) = \{(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}) : \mathrm{sep}_H(\mathbf{X}; \mathbf{Y} | \mathbf{Z})\}$

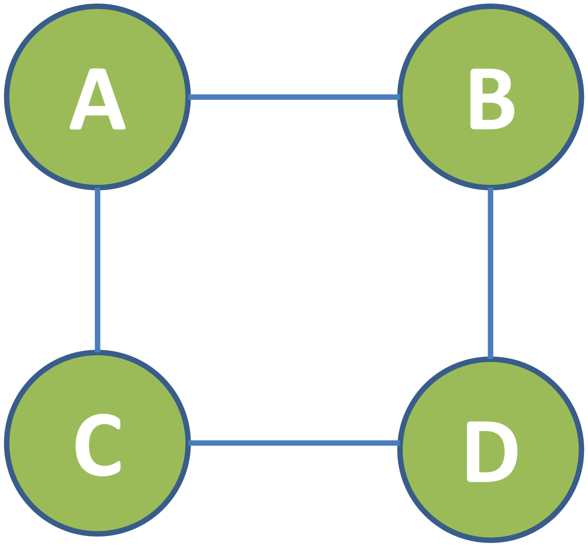
# Compare with Bayesian Networks

- Can all independencies encoded by Markov networks be perfectly encoded by Bayesian networks?
  - No, polygons with more than three nodes
  - See Ind(A;B | C,D) and Ind(C;D | A,B) example
- Can all independencies encoded by Bayesian networks be perfectly encoded by Markov networks?
  - No, immoral v-structures (explaining away)

*Try to prove these statements in the Assignment #4*

# Markov Network Factors: A Simple Example

- Nodes correspond to random variables
- **Local factors** are attached to <u>sets of nodes</u>
  - Factor elements are positive
  - Do not have to sum to 1
  - Represent affinities




| A | C | $\pi_1[A,C]$ |
|---|---|---|
| $a^0$ | $c^0$ | 20 |
| $a^0$ | $c^1$ | 3 |
| $a^1$ | $c^0$ | 5 |
| $a^1$ | $c^1$ | 35 |

| A | B | $\pi_2[A,B]$ |
|---|---|---|
| $a^0$ | $b^0$ | 15 |
| $a^0$ | $b^1$ | 8 |
| $a^1$ | $b^0$ | 6 |
| $a^1$ | $b^1$ | 10 |

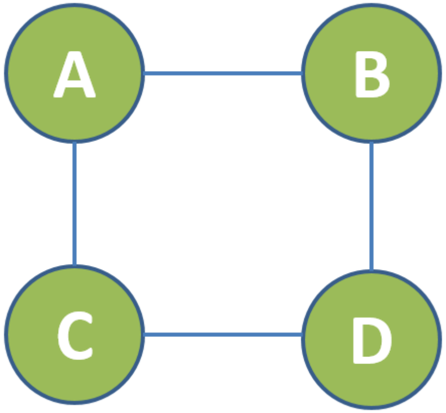
| C | D | $\pi_3[C,D]$ |
|---|---|---|
| $c^0$ | $d^0$ | 50 |
| $c^0$ | $d^1$ | 1 |
| $c^1$ | $d^0$ | 2 |
| $c^1$ | $d^1$ | 20 |

| B | D | $\pi_4[B,D]$ |
|---|---|---|
| $b^0$ | $d^0$ | 100 |
| $b^0$ | $d^1$ | 1 |
| $b^1$ | $d^0$ | 1 |
| $b^1$ | $d^1$ | 1000 |

How to represent the joint probability $P(ABCD)$?

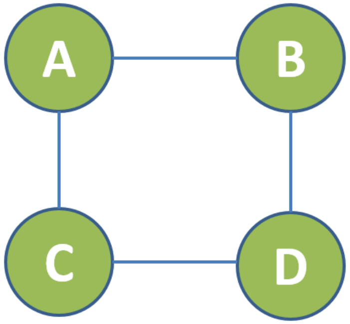
# Gibbs Distribution: A Simple Example

- Represents joint distribution
  - Un-normalized **factor** product
    - $F(a,b,c,d) = \pi_1[a,b]\pi_2[a,c]\pi_3[b,d]\pi_4[c,d]$
  - Partition function (**normalization function**)
    - $Z = \sum_{a,b,c,d} F(a,b,c,d)$
  - Probability (**Gibbs Distribution**)
    - $P(a,b,c,d) = \frac{1}{Z} F(a,b,c,d)$



# General: Markov Network Factors

- A **factor** is a function from value assignments of a set of random variables ***D*** to real positive numbers $\Re^+$. The set of variables ***D*** is the scope of the factor.
- The set of variables ***D*** in the same factor are **directly dependent** with each other
- Factors generalize the notion of CPDs
  - Every CPD is a factor



