

# Data Mining 2020

## Bayesian Networks (1)

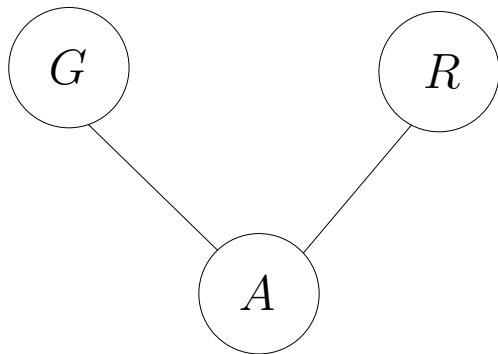
Ad Feelders

Universiteit Utrecht

# Do you like noodles?

		Do you like noodles?	
Race	Gender	Yes	No
Black	Male	10	40
	Female	30	20
White	Male	100	100
	Female	120	80

## Do you like noodles? Undirected



$$G \perp\!\!\!\perp R \mid A$$

Strange: Gender and Race are prior to Answer, but this model says they are independent *given* Answer!

# Do you like noodles?

Marginal table for Gender and Race:

Gender	Race	
	Black	White
Male	50	200
Female	50	200

From this table we conclude that Race and Gender are independent in the data. because both have a 50/50 split.

$$\text{cpr}(G,R)=1$$

# Do you like noodles?

Table for Gender and Race given Answer=yes:

Gender	Race	
	Black	White
Male	10	100
Female	30	120

$$\text{cpr}(G,R) = 0.4$$

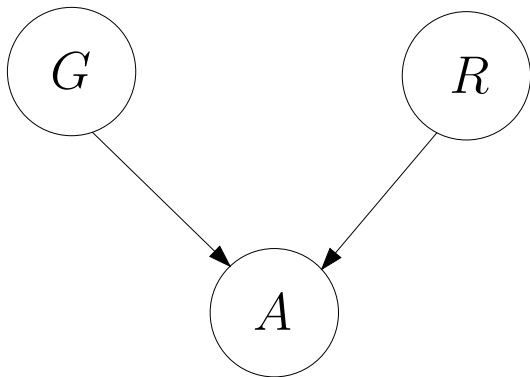
Table for Gender and Race given Answer=no:

Gender	Race	
	Black	White
Male	40	100
Female	20	80

$$\text{cpr}(G,R)=1.6$$

From these tables we conclude that Race and Gender are dependent given Answer.

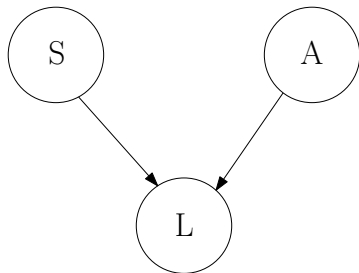
## Do you like noodles? Directed



$$G \perp\!\!\!\perp R, \quad G \not\perp\!\!\!\perp R | A$$

Gender and Race are marginally independent  
(but *dependent* given Answer).

## Explaining away



- Smoking (S) and asbestos exposure (A) are independent, but become dependent if we observe that someone has lung cancer (L).
- If we observe L, this raises the probability of both S and A.
- If we subsequently observe S, then the probability of A drops (explaining away effect).

# Directed Independence Graphs

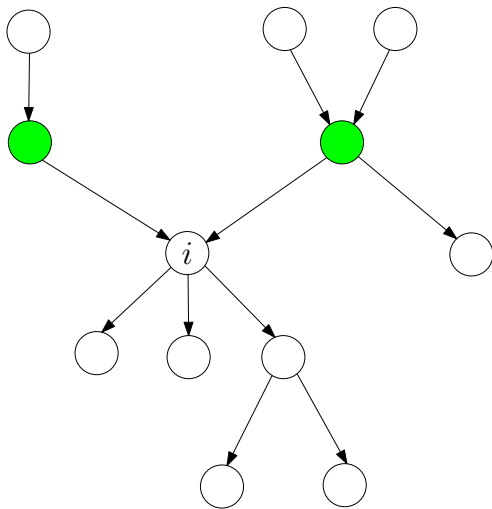
$G = (K, E)$ ,  $K$  is a set of vertices and  $E$  is a set of edges with *ordered* pairs of vertices.

