EECS 545: Machine Learning

Lecture 13. Bayesian Networks

Honglak Lee

2/21/2011

Outline

- Overview of graphical models
- Bayesian networks (Directed graphical models)
  - Representation
  - Examples
  - Parameterization
  - Conditional Independence
  - D-separation

The Joint Probability Table

- Given a set of random variables $\{x_1 \ldots x_K\}$, the Joint Probability Table (JPT) $p(x_1 \ldots x_K)$ lets you answer any probabilistic question that can be asked.
- But: the JPT has size exponential in K.
- And: many of the entries are difficult to fill.

- Q1: How can we express the JPT concisely?
- Q2: How do we infer answers to questions?

Decomposing the JPT

- The product rule decomposes any joint probability table into conditional probabilities:

$$\begin{aligned} p(a,b,c) &= p(c|a,b)p(a,b) \\ &= p(c|a,b)p(b|a)p(a) \end{aligned}$$

- More generally,

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|x_1,\ldots,x_{k-1}) \text{ where } \mathbf{x} = \{x_1,\ldots,x_K\}$$

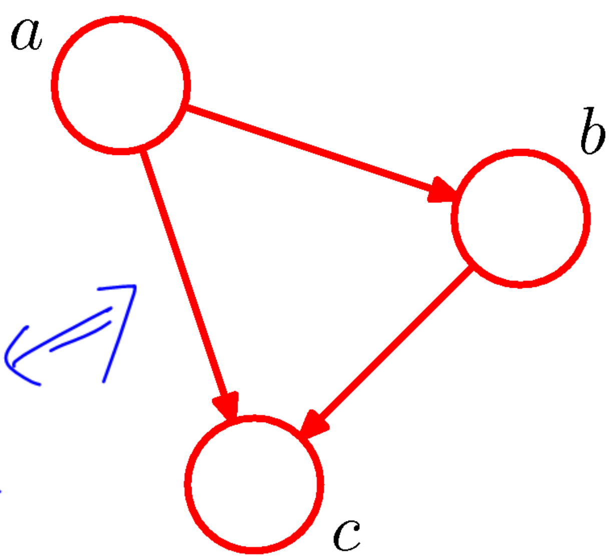
$X = (X_1, X_2, \ldots, X_K)$

- By itself, this is not more concise.

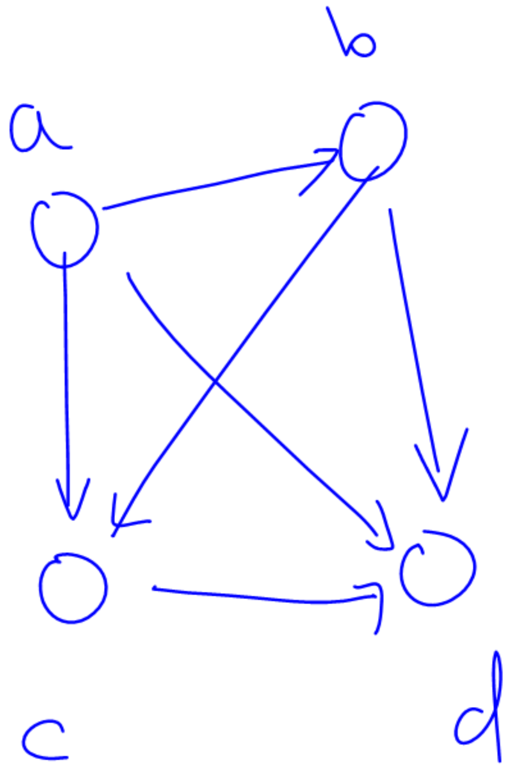
Graphical Representation

- Create a node for each random variable.
- Link each variable to those it is conditionally dependent on.
  - Directed Acyclic Graph
    - (DAG)

$$
\begin{aligned}
p(a,b,c) &= p(c|a,b)p(a,b) \\
&= p(c|a,b)p(b|a)p(a)
\end{aligned}
$$



  - Annotate the graph with the conditional probability tables.

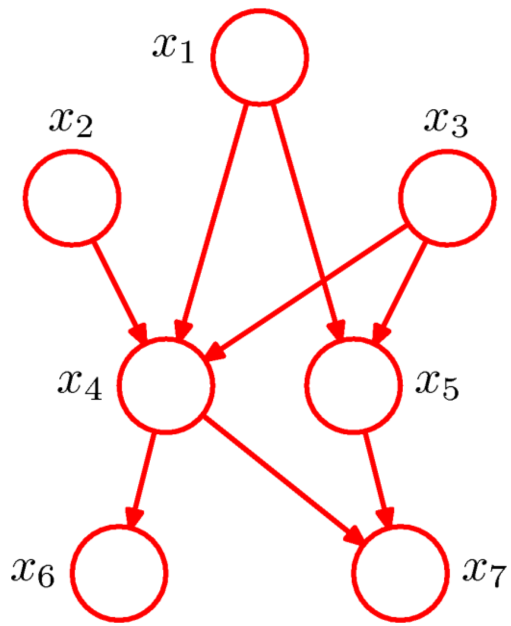


$$\iff$$

$$P(a,b,c,d) = P(a)\,P(b|a)\,P(c|a,b)\,P(d|a,b,c)$$

Missing Links

- The fully-connected graph represents the fully general JPT.
- In most domains, some variables will be *independent* (or *conditionally independent*) so some links will be missing.



$P(X_1)\, P(X_2|X_1)\, P(X_3|X_1,X_2) \ldots \quad P(X_7|X_1,\ldots,X_6)$

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$

X_i : M choices

$P(X_2) \leftrightarrow$ free parameters? $M-1$

$P(X_2|X_1) \leftrightarrow$ $M(M-1)$

Independence

- Random variables x and y are *independent* if $p(x,y) = p(x)\, p(y)$. [So: $p(y \mid x) = p(y)$.]
  - Information about x tells us nothing about y, and vice versa.

- x and y are *conditionally independent* given z if
  ① $p(x,y \mid z) = p(x \mid z)\, p(y \mid z)$.
  - Once z is known, information about x tells us nothing about y, and vice versa.

$P(x,y \mid z) = P(x \mid z)\, P(y \mid x,z)$

② $P(y \mid z) = P(y \mid x,z)$

③ $P(x \mid z) = P(x \mid y,z)$

To build a Bayesian network

- Specify the variables x = $\{x_1 \ldots x_K\}$
- Identify what each variable depends on:
  - Variable x_k depends on its parent variables pa_k
- Add links to x_k from its parents.
  - These are often causal relations in the domain.
- Annotate x_k with $p(x_k \mid \text{pa}_k)$.
- Then the joint probability table is:

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\text{pa}_k)$$

= Parent of X_k.

Graph

cf. $P(X) = \prod_{k=1} P(X_k \mid X_1, \ldots, X_{k-1})$

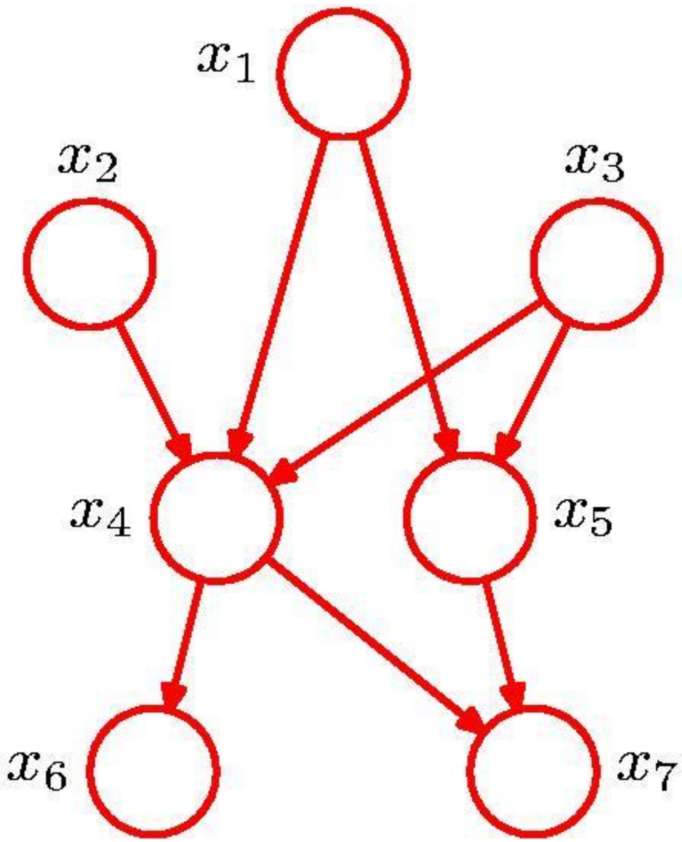
Bayesian Networks

Given ordering, $P(x_k | x_1, \ldots, x_{k-1}) = P(x_k | \mathrm{Pa}\, x_k)$

$P(x_1) P(x_2) P(x_3) \; P(x_4 | \underbrace{x_1, x_2, x_3}_{\mathrm{Pa}\, x_4})$