# General: Markov Network Factors

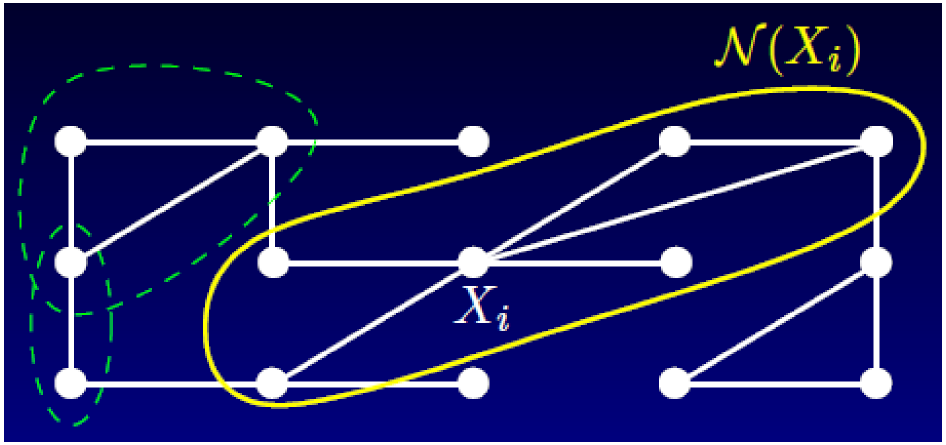
- Are there constraints imposed on the network structure ***H*** by a factor whose scope is ***D***?
  - Think of the independencies that must be satisfied
  - Generalize from the basic case of |***D***|=2

The induced subgraph over ***D*** must be a **clique** (fully connected)

(otherwise two unconnected variables may be independent by blocking the active path between them, contradicting the direct dependency between them in the factor over ***D***)

# Cliques: Closed Sets on the Graph

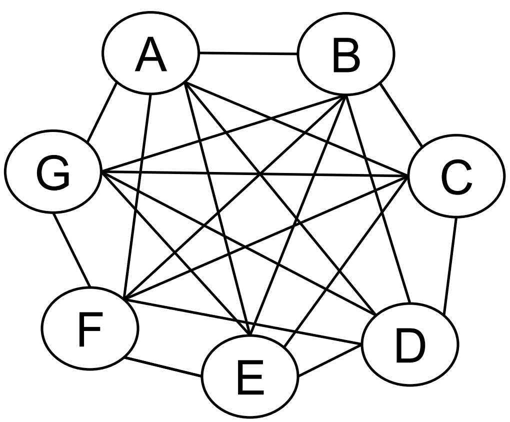
- Define a Markov network $\{V, E\}$, s.t. $\{X_i, X_j\} \in E$ iff $P(x_i|\boldsymbol{x}\backslash x_i) \neq P\left(x_i|\boldsymbol{x}\backslash\left(x_i, x_j\right)\right)$
  - Define *the **equivalence** between Graph & Distribution*
- Define $\mathcal{N}(X_i)$ s.t. $X_j \in \mathcal{N}(X_i)$ iff $\{X_i, X_j\} \in E$
- <u>$C \subseteq V$ is a clique, iff $C \subseteq \{X, \mathcal{N}(X)\}$ for $\forall\, X \in C$</u>




Basic local structures in MN are ***cliques***!!

# Required Markov Network Factors

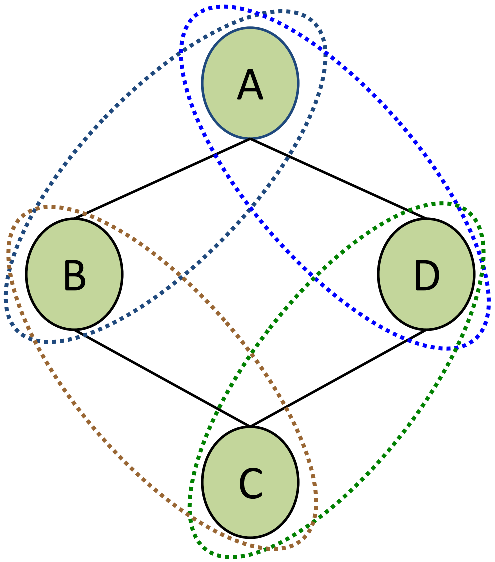
- Can we represent any joint distribution by using only factors that are defined on edges?
  - No! Full connected binary variables
    - Joint distribution has $2^n$-1 independent parameters
    - Markov network with edge factors has $4\binom{2}{n}$ parameters

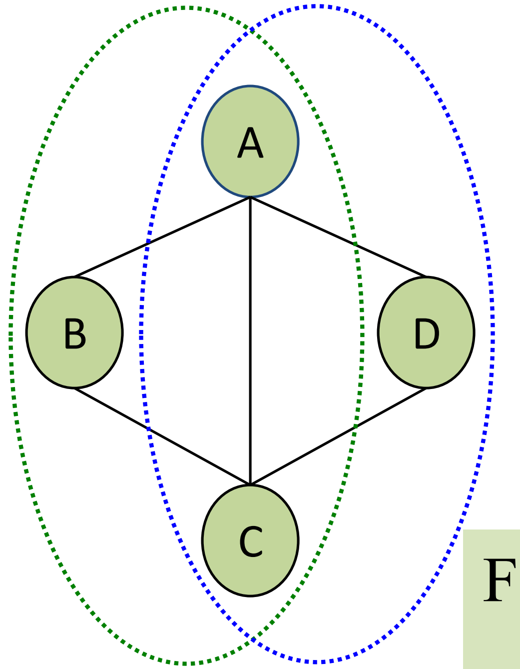



# Edge Parameters: 4*21=**84**

# Required Parameters: $2^7$-1=**127**

# Required Factors: Maximal Cliques




Maximal cliques
- {A,B}
- {B,C}
- {C,D}
- {A,D}

Maximal cliques
- {A,B,C}
- {A,C,D}

For example:

$\pi_1[ABC]\ \pi_1[AB]\ \pi_1[BC]$

$\leftrightarrow \pi[ABC]$

The non-maximal cliques can be represented by the corresponding maximal cliques

# Factorization of Markov Networks

## **Hammersley-Clifford Theorem** (1971)

***$D_i$*** encodes the ***cliques*** in the graph ***H***

- *I*-Map to Factorization

*I*-Map: $I(H) \subseteq I(P)$

  - Given a undirected graph *H*, **if *H* is an *I*-Map of *P*, *P* can be factorized as $P(X) = \frac{1}{Z}\prod \pi_i[D_i]$**
- Factorization to *I*-Map
  - Give a undirected graph *H*, **if *P* can be factorized as $P(X) = \frac{1}{Z}\prod \pi_i[D_i]$, *H* is an *I*-Map of *P***

Please read the proofs in the Textbook 4.3

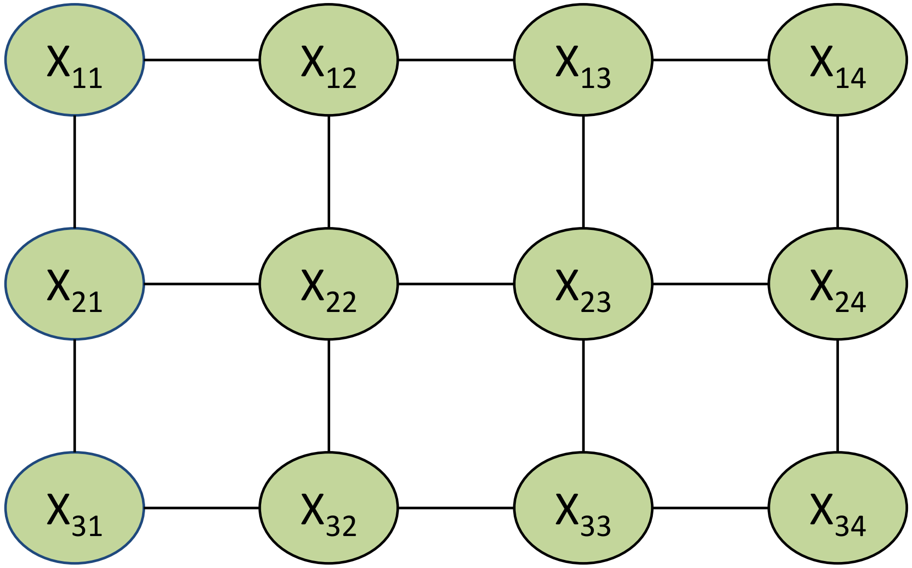
# Gibbs Distribution: General

- A distribution $P$ *factorizes* over $H$:
  - Cliques $\boldsymbol{D}_1, \cdots, \boldsymbol{D}_m$ in $H$
  - Factors $\pi_1[\boldsymbol{D}_1], \cdots, \pi_m[\boldsymbol{D}_m]$ defined on cliques
    - $P(X_1, \cdots, X_n) = \frac{1}{Z} f(X_1, \cdots, X_n)$
      - $f(X_1, \cdots, X_n) = \prod_{i=1}^{m} \pi_i[\boldsymbol{D}_i]$
      - $Z = \sum_{X_1, \cdots, X_n} f(X_1, \cdots, X_n) = \sum_{X_1, \cdots, X_n} \prod_{i=1}^{m} \pi_i[\boldsymbol{D}_i]$
  - $Z$ is called the ***partition function*** *(total energy)*

The distribution $P$ in factorized format is the corresponding **Gibbs distribution** over $H$

# Example: Pairwise Markov Networks

- A pairwise Markov network over a graph *H* has:
  - A set of *node factors* or potentials $\{\pi[X_i]:i=1,\ldots n\}$
  - A set of *edge factors or* potentials $\{\pi[X_i,X_j]: X_i,X_j \in H\}$
  - Example:



# Logarithmic Representation

- We represent energy potentials by applying a log transformation to the original distribution
  - $\pi[\boldsymbol{D}] = \exp(-\varepsilon[\boldsymbol{D}])$ where $\varepsilon[\boldsymbol{D}] = -\ln\left[\pi[\boldsymbol{D}]\right]$
  - Any Markov Network parameterized with factors can be converted to a logarithmic representation
  - The log-transformed potentials can take on any real value $P(X_1, \cdots, X_n) = \frac{1}{Z}\exp[-\sum_{i=1}^{m} \varepsilon_i[\boldsymbol{D}_i]]$
  - Change PRODUCT to **linear combination**!

