ECE 6504: Advanced Topics in Machine Learning

Probabilistic Graphical Models and Large-Scale Learning

Topics

- Bayes Nets
  - (Finish) Structure Learning

Readings: KF 18.4; Barber 9.5, 10.4

Dhruv Batra

Virginia Tech

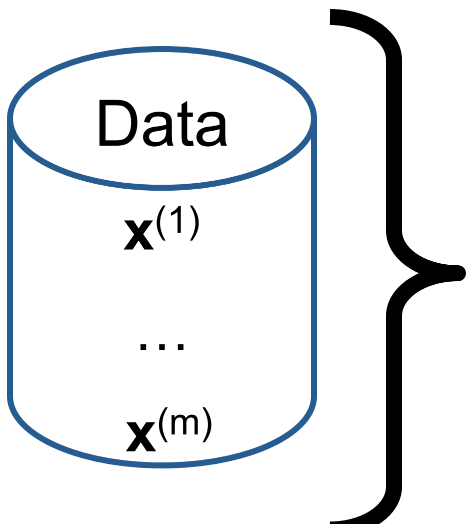
Administrativia

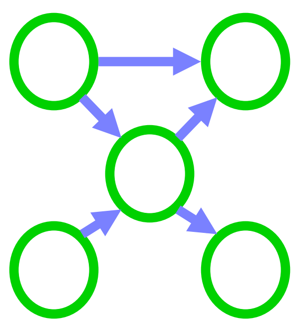
- HW1
  - Out
  - Due in 2 weeks: ~~Feb 17~~, Feb 19, 11:59pm
  - Please please please please start early
  - Implementation: TAN, structure + parameter learning
  - Please post questions on Scholar Forum.

Recap of Last Time

Learning Bayes nets

	Known structure	Unknown structure
Fully observable data	Very easy	Hard
Missing data	Somewhat easy (EM)	Very very hard