- No directed cycles (DAG)
- parent/child
- ancestor/descendant
- ancestral set

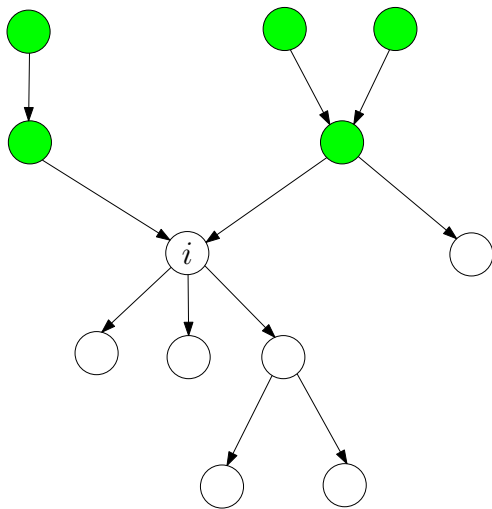
Because  $G$  is a DAG, there exists a *complete ordering* of the vertices that is respected in the graph (edges point from lower ordered to higher ordered nodes).



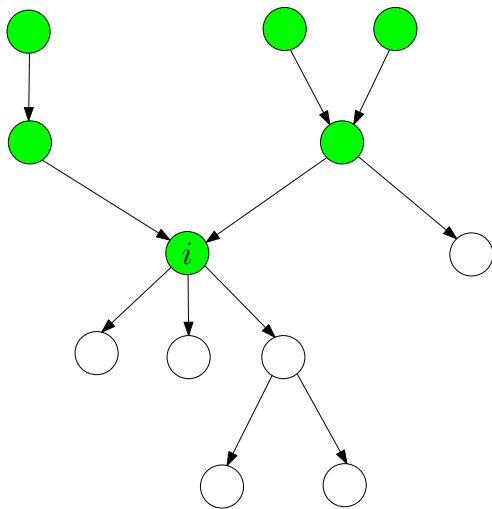
## Parents Of Node $i$ : $pa(i)$



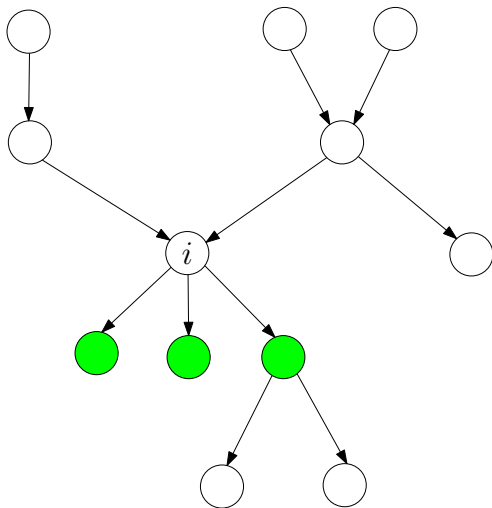
## Ancestors Of Node $i$ : $an(i)$



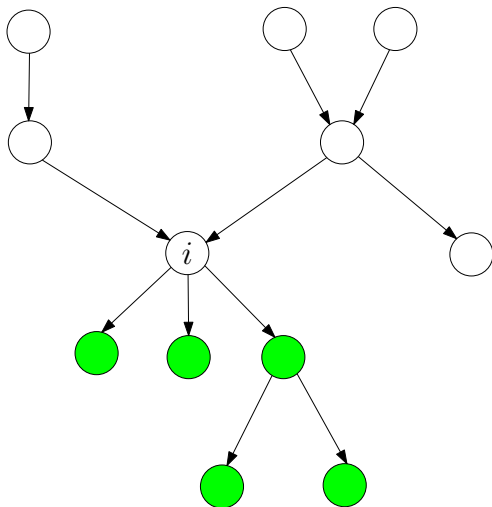
## Ancestral Set Of Node $i$ : $an^+(i)$