# MN Representation Revisited

- For an undirected graph $\boldsymbol{H}$, define $Q$ function:
  - $Q(x|\mathrm{NB}(x)) = \log\left[\frac{P(x|\mathrm{NB}(x))}{P(x=0|\mathrm{NB}(x))}\right]$ **Ground State!**
- $\boldsymbol{H}$ is $I$-Map of positive distribution $P \Leftrightarrow$ This exists a unique expansion of $Q$
  - $Q(\boldsymbol{x}) = \sum_i x_i \psi_i(x_i) + \sum_{i,j} x_i x_j \psi_{i,j}(x_i, x_j) + \cdots + x_1 x_2 \cdots x_n \psi_{1,2,\cdots,n}(x_1, x_2, \cdots, x_n)$
  - $\psi_S \neq 0$ only if $v \in S$ form a clique in $\boldsymbol{H}$

HC Theorem, a constructive proof by ***Julian Besag*** in 1974

# A Demo Example with Three Binary Variables

- Three binary variables {0,1} have
  - $Q(x) = \ln\left[\frac{P(\boldsymbol{x})}{P(\boldsymbol{x}=\boldsymbol{0})}\right] = \sum_{i=1,2,3} \alpha_i x_i + \sum_{i<j=1,2,3} \alpha_{ij} x_i x_j + \alpha_{123} x_1 x_2 x_3$

- We can derive
  - $Q(x_1|x_2,x_3) = \ln\left[\frac{P(x_1|x_2,x_3)}{P(x_1=0|x_2,x_3)}\right] = \ln\left[\frac{P(x_1,x_2,x_3)}{P(x_1=0,x_2,x_3)}\right] = \ln\left[\frac{\frac{P(x_1,x_2,x_3)}{P(0,0,0)}}{\frac{P(x_1=0,x_2,x_3)}{P(0,0,0)}}\right]$

$$= Q(x_1,x_2,x_3) - Q(x_1=0,x_2,x_3)$$
$$= \alpha_1 x_1 + \alpha_{12} x_1 x_2 + \alpha_{13} x_1 x_3 + \alpha_{123} x_1 x_2 x_3$$

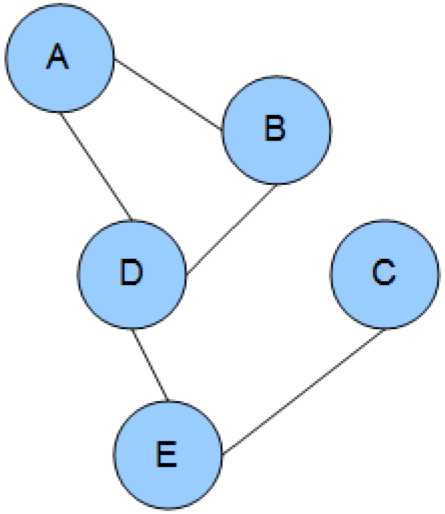
- If they form a linear chain $X_1 - X_2 - X_3$ and it is an *I*-Map
- The graph encodes $X_1 \perp X_3 | X_2$, above items with $x_3$ should be equal to zero $\Longrightarrow \alpha_{13}, \alpha_{123} = 0$ (*Note: a unique expansion*)

*The reverse statement is also correct. Please prove by yourself.*

# MN Representation Revisited

- Generally, a Markov network (undirected graph) encode a 0/1 binary probability as below:

$$P(X) = \frac{1}{Z}\exp(-U)$$




*U*: Potential Function

$\beta_i > 0$

$$U(X = x) = -\sum_i \left( \beta_i \prod_{x_j \in C_i} x_j \right)$$

*Z*: Partition Function

$$Z = \sum_X \exp(-U(X))$$

Could you directly write down its distribution?

# Comments on HC Theorem by Besag

- Ground state: when a variable is in its ground state, all the potential(s) associated with this variable were reduced to zero
- $Q$ function describes a relative probability compared to the corresponding ground state
- Independences encoded by undirected graphs are equal to the independences encoded by distributions (through $Q$ functions)
- A Markov network $\{H, P\}$, can be easily represented in log-linear formats by giving the potentials $\psi(C)$ over **cliques** in the graph

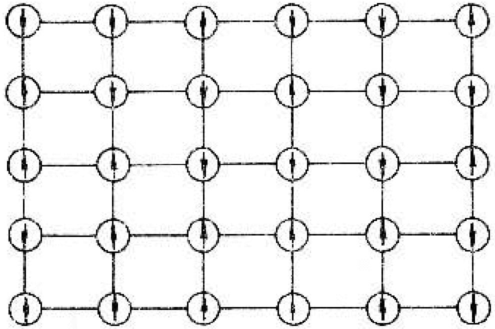
# Comments on HC Theorem by Besag

- He gives a constructive proof of Hammersley-Clifford theorem with lattice systems

## Big contribution!

You can directly write down the **Gibbs distribution** of any Markov network (binary variables) in log-linear format by enumerating all the cliques

# Ising Models



- A mathematical model of ferromagnetism in statistical mechanics.
- The model consists of discrete variables that represent magnetic dipole moments of atomic spins that can be in one of two states (+1 or −1).
- The spins are arranged in a graph, usually a lattice, allowing each spin to interact with its neighbors with interacting affinity $w_{i,j}$
- The total energy can be represented as

