Stat 521A
Lecture 5

Outline

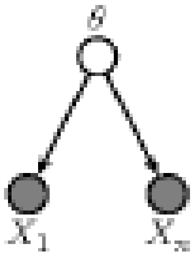
- Template models (6.3-6.5)
- Structural uncertainty (6.6)
- Multivariate Gaussians (7.1)
- Gaussian DAGs (7.2)
- Gaussian MRFs (7.3)

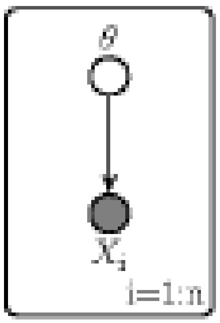
Parameter tying

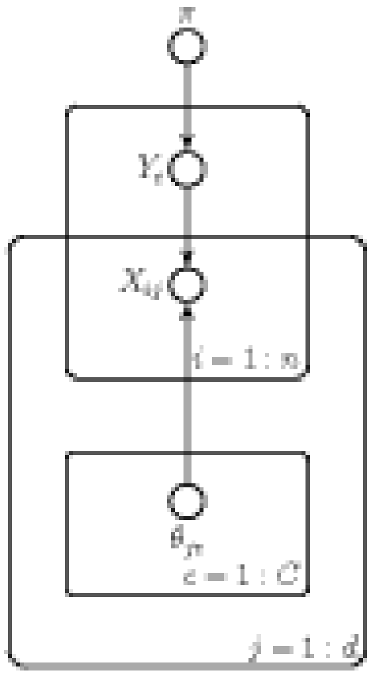
- A DBN defines a distribution over an unboundedly large number of variables by assuming that they all share the same CPDs.
- This is called parameter tying (weight sharing).
- It is useful even for fixed sized models in order to help learning (pool the sufficient statistics).
- We now discuss notational conventions ("syntactic sugar") for representing large "unrolled" networks with shared parameters.

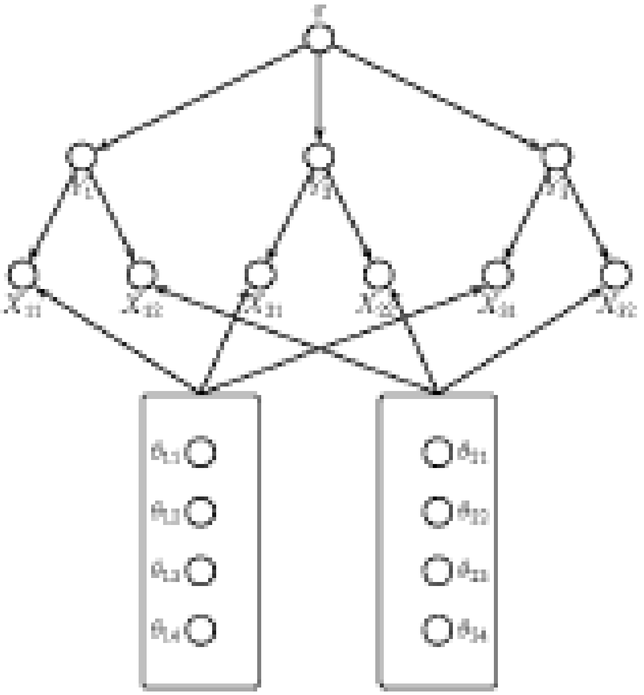
Plates

- Plates are useful for specifying simple repetitive patterns, as frequently arise in hierarchical Bayesian models

