Royal Institute of Technology

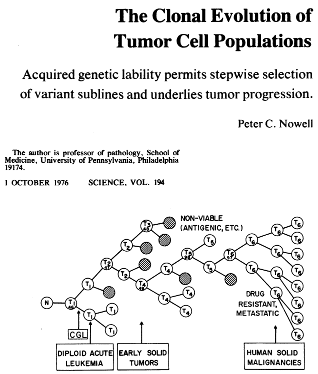
# STATISTICAL METHODS IN CS, CH 10

## Lecture 5

# REPRESENTING AND WORKING WITH DISTRIBUTIONS

* For all but the smallest n, the explicit representation of the joint distribution is *unmanageable from every perspective*.
* Computationally, it is very *expensive to manipulate* and generally *too large to store* in memory.
* Cognitively, it is *impossible to acquire so many numbers* from a human expert; moreover, the numbers are very small and *do not correspond to events that people can reasonably contemplate*.
* Statistically, if we want to learn the distribution from data, we would *need ridiculously large amounts of data to estimate* this many parameters robustly.
* These problems were the *main barrier* to the adoption of probabilistic methods for expert systems *until the development of the methodologies we now will consider.*

# SOMATIC EVOLUTION



**The Clonal Evolution of Tumor Cell Populations**

Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression.

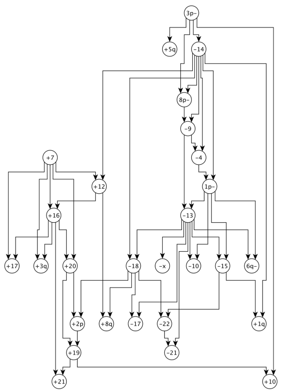
Peter C. Nowell

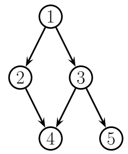
The author is professor of pathology, School of Medicine, University of Pennsylvania, Philadelphia 19174.

1 OCTOBER 1976 SCIENCE, VOL. 194



Oncogenetic network

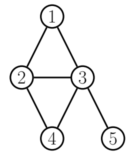






Directed graphical model

* DAG
* vertices r.v.s
* equipped with local CPDs



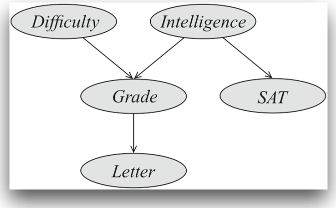


Undirected graphical model

* graph
* vertices r.v.s
* equipped with local "factors"

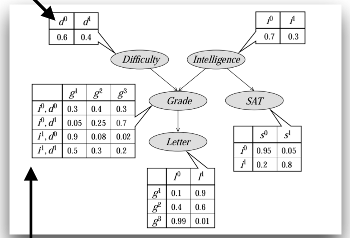
# GRAPHICAL MODELS

# DGM - GRAPH AND CPDS VS JOINT




P(D,I,G,S,L)

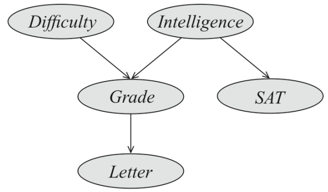
d has value 0

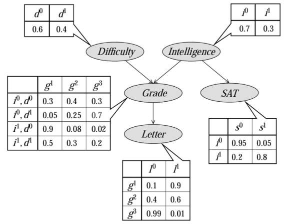


CPD

- CPT - table, i.e., categorical
- Gaussian

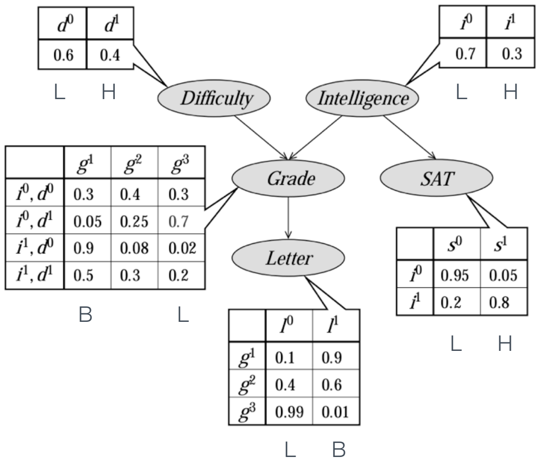
# THREE LEVELS OF COMPUTATIONAL PROBLEMS