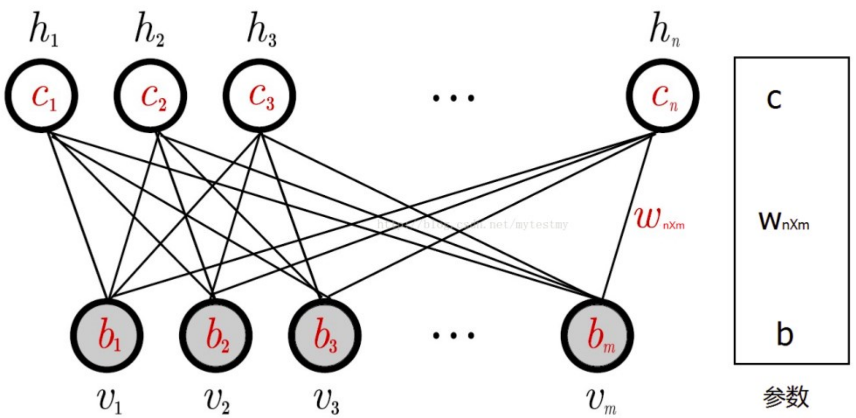
$$-\ \xi = -\sum_{i<j} w_{i,j} x_i x_j - \sum_i u_i x_i$$

The probability of a state configuration can be represented as the following Gibbs Distribution $P(\xi) = exp(-\xi)/Z$

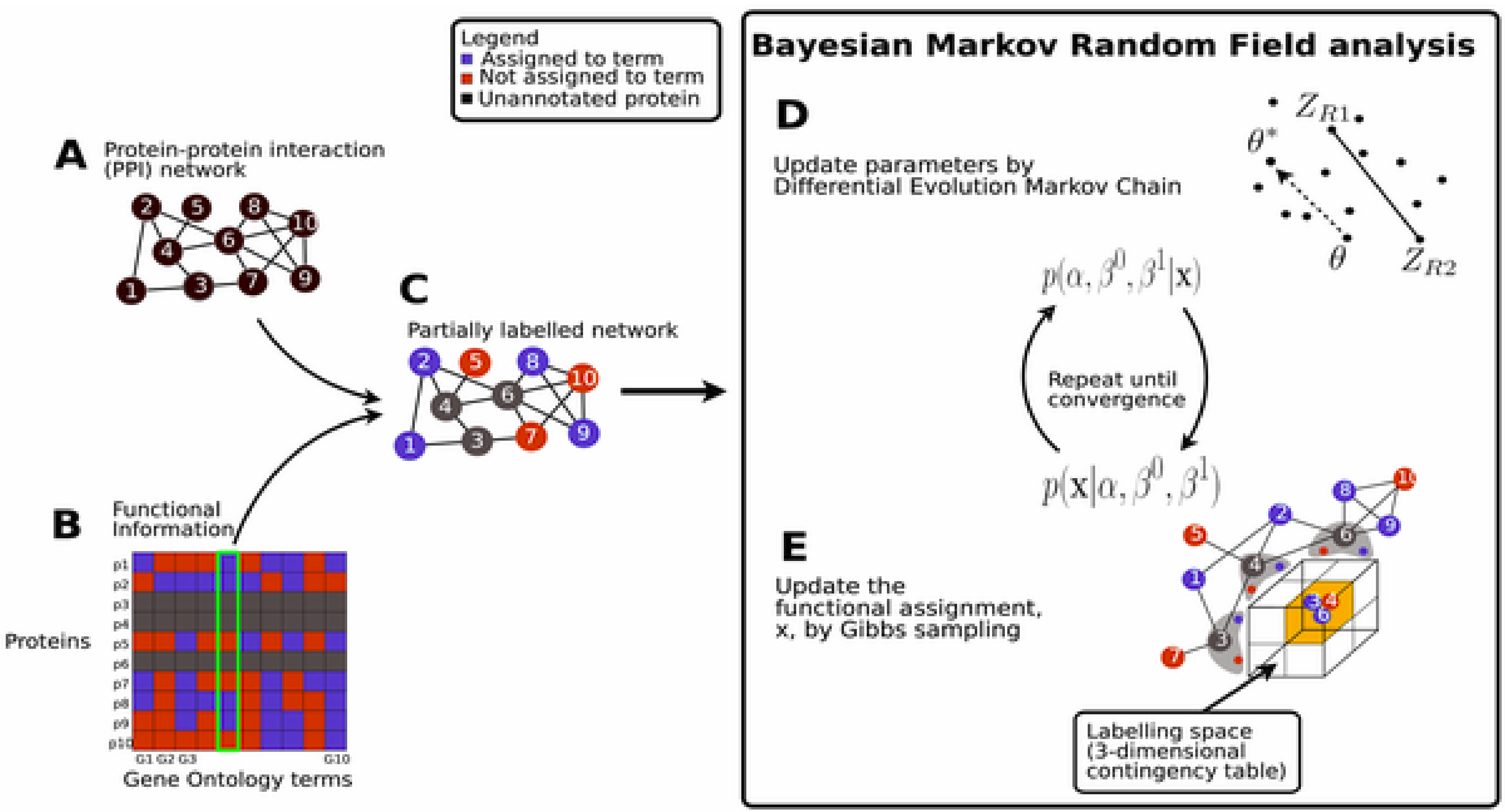
# Restricted Boltzmann Machines

- The variables are usually taken to have values {0, 1} but with the similar energy form of Ising models: $\xi = -\sum_{i,j} w_{i,j} b_i c_j - \sum_i v_i b_i - \sum_i h_i c_i$
- A nonzero contribution only takes effect when the interacted variables both equal 1