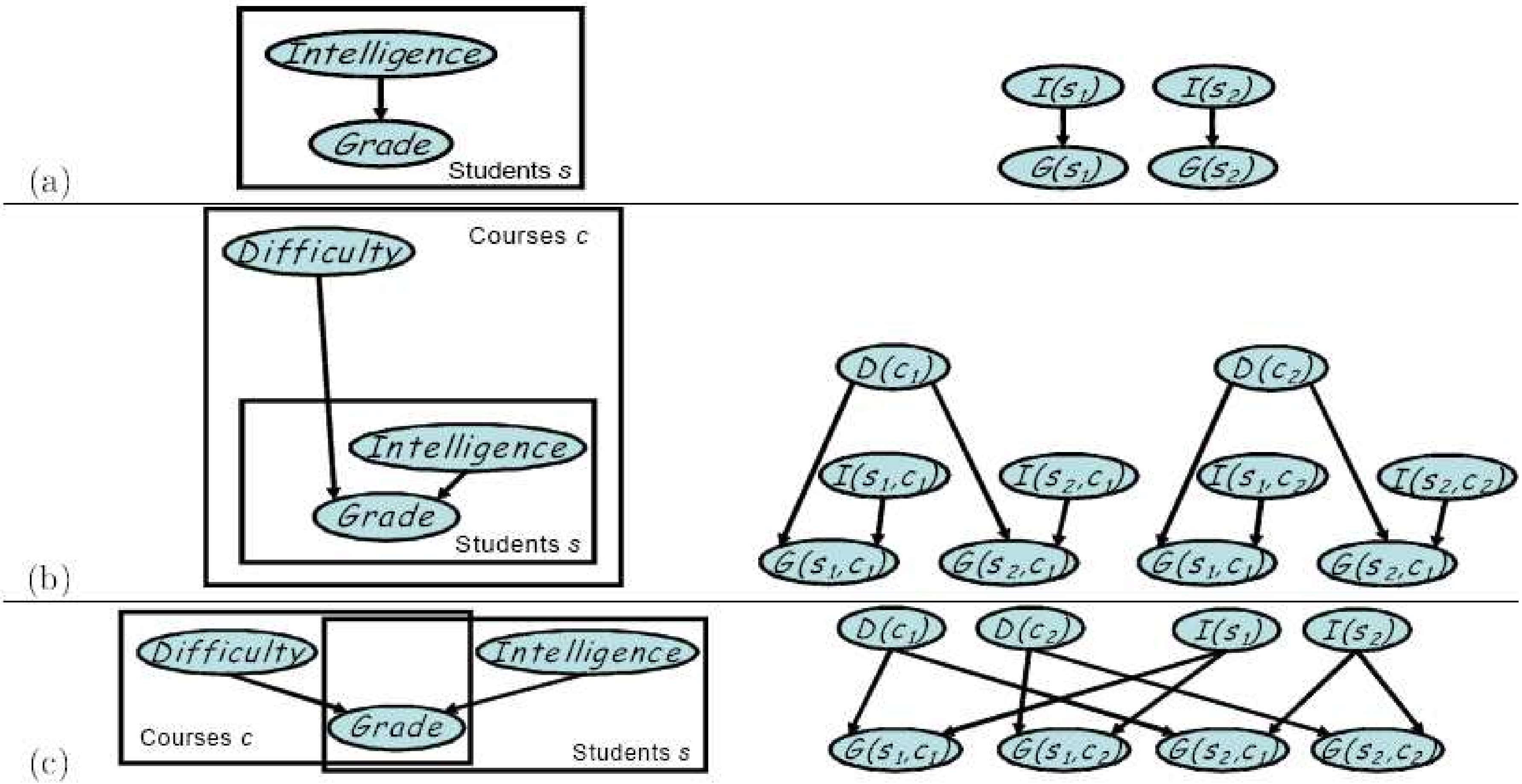






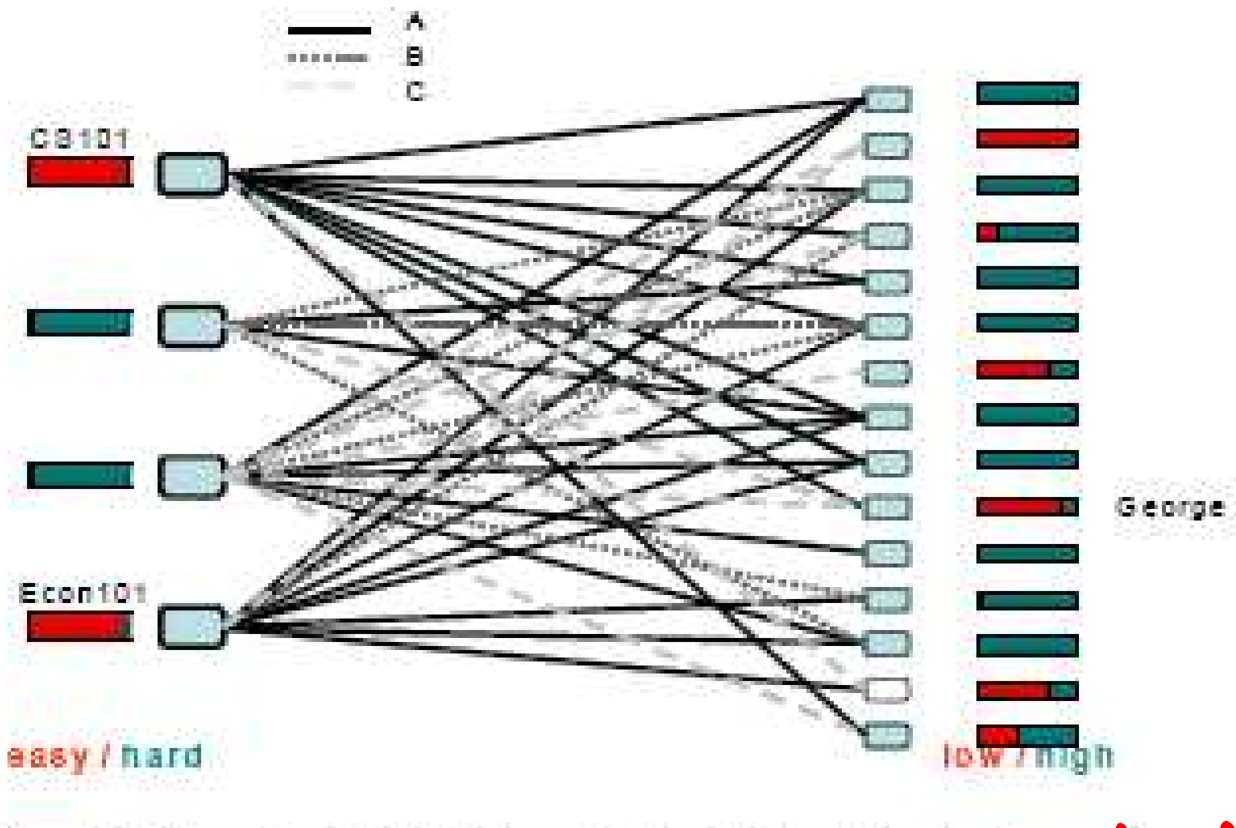


Plates



(a)
Intelligence
Grade
Students s
I(s1)
I(s2)
G(s1)
G(s2)
(b)
Difficulty
Courses c
Intelligence
Grade
Students s
D(c1)
D(c2)
I(s1,c1)
I(s2,c1)
I(s1,c2)
I(s2,c2)
G(s1,c1)
G(s2,c1)
G(s1,c1)
G(s2,c1)
(c)
Difficulty
Grade
Intelligence
Courses c
Students s
D(c1)
D(c2)
I(s1)
I(s2)
G(s1,c1)
G(s1,c2)
G(s2,c1)
G(s2,c2)

Unrolled network



Grade(s,c) in {A,B,C} is encoded on edges.
Cf discrete probabilistic matrix factorization

$$\begin{array}{cc} & D_1 \cdots D_4 \\ \begin{array}{c} I_1 \\ \vdots \\ I_{14} \end{array} & \begin{pmatrix} \\ G \\ \\ \end{pmatrix} \end{array}$$

Limitations of plates

- There are various structures that plates cannot represent
- Eg DBNs
- Eg genotype(x1) depends on genotype(x2), where x2=parent(x1)
- We can write programs to generate graphs of specified structure, but we would like a declarative representation language for such repetitive patterns so that no new code has to be written

