ompact representations of distributions

irected Acyclic Graphical models

Bayes-ball algorithm

✓

A6:- typo in Fenchel dual:

$$D(y) = -f^*(-y^*) - g^*(A^T y)$$

- extra hints now included

Midterm (a modest proposal) change?

- Cancel class Monday? (you can go to Bertsekas talk)
- midterm November 12, 17, 19?
- A6 due Monday?
- tutorials moved?

3: code red

2: major inconvenience

1: minor inconvenience

Compact representations of joint distributions

Notation for today:

- we'll use x_j for $(x_i)_j$

Consider $x \in \{0,1\}^d$ (binary vectors)

We want to model $p(x)$ (joint distribution)

- why? scientific discovery, outlier detection, $p(x_i|y_i)$ in generative model, $p(x_A|x_B)$
- Previous: naive Bayes, Gaussian, mixture models
- Today: graphical models.

(acyclic)

Basic idea behind directed graphical models:

- use product rule repeatedly:

$$p(x) = p(x_1)p(x_2|x_1)p(x_3|x_2,x_1)p(x_4|x_3,x_2,x_1)\cdots p(x_d|x_{1:(d-1)})$$

$p(x_d \mid x_{1:(d-1)})$
2^{d-1}

$$= \prod_{j=1}^{d} p(x_j|x_{1:(j-1)})$$

too many parameters!

Solutions:

- "parsimonious" parameterization $p(x_j|x_{1:(j-1)}) = f(w^T x_{1:(j-1)})$ — d parameters

- conditional independence

  e.g. naive Bayes: $x_j \perp x_{1:(j-1)} | y \Rightarrow p(x_j|x_{1:(j-1)}) = p(x_j|y)$

  Markov chain: $x_j \perp x_{1:(j-2)} | x_{j-1}$ $p(x_j|x_{1:(j-1)}) = p(x_j|x_{j-1})$

  General: $x_j \perp x_{1:(j-1)} \setminus \pi(j) | \pi(j)$ "parents"

- can also do both.

Directed acyclic graphical (DAG) models

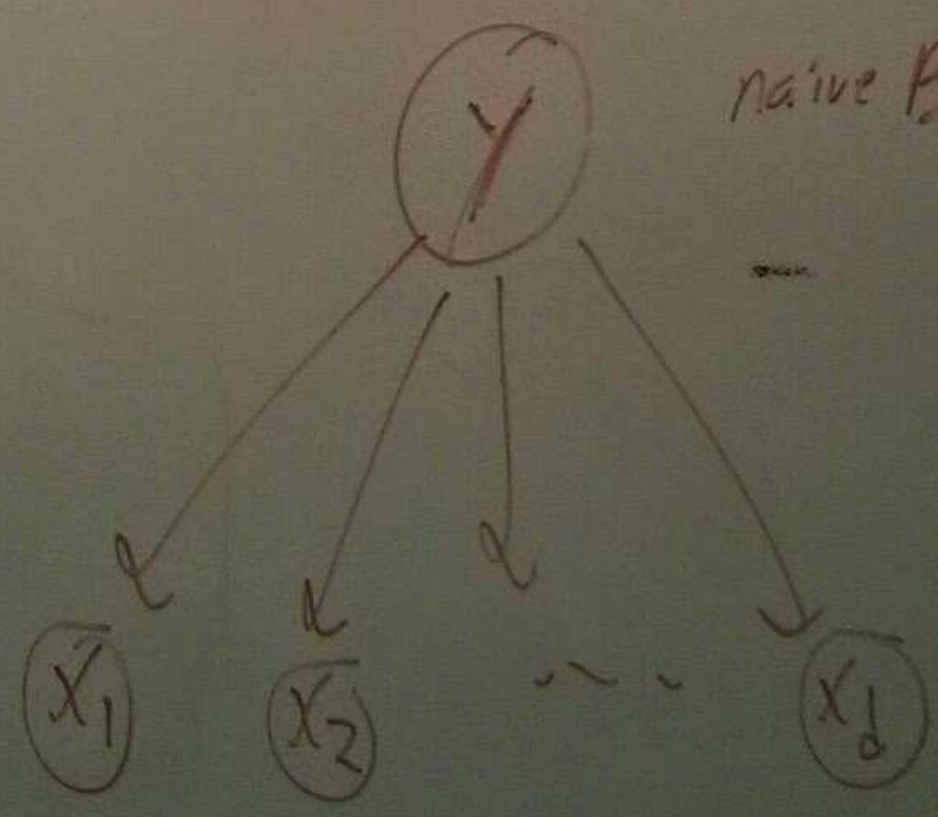
Graph $G = (V, E)$

"vertices" "edges"