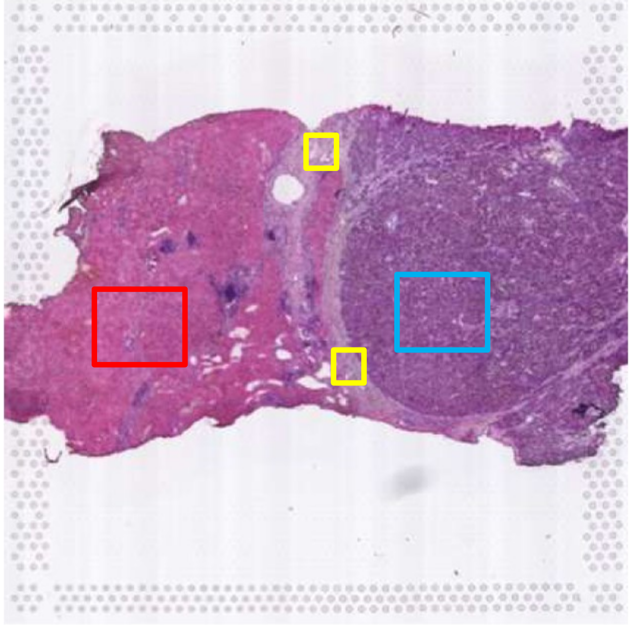
Restricted Boltzmann Machines (RBMs) in DL

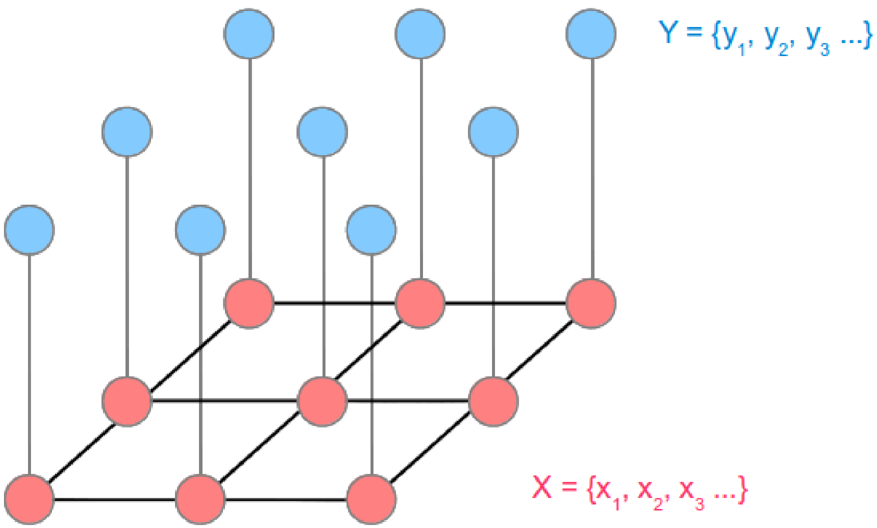


# Example: Protein Function Prediction



# Example: Image Segmentation



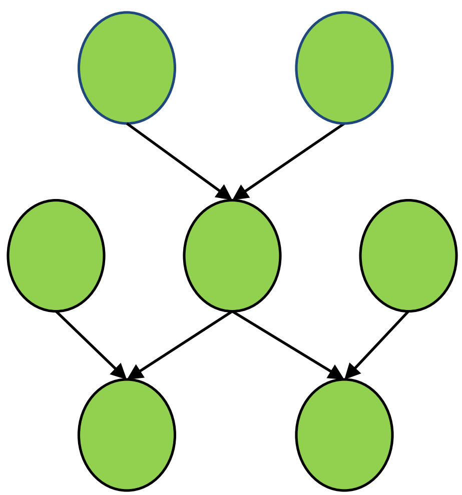


# The Independences in MNs

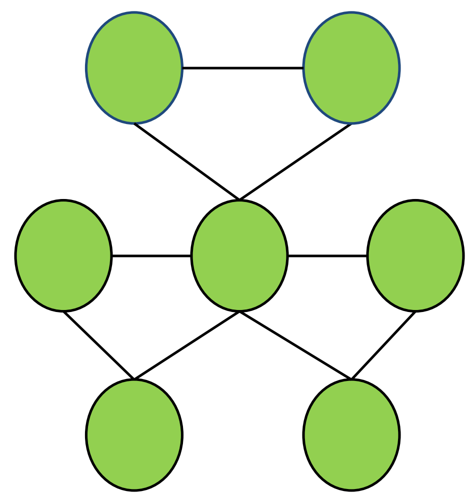
- **Markov blanket**
  - For $\boldsymbol{X}$, all the directly connected variables form the Markov blanket of $\boldsymbol{X}$, denoted as $\mathrm{MB}_H(\boldsymbol{X})$
- $X$ are independent with other variables if the corresponding $\mathrm{MB}_H(\boldsymbol{X})$ is given
- All the other independences can be derived from above basic independences

# Bayesian Networks to Markov Networks

- The moralized graph $H$ of a Bayesian network $G$ is the undirected graph that contains an undirected edge between $X$ and $Y$ if
  - $X$ and $Y$ are directly connected in G
  - $X$ and $Y$ have a common child in G