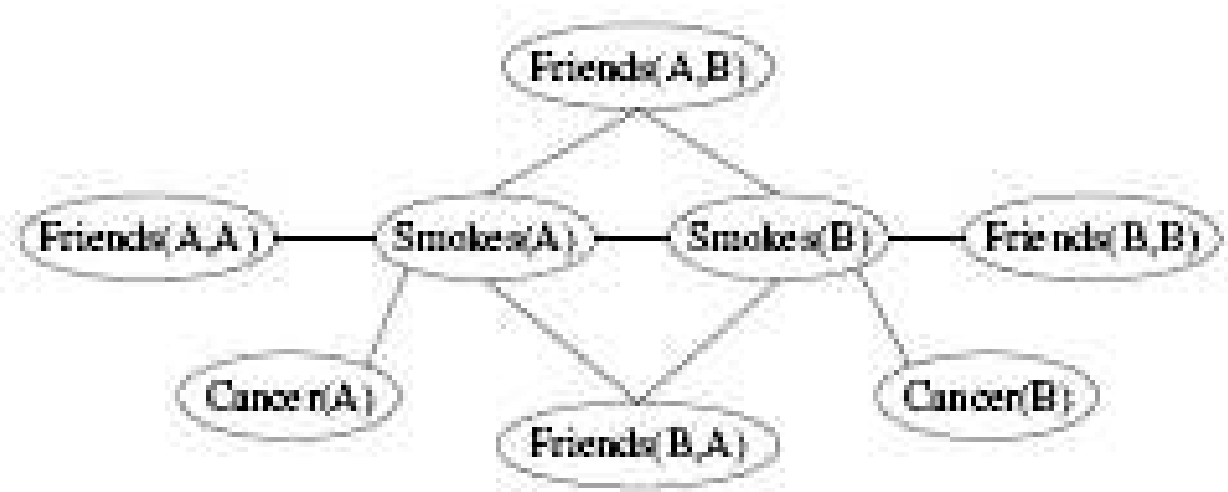
Beyond plates

- Probabilistic Relational Models (PRMs) encode large DAG models with tied CPDs
- Relational Markov Networks encode large MRFs with tied factors
- Markov Logic Networks are like RMNs, except the factors are represented in log-linear form, and the features are represented as logical expressions

Markov Logic Networks

Table I. Example of a first-order knowledge base and MLN. Fr() is short for Friends(), Sm() for Smokes(), and Ca() for Cancer().

English	First-Order Logic	Clausal Form	Weight
Friends of friends are friends.	$\forall x \forall y \forall z\ \mathtt{Fr}(x,y) \wedge \mathtt{Fr}(y,z) \Rightarrow \mathtt{Fr}(x,z)$	$\neg\mathtt{Fr}(x,y) \vee \neg\mathtt{Fr}(y,z) \vee \mathtt{Fr}(x,z)$	0.7
Friendless people smoke.	$\forall x\ (\neg(\exists y\ \mathtt{Fr}(x,y)) \Rightarrow \mathtt{Sm}(x))$	$\mathtt{Fr}(x, g(x)) \vee \mathtt{Sm}(x)$	2.3
Smoking causes cancer.	$\forall x\ \mathtt{Sm}(x) \Rightarrow \mathtt{Ca}(x)$	$\neg\mathtt{Sm}(x) \vee \mathtt{Ca}(x)$	1.5
If two people are friends, either both smoke or neither does.	$\forall x \forall y\ \mathtt{Fr}(x,y) \Rightarrow (\mathtt{Sm}(x) \Leftrightarrow \mathtt{Sm}(y))$	$\neg\mathtt{Fr}(x,y) \vee \mathtt{Sm}(x) \vee \neg\mathtt{Sm}(y)$,	1.1
		$\neg\mathtt{Fr}(x,y) \vee \neg\mathtt{Sm}(x) \vee \mathtt{Sm}(y)$	1.1



Directed vs undirected models

- Undirected models are simpler: no need to worry about cycles, lots of freedom in defining factors
- However, in a UG, the probability of a node depends on the *size* of the graph and/or its connectivity, even if all the other nodes are hidden.
- This may not be desirable.

$X1 \to X2 \to X3$

$X1 - X2 - X3$

$X1 \to X2 \to \cdots \to X10$

$X1 - X2 - \cdots - X10$

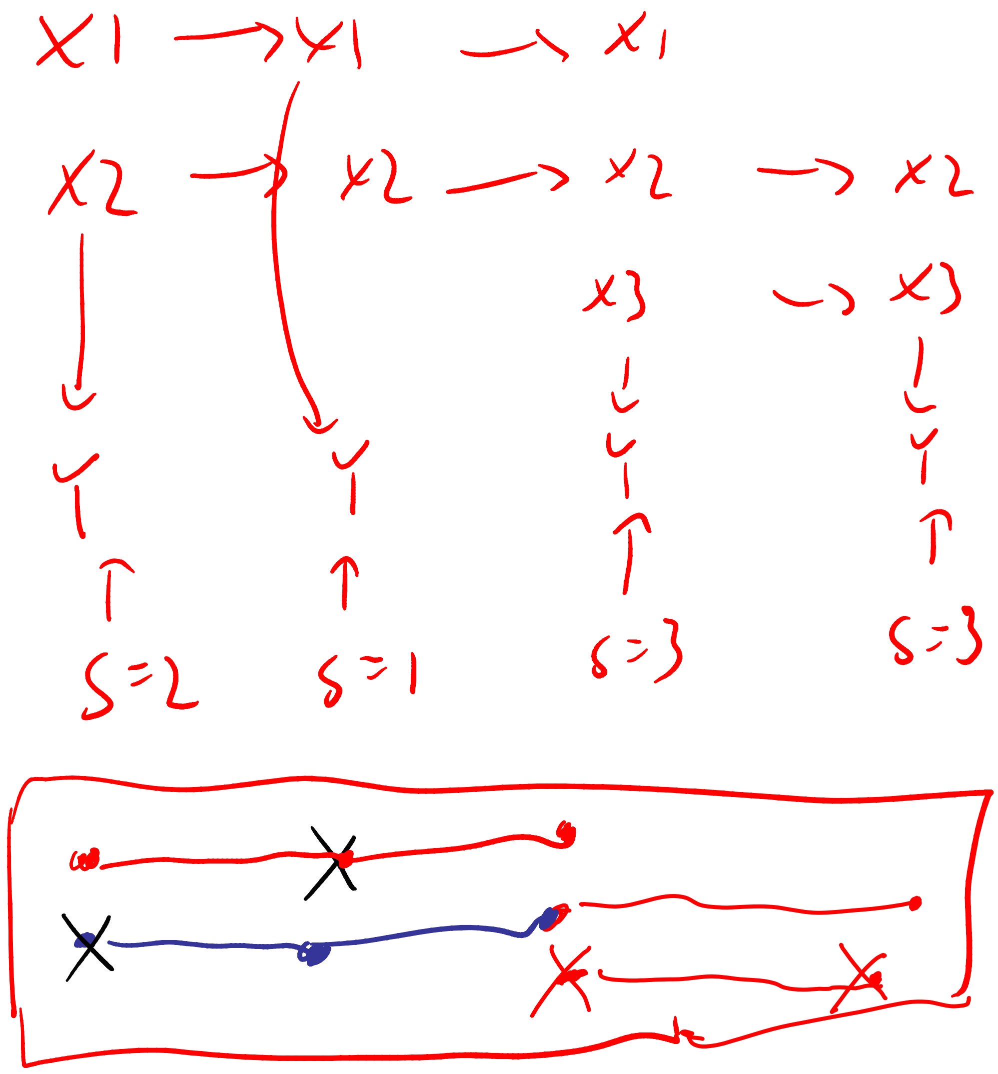
$p(X_2)$ same

$p(X2)$ different

Structural uncertainty

- For a fixed domain, if we do not know the graph structure, we may estimate it using model selection.
- But for relational domains, the structure may change depending on the values of the nodes
- Eg. Genotype(x1) -> genotype(x2) is only active if parent(x1,x2)=true
- In addition, we may be uncertain about how many objects exist in the world
- Eg. In tracking, 3 blips on the radar is consistent with {0,1,…, infty} objects in the world!