V: random variables X_j

E: $X_{\pi(j)} \to X_j$

$$p(x) = \prod_{j=1}^{d} p(x_j \mid x_{\pi(j)})$$



naive Bayes

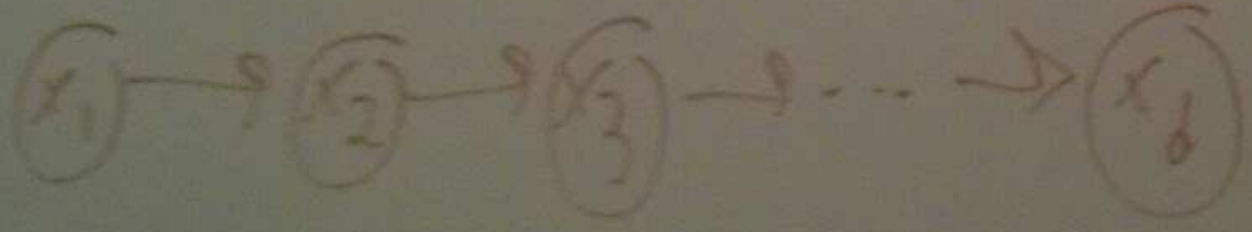
models "Bayesian networks"

"belief networks"

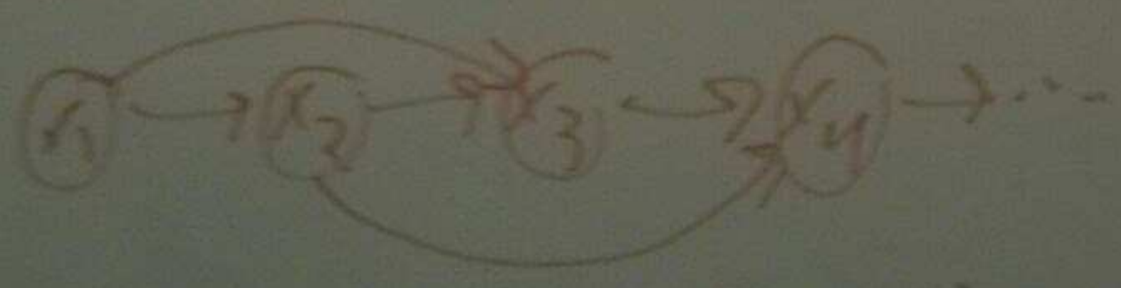
"causal networks" → (you need a justification to interpret edges causally)