- Inference: given G and θ, compute probabilities or marginalize
- Parameter learning: given G and D, learn θ
- Structure learning: given D, learn G and θ

# EXTENDED STUDENT EXAMPLE




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

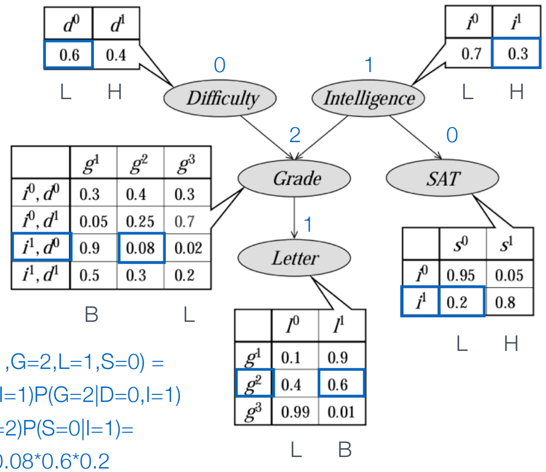
| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

B - better
H - higher
L - less

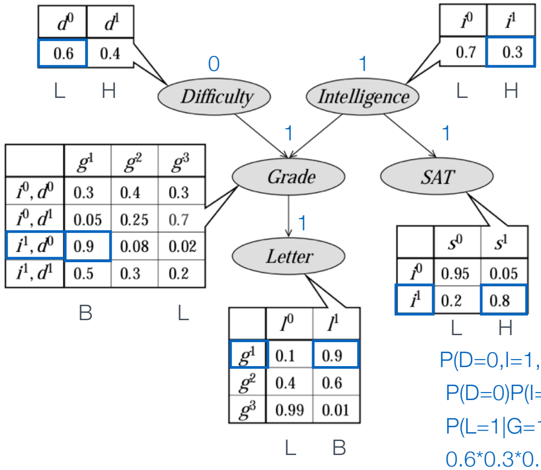
# EXTENDED STUDENT EXAMPLE




P(D=0,I=1,G=2,L=1,S=0) =
P(D=0)P(I=1)P(G=2|D=0,I=1)
P(L=1|G=2)P(S=0|I=1)=
0.6*0.3*0.08*0.6*0.2

B - better
H - higher
L - less

# EXTENDED STUDENT EXAMPLE




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

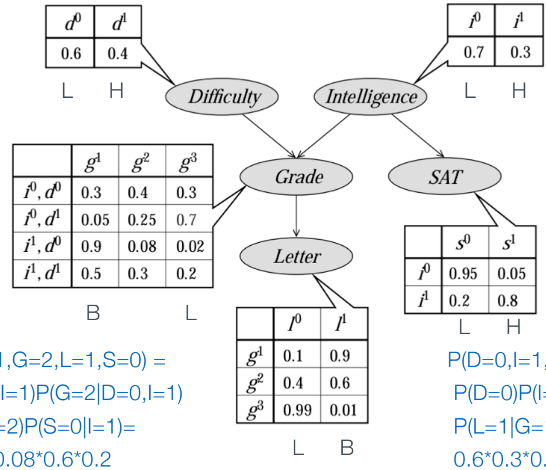
| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