Data association ambiguity



Citation matching

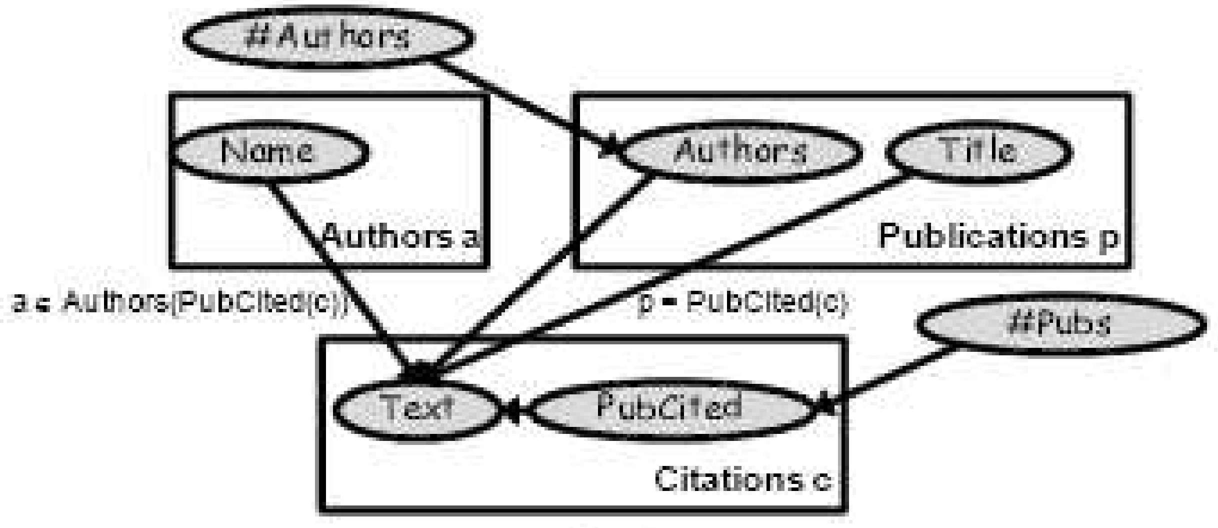
Are these the same article?
Huge industry concerned with database merging

Elston R, Stewart A. A General Model for the Genetic Analysis of Pedigree Data. Hum. Hered. 1971;21:523-542.

Elston RC, Stewart J (1971): A general model for the analysis of pedigree data. Hum Hered 21523-542.

DAG model

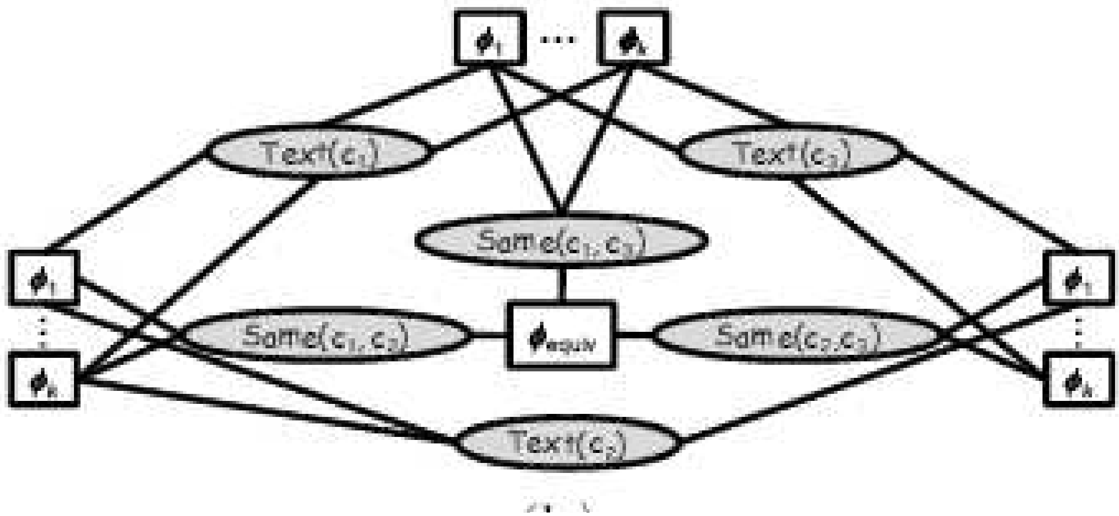
- Assumes there is an unknown number of authors and papers, which generates the observed set of citation strings.



(a)

UG model

- No unknown objects. Just enforce that citations are the same.
- Need 3 way factor to encode transitivity of sameness relation: S(c1,c2), and S(c2,c3) => S(c1,c3)
- And if 2 docs are same, text should be similar: Factor(s(c1,c2), T(c1), T(c2))



MVN: 2 parameterizations

- Moment form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \stackrel{\text{def}}{=} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

- Information (canonical) form

$$\boldsymbol{\Lambda} \stackrel{\text{def}}{=} \boldsymbol{\Sigma}^{-1}$$

precision (information) matrix

$$\begin{aligned}
\boldsymbol{\eta} &\stackrel{\text{def}}{=} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\
\mathcal{N}(\mathbf{x}|\boldsymbol{\eta}, \boldsymbol{\Lambda}) &= \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{d/2}} \exp[-\tfrac{1}{2}(\mathbf{x}^T\boldsymbol{\Lambda}\mathbf{x} + \boldsymbol{\eta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\eta} - 2\mathbf{x}^T\boldsymbol{\eta})] \\
&= \exp[c - \tfrac{1}{2}\mathbf{x}^T\boldsymbol{\Lambda}\mathbf{x} + \mathbf{x}^T\boldsymbol{\eta}]
\end{aligned}$$

Moment and anonical form

- Canonical form is denoted

$$\mathbf{x} \sim \mathcal{N}_C(\mathbf{b}, \mathbf{Q}) \iff p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{b}^T\mathbf{x}\right)$$

- Moment form

$$\mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1}) = \mathcal{N}_C(\mathbf{Q}\boldsymbol{\mu}, \mathbf{Q})$$