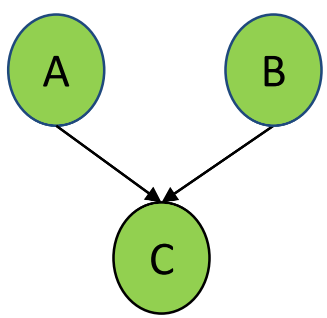
Bayesian network

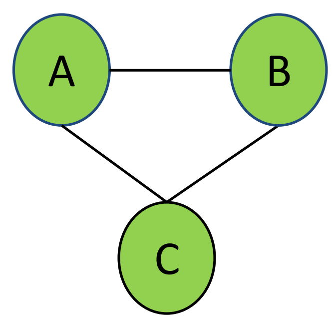


Moralized graph

# Parameterizing Moralized Graphs

- Moralized graph contains a full clique for every $X_i$ and its parents $Pa(X_i)$
  - → We can associate CPDs with a clique
- Do we lose independence assumptions implied by the graph structure?
  - Yes, immoral v-structures



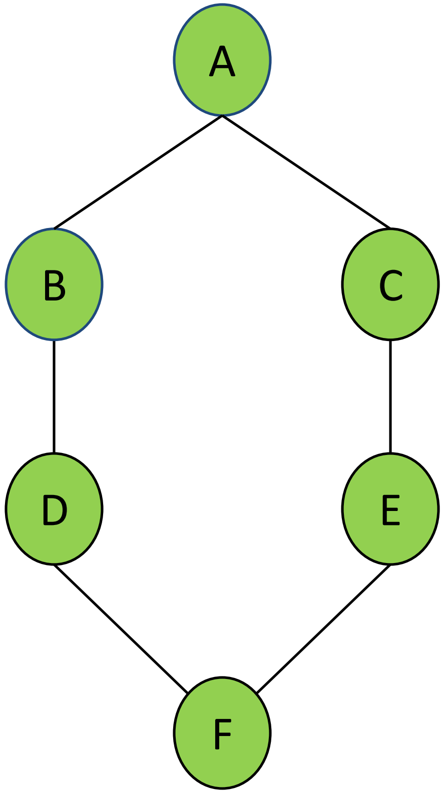

Ind(A;B)

# Markov Networks to Bayesian Networks

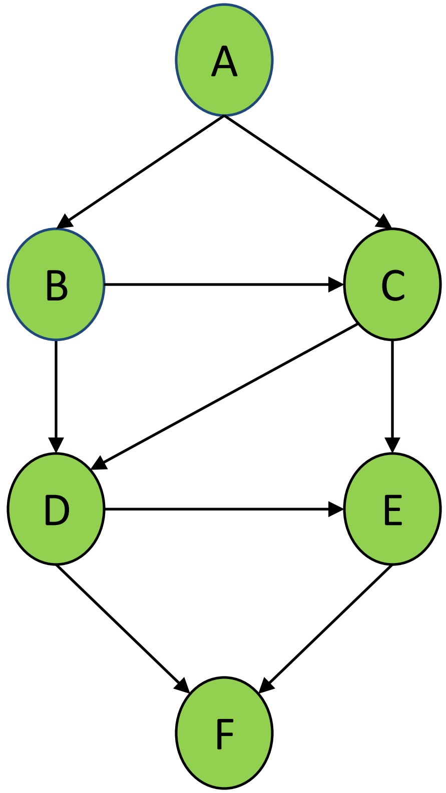
- Transformation is more difficult and the resulting network can be much larger than the Markov network

- Construction algorithm
  - Use Markov network as template for independencies
  - Fix ordering of nodes
  - Add each node along with its minimal parent set according to the independencies defined in the distribution

# Markov Networks to Bayesian Networks




Order: A,B,C,D,E,F



# Chordal Graphs

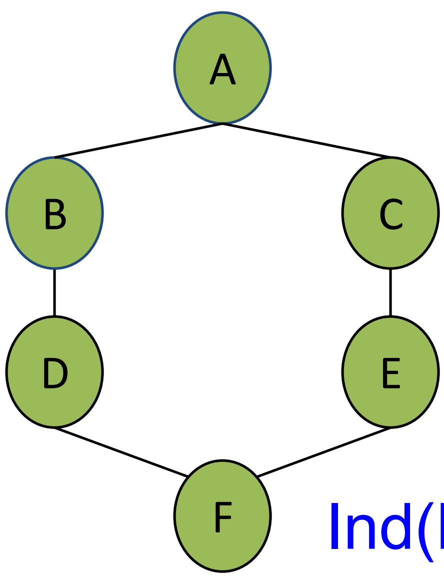
- Let $X_1$—$X_2$—...—$X_k$—$X_1$ be a loop in the graph
- A chord in the loop is an edge connecting $X_i$ and $X_j$ for two nonconsecutive nodes $X_i$ and $X_j$

- An undirected graph is chordal if any loop $X_1$—$X_2$—...—$X_k$—$X_1$ for $k \geq 4$ has a chord
  - That is, longest minimal loop is a triangle
  - Chordal graphs are often called triangulated