$P(x_5 | x_1, x_3) \; P(x_6 | x_4)$

$P(x_7 | x_4, x_5)$

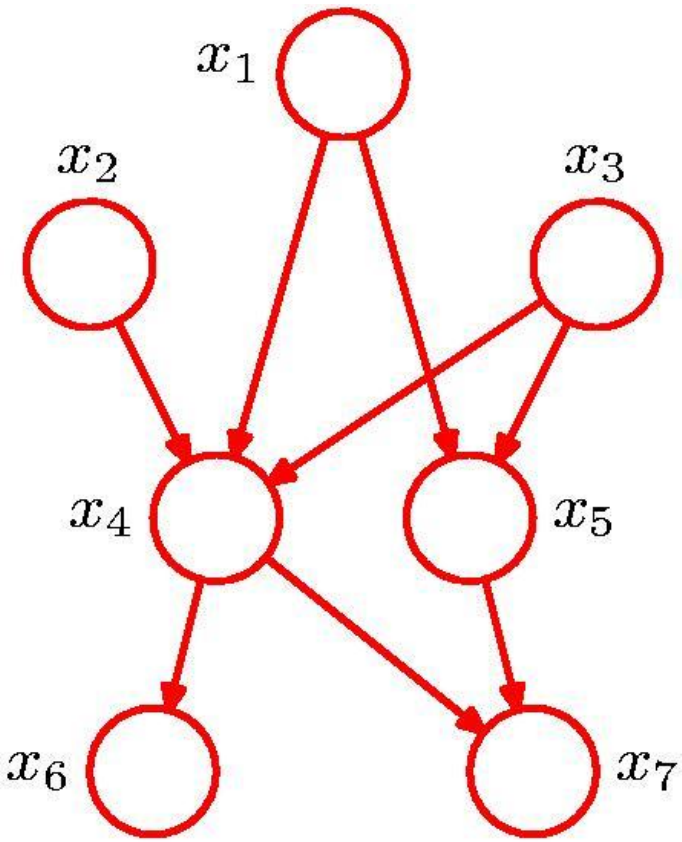


General Factorization for Bayesian Networks

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k | \mathrm{pa}_k)$$

Bayesian Networks

$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

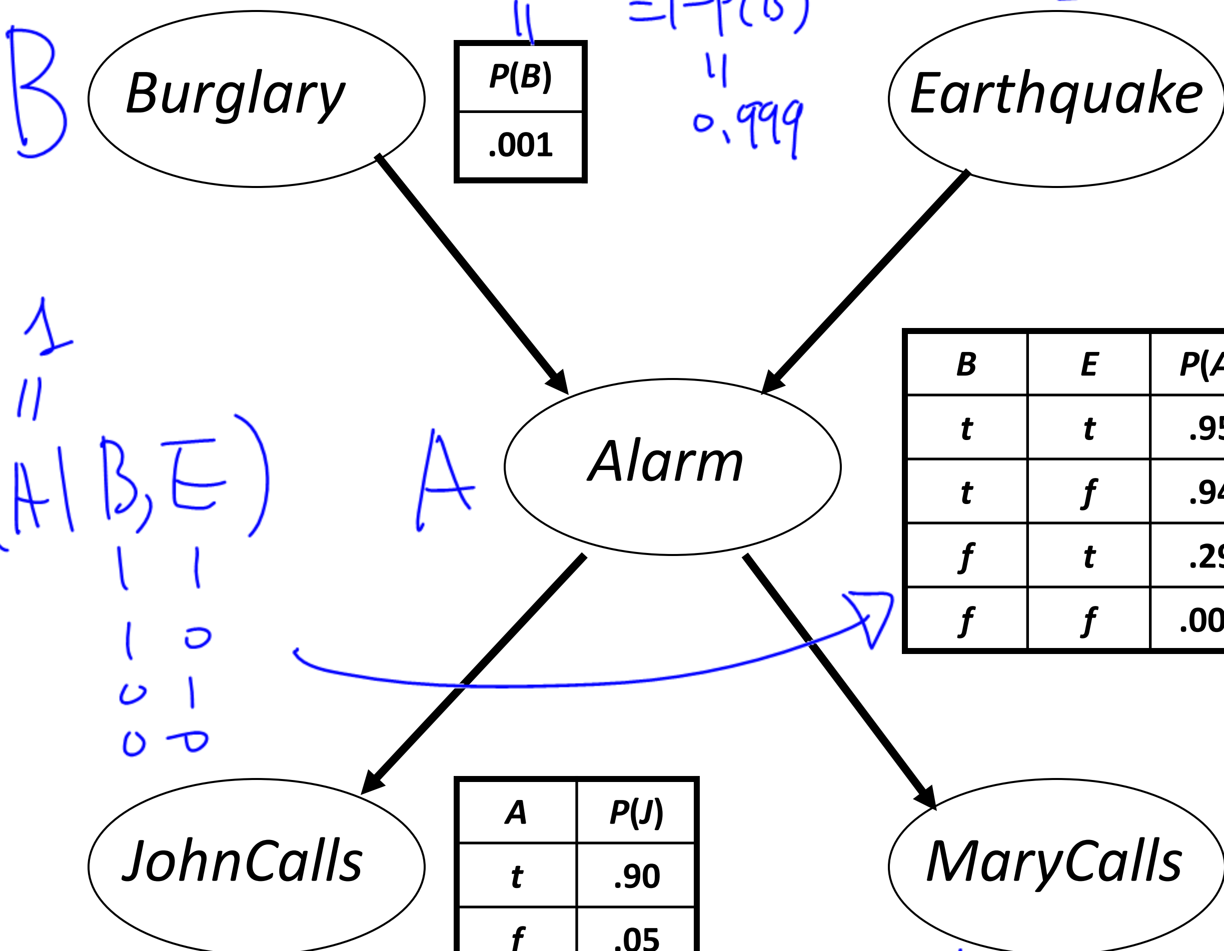


General Factorization for Bayesian Networks

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

Examples of Bayesian Networks

Classic Bayes Net Example



$P(B=1) = P(B)$

$P(B=0) = 1 - P(B) = 0.999$

P(B)
.001

P(E)
.002

$1 = P(A \mid B, E)$

B	E
1	1
1	0
0	1
0	0

B	E	P(A)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	P(J)
t	.90
f	.05

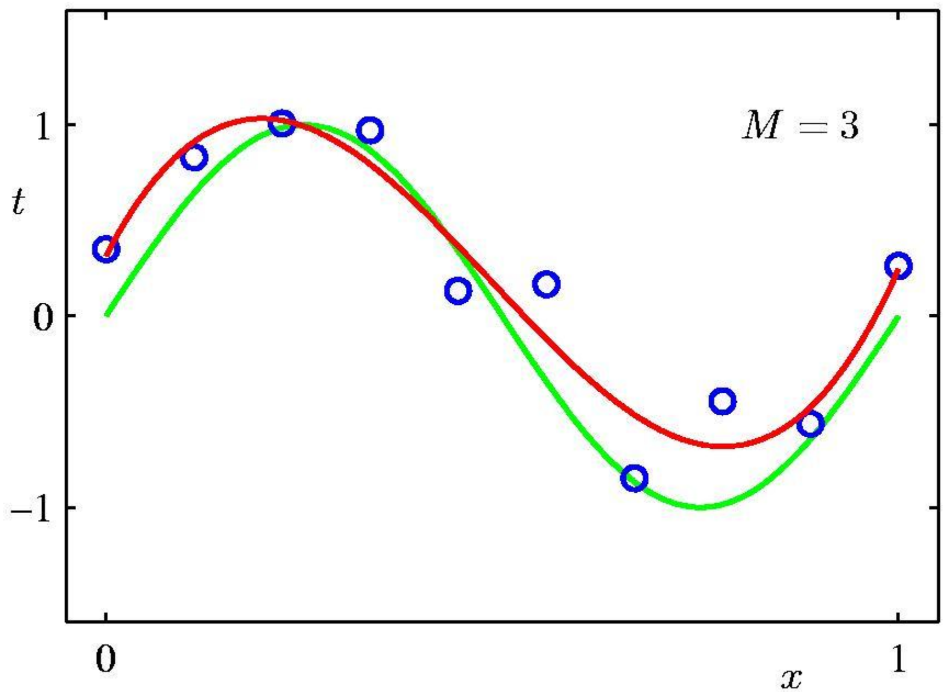
$P(J \mid A)$

$A = 1$, 0

A	P(M)
t	.70
f	.01

- From Pearl (1986)

Bayesian Curve Fitting



Polynomial

$$y(x,\mathbf{w}) = \sum_{j=0}^{M} w_j x^j$$

$$\mathcal{N}(t_n | w^T \phi(x_n), \sigma^2)$$

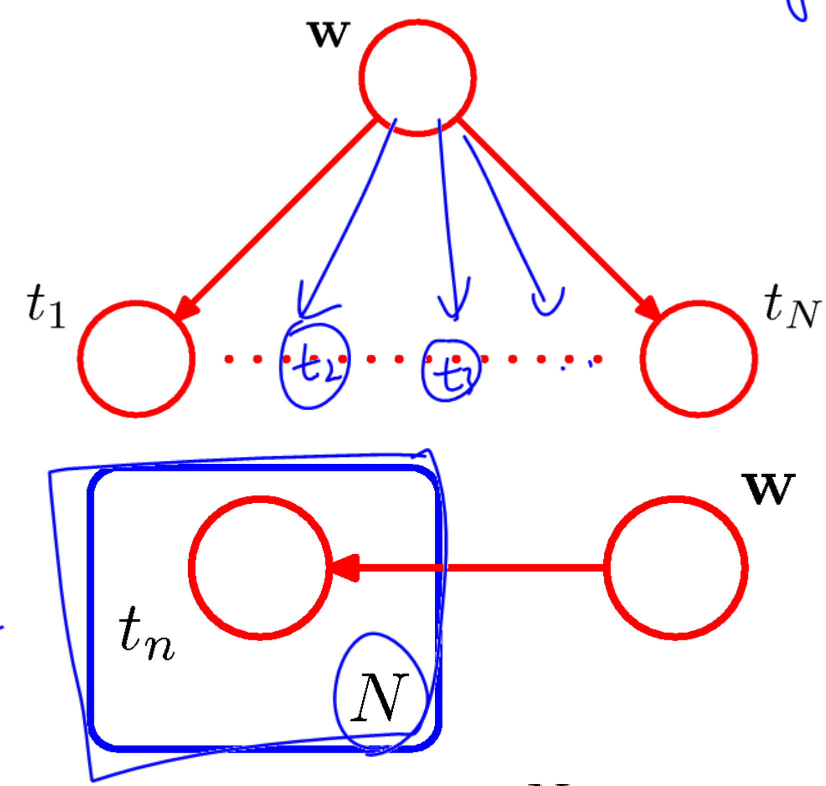
$$\parallel$$

$$p(\mathbf{t},\mathbf{w}) = \underbrace{p(\mathbf{w})}_{\text{prior for } w} \underbrace{\prod_{n=1}^{N} p(t_n|y(\mathbf{w},x_n))}_{\text{likelihood.}}$$