## Children Of Node $i$ : $ch(i)$



## Descendants Of Node $i$ : $de(i)$



# Construction of DAG

Suppose that *prior knowledge* tells us the variables can be labeled  $X_1, X_2, \dots, X_k$  such that  $X_i$  is prior to  $X_{i+1}$ .  
(for example: causal or temporal ordering)

Corresponding to this ordering we can use the product rule to factorize the joint distribution of  $X_1, X_2, \dots, X_k$  as

$$P(X) = P(X_1)P(X_2 \mid X_1) \cdots P(X_k \mid X_{k-1}, X_{k-2}, \dots, X_1)$$

Note that:

- 1 This is an identity of probability theory, no independence assumptions have been made yet!
- 2 The joint probability of any initial segment  $X_1, X_2, \dots, X_j$  ( $1 \leq j \leq k$ ) is given by the corresponding initial segment of the factorization.

# Constructing a DAG from pairwise independencies

Starting from the complete graph (containing arrows  $i \rightarrow j$  for all  $i < j$ ) an arrow from  $i$  to  $j$  is removed if  $P(X_j \mid X_{j-1}, \dots, X_1)$  does not depend on  $X_i$ , in other words, if

$$j \perp\!\!\!\perp i \mid \{1, \dots, j\} \setminus \{i, j\}$$

More loosely

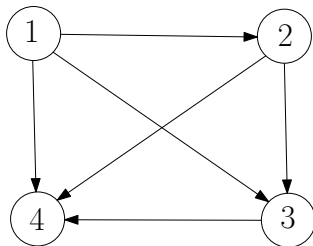
$$j \perp\!\!\!\perp i \mid \text{prior variables}$$

Compare this to pairwise independence

$$j \perp\!\!\!\perp i \mid \text{rest}$$

in undirected independence graphs.

# Construction Of DAG



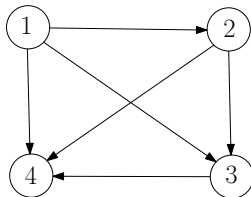
$$P(X) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)$$

Suppose the following independencies are given:

- 1  $X_1 \perp\!\!\!\perp X_2$
- 2  $X_4 \perp\!\!\!\perp X_3 | (X_1, X_2)$
- 3  $X_1 \perp\!\!\!\perp X_3 | X_2$



# Construction Of DAG

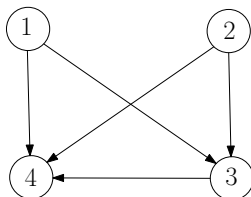


$$P(X) = P(X_1) \underbrace{P(X_2|X_1)}_{P(X_2)} P(X_3|X_1, X_2) P(X_4|X_1, X_2, X_3)$$

❶ If  $X_1 \perp\!\!\!\perp X_2$ , then  $P(X_2|X_1) = P(X_2)$ .

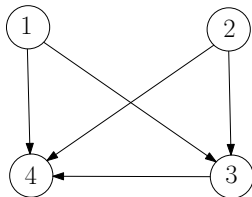
The edge  $1 \rightarrow 2$  is removed.

# Construction Of DAG



$$P(X) = P(X_1)P(X_2)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)$$

# Construction Of DAG

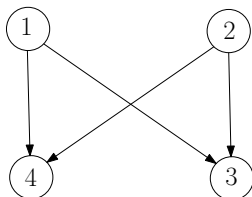


$$P(X) = P(X_1)P(X_2)P(X_3|X_1, X_2) \underbrace{P(X_4|X_1, X_2, X_3)}_{P(X_4|X_1, X_2)}$$

② If  $X_4 \perp\!\!\!\perp X_3 | (X_1, X_2)$ , then  $P(X_4|X_1, X_2, X_3) = P(X_4|X_1, X_2)$ .

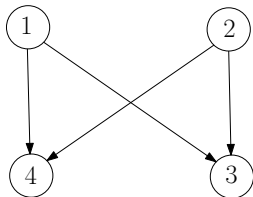
The edge  $3 \rightarrow 4$  is removed.