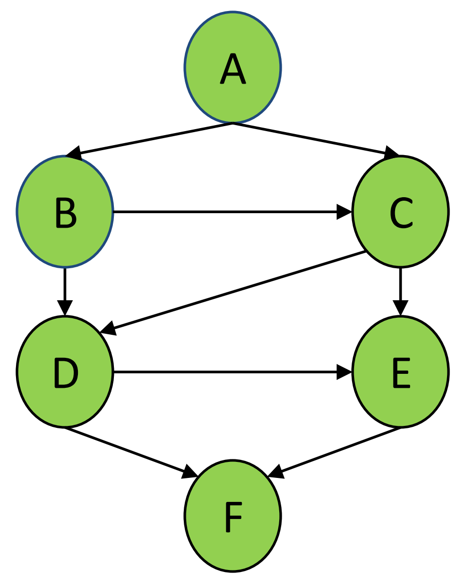
- A directed graph is chordal if its underlying undirected graph is chordal

# Markov Networks to Bayesian Networks

- Theorem: Let *H* be a Markov network and *G* be any minimal *I*-map for *H*, then *G* is chordal
- The process of turning a Markov network into a Bayesian network is called triangulation
  - The process loses independencies




Ind(B;C|A,F)



# Extended to Continuous Models

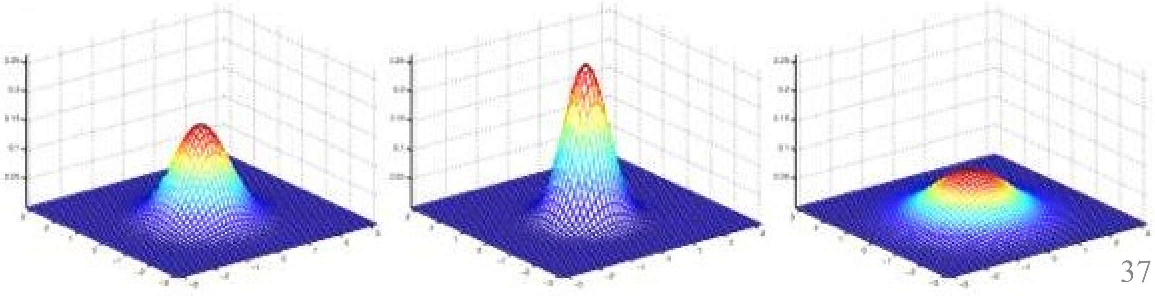
- Gaussian Random Fields
  - Multivariate Gaussians
  - Independences in Information Matrix
  - Gaussian Random Fields
- *Exponential Families*
  - *General Models*
  - *Linear and Factored Exponential Families*
  - *Entropy and Relative Entropy*
  - *Projections*

# Basic Parameters of Gaussians

- A multivariate Gaussian distribution over $X_1, \dots, X_n$ has the mean vector $\mu$ and covariance matrix $\Sigma$ $\left(\Sigma_{ij} = cov\left(X_i, X_j\right)\right)$. The density function of the distribution is

$$p(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

- The information matrix $J = \Sigma^{-1}$
- We get: $p(x) \propto exp\left(-\frac{1}{2}x^T J x + (J\mu)^T x\right)$



# Independences in Gaussians

- For a multivariate Gaussian, $X_i$ and $X_j$ are independent if their *covariance* is equal to zero $\Sigma_{ij} = 0$.
- The *information matrix* $J_{ij} = 0$, if and only if $p \vDash \left(X_i \perp X_j | X / \{X_i, X_j\}\right)$
- Information matrix clearly defines a set of pairwise independences. We can directly construct a **minimal I-map Markov network** with edges on the nonzero elements.

# Gaussian Random Fields

- According to the distribution
  - $p(x) \propto exp\left(-\frac{1}{2}x^T J x + (J\mu)^T x\right)$
- We can split it into two terms
  - $-\frac{1}{2}J_{ii}x_i x_i + h_i x_i$ ➔ single node
  - $-J_{ij}x_i x_j$ ➔ pairwise interaction
- Any multivariate Gaussian can be represented by a pairwise Markov network. This network is called a *Gaussian Markov Random Field*.
- But if the information matrix is not positive definite matrix, the GMRF may be illegal.

# Generalized to Exponential Families

- Each vector of parameters $\theta$ specifies a distribution in the exponential families as
  - $P_\theta(\xi) = \frac{1}{Z} A(\xi) exp(\langle t(\theta), \tau(\xi) \rangle)$
- Consider a simple Bernoulli distribution
  - $t(\theta) = \langle ln\theta, ln(1-\theta) \rangle$
  - $\tau(\xi) = \langle I\{\xi = 1\}, I\{\xi = 0\} \rangle$
- So for $X = 0$, we have
  - $exp(\langle t(\theta), \tau(\xi) \rangle) = exp\big(0 \cdot ln\theta + 1 \cdot \ln(1-\theta)\big)$

# Summary

- Hammersley-Clifford theorem: For a Markov network $\{H, P\}$, the positive distribution $P$ can factorize as factor products over cliques iff $H$ is $I$-map of $P$
- Besag's gives a constructive proof of HC theorem by a logarithmic representation over cliques
- A Markov network $\{H, P\}$, can be simply represented in log-linear formats by giving the potentials $\psi(C)$ over **cliques** in the graph

- Polygons (over triangles) in MN and v-structures in BN cannot be perfectly represented by each other