Bayesian Curve Fitting

$t_n \sim N(w^T \phi(x_n), \sigma^2)$

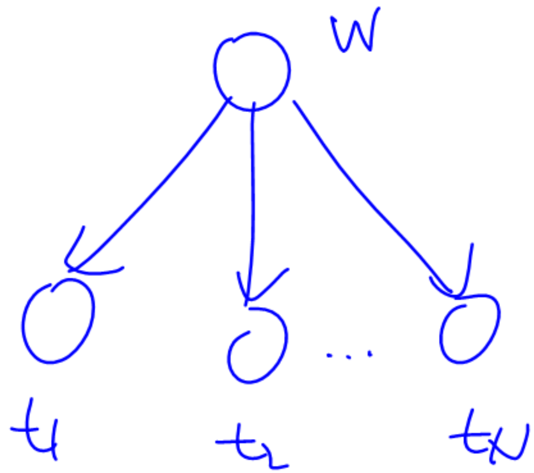
- Describe with random variables for target values t_i and weight vector w.
- Not shown: parameters like input data x, variance, and the hyperparameter on w.
- *Plate* models multiple iterated nodes.



$\Leftrightarrow$ Plate

$$p(\mathbf{T}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n | \mathbf{w})$$

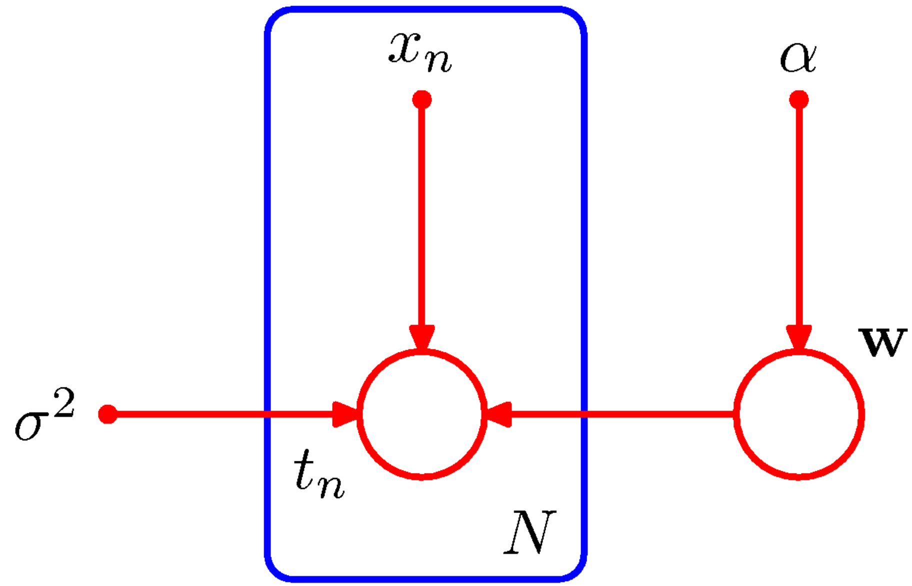
$$p(w)\, P(t_1 \ldots t_N | w) = \prod_{n=1} p(t_n | w)\, p(w)$$



$$P(W)\, P(t_1 \mid W)\, P(t_2 \mid W) \cdots P(t_N \mid W)$$

Bayesian Curve Fitting

- Open circles represent random variables.
- Small filled circles represent constant parameters: input, variance, and hyperparameter.

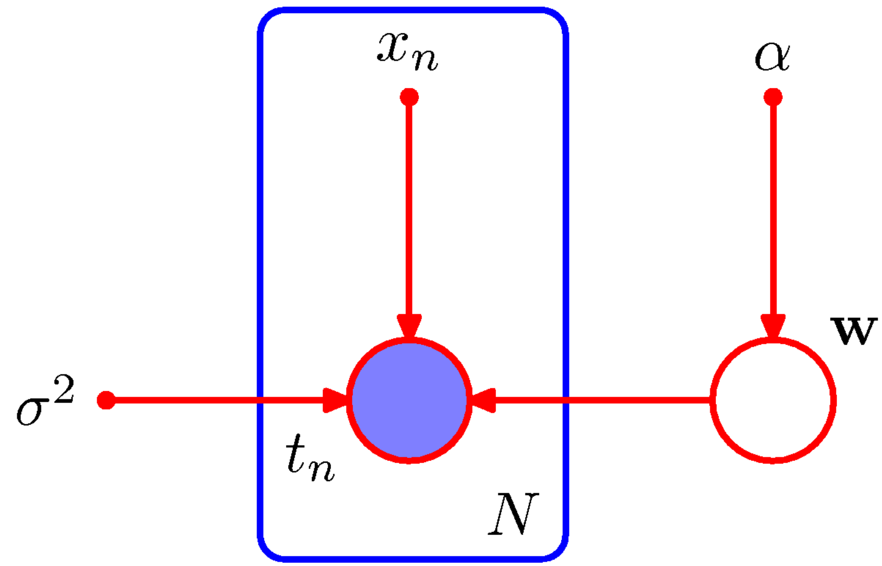


$$p(\mathbf{T}, \mathbf{w} | \mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n | \mathbf{w}, x_n, \sigma^2)$$

$\mathcal{N}(0, \alpha^{-1} I)$ $\mathcal{N}(w^T \phi(x_n), \sigma^2)$

Bayesian Curve Fitting—Learning

- Shaded circles represent random variables with observed values.



MAP

$\max_{w} P(w|T)$

Bayesian

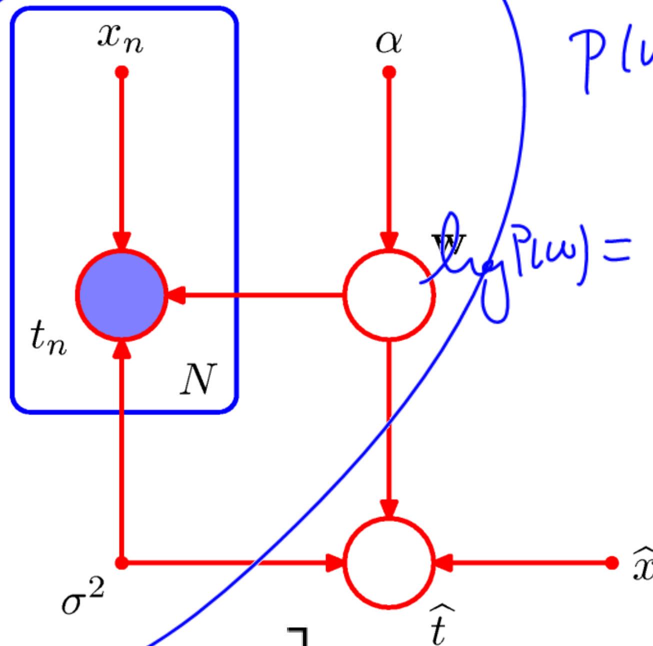
$P(w|T)$

$$p(\mathbf{w}|\mathbf{T}) \propto p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$

$$\propto P(w) \; P(T|w)$$

Bayesian Curve Fitting—Prediction

- Add new variables to represent a query and the estimated target value.



$P(w) \sim N(0, \alpha^{-1} I)$

$\log P(w) = \frac{\alpha}{2} w^T w$

$$p(\hat{t}, \mathbf{T}, \mathbf{w} | \hat{x}, \mathbf{x}, \alpha, \sigma^2) = \left[\prod_{n=1}^{N} p(t_n | x_n, \mathbf{w}, \sigma^2) \right] p(\mathbf{w}|\alpha) p(\hat{t} | \hat{x}, \mathbf{w}, \sigma^2)$$

$$p(\hat{t} | \hat{x}, \mathbf{x}, \mathbf{T}, \alpha, \sigma^2) \propto \int p(\hat{t}, \mathbf{T}, \mathbf{w} | \hat{x}, \mathbf{x}, \alpha, \sigma^2)\, d\mathbf{w}$$

Sampling the Joint Distribution

- To sample the JPT p($x_1 \ldots x_K$).
- In p(x_k | pa$_k$), the variables in pa$_k$ come before x_k in the ordering of variables, so we can sample in sequence.

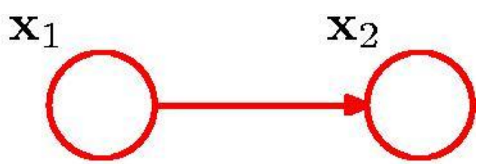
1

- For example: $p(a,b,c) = p(c|a,b)\, p(b|a)\, p(a)$
  - Draw a from distribution $p(a)$.
  - Given that value of a, draw b from $p(b|a)$. $\begin{cases} p(a=1) & 1 \\ p(a=0) & 0 \end{cases}$
  - Given a and b, draw c from $p(c|a,b)$.
  - (a,b,c) is drawn from $p(a,b,c)$.