P(D=0,I=1,G=1,L=1,S=1) =
P(D=0)P(I=1)P(G=1|D=0,I=1)
P(L=1|G=1)P(S=1|I=1)=
0.6*0.3*0.9*0.9*0.8

# EXTENDED STUDENT EXAMPLE




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

P(D=0,I=1,G=2,L=1,S=0) =
P(D=0)P(I=1)P(G=2|D=0,I=1)
P(L=1|G=2)P(S=0|I=1)=
0.6*0.3*0.08*0.6*0.2

P(D=0,I=1,G=1,L=1,S=1) =
P(D=0)P(I=1)P(G=1|D=0,I=1)
P(L=1|G=1)P(S=1|I=1)=
0.6*0.3*0.9*0.9*0.8

# EXTENDED STUDENT EXAMPLE

P(D=0,I=1,G=2,L=1,S=0) P(D=0,I=1,G=1,L=1,S=0)
= P(D=0)P(I=1)P(G=1|D=0,I=1) P(L=1|G=1)P(S=1|I=1)
P(D=0)P(I=1)P(G=2|D=0,I=1)P(L=1|G=2)P(S=0|I=1)
= P(D=0)$^2$P(I=1)$^2$P(G=1|D=0,I=1)P(G=2|D=0,I=1)
P(L=1|G=1)P(L=1|G=2)P(S=1|I=1)P(S=0|I=1)

P(D=0,I=1,G=2,L=1,S=0) =
P(D=0)P(I=1)P(G=2|D=0,I=1)
P(L=1|G=2)P(S=0|I=1)=
0.6*0.3*0.08*0.6*0.2

P(D=0,I=1,G=1,L=1,S=1) =
P(D=0)P(I=1)P(G=1|D=0,I=1)
P(L=1|G=1)P(S=1|I=1)=
0.6*0.3*0.9*0.9*0.8

# INFERENCE – THE CHAIN RULE

$$p(\underbrace{\boldsymbol{x}_{[V]}}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_V}) = p(\boldsymbol{x}_1)p(\boldsymbol{x}_2|\boldsymbol{x}_1)p(\boldsymbol{x}_3|\boldsymbol{x}_1,\boldsymbol{x}_2)\cdots p(\boldsymbol{x}_V|\boldsymbol{x}_{[V-1]})$$

- Assuming binary r.v., p($X_V$ | $X_{[V-1]}$) has $2^{V-1}$ parameters
- Total # parameters $\sum_{1 \le i \le V} 2^{i-1} = 2^V - 1$

# EX. WHERE IND. OBVIOUSLY FACILITATES

$$p(\underbrace{\boldsymbol{x}_{[V]}}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_V}) = p(\boldsymbol{x}_1)p(\boldsymbol{x}_2|\boldsymbol{x}_1)p(\boldsymbol{x}_3|\boldsymbol{x}_1,\boldsymbol{x}_2)\cdots p(\boldsymbol{x}_V|\boldsymbol{x}_{[V-1]})$$

- ★Assume first order Markov property $\boldsymbol{x}_t \perp \boldsymbol{x}_{[t-2]}|\boldsymbol{x}_{t-1}$

  i.e., if time ordered, future independent of past given present
- ★Then $$p(\boldsymbol{x}_{[V]}) = p(\boldsymbol{x}_1)\prod_{t=1}^{V-1} p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$$

# FACTORIZATION OVER G

$$p(x_1,\ldots,x_N) = \prod_{n=1}^{N} p(x_n|\boldsymbol{x}_{\mathrm{pa}(x_n)})$$

p can be factorized over G if it can be expressed as above

# CAT – NOTATION

- ★ For a v ∈ [V],

  values $k \in [K_v]$

  combined values $c \in C_v = \prod_{s\in \mathrm{pa}(v)} [K_s]$ ← Cartesian product
- ★ Cat CPDs

  where $P(x_v|x_{\mathrm{pa}(v)} = c) = \mathrm{Cat}(\boldsymbol{\theta}_{vc})$

  and $\boldsymbol{\theta}_{vck} = P(x_v = k|x_{\mathrm{pa}(v)} = c)$

# THE LIKELIHOOD FACTORIZES

- ★ Complete data $\mathcal{D} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$