# Construction Of DAG



$$P(X) = P(X_1)P(X_2)P(X_3|X_1, X_2)P(X_4|X_1, X_2)$$

# Construction Of DAG



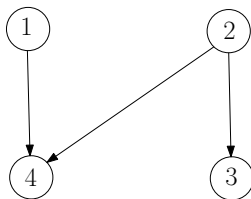
$$P(X) = P(X_1)P(X_2) \underbrace{P(X_3|X_1, X_2)}_{P(X_3|X_2)} P(X_4|X_1, X_2)$$

③ If  $X_1 \perp\!\!\!\perp X_3|X_2$ , then  $P(X_3|X_1, X_2) = P(X_3|X_2)$

The edge  $1 \rightarrow 3$  is removed.

# Construction Of DAG

We end up with this independence graph and corresponding factorization:



$$P(X) = P(X_1)P(X_2)P(X_3|X_2)P(X_4|X_1, X_2)$$

# Joint probability distribution of Bayesian Network

We can write the joint probability distribution more elegantly as

$$P(X_1, \dots, X_k) = \prod_{i=1}^k P(X_i \mid X_{pa(i)})$$

# Independence Properties of DAGs: d-separation and Moral Graphs

Can we infer other/stronger independence statements from the directed graph like we did using separation in the undirected graphical models?

Yes, the relevant concept is called d-separation.

- establishing d-separation directly (Pearl)
- establishing d-separation via the moral graph and “normal” separation

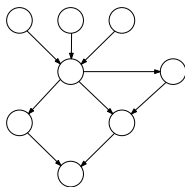
We discuss the second approach.



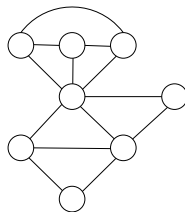
# Independence Properties of DAGs: Moral Graph

Given a DAG  $G = (K, E)$  we construct the moral graph  $G^m$  by marrying parents, and deleting directions, that is,

- 1 For each  $i \in K$ , we connect all vertices in  $\text{pa}(i)$  with undirected edges.
- 2 We replace all directed edges in  $E$  with undirected ones.



DAG



Moral Graph

# Independence Properties of DAGs: Moral Graph

The directed independence graph  $G$  possesses the conditional independence properties of its associated moral graph  $G^m$ . Why?

We have the factorisation:

$$\begin{aligned} P(X) &= \prod_{i=1}^k P(X_i \mid X_{pa(i)}) \\ &= \prod_{i=1}^k g_i(X_i, X_{pa(i)}) \end{aligned}$$

by setting  $g_i(X_i, X_{pa(i)}) = P(X_i \mid X_{pa(i)})$ .

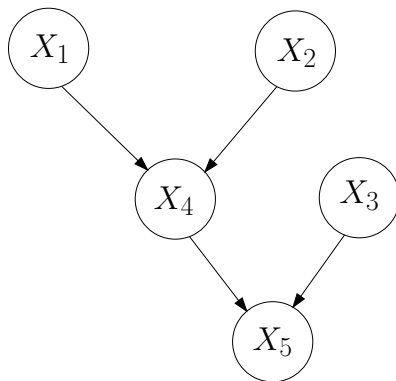
# Independence Properties of DAGs: Moral Graph

We have the factorisation:

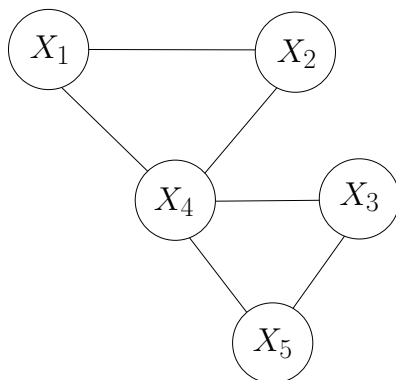
$$P(X) = \prod_{i=1}^k g_i(X_i, X_{pa(i)})$$

- We thus have a factorisation of the joint probability distribution in terms of functions  $g_i(X_{a_i})$  where  $a_i = \{i\} \cup pa(i)$ .
- By application of the factorisation criterion the sets  $a_i$  become cliques in the undirected independence graph.
- These cliques are formed by moralization.