Hidden Markov Models

Markov chain



2nd-order Markov chain



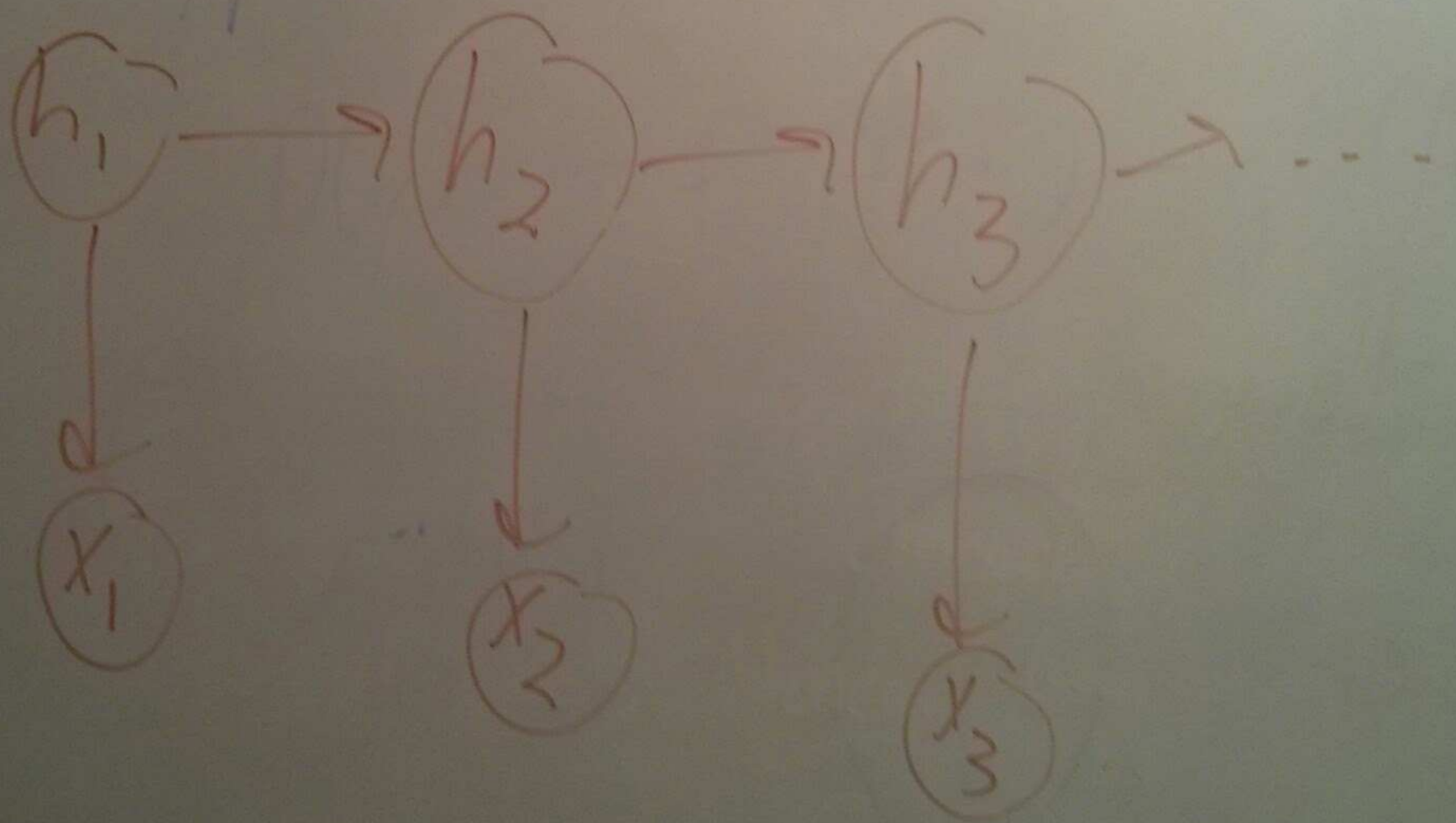
$p(x_j | x_{j-1}, x_{j-2})$

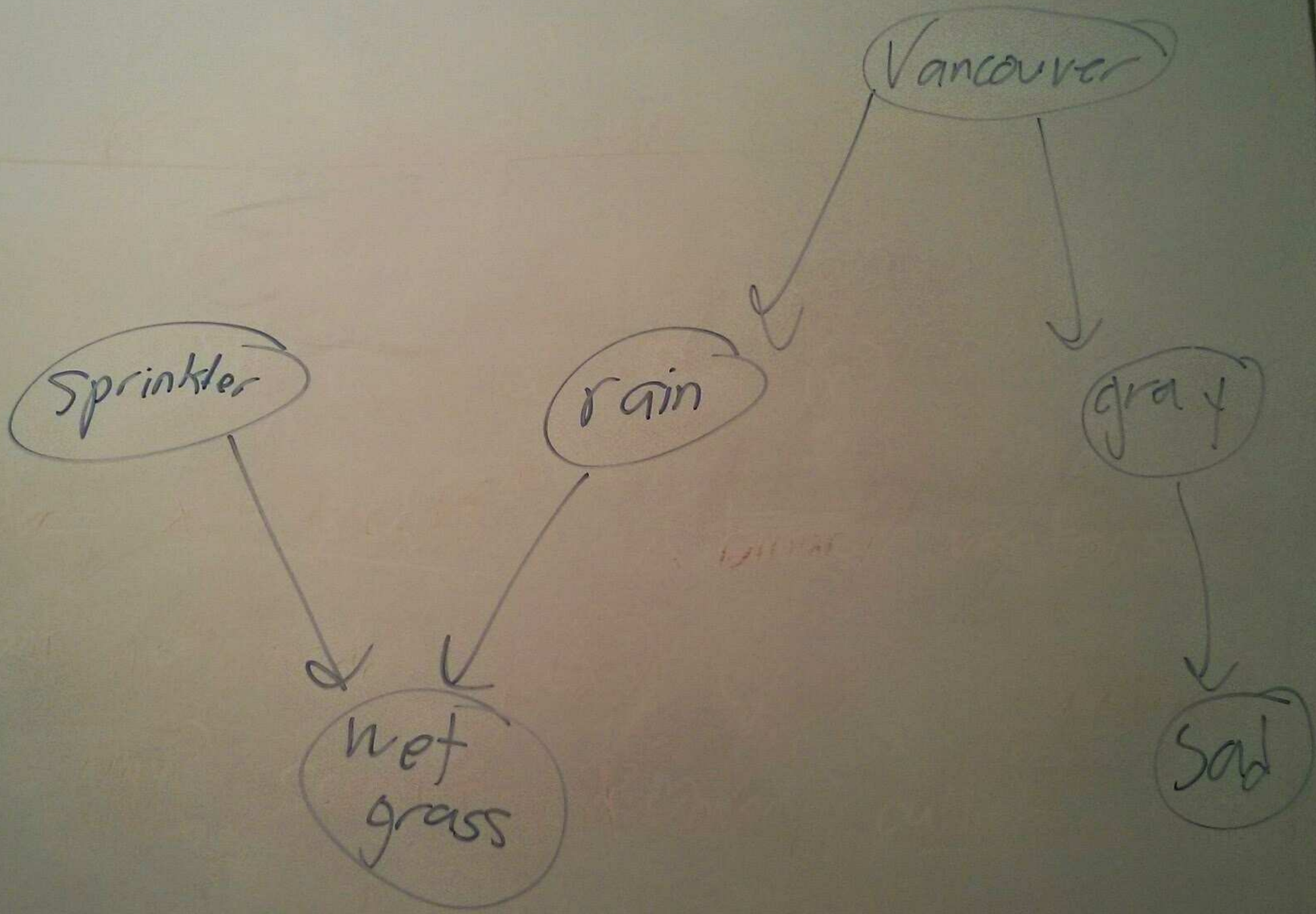
Directed Acyclic Graphical models

Bayes-ball algorithm

✓

HMM (Kalman filtering)




Vancouver
Sprinkler
rain
gray
wet grass
sad

Learning:

- fit each $p(x_j \mid x_{\pi(j)})$ independently

- "inference": computing $p(x)$
  computing $p(x_j \mid x_{\pi(j)})$ } easy
  computing $p(x_j \mid x_{j+1})$ } #P-hard

Conditional Independence Properties

- We can use the graph to check whether

$$X_A \perp X_B \mid X_E$$

follows from CI assumptions.

$X_A \perp X_B \mid X_E \iff$ A and B are "d-separated" given E.

A and B are d-separated if for all paths P between A and B, at least one of the following

1. P includes a "chain" A → E → B
2. P includes a "fork" ← E →
3. P contains a "collider" → C ←, where C and all its descendants are unobserved