  $\boldsymbol{x}_n = \{\boldsymbol{x}_{n1},\ldots,\boldsymbol{x}_{nV}\}$
- ★ Likelihood

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta}) = \prod_{n=1}^{N}\prod_{v=1}^{V} p(\boldsymbol{x}_{nv}|\boldsymbol{x}_{n,\mathrm{pa}(v)},\boldsymbol{\theta})$$

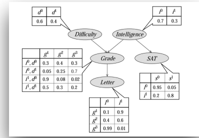
$$= \prod_{v=1}^{V}\prod_{n=1}^{N} p(\boldsymbol{x}_{nv}|\boldsymbol{x}_{n,\mathrm{pa}(v)},\boldsymbol{\theta}) = \prod_{v=1}^{V} p(\mathcal{D}_v|\boldsymbol{\theta}_v)$$

  where $D_v$ is values of v together with its parents and $\theta_v$ is v's CPD
- ★ Called: decomposable likelihood (factorizes into family-factors)

# THE LIKELIHOOD FACTORIZES

- Complete data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$\boldsymbol{x}_n = \{\boldsymbol{x}_{n1}, \ldots, \boldsymbol{x}_{nV}\}$$



- Likelihood

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{v=1}^{V} p(\boldsymbol{x}_{nv}|\boldsymbol{x}_{n,\mathrm{pa}(v)}, \boldsymbol{\theta})$$

$$= \prod_{v=1}^{V} \prod_{n=1}^{N} p(\boldsymbol{x}_{nv}|\boldsymbol{x}_{n,\mathrm{pa}(v)}, \boldsymbol{\theta}) = \prod_{v=1}^{V} p(\mathcal{D}_v|\boldsymbol{\theta}_v)$$

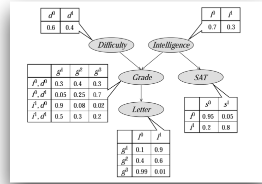
where $D_v$ is values of v together with its parents and $\theta_v$ is v's CPD

- Called: decomposable likelihood (factorizes into family-factors)

# MLE FOR CAT CPDS

- Each $P(\mathcal{D}_v|\boldsymbol{\theta}_v)$ , i.e., here each $\boldsymbol{\theta}_{vc} = (\theta_{vc1}, \ldots, \theta_{vcK_v})$

  can be maximized independently



- So, MLE is

$$\boldsymbol{\theta}_{\boldsymbol{vck}} = N_{\boldsymbol{vck}}/N_{\boldsymbol{vc}}$$

- where

$$N_{\boldsymbol{vck}} = \sum_{n=1}^{N} I(x_{nv} = k, x_{n,\mathrm{pa}(v)} = \mathbf{c})$$

$$N_{\boldsymbol{vc}} = \sum_{n=1}^{N} I(x_{n,\mathrm{pa}(v)} = \mathbf{c})$$

c

# BAYESIAN PARAMETER LEARNING

- Decomposable prior

$$p(\boldsymbol{\theta}) = \prod_{v=1}^{V} p(\boldsymbol{\theta}_v) \quad \text{where} \quad \boldsymbol{\theta}_v = (\boldsymbol{\theta}_{v_1}, \ldots, \boldsymbol{\theta}_{v\mathbf{K}_{\mathrm{pa}(v)}})$$

- Gives decomposable posterior

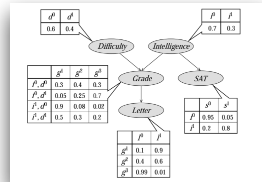
$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \prod_{v=1}^{V} p(\mathcal{D}_v|\boldsymbol{\theta}_v)p(\boldsymbol{\theta}_v)$$