Parameterization

Discrete Variables (1)

- General joint distribution: K^2 - 1 parameters



$$p(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}) = \prod_{k=1}^{K} \prod_{l=1}^{K} \mu_{kl}^{x_{1k} x_{2l}}$$

$p(x_1)$ $K-1$ $p(x_2|x_1) \rightarrow K(K-1)$

$\uparrow K$

$K^2-K+K-1=K^2-1$

- Independent joint distribution: 2(K-1) parameters

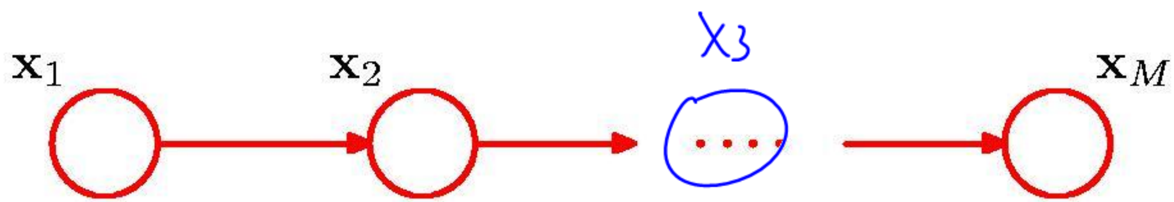
$$\hat{p}(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_{1k}^{x_{1k}} \prod_{l=1}^{K} \mu_{2l}^{x_{2l}}$$

$P(x_1)$ $K-1$ $P(x_2)$ $K-1$

Discrete Variables (2)

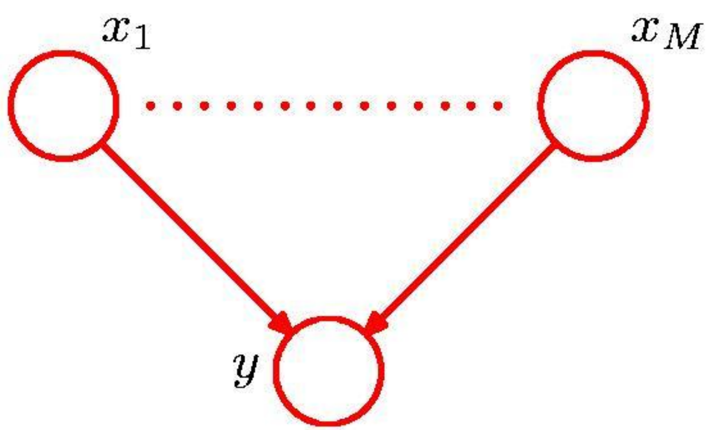
General joint distribution over M variables:
$K^M - 1$ parameters

M -node Markov chain: K - 1 + (M-1) K (K-1) $\sim O(MK^2)$
parameters



$$\underset{\Updownarrow \; K-1}{P(x_1)} \quad \underbrace{P(x_2|x_1) \quad P(x_3|x_2) \quad \cdots P(x_M|x_{M-1})}_{K(K-1)\ (M-1)}$$

Parameterized Conditional Distributions



If $x_1, \ldots, x_M$ are discrete, K-state variables, K^M $p(y = 1|x_1, \ldots, x_M)$ in general has $O(K^M)$ parameters.

The parameterized form (logistic regression)

$$p(y = 1|x_1, \ldots, x_M) = \sigma\left(w_0 + \sum_{i=1}^{M} w_i x_i\right) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

requires only M+1 parameters

Data Complexity

- How many parameters specify the distribution of *M* discrete variables?
  - If graph is disconnected, $o(M)$. ← all variables are indep.
  - If graph is fully connected, $o(2^M)$. ← no independence
    binary
  - Partial structure gives intermediate complexity.

- For M Gaussian variables?
  - If graph is disconnected, $o(M)$.
  - If graph is fully connected, $o(M^2)$.

Conditional Independence

Conditional Independence

- Suppose ① $p(a|b, c) = p(a|c)$ a and b are cnd. indp given c

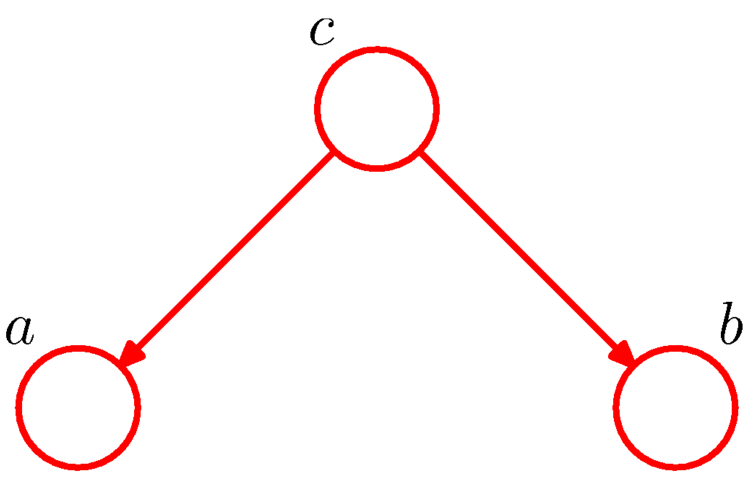
chain rule

- Then ② $$\begin{aligned} p(a, b|c) &= p(a|b, c)p(b|c) \\ &= p(a|c)p(b|c) \end{aligned}$$

- And *a* and *b* are *conditionally independent*, given *c*.
  - Notation: $a \perp\!\!\!\perp b \mid c$
- This property is very useful, and can be inferred from the graph structure alone.

Conditional Independence

- For this graphical model, without knowledge of *c*,
- There is a path between *a* and *b*,
- And they are not independent.



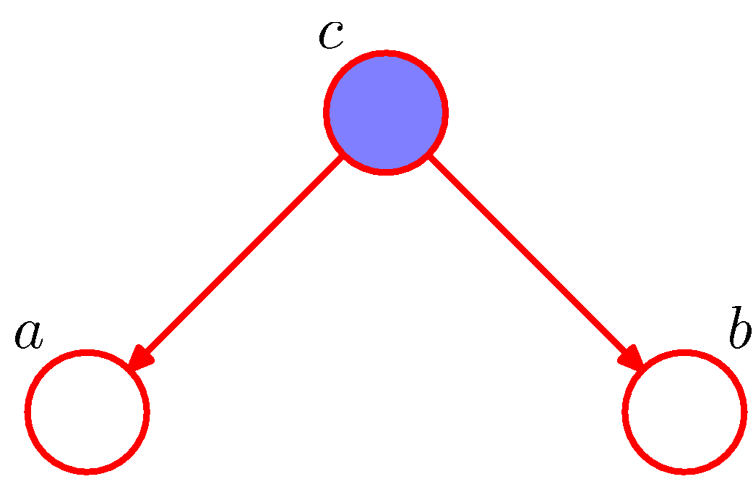
$$p(a, b, c) = p(a|c)p(b|c)p(c)$$

$$p(a, b) = \sum_c p(a|c)p(b|c)p(c)$$

$$a \not\perp\!\!\!\perp b \mid \emptyset$$

The Tail-to-Tail Connection

- But if we condition on c,
- then *a* and *b* are conditionally independent.
- A tail-to-tail connection is blocked by knowledge of the connecting value.

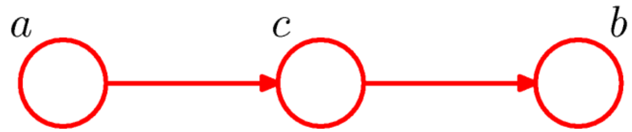


$$\begin{aligned} p(a,b|c) &= p(a,b,c)/p(c) \\ &= p(a|c)p(b|c) \end{aligned}$$

$$a \perp\!\!\!\perp b \mid c$$

The Head-to-Tail Connection

- For the head-to-tail model



$$p(a,b,c) = p(a)p(c|a)p(b|c)$$

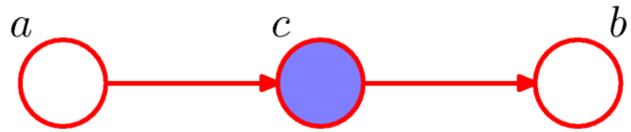
- Without knowledge of *c*, *a* and *b* are not independent.

$$p(a,b) = p(a)\sum_c p(c|a)p(b|c) = p(a)p(b|a)$$

$$a \not\perp\!\!\!\perp b \mid \emptyset$$

The Head-to-Tail Connection

- For the head-to-tail model



$$p(a,b,c) = p(a)p(c|a)p(b|c)$$

- But knowledge of *c* blocks the path, so they are conditionally independent.