# Moralisation: Example



# Moralisation: Example



$\{i\} \cup pa(i)$  becomes a complete subgraph in the moral graph  
(by marrying all unmarried parents).

# Moralisation Continued

Warning: the complete moral graph can obscure independencies!

To verify

$$i \perp\!\!\!\perp j \mid S$$

construct the moral graph of the induced subgraph on:

$$A = \text{an}^+(\{i, j\} \cup S),$$

that is,  $A$  contains  $i, j, S$  and all their ancestors.

Let  $G = (K, E)$  and  $A \subseteq K$ . The induced subgraph  $G_A$  contains nodes  $A$  and edges  $E'$ , where

$$i \rightarrow j \in E' \Leftrightarrow i \rightarrow j \in E \text{ and } i \in A \text{ and } j \in A.$$

# Moralisation Continued

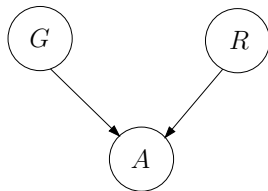
Since for  $\ell \in A$ ,  $pa(\ell) \in A$ , we know that the joint distribution of  $X_A$  is given by

$$P(X_A) = \prod_{\ell \in A} P(X_\ell \mid X_{pa(\ell)})$$

which corresponds to the subgraph  $G_A$  of  $G$ .

- 1 This is a product of factors  $P(X_\ell \mid X_{pa(\ell)})$ , involving the variables  $X_{\{\ell\} \cup pa(\ell)}$  only.
- 2 So it factorizes according to  $G_A^m$ , and thus the independence properties for undirected graphs apply.
- 3 Hence, if  $S$  separates  $i$  from  $j$  in  $G_A^m$ , then  $i \perp\!\!\!\perp j \mid S$ .

## Full moral graph may obscure independencies: example



$$P(G, R, A) = P(G)P(R)P(A \mid G, R)$$

Does  $G \perp\!\!\!\perp R$  hold? Summing out  $A$  we obtain:

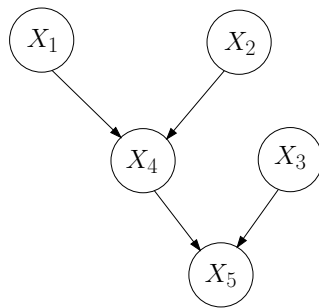
$$P(G, R) = \sum_a P(G, R, A = a) \quad (\text{sum rule})$$

$$= \sum_a P(G)P(R)P(A = a \mid G, R) \quad (\text{BN factorisation})$$

$$= P(G)P(R) \sum_a P(A = a \mid G, R) \quad (\text{rule of summation})$$

$$= P(G)P(R) \quad (\sum_a P(A = a \mid G, R) = 1)$$



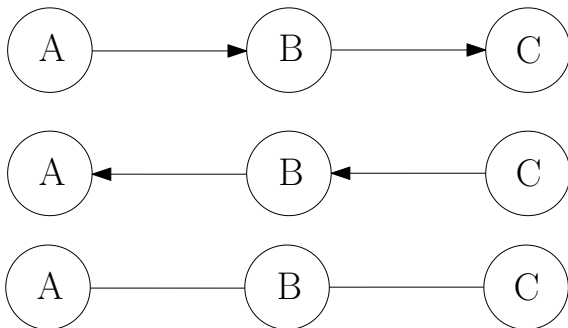


- 1 Are  $X_3$  and  $X_4$  independent?
- 2 Are  $X_1$  and  $X_3$  independent?
- 3 Are  $X_3$  and  $X_4$  independent given  $X_5$ ?
- 4 Are  $X_1$  and  $X_3$  independent given  $X_5$ ?

# Equivalence

When no marrying of parents is required (there are no “immoralities” or “v-structures”), then the independence properties of the directed graph are identical to those of its undirected version.

These three graphs express the same independence properties:



# Learning Bayesian Networks

- ① Parameter learning: structure known/given; we only need to estimate the conditional probabilities from the data.
- ② Structure learning: structure unknown; we need to learn the networks structure as well as the corresponding conditional probabilities from the data.