# POSTERIOR

$$\boldsymbol{\theta}_{vc} = (\theta_{vc1}, \ldots, \theta_{vcK_v})$$

- $\alpha_{vc}$ is a vector of hyperparameters, prior

$$\boldsymbol{\theta}_{\boldsymbol{vc}} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{\boldsymbol{vc}})$$



- The posterior is

$$\boldsymbol{\theta}_{\boldsymbol{vc}}|\mathcal{D} \sim \mathrm{Dir}(\boldsymbol{N}_{\boldsymbol{vc}} + \boldsymbol{\alpha}_{\boldsymbol{vc}})$$

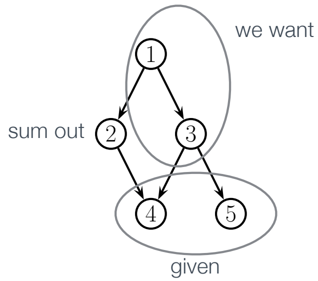
# MARGINALIZE

$X, X'$ two hidden variables

$X_h$ the other hidden variables

$X_v$ the visible variables

$$
\begin{aligned}
&p(X = k, X' = k' | \boldsymbol{x}_v, \boldsymbol{\theta}) \\
&\quad = \frac{p(X = k, X' = k', \boldsymbol{x}_v | \boldsymbol{\theta})}{P(\boldsymbol{x}_v | \boldsymbol{\theta})} \\
&\quad = \frac{\sum_{x_h} p(X = k, X' = k', \boldsymbol{x}_h, \boldsymbol{x}_v | \boldsymbol{\theta})}{\sum_{x, x', x_h} p(x, x', \boldsymbol{x}_h, \boldsymbol{x}_v | \boldsymbol{\theta})}
\end{aligned}
$$




# MARGINALIZE

$$
p(\boldsymbol{X}_m | \boldsymbol{x}_e, \boldsymbol{\theta}) = \frac{p(\boldsymbol{X}_m, \boldsymbol{x}_e | \boldsymbol{\theta})}{p(\boldsymbol{x}_e | \boldsymbol{\theta})} = \frac{\sum_{\boldsymbol{x}_{V \setminus (m \cup e)}} p(\boldsymbol{X}_m, \boldsymbol{x}_e | \boldsymbol{\theta})}{\sum_{\boldsymbol{x}_{V \setminus e}} p(\boldsymbol{x}_e | \boldsymbol{\theta})}
$$

- The denominator contains a marginal likelihood
- Summing out V binary hidden variables – $O(2^V)$
- K values – $O(K^V)$

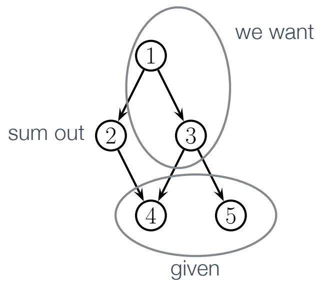
# EXPECTED SUFFICIENT STATISTICS - ESS

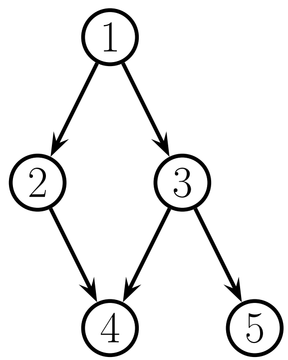
$X, X'$ two hidden variables

$X_h$ the other hidden variables

$X_v$ the visible variables

$$
\begin{aligned}
&E[N_{k,k'}] \\
&\quad = \sum_{\boldsymbol{x} \in \mathcal{D}} p(X = k, X' = k' | \boldsymbol{x}_v, \boldsymbol{\theta})
\end{aligned}
$$