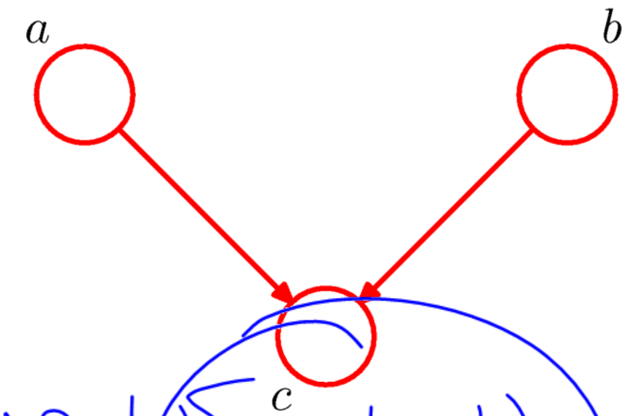
$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

$$a \perp\!\!\!\perp b \mid c$$

The Head-to-Head Connection

- The Head-to-Head model:

$$p(a,b,c) = p(a)p(b)p(c|a,b)$$



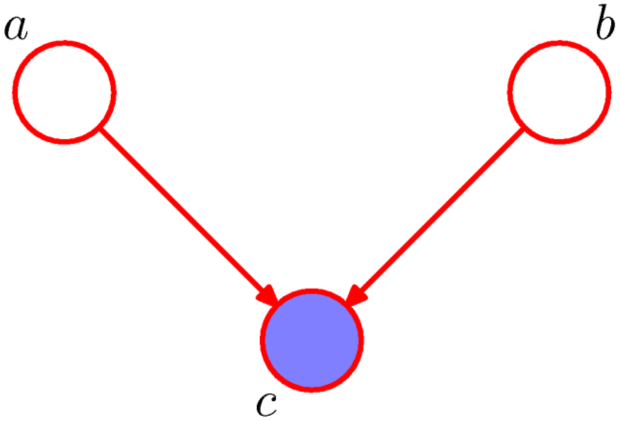
- Without knowledge of *c*, *a* and *b* are independent!
  - (even with an undirected connection) $\sum_c p(a,b,c) = p(a)p(b)\underbrace{\sum_c p(c|a,b)}_{=1}$
  - (marginalize over *c*)

$$p(a,b) = p(a)p(b) \qquad a \perp\!\!\!\perp b \mid \emptyset$$

Note: this is the opposite of Example 1, with c unobserved.

The Head-to-Head Connection

- Without knowledge of *c*, *a* and *b* are independent.
- But knowledge of *c* creates a dependency between *a* and *b*!
- "Explaining away"



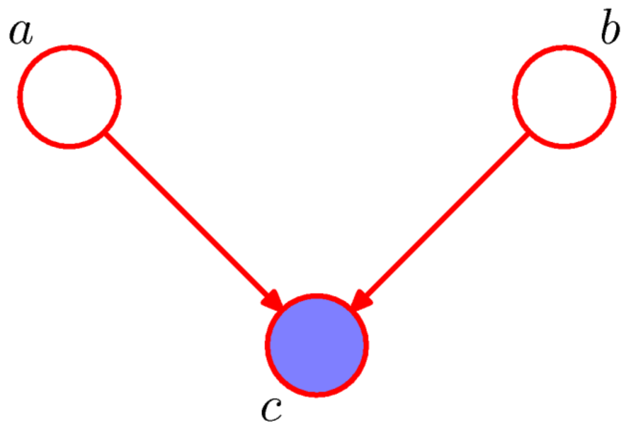
$$
\begin{aligned}
p(a,b|c) &= p(a,b,c)/p(c) \\
&= p(a)p(b)p(c|a,b)/p(c) \\
&\neq p(a|c)p(b|c)
\end{aligned}
$$

$$a \not\perp\!\!\!\perp b \mid c$$

Note: this is the opposite of Example 1, with c unobserved.

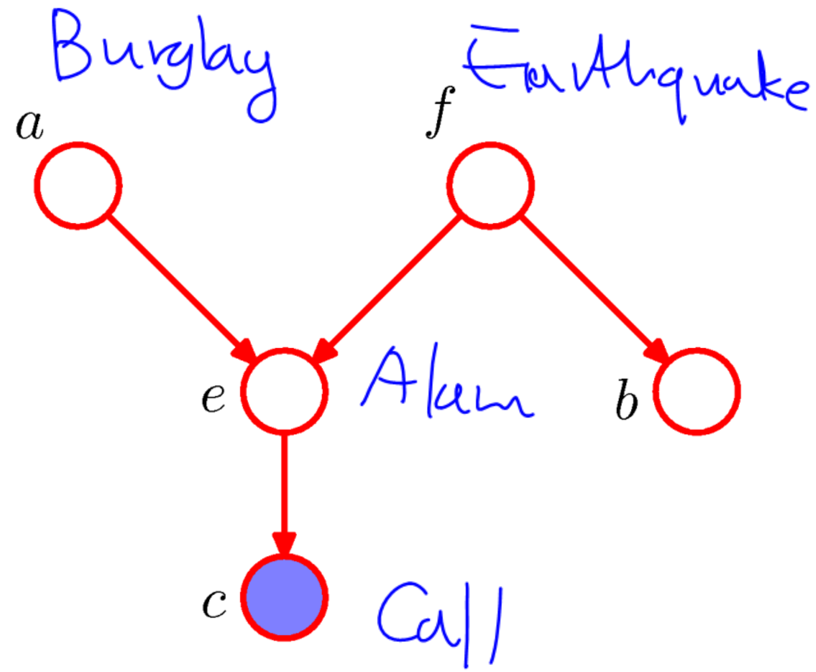
"Explaining Away"

- Given evidence *c*, hypotheses *a* and *b* can both explain it.
  - Knowing that *a* is true makes *b* less likely.
  - Knowing that *b* is true makes *a* less likely.
- Therefore, *a* and *b* are dependent.



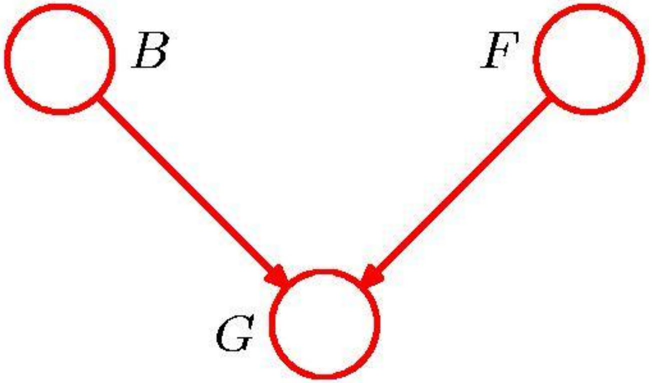
“Explaining Away”

- A burglar (a) might set off your alarm (e), and your neighbor would call (c).
- But an earthquake (f) could also cause the alarm (d), and might be reported on the radio (b).



“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8 \\
p(G=1|B=1,F=0) &= 0.2 \\
p(G=1|B=0,F=1) &= 0.2 \\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$



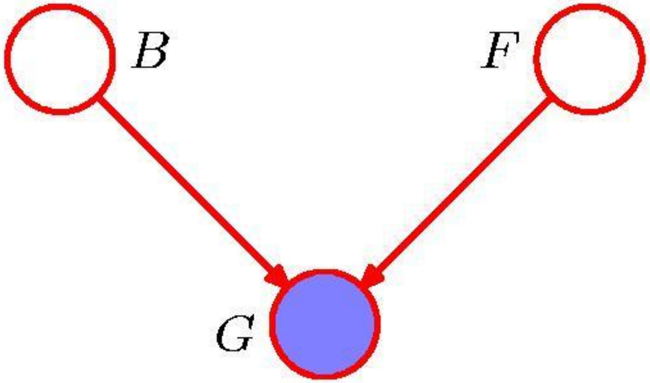
$$
\begin{aligned}
p(B=1) &= 0.9 \\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

$$
p(F=0) = 0.1
$$

B = Battery (0=flat, 1=fully charged)
F = Fuel Tank (0=empty, 1=full)
G = Fuel Gauge Reading
(0=empty, 1=full)

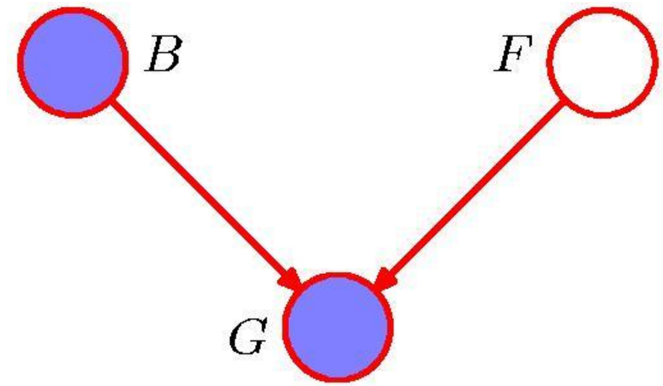
“Am I out of fuel?”



$$
\begin{aligned}
p(F=0|G=0) &= \frac{p(G=0|F=0)p(F=0)}{p(G=0)} \\
&\simeq 0.257
\end{aligned}
$$

Probability of an empty tank increased by observing G = 0.

“Am I out of fuel?”



$$p(F=0|G=0,B=0) = \frac{p(G=0|B=0,F=0)p(F=0)}{\sum_{F\in\{0,1\}} p(G=0|B=0,F)p(F)} \simeq 0.111$$

Probability of an empty tank reduced by observing B = 0.
This referred to as “explaining away”.

D-separation

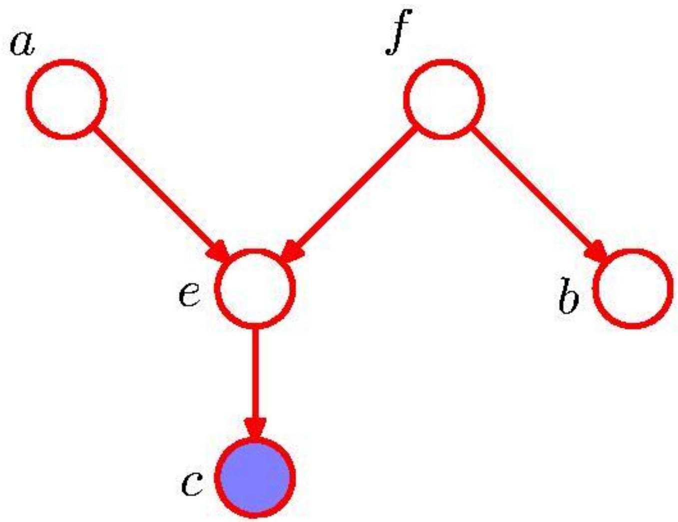
- A, B, and C are non-intersecting subsets of nodes in a directed graph.
- A path from A to B is blocked if it contains a node such that either
  a) the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C, or
  b) the arrows meet head-to-head at the node, and neither the node, nor **any** of its descendants, are in the set C.
- If all paths from A to B are blocked, A is said to be d-separated from B by C.
- If A is d-separated from B by C, the joint distribution over all variables in the graph satisfies $A \perp\!\!\!\perp B \mid C$.