Independencies in MVN

- Thm 7.1.3. Let X ~ MVN. $X_i \perp X_j$ iff $\Sigma_{i,j}=0$
- Thm 7.1.4. let X ~ MVN with info matrix J. Then $J_{i,j}=0$ iff $X_i \perp X_j \mid X_{-ij}$
- Factorization thm.

$$\mathbf{x} \perp \mathbf{y} | \mathbf{z} \iff p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f(\mathbf{x}, vz) g(\mathbf{y}, vz)$$

Indep => uncorrelated

- Ex 7.2.1. For any p(X,Y), if X $\perp$ Y then Cov[X,Y]=0.

$$
\begin{aligned}
\text{Cov}[x,y] &= \int\int p(x,y)(x-\overline{x})(y-\overline{y})dxdy \\
&= (\int p(x)(x-\overline{x})dx)(\int p(y)(y-\overline{y})dy) \\
&= (\overline{x}-\overline{x})(\overline{y}-\overline{y}) = 0
\end{aligned}
$$

Uncorrelated & MVN => indep

- Ex 7.2.2. If p(X,Y) is Gaussian, and Cov[X,Y]=0, then X $\perp$ Y.
- Pf. The bivariate Gaussian can be written as

$$\begin{aligned} p(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp[-\frac{1}{2(1-\rho^2)}(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \\ &\quad -2\rho\frac{(x_1-\mu_1)}{\sigma_1}\frac{(x_2-\mu_2)}{\sigma_2})] \end{aligned}$$

- If \rho=0, then

$$\begin{aligned} p(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp[-\frac{1}{2}(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2})] \\ &= f(x_1)g(x_2) \end{aligned}$$

- Hence by factorization thm, x1 \perp x2.

Uncorrelated not imply independent

- Ex 7.2.3. Find an example where Cov[X,Y]=0 yet not $X \perp Y$.
- Let X ~ U(-1,1) and Y=X^2. Clearly Y is dependent on X yet one can show (exercise) that Cov(X,Y)=0.
- Let X ~ N(0,1) and Y= W X, p(W=-1)=p(W=1)=0.5. Clearly Y is dependent on X, yet one can show (exercise) that Y ~ N(0,1) and Cov[X,Y]=0.

Independencies in MVN

- Thm 7.1.3. Let X ~ MVN. $X_i \perp X_j$ iff $\Sigma_{i,j}=0$
- Pf. By ex 7.2.1, we have => direction.
- By ex 7.2.2, we have that <= direction.
- By ex 7.2.3, we have that X ~ MVN is necessary for <= direction to work.

Conditional Independencies in MVN

- Thm 7.1.4. let X ~ MVN with info matrix J. Then $J_{i,j}=0$ iff $X_i \perp X_j \mid X_{-ij}$
- Pf. Let mu=0.

$$
\begin{aligned}
p(x_i, x_j, \mathbf{x}_{-ij}) &\propto \exp(-\tfrac{1}{2}\sum_{k,l} x_k Q_{kl} x_l) \\
&\propto \exp\left(-\tfrac{1}{2} x_i x_j (Q_{ij} + Q_{ji}) - \tfrac{1}{2} \sum_{\{k,l\} \neq \{i,j\}} x_k Q_{kl} x_l\right)
\end{aligned}
$$

- The second term does not involve x_i x_j, and nor does the first iff $Q_{ij}=0$. Hence this factorizes into $f(x_i,x_{-ij})$ $g(x_j,x_{-ij})$ iff $Q_{ij}=0$. QED.

Structural zeros

- Zeros in the precision matrix correspond to missing edges in the UGM

$$\mathbf{\Sigma} = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 5 & -5 \\ -2 & -5 & 8 \end{pmatrix}, \quad \mathbf{\Lambda} = \mathbf{\Sigma}^{-1} = \begin{pmatrix} 0.3125 & -0.125 & 0 \\ -0.125 & 0.5833 & 0.3333 \\ 0 & 0.3333 & 0.3333 \end{pmatrix}$$

$X_1 - X_2 - X_3$

Marginals and conditionals

	Marginal $p(\mathbf{x}_2)$
Moment	$\mathcal{N}(\mathbf{x}_2 \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$
Info	$\mathcal{N}(\mathbf{x}_2 \mid \boldsymbol{\eta}_2 - \boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\eta}_1, \boldsymbol{\Lambda}_{22} - \boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12})$

	Conditional $p(\mathbf{x}_2 \mid \mathbf{x}_1)$
Moment	$\mathcal{N}(\mathbf{x}_1 \mid \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$
Info	$\mathcal{N}(\mathbf{x}_2 \mid \boldsymbol{\eta}_1 - \boldsymbol{\Lambda}_{12}\mathbf{x}_2, \boldsymbol{\Lambda}_{11})$

Marginalization easy in moment form.
Conditioning easy in canonical form.

Conditioning in canonical form

- Thm (Conditioning).

$$\mathbf{x} \sim \mathcal{N}_C(\mathbf{b}, \mathbf{Q}) \Rightarrow \qquad \mathbf{x}_A | \mathbf{x}_B \sim \mathcal{N}_C(\mathbf{b}_A - \mathbf{Q}_{AB}\mathbf{x}_B, \mathbf{Q}_{AA})$$

- Thm (soft conditioning) .

$$\mathbf{x} \sim \mathcal{N}_C(\mathbf{b}, \mathbf{Q}) \quad \text{and} \quad \mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{x}, \mathbf{P}^{-1})$$

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}_C(\mathbf{b} + \mathbf{P}\mathbf{y}, \mathbf{Q} + \mathbf{P})$$

Precisions add

- We can accumulate evidence by addition of matrix-vector products, and then compute posterior mean at end by solving Qb = mu.