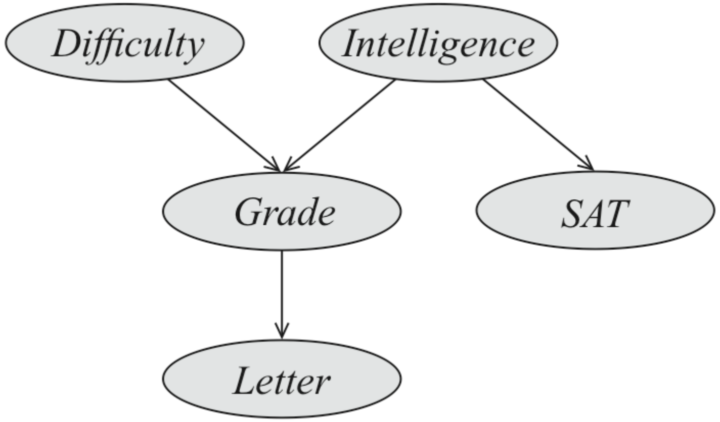

# DGM

★What is the meaning of the underlying DAG? what is the semantics?

★What does a DGM mean? what is the semantics?

★Which DGMs represent a given distribution?

# EXTENDED STUDENT EXAMPLE




# INDEPENDENCE I-MAP

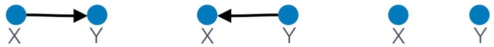
* I(G) independences implied by G (not yet defined)
* I(P) independences in the distribution P
* G I-map for P in I(G) ⊆ I(P)

p

| $X$ | $Y$ | $P(X,Y)$ |
|---|---|---|
| $x^0$ | $y^0$ | 0.08 |
| $x^0$ | $y^1$ | 0.32 |
| $x^1$ | $y^0$ | 0.12 |
| $x^1$ | $y^1$ | 0.48 |

q

| $X$ | $Y$ | $P(X,Y)$ |
|---|---|---|
| $x^0$ | $y^0$ | 0.4 |
| $x^0$ | $y^1$ | 0.3 |
| $x^1$ | $y^0$ | 0.2 |
| $x^1$ | $y^1$ | 0.1 |




# INDEPENDENCE I-MAP

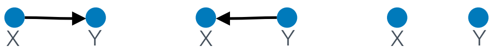
* I(G) independences implied by G (not yet defined)
* I(P) independences in the distribution P
* G I-map for P in I(G) ⊆ I(P)

p

| $X$ | $Y$ | $P(X,Y)$ |
|---|---|---|
| $x^0$ | $y^0$ | 0.08 |
| $x^0$ | $y^1$ | 0.32 |
| $x^1$ | $y^0$ | 0.12 |
| $x^1$ | $y^1$ | 0.48 |

q

| $X$ | $Y$ | $P(X,Y)$ |
|---|---|---|
| $x^0$ | $y^0$ | 0.4 |
| $x^0$ | $y^1$ | 0.3 |
| $x^1$ | $y^0$ | 0.2 |
| $x^1$ | $y^1$ | 0.1 |




* p: X and Y ind. ex. p(X=1) = 0.48 + 0.12 =0.6, p(Y=1) = 0.8, and p(X=1,Y=1) = 0.48
* q: X and Y are dependent

# INDEPENDENCE I-MAP

* I(G) independences implied by G (not yet defined)
* I(P) independences in the distribution P
* G I-map for P in I(G) ⊆ I(P)

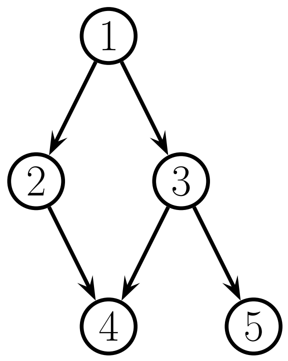
p

| $X$ | $Y$ | $P(X,Y)$ |
|---|---|---|
| $x^0$ | $y^0$ | 0.08 |
| $x^0$ | $y^1$ | 0.32 |
| $x^1$ | $y^0$ | 0.12 |
| $x^1$ | $y^1$ | 0.48 |