# Maximum Likelihood Estimation

Find value of unknown parameter(s) that maximize the probability of the observed data.

$n$  independent observations on binary variable  $X \in \{1, 2\}$ . We observe  $n(1)$  outcomes  $X = 1$  and  $n(2) = n - n(1)$  outcomes  $X = 2$ .

What is the maximum likelihood estimate of  $p(1)$ ?

The likelihood function (probability of the data) is given by:

$$L = p(1)^{n(1)}(1 - p(1))^{n - n(1)}$$

Taking the log we get

$$\mathcal{L} = n(1) \log p(1) + (n - n(1)) \log(1 - p(1))$$

# Maximum Likelihood Estimation

Take derivative with respect to  $p(1)$ , equate to zero, and solve for  $p(1)$ .

$$\frac{d\mathcal{L}}{dp(1)} = \frac{n(1)}{p(1)} - \frac{n - n(1)}{1 - p(1)} = 0,$$

since  $\frac{d \log x}{dx} = \frac{1}{x}$  (where log is the natural logarithm).

Solving for  $p(1)$ , we get

$$p(1) = \frac{n(1)}{n}.$$

This is just the fraction of one's in the sample!

# ML Estimation of Multinomial Distribution

Let  $X \in \{1, 2, \dots, J\}$ .

Estimate the probabilities  $p(1), p(2), \dots, p(J)$  of getting outcomes  $1, 2, \dots, J$ . If in  $n$  trials, we observe  $n(1)$  outcomes of 1,  $n(2)$  of 2,  $\dots$ ,  $n(J)$  of  $J$ , then the obvious guess is to estimate

$$p(j) = \frac{n(j)}{n}, \quad j = 1, 2, \dots, J.$$

This is indeed the maximum likelihood estimate.

# BN-Factorisation

For a given BN-DAG, the joint distribution factorises according to

$$P(X) = \prod_{i=1}^k p(X_i \mid X_{pa(i)})$$

So to specify the distribution we have to estimate the probabilities

$$p(X_i \mid X_{pa(i)}) \quad i = 1, 2, \dots, k$$

for the conditional distribution of each variable given its parents.

# ML Estimation of BN

The joint probability for  $n$  independent observations is

$$\begin{aligned} P(X^{(1)}, \dots, X^{(n)}) &= \prod_{j=1}^n P(X^{(j)}) \\ &= \prod_{j=1}^n \prod_{i=1}^k p(X_i^{(j)} \mid X_{pa(i)}^{(j)}), \end{aligned}$$

where  $X^{(j)}$  denotes the  $j$ -th row in the data table.

The likelihood function is therefore given by

$$L = \prod_{i=1}^k \prod_{x_i, x_{pa(i)}} p(x_i \mid x_{pa(i)})^{n(x_i, x_{pa(i)})}$$

where  $n(x_i, x_{pa(i)})$  is a count of the number of records with  $X_i = x_i$ , and  $X_{pa(i)} = x_{pa(i)}$ .



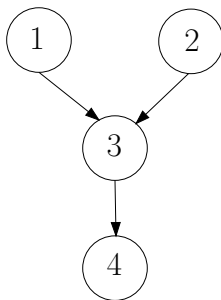
# ML Estimation of BN

Taking the log of the likelihood function, we get

$$\mathcal{L} = \sum_{i=1}^k \sum_{x_i, x_{pa(i)}} n(x_i, x_{pa(i)}) \log p(x_i \mid x_{pa(i)})$$

- Maximize the log-likelihood function with respect to the unknown parameters  $p(x_i \mid x_{pa(i)})$ .
- This decomposes into a collection of independent multinomial estimation problems.
- Separate estimation problem for each  $X_i$  and configuration of  $X_{pa(i)}$ .