Partial correlation coefficient

- Let X ~ Mvn with precision matrix

$$\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = \begin{pmatrix} \omega_{11} & \cdots & \omega_{1d} \\ \vdots & \ddots & \ddots \\ \omega_{d1} & \cdots & \omega_{dd} \end{pmatrix}$$

- The conditional distribution p(x1,x2|x3,…,xd) is bivariate Gaussian with covariance

$$\begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix}^{-1} = \frac{1}{\omega_{11}\omega_{22} - (\omega_{12})^2} \begin{pmatrix} \omega_{22} & -\omega_{12} \\ -\omega_{21} & \omega_{11} \end{pmatrix}$$

- The partial correlation coefficient is given by

$$\rho_{1,2|3,\ldots,d} \stackrel{\text{def}}{=} \frac{Cov[X_1, X_2|X_{3:d}]}{\sqrt{\text{Var}\,[X_1|X_{3:d}]\text{Var}\,[X_2|X_{3:d}]}} = \frac{-\omega_{21}}{\sqrt{\omega_{11}\omega_{22}}}$$

Conditioning in moment form

- Thm (Rue&Held p26).

$$\begin{aligned}
\mathbf{x} &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1}) \Rightarrow \\
\mathbf{x}_A|\mathbf{x}_B &\sim \mathcal{N}(\boldsymbol{\mu}_{A|B}, \mathbf{Q}_{AA}^{-1}) \\
\boldsymbol{\mu}_{A|B} &= \boldsymbol{\mu}_A - \mathbf{Q}_{AA}^{-1}\mathbf{Q}_{AB}(\mathbf{x}_B - \boldsymbol{\mu}_B)
\end{aligned}$$

- Thus to find the mean we need to solve the linear system

$$\mathbf{Q}_{AA}\boldsymbol{\mu}_{A|B} = \mathbf{Q}_{AA}\boldsymbol{\mu}_A - \mathbf{Q}_{AB}\mathbf{x}_B + \mathbf{Q}_{AB}\boldsymbol{\mu}_B$$

- Eg if A={i} we have

$$\begin{aligned}
E[x_i|\mathbf{x}_{-i}] &= \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\neq i} Q_{ij}(x_j - \mu_j) \\
\text{prec}(x_i|\mathbf{x}_{-i}) &= Q_{ii}
\end{aligned}$$

Proof

- Assume \mu=0 for simplicity. Then

$$
\begin{aligned}
p(\mathbf{x}_A|\mathbf{x}_B) &\propto \exp\left(-\frac{1}{2}\begin{pmatrix}\mathbf{x}_A & \mathbf{x}_B\end{pmatrix}\begin{pmatrix}\mathbf{Q}_{AA} & \mathbf{Q}_{AB} \\ \mathbf{Q}_{BA} & \mathbf{Q}_{BB}\end{pmatrix}\begin{pmatrix}\mathbf{x}_A \\ \mathbf{x}_B\end{pmatrix}\right) \\
&\propto \exp\left(-\frac{1}{2}\mathbf{x}_A^T\mathbf{Q}_{AA}\mathbf{x}_A - (\mathbf{Q}_{AB}\mathbf{x}_B)^T\mathbf{x}_A\right)
\end{aligned}
$$

- Compare this to a Gaussian with precision K and mean m

$$
p(\mathbf{z}) \propto \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{K}\mathbf{z} + (\mathbf{K}\mathbf{m})^T\mathbf{z}\right)
$$

- We see that Q_{AA} is the conditional precision and the conditional mean is given by

$$
\mathbf{Q}_{AA}\boldsymbol{\mu}_{A|B} = -\mathbf{Q}_{AB}\mathbf{x}_B
$$

QED

Soft conditioning in moment form

$$
\begin{aligned}
\mathbf{x} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\mathbf{y}|\mathbf{x} &\sim \mathcal{N}(\mathbf{x}, \mathbf{S}) \\
\mathbf{x}|\mathbf{y} &\sim \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \\
\boldsymbol{\Sigma}_{x|y}^{-1} &= \boldsymbol{\Sigma}^{-1} + \mathbf{S}^{-1} \\
\boldsymbol{\Sigma}_{x|y}^{-1}\boldsymbol{\mu}_{x|y} &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{S}^{-1}\mathbf{y}
\end{aligned}
$$

Bayes rule for linear Gaussian systems

Linear Gaussian DGMs

- A CPD is linear Gaussian if

$$p(x_i|x_{\pi_i}) = \mathcal{N}(x_i|\sum_{j\in\pi_i} w_{ij}x_j + b_i, v_i)$$

- A DGM is linear Gaussian if all CPDs are LG.
- Such networks define a joint Gaussian. Each node is given by

$$x_i = \sum_{j\in\pi_i} w_{ij}x_j + b_i + \sqrt{v_i}\epsilon_i$$

where $\varepsilon_i \sim N(0,1)$ and $E[\varepsilon_i\, \varepsilon_j] = I_{i,j}$.

- W is lower triangular matrix: w_{i,j} = weights into i from j.

LG DGM to MVN

- We can compute the global mean and covariance recursively, in topological order

$$x_i = \sum_{j \in \pi_i} w_{ij} x_j + b_i + \sqrt{v_i} \epsilon_i$$

$$E[x_i] = \sum_{j \in \pi_i} w_{ij} E[x_j] + b_i$$

$$\begin{aligned}
\text{Cov}[x_i, x_j] &= E[(x_i - E[x_i])(x_j - E[x_j])] \\
&= E\left[(x_i - E[x_i]) \left\{ \sum_{k \in \pi_j} w_{jk}(x_k - E[x_k]) + \sqrt{v_j} \epsilon_j \right\} \right] \\
&= \sum_{k \in \pi_j} w_{jk} \text{Cov}[x_i, x_k] + I_{i,j} v_j
\end{aligned}$$

Bishop p371

LG DGM to MVN

- Consider a chain x1 -> x2 -> x3

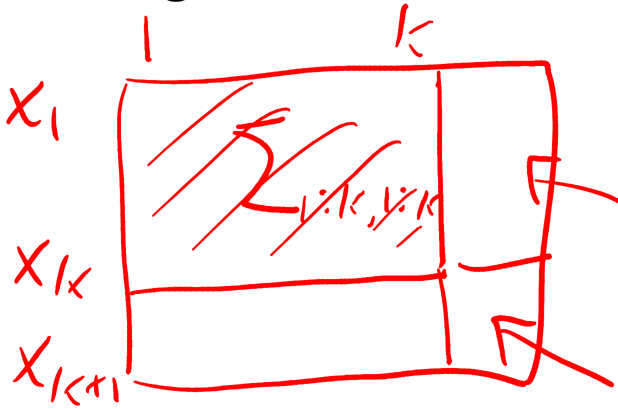
$$
W: \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & \times & \times & \times \\ 2 & w_{21} & \times & \times \\ 3 & 0 & w_{32} & \times \end{array}
$$

$$
\boldsymbol{\mu} = (b_1, b_2 + w_{21} b_1, b_3 + w_{32} b_2 + w_{32} w_{21} b_1)
$$

$$
\boldsymbol{\Sigma} = \begin{pmatrix} v_1 & w_{21} v_1 & w_{32} w_{31} v_1 \\ w_{21} v_1 & v_2 + w_{21}^2 v_1 & w_{32}(v_2 + w_{21}^2 v_1) \\ w_{32} w_{21} v_1 & w_{32}(v_2 + w_{21}^2 v_1) & v_3 + w_{32}^2 (v_2 + w_{21}^w v_1) \end{pmatrix}
$$