q

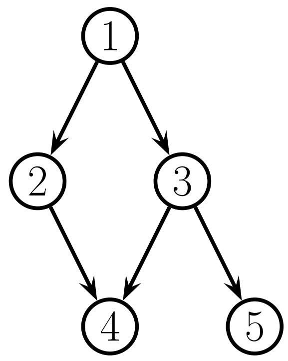
| $X$ | $Y$ | $P(X,Y)$ |
|---|---|---|
| $x^0$ | $y^0$ | 0.4 |
| $x^0$ | $y^1$ | 0.3 |
| $x^1$ | $y^0$ | 0.2 |
| $x^1$ | $y^1$ | 0.1 |


* All three graphs are I-maps for p
* $G_1$ and $G_2$ are I-maps for q, but $G_3$ is not



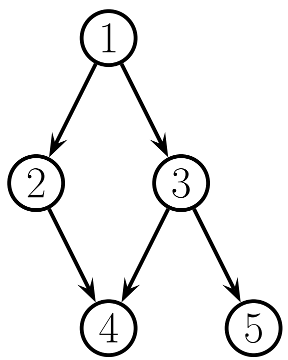

# TERMINOLOGY

- ★ Parent
- ★ Child
- ★ Family
- ★ Root
- ★ Leaf
- ★Neighbor



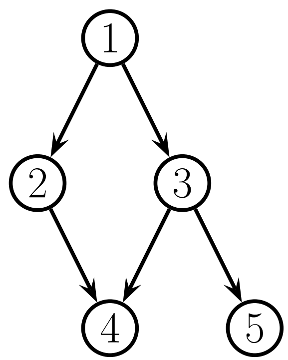

# TERMINOLOGY

- ★ Degree (in and out)
- ★ Cycle (directed or not)
- ★ Directed Acyclic Graph (DAG)
- ★ Topological order (parents < child)
- ★ Path (directed or not)
- ★ Ancestors




# TERMINOLOGY

- ★ Tree
- ★ Polytree – directed tree with multiple parents for some vertices
- ★ Forest
- ★ Subgraph
- ★ Clique
- ★ Maximal clique



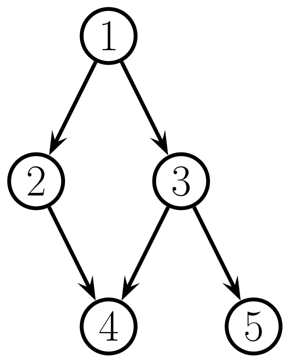

# ORDERED MARKOV PROPERTY

★ The directed local Markov property.

$$\boldsymbol{x}_t \perp \boldsymbol{x}_{V\setminus \mathrm{desc}(t)} | \boldsymbol{x}_{\mathrm{pa}(t)}$$

★ In this case

$$\begin{aligned} p(\boldsymbol{x}_{[5]}) &= p(\boldsymbol{x}_1)p(\boldsymbol{x}_2|\boldsymbol{x}_1)p(\boldsymbol{x}_3|\boldsymbol{x}_1,\boldsymbol{x}_2) \\ &\quad p(\boldsymbol{x}_4|\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)p(\boldsymbol{x}_5|\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4) \\ &= p(\boldsymbol{x}_1)p(\boldsymbol{x}_2|\boldsymbol{x}_1)p(\boldsymbol{x}_3|\boldsymbol{x}_1) \\ &\quad p(\boldsymbol{x}_4|\boldsymbol{x}_2,\boldsymbol{x}_3)p(\boldsymbol{x}_5|\boldsymbol{x}_3) \end{aligned}$$



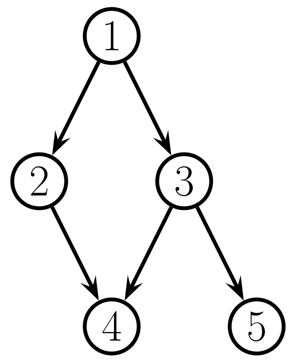

## ORDERED MARKOV PROPERTY

★ The directed local Markov property.