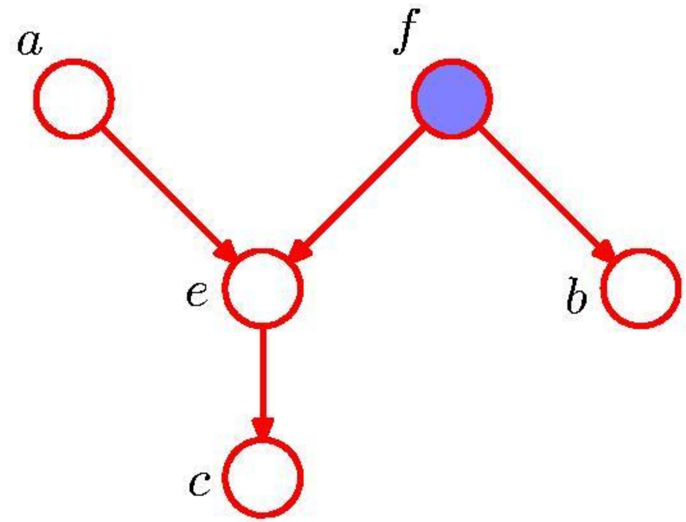
Conditional Independence

- We can use D-Separation to determine whether any sets *A* and *B* of variables are conditionally independent, given knowledge of the values of variables in *C*.

- The graphical model contains everything needed to infer conditional independence.

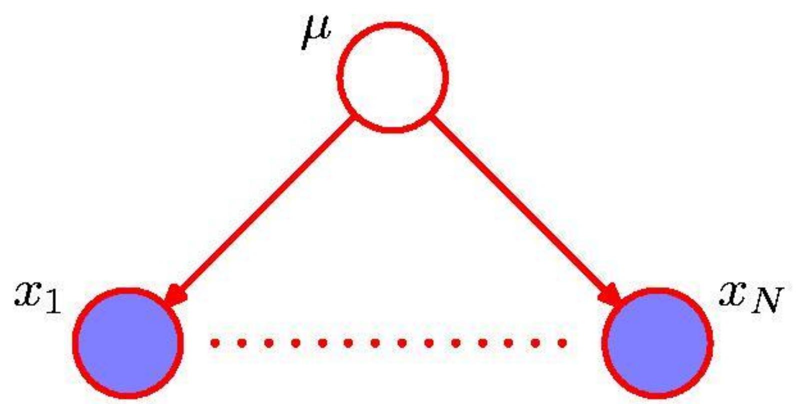
D-separation: Example





$$a \not\perp\!\!\!\perp b \mid c \qquad\qquad a \perp\!\!\!\perp b \mid f$$

D-separation: I.I.D. Data



$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$

$$p(\mathcal{D}) = \int_{-\infty}^{\infty} p(\mathcal{D}|\mu)p(\mu)\,\mathrm{d}\mu \neq \prod_{n=1}^{N} p(x_n)$$

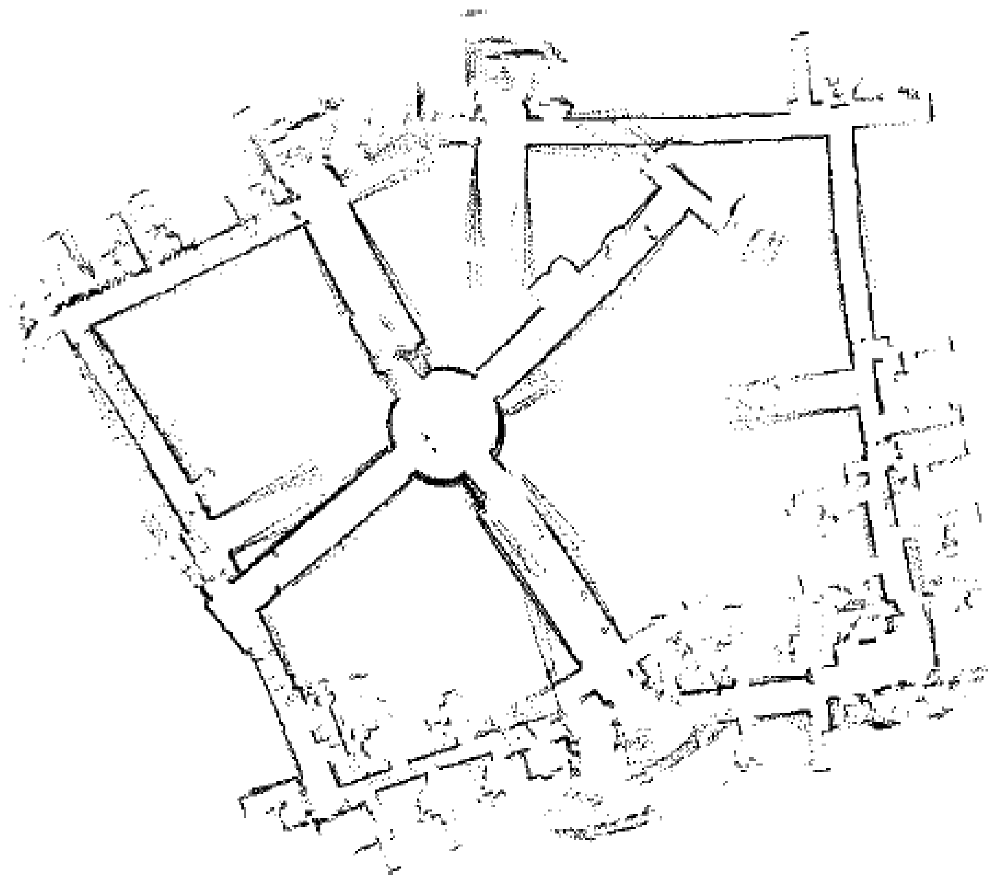
More examples

Example from Robot Mapping

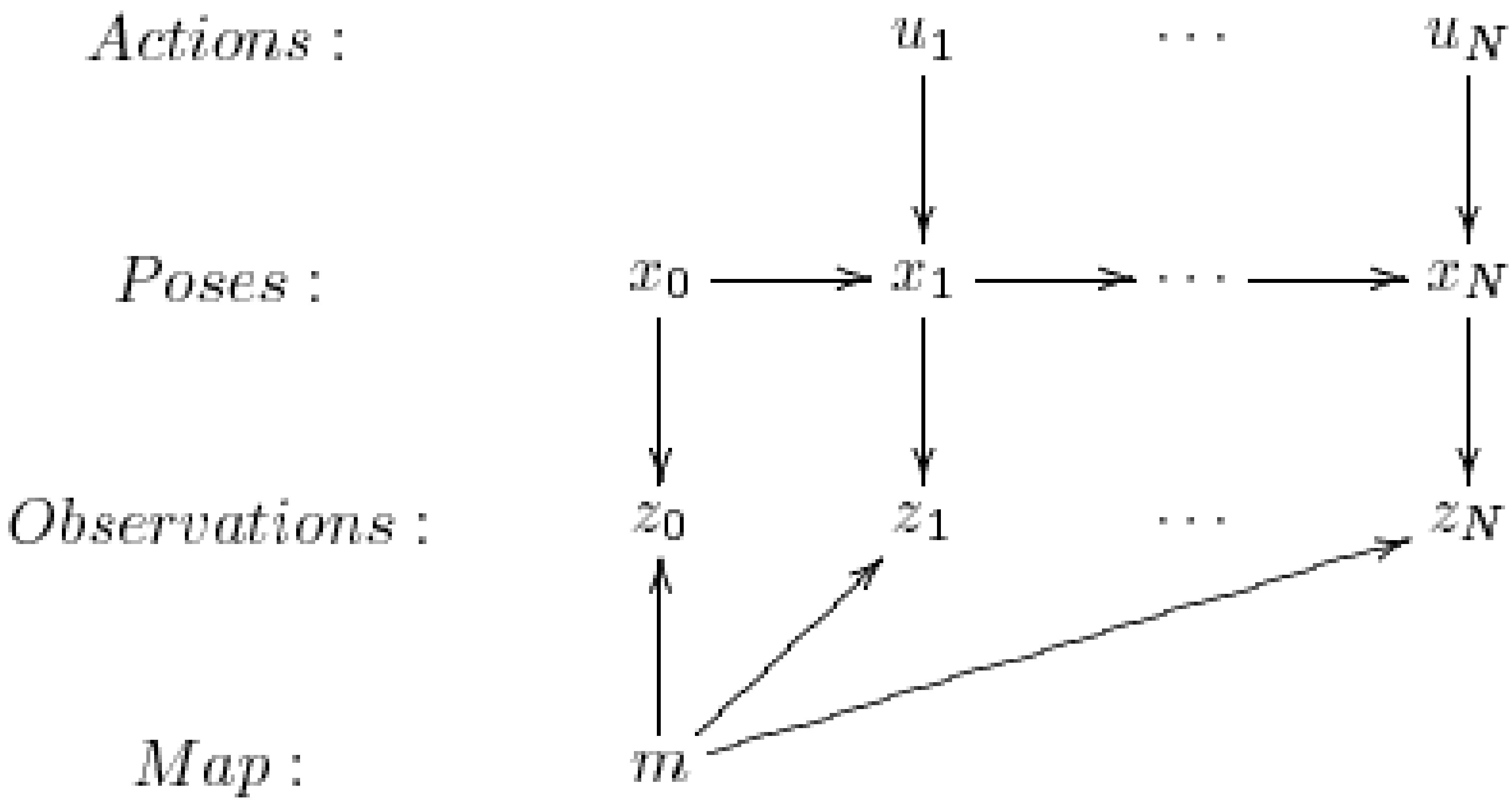
- Local metrical maps
  - Given local maps of each place…
- Global topological maps
  - Given a single best structural hypothesis …
- Global metrical map
  - Displacement along each travel segment
  - Global layout of places
  - All robot poses in the global frame of reference

Building large metrical maps

- When closing large loops, cumulative error can lead to incoherent maps.