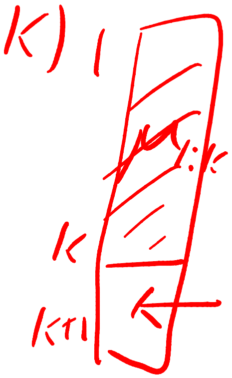
- In general, when adding node (k+1)

$\beta \triangleq W(k+1, 1:k)$



$\Sigma \beta$

$v_{k+1} + \beta^T \Sigma \beta$



$b_k + \beta^T \mu_{1:k}$

K&F Thm 7.2.2

Alternative parameterization

- The results are much “prettier” if we write

$$X_j = \mu_j + \sum_{k \in \pi_j} w_{jk}(X_k - \mu_k) + \sqrt{v_j} Z_j$$

where the offset is given by

$$w_j^{(0)} = \mu_j - \sum_{k \in \pi_j} w_{jk} \mu_k$$

- Then we have

$$\begin{aligned} (\mathbf{x} - \boldsymbol{\mu}) &= \mathbf{W}(\mathbf{x} - \boldsymbol{\mu}) + \mathbf{S}^T \mathbf{z} = \mathbf{W}(\mathbf{x} - \boldsymbol{\mu}) + \mathbf{e} \\ \mathbf{e} &= \mathbf{S}^T \mathbf{z} = (\mathbf{I} - \mathbf{W})(\mathbf{x} - \boldsymbol{\mu}) \end{aligned}$$

$$\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_d \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ -w_{21} & 1 & & & \\ -w_{32} & -w_{31} & 1 & & \\ \vdots & & & \ddots & \\ -w_{d1} & -w_{d2} & \dots & -w_{d,d-1} & 1 \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_d - \mu_d \end{pmatrix}$$

DAG weights = Cholesky Decomposition

$$
\begin{aligned}
\mathbf{x} - \boldsymbol{\mu} &= (\mathbf{I} - \mathbf{W})^{-1}\mathbf{e} \stackrel{\text{def}}{=} \mathbf{U}\mathbf{e} = \mathbf{U}\mathbf{S}^T\mathbf{z} \stackrel{\text{def}}{=} \mathbf{A}^T\mathbf{z} \\
\boldsymbol{\Sigma} &= \text{Var}\left[\mathbf{x}\right] = \text{Var}\left[\mathbf{x} - \boldsymbol{\mu}\right] \\
&= \text{Var}\left[\mathbf{A}^T\mathbf{z}\right] = \mathbf{A}^T\text{Var}\left[\mathbf{z}\right]\mathbf{A} = \mathbf{A}^T\mathbf{A} \\
&= \mathbf{U}\mathbf{S}^T\mathbf{S}\mathbf{U}^T = \mathbf{U}\mathbf{D}\mathbf{U}^T \\
\boldsymbol{\Sigma}^{-1} &= \mathbf{U}^{-T}\mathbf{D}^{-1}\mathbf{U}^{-1} = (\mathbf{I} - \mathbf{W})^T\mathbf{D}^{-1}(\mathbf{I} - \mathbf{W}) \stackrel{\text{def}}{=} \mathbf{T}^T\mathbf{D}^{-1}\mathbf{T}
\end{aligned}
$$

$$
\mathbf{T} = \begin{pmatrix}
1 & & & & \\
-w_{21} & 1 & & & \\
-w_{32} & -w_{31} & 1 & & \\
\vdots & & & \ddots & \\
-w_{d1} & -w_{d2} & \dots & -w_{d,d-1} & 1
\end{pmatrix}
$$

Chains

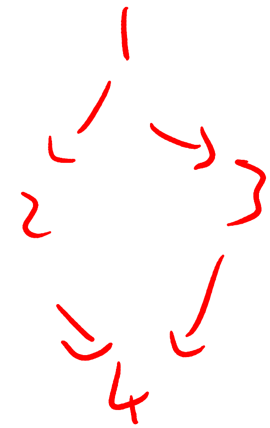
- Consider a chain X1 -> X2 -> … -> X5
- The DAG and UG are both sparse (same CI)

```
n = 5;
w=randn(n,1);
W = spdiags([w zeros(n,1) zeros(n,1)], -1:1, n, n);
T = eye(n)-W;
D = diag(ones(n,1));
K = T'*D*T;

>> full(W)
ans =
         0         0         0         0         0
    1.1909         0         0         0         0
         0    1.1892         0         0         0
         0         0   -0.0376         0         0
         0         0         0    0.3273         0
>> K
K =
    2.4183   -1.1909         0         0         0
   -1.1909    2.4141   -1.1892         0         0
         0   -1.1892    1.0014    0.0376         0
         0         0    0.0376    1.1071   -0.3273
         0         0         0   -0.3273    1.0000
```

Diamond

- DAG is sparse, Sigma and SigmaInv are dense

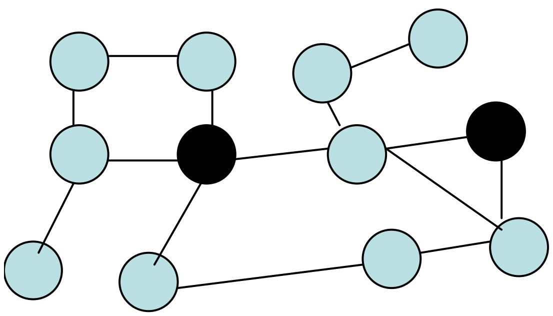



```
W =
         0         0         0         0
    0.5488         0         0         0
    0.7152         0         0         0
         0    0.6028    0.5449         0
>> K
K =
    1.8127   -0.5488   -0.7152         0
   -0.5488    1.3633    0.3284   -0.6028
   -0.7152    0.3284    1.2969   -0.5449
         0   -0.6028   -0.5449    1.0000
>> inv(K)
ans =
    1.0000    0.5488    0.7152    0.7205
    0.5488    1.3012    0.3925    0.9982
    0.7152    0.3925    1.5115    1.0602
    0.7205    0.9982    1.0602    2.1793
```