$$\boldsymbol{x}_t \perp \boldsymbol{x}_{V \setminus \mathrm{desc}(t)} | \boldsymbol{x}_{\mathrm{pa}(t)}$$

★In general, if 1,…,V topological order, the likelihood is decomposable (factorizes)

$$p(\boldsymbol{x}_{[V]}|G) = \prod_{t=1}^{V} p(\boldsymbol{x}_t|\boldsymbol{x}_{\mathrm{pa}(t)})$$




## ORDERED MARKOV PROPERTY

★ The ordered Markov property.

$$\boldsymbol{x}_t \perp \boldsymbol{x}_{\mathrm{pred}(t)} | \boldsymbol{x}_{\mathrm{pa}(t)}$$

has path to t, does not include t

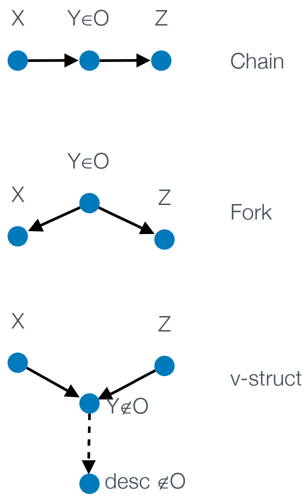
★In general, if 1,…,V topological order, the likelihood is decomposable (factorizes)

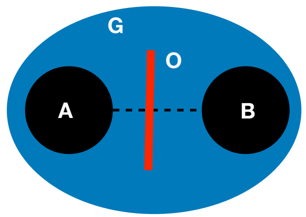
$$p(\boldsymbol{x}_{[V]}|G) = \prod_{t=1}^{V} p(\boldsymbol{x}_t|\boldsymbol{x}_{\mathrm{pa}(t)})$$

## D-SEPARATION

★ A path is d-separated by O if it has

- a chain X → Y → Z where Y ∊ O
- a fork X ← Y → Z where Y ∊ O
- a v-structure X → Y ← Z where (Y ∪ desc(Y)) ∩ O = ∅




## D-SEPARATION SETS AND CI OF DAGS

★ A is d-separated from B given O if every undirected path between A and B is d-separated by O

★ In a DAG G,

$$\boldsymbol{x}_A \perp_G \boldsymbol{x}_B | \boldsymbol{x}_O$$

$$\Updownarrow$$

A is d-separated from B given O

- Global (G): d-separation
- Local (L): $\boldsymbol{X}_t \perp \boldsymbol{X}_{V\setminus \text{desc}(t)} | \boldsymbol{X}_{\text{pa}(t)}$
- Ordered (O): $\boldsymbol{X}_t \perp \boldsymbol{X}_{\text{pred}(t)} | \boldsymbol{X}_{\text{pa}(t)}$

  where pred is according to a topological order
- Factorized (F): can be family-factorized
- Theorem: G ⇔ L ⇔ O ⇔ F

# EQUIVALENCE OF INDEPENDENCE DEFINITIONS

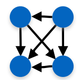
# SOUNDNESS AND COMPLETENESS

- Theorem

  If a distribution P factorizes according to G, then I(G) ⊆ I(P)
- Theorem

  If X and Y are not d-separated given Z in G, then X and Y are dependent given Z in some distribution P that factorize over G.

We cannot have all. Ex. clique and independent distribution



# SKELETON AND EQUIVALENCE

- The skeleton is the underlying undirected graph
- Immorality is a pair of unmarried parents
- Theorem

  Let $G_1$ and $G_2$ be two graphs over X. Then $G_1$ and $G_2$ have the same skeleton and the same set of immoralities if and only if $I(G_1)=I(G_2)$

# THE END