## Example BN and Factorisation



$$P(X_1, X_2, X_3, X_4) = p_1(X_1)p_2(X_2)p_{3|12}(X_3|X_1, X_2)p_{4|3}(X_4|X_3)$$

## Example BN: Parameters

$$P(X_1, X_2, X_3, X_4) = p_1(X_1)p_2(X_2)p_{3|12}(X_3|X_1, X_2)p_{4|3}(X_4|X_3)$$

Now we have to estimate the following parameters ( $X_4$  ternary, rest binary):

inferred parameters

$$p_1(1) \quad p_1(2) = 1 - p_1(1)$$

$$p_2(1) \quad p_2(2) = 1 - p_2(1)$$

$$p_{3|1,2}(1|1, 1) \quad p_{3|1,2}(2|1, 1) = 1 - p_{3|1,2}(1|1, 1)$$

$$p_{3|1,2}(1|1, 2) \quad p_{3|1,2}(2|1, 2) = 1 - p_{3|1,2}(1|1, 2)$$

$$p_{3|1,2}(1|2, 1) \quad p_{3|1,2}(2|2, 1) = 1 - p_{3|1,2}(1|2, 1)$$

$$p_{3|1,2}(1|2, 2) \quad p_{3|1,2}(2|2, 2) = 1 - p_{3|1,2}(1|2, 2)$$

$$p_{4|3}(1|1) \quad p_{4|3}(2|1) \quad p_{4|3}(3|1) = 1 - p_{4|3}(1|1) - p_{4|3}(2|1)$$

$$p_{4|3}(1|2) \quad p_{4|3}(2|2) \quad p_{4|3}(3|2) = 1 - p_{4|3}(1|2) - p_{4|3}(2|2)$$

# Example Data Set

obs	$X_1$	$X_2$	$X_3$	$X_4$
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
4	1	2	2	1
5	1	2	2	2
6	2	1	1	2
7	2	1	2	3
8	2	1	2	3
9	2	2	2	3
10	2	2	1	3

# Maximum Likelihood Estimation

obs	$X_1$	$X_2$	$X_3$	$X_4$
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
4	1	2	2	1
5	1	2	2	2
6	2	1	1	2
7	2	1	2	3
8	2	1	2	3
9	2	2	2	3
10	2	2	1	3

$$\hat{p}_1(1) = \frac{n(x_1 = 1)}{n} = \frac{5}{10} = \frac{1}{2}$$

# Maximum Likelihood Estimation

obs	$X_1$	$X_2$	$X_3$	$X_4$
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
4	1	2	2	1
5	1	2	2	2
6	2	1	1	2
7	2	1	2	3
8	2	1	2	3
9	2	2	2	3
10	2	2	1	3

$$\hat{p}_2(1) = \frac{n(x_2 = 1)}{n} = \frac{6}{10}$$

# Maximum Likelihood Estimation

obs	$X_1$	$X_2$	$X_3$	$X_4$
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
4	1	2	2	1
5	1	2	2	2
6	2	1	1	2
7	2	1	2	3
8	2	1	2	3
9	2	2	2	3
10	2	2	1	3

$$\hat{p}_{3|1,2}(1|1,1) = \frac{n(x_1 = 1, x_2 = 1, x_3 = 1)}{n(x_1 = 1, x_2 = 1)} = \frac{2}{3}$$

# Maximum Likelihood Estimation

obs	$X_1$	$X_2$	$X_3$	$X_4$
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
4	1	2	2	1
5	1	2	2	2
6	2	1	1	2
7	2	1	2	3
8	2	1	2	3
9	2	2	2	3
10	2	2	1	3

$$\hat{p}_{3|1,2}(1|1,1) = \frac{n(x_1 = 1, x_2 = 1, x_3 = 1)}{n(x_1 = 1, x_2 = 1)} = \frac{2}{3}$$



# ML Estimation of BN

The maximum likelihood estimate of  $p(x_i \mid x_{pa(i)})$  is given by:

$$\hat{p}(x_i \mid x_{pa(i)}) = \frac{n(x_i, x_{pa(i)})}{n(x_{pa(i)})},$$

how many times  
the parent configuration occur

where

- $n(x_i, x_{pa(i)})$  is the number of records in the data with  $X_i = x_i$  and  $X_{pa(i)} = x_{pa(i)}$ , and
- $n(x_{pa(i)})$  is the number of records in the data with  $X_{pa(i)} = x_{pa(i)}$ .