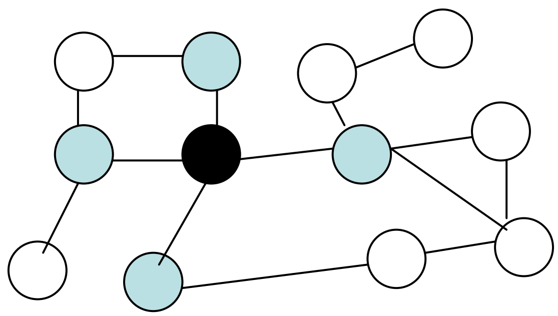
Gaussian MRFs

- Defn. A GMRF is a Gaussian of the form $N(\mu, Q^{-1})$ where $Q_{ij} \neq 0$ iff $G_{ij} \neq 0$ (Q=precision matrix)
- Thm. For a GMRF, the following properties are equivalent.
- Pairwise Markov: $x_i \perp x_j | \mathbf{x}_{-ij}$ if $G_{i,j} = 0$ and $i \neq j$
- Local Markov: $x_i \perp \mathbf{x}_{-i,ne(i)} | \mathbf{x}_{ne(i)}$
- Global Markov: $x_A \perp x_B | x_C$

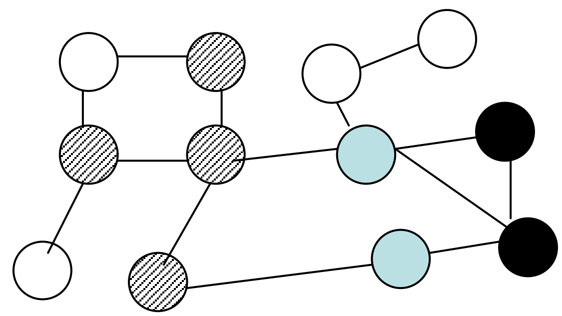
Rue&Held p25



Blacks indep given gray



Black indep of white given gray



Black indep striped given gray

MVN to Gaussian UGM

- We can convert any MVN into a UGM with pairwise potentials which are quadratics

$$
\begin{aligned}
\mathbf{J} &\stackrel{\text{def}}{=} \boldsymbol{\Sigma}^{-1} \\
\mathbf{h} &\stackrel{\text{def}}{=} \mathbf{J}\boldsymbol{\mu} \\
\mathcal{N}(\mathbf{x}|\mathbf{h},\mathbf{J}) &= \exp[c - \tfrac{1}{2}\mathbf{x}^T\mathbf{J}\mathbf{x} + \mathbf{x}^T\mathbf{h}] \\
\log p(\mathbf{x}) &= c - \tfrac{1}{2}\sum_i [J_{i,i}x_i^2 + h_i x_i] - \tfrac{1}{2}\sum_i \sum_j J_{i,j} x_i x_j \\
&= c + \sum_i \phi_i(x_i) + \sum_i \sum_{j>i} \phi_{i,j}(x_i, x_j) \\
\phi_i(x_i) &= -\tfrac{1}{2}[J_{i,i}x_i^2 + h_i x_i] \\
\phi_{i,j}(x_i, x_j) &= -J_{i,j} x_i x_j
\end{aligned}
$$

Pairwise UGM to MVN

- Consider a UGM in which the node and edge potentials are quadratics

$$
\begin{aligned}
\epsilon_i(x_i) &= d_0^i + d_1^i x_1 + d_2^i x_i^2 \\
\epsilon_{ij}(x_i, x_j) &= a_{00}^{i,j} + a_{01}^{i,j} x_i + a_{10}^{ij} x_j + a_{11}^{ij} x_i x_j + a_{02}^{ij} x_i^2 + a_{20}^{ij} x_j^2
\end{aligned}
$$

- We can always rewrite the corresponding unnormalized distribution as

$$
p'(\mathbf{x}) = \exp[-\tfrac{1}{2}\mathbf{x}^T \mathbf{J} \mathbf{x} + \mathbf{x}^T \mathbf{h}]
$$

- But the normalization constant Z will only be finite if J is positive definite.

Sufficient conditions on info matrix

- Def 7.3.1. A matrix J is attractive if, for all i \neq j, we have that all partial correlations are non-neg

$$-\frac{J_{i,j}}{\sqrt{J_{i,i}J_{j,j}}} \geq 0$$

- Thm 7.3.2. If J is attractive, then p is a valid MVN.
- Def 7.3.1b. A matrix J is diagonally dominant if, for all rows i, $J_{ii} > \sum_{j \neq i} |J_{i,j}|$
- Thm 7.3.2b. If J is diagonally dominant, then p is a valid MVN.

Pairwise normalizable

- Def 7.3.3. A pairwise MRF with energies of the form

$$\begin{aligned} \epsilon_i(x_i) &= d_0^i + d_1^i x_1 + d_2^i x_i^2 \\ \epsilon_{ij}(x_i, x_j) &= a_{00}^{i,j} + a_{01}^{i,j} x_i + a_{10}^{ij} x_j + a_{11}^{ij} x_i x_j + a_{02}^{ij} x_i^2 + a_{20}^{ij} x_j^2 \end{aligned}$$

is called pairwise normalizable if

$d_2^i > 0, \forall i$ and $\begin{pmatrix} a_{02}^{ij} & a_{11}^{ij}/2 \\ a_{11}^{ij}/2 & a_{20}^{ij} \end{pmatrix}$ is psd for all i,j

- Thm 7.3.4. If the MRF is pairwise normalizable, then it defines a valid Gaussian.
- Sufficient but not necessary eg.

$$\begin{pmatrix} 1 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 1 \end{pmatrix}$$

May be able to reparameterize the node/ edge potentials to ensure pairwise normalized.

Conditional autoregressions (CAR)

- We can parameterize a GMRF in terms of its full conditionals

$$
\begin{aligned}
E[x_i|\mathbf{x}_{-i}] &= \mu_i - \sum_{j:j\sim i} \beta_{ij}(x_j - \mu_j) \\
\text{prec}[x_i|\mathbf{x}_{-i}] &= \kappa_i > 0
\end{aligned}
$$

- From before, we have

$$
\begin{aligned}
E[x_i|\mathbf{x}_{-i}] &= \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\neq i} Q_{ij}(x_j - \mu_j) \\
\text{prec}(x_i|\mathbf{x}_{-i}) &= Q_{ii}
\end{aligned}
$$

- To be a valid MVN we must set

$$
\begin{aligned}
\kappa_i &= Q_{ii}, \beta_{ij} = \frac{Q_{ij}}{\kappa_i}, \kappa_i\beta_{ij} = \kappa_j\beta_{ji} \\
\mathbf{Q} &= \text{diag}(\boldsymbol{\kappa})(\mathbf{I} + \boldsymbol{\beta})
\end{aligned}
$$

Rue&Held p29