Dynamic Bayesian Networks



Given the Topological Map …

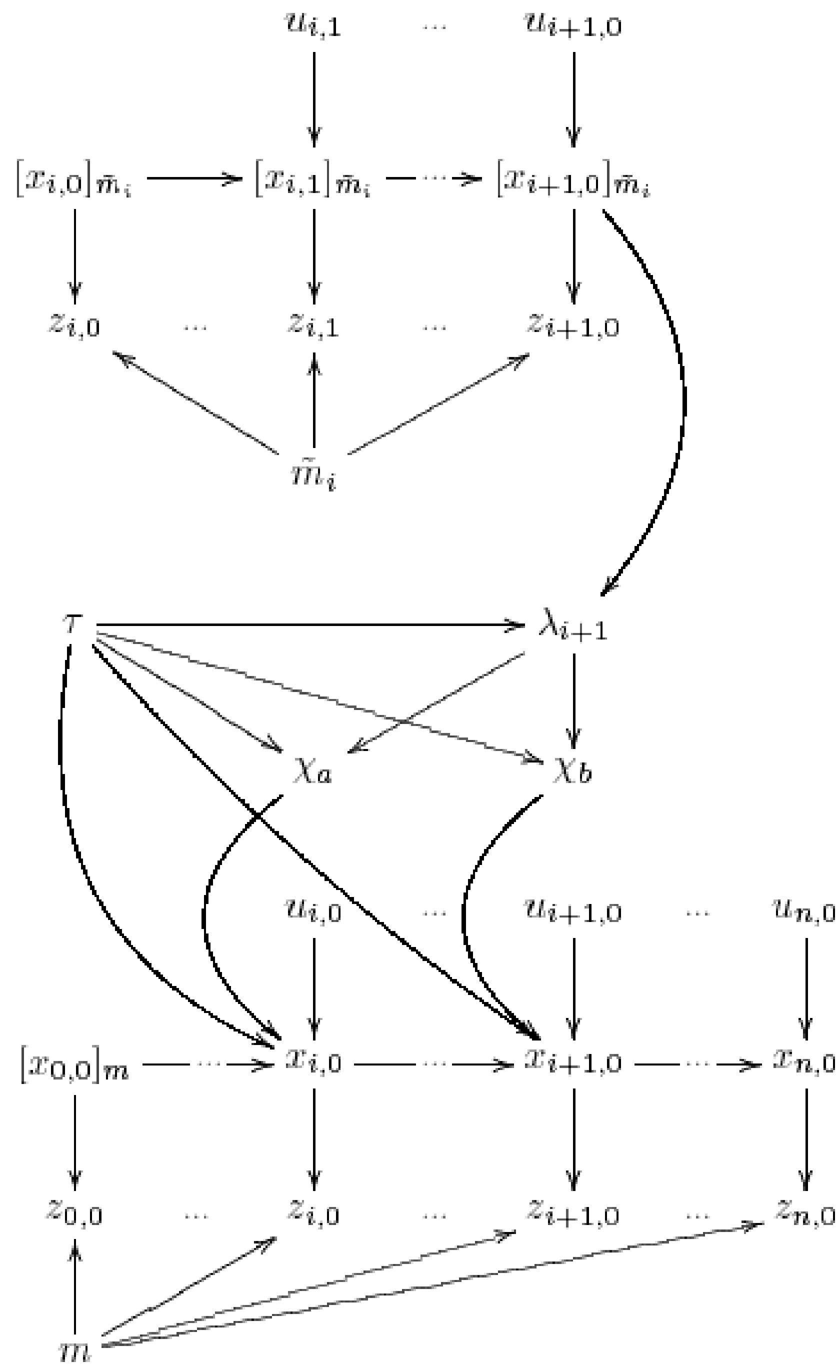
- The loop-closing problem is solved.
  - The topological map specifies which loops close, and where.
- Each place has an accurate local metrical map in its own local frame of reference.
- Continuous behavior divides into segments at distinctive place neighborhoods
- The global metrical map combines information from separate local maps.

The Global Metrical Map: Factoring the Problem

- Displacements: the pose of each place in the frame of reference of its predecessor.
- Layout: the pose of each place in the global frame of reference.
- Robot poses: the robot pose at each timestep in the global frame of reference.

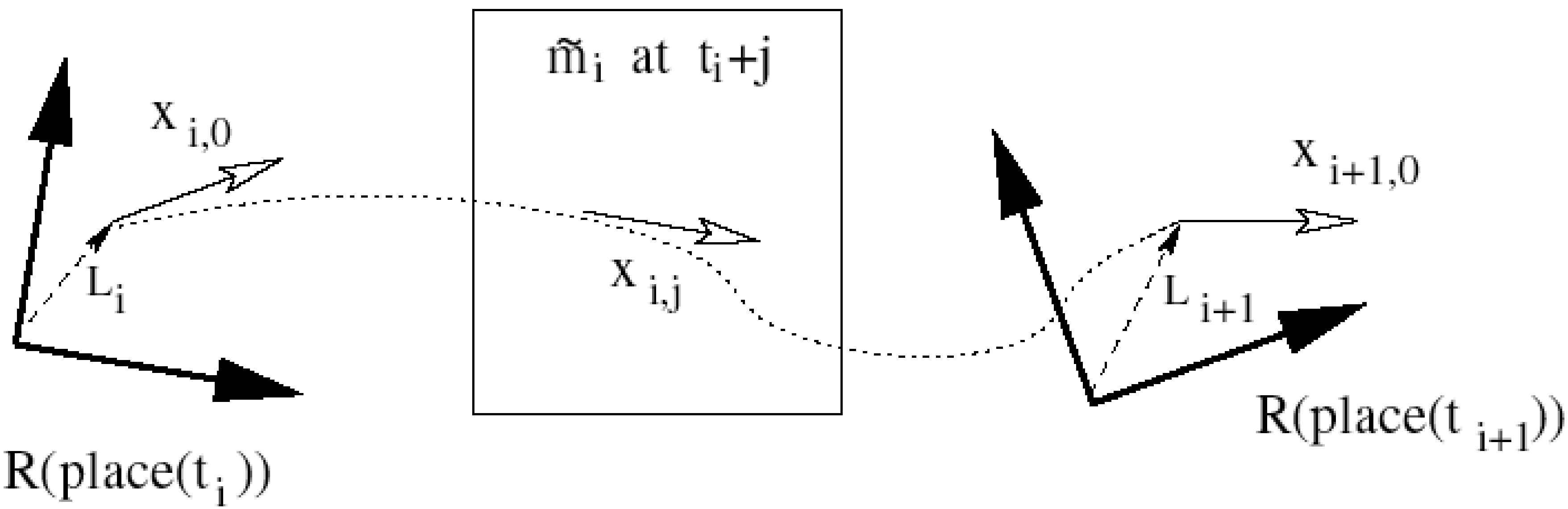
Factored DBN

- For building the global metrical map on the topological skeleton τ.
  - Local maps m_i
  - Displacements λ
  - Place layout χ
  - Global poses x
  - Global map m