Conditional Independance Properties

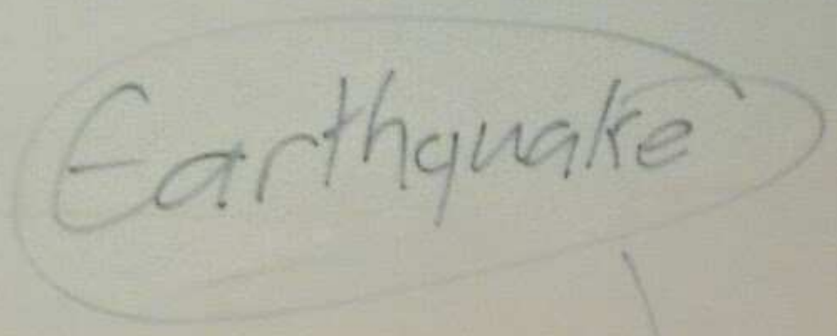
- We can use the graph to check whether

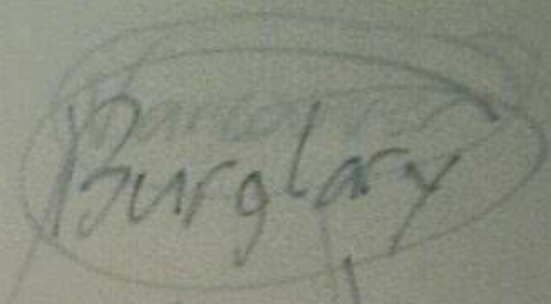
$$X_A \perp X_B \mid X_E$$

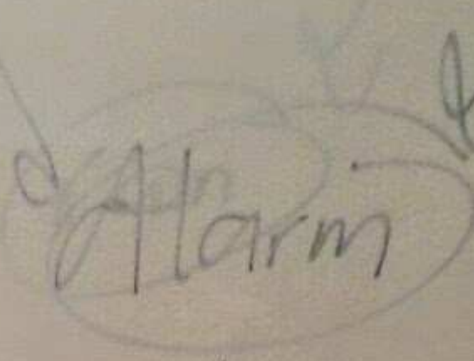
follows from the assumptions.

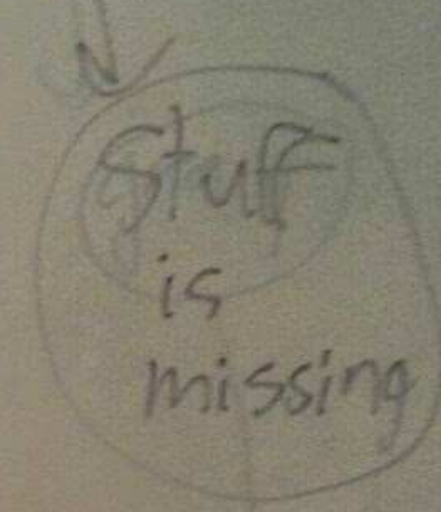
$X_A \perp X_B \mid X_E \iff$ A and B are "d-separated" given E.

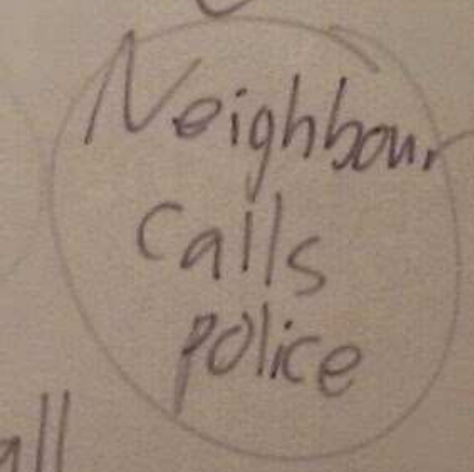
A and B
— are d-separated if for all paths P between A and B,
1. P includes a "chain"
2. P includes a "fork"
3. P contains a "collider"











✓

Burglary ⊥ Call

Earthquake ⊥ Call

Earthquake ⊥ Call | Alarm

Alarm ⊥ Stuff is missing

Alarm ⊥ Stuff is missing | Burglary

Earthquake ⊥ Burglary

Earthquake ⊥ Burglary | Alarm

Earthquake ⊥ Burglary | Call

Call ⊥ "explaining away"

Stuff is missing

Gaussian DAGs

$$p(x_j \mid x_{\pi(j)}) \sim N(\mu_j + \vec{w}_j^T x_{\pi(j)}, \sigma_j^2)$$

$\Leftrightarrow$

$$x \sim N(\mu, \Sigma)$$

$$\Sigma^{-1} = LL^T,$$

$\therefore$ non-zero pattern of L comes from G.

Plate Notation