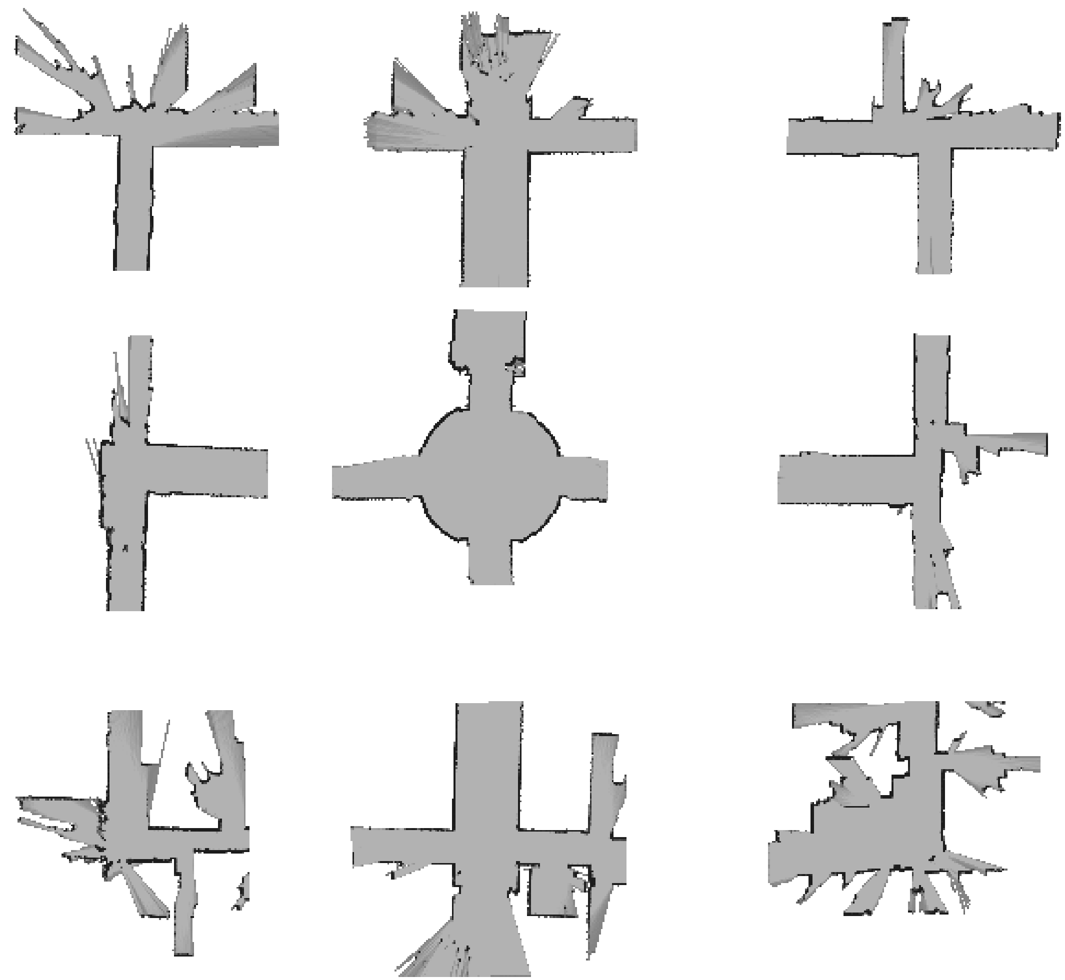
Estimating Displacements

- Use incremental SLAM to estimate pose $x_{i+1,0}$ in the frame of reference of m_i.
- Localize to get $x_{i+1,0}$ in frame m_{i+1}.
- Derive displacement λ_i between the two place poses.



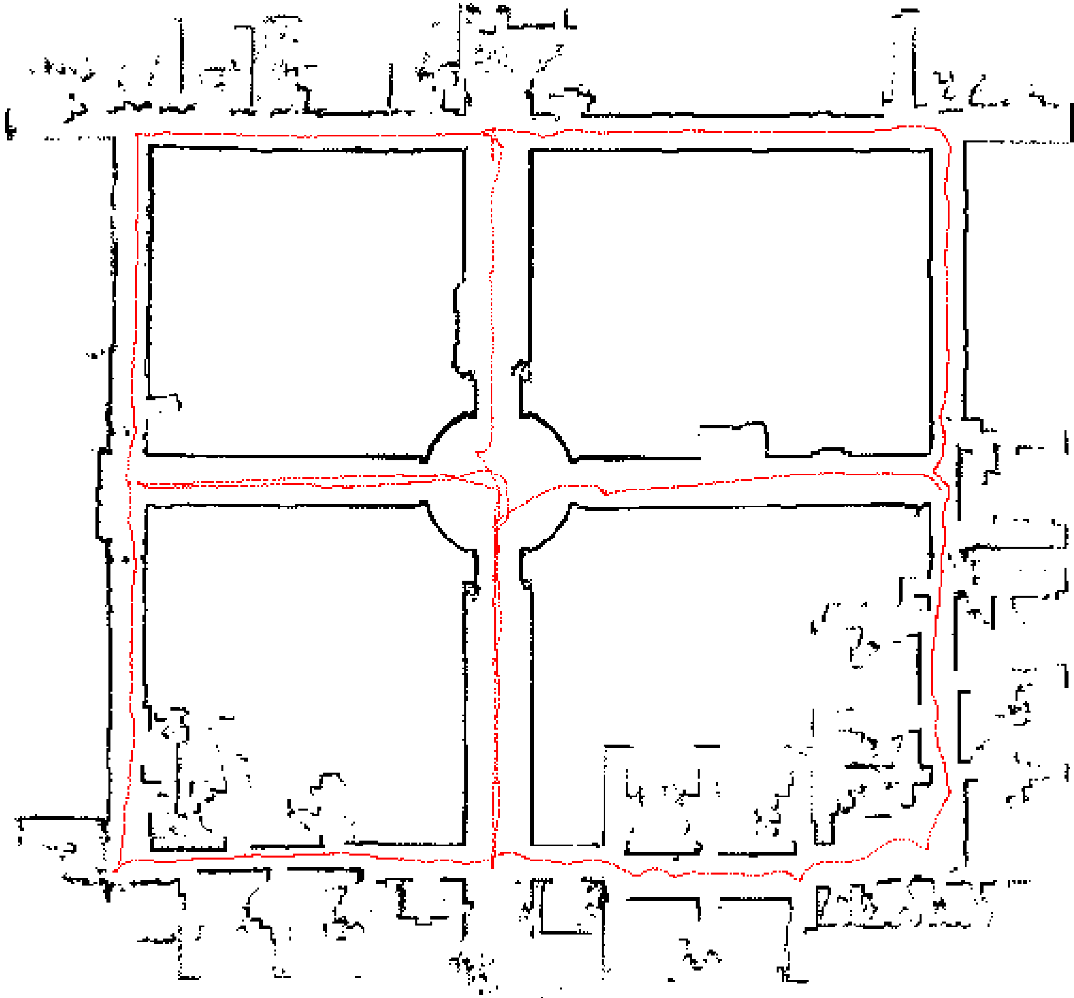
Estimating Place Layout

- Local displacement propagate to global place layout.
  - Loop-closings are especially helpful.
- Greedy hill-climbing search converges quickly to a maximum likelihood layout.



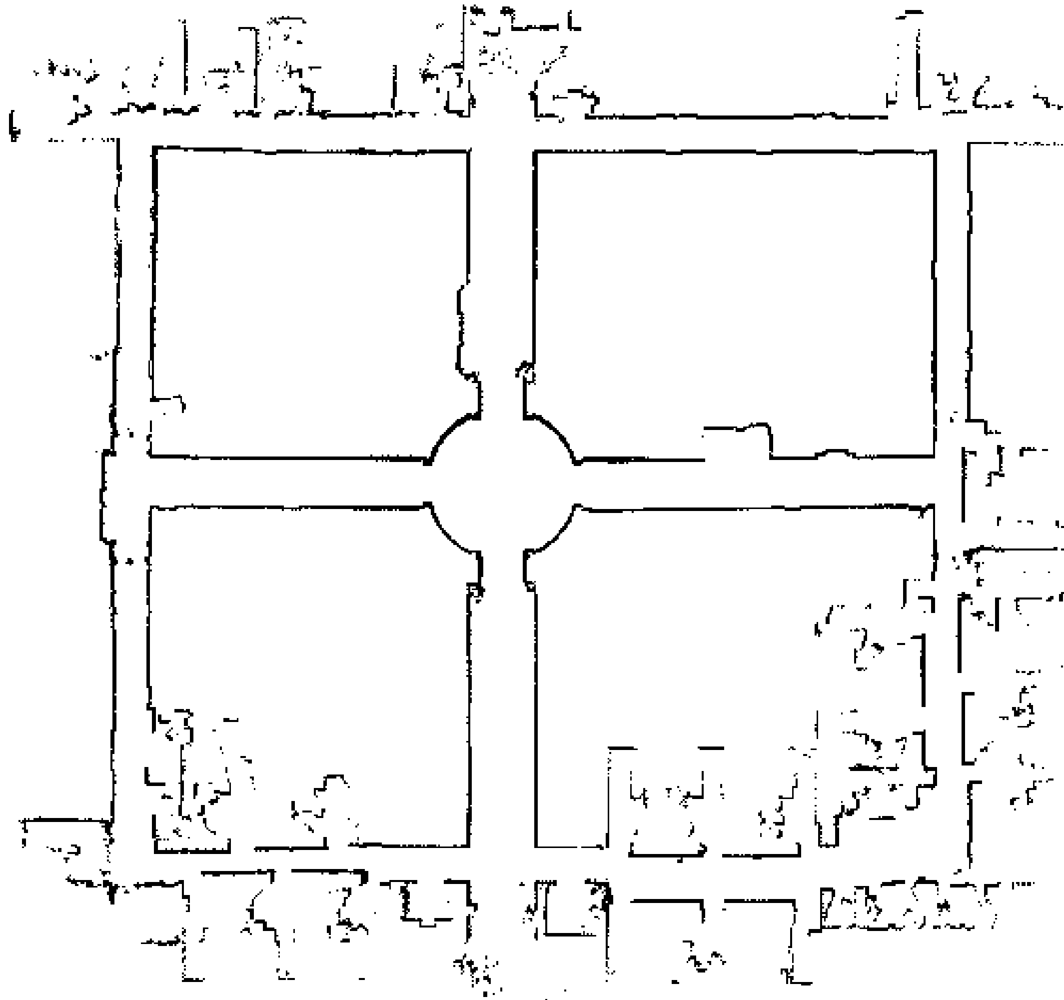
Estimating Robot Poses

- Given a max likelihood layout
- Use SLAM-corrected odometry between poses in each segment.
- Interpolate poses $x_{i,j}$ between fixed anchors in place neighborhoods.
- Uncertainty increases with distance from anchor poses.



Estimating the Global Map

- The pose distribution is a highly accurate proposal distribution.
- Treat it as providing corrected odometry.
- Now do SLAM in the global frame of reference.



Next

- Undirected graphical models
  - Markov Random Fields
- Inference in graphical models