structure

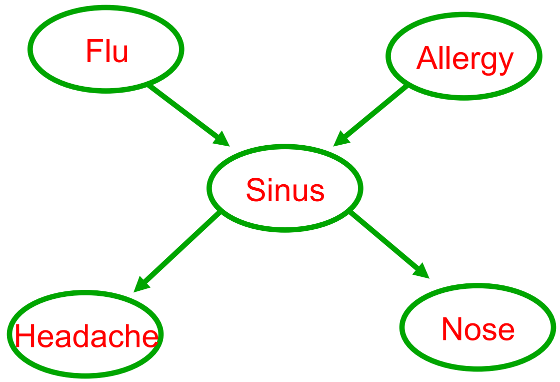
+

CPTs – $P(X_i \mid \mathbf{Pa}_{X_i})$

parameters

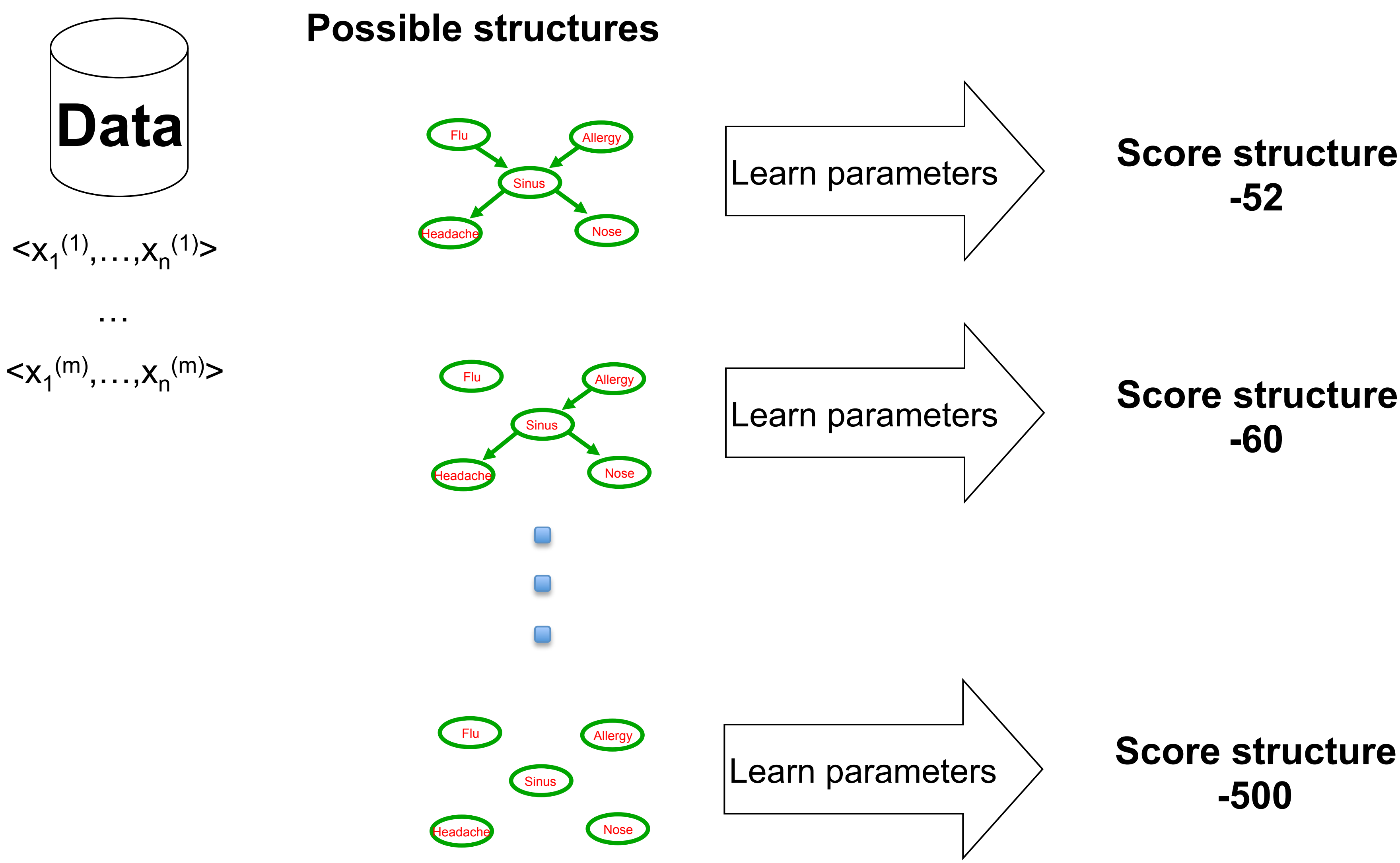
Types of Errors

- Truth:



- Recovered:

Score-based approach



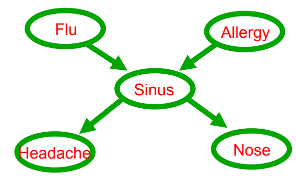
How many graphs?

- N vertices.
- How many (undirected) graphs?
- How many (undirected) trees?

What's a good score?

- Score(G) = log-likelihood(G : D, θ_{MLE})
  = logP(D | θ_{MLE}, G)

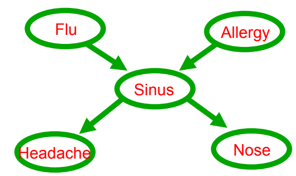
Information-theoretic interpretation of Maximum Likelihood Score



$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \mathbf{Pa}_{X_i}) - m \sum_i \hat{H}(X_i)$$

- Implications:
  - Intuitive: higher mutual info → higher score
  - Decomposes over families in BN (node and it's parents)
  - Same score for I-equivalent structures!

Log-Likelihood Score Overfits



$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \mathbf{Pa}_{X_i}) - m \sum_i \hat{H}(X_i)$$

- Adding an edge only improves score!
  - Thus, MLE = complete graph

- Two fixes:
  - Restrict space of graphs
    - say only d parents allowed (d=1 → trees)
  - Put priors on graphs
    - Prefer sparser graphs

Chow-Liu tree learning algorithm 1

- For each pair of variables X_i, X_j
  - Compute empirical distribution:

$$\widehat{P}(x_i, x_j) = \frac{\text{Count}(x_i, x_j)}{m}$$

  - Compute mutual information:

$$\widehat{I}(X_i, X_j) = \sum_{x_i, x_j} \widehat{P}(x_i, x_j) \log \frac{\widehat{P}(x_i, x_j)}{\widehat{P}(x_i)\widehat{P}(x_j)}$$

- Define a graph
  - Nodes $X_1, \ldots, X_n$
  - Edge (i,j) gets weight $\widehat{I}(X_i, X_j)$

Chow-Liu tree learning algorithm 2

- Optimal tree BN
  - Compute maximum weight spanning tree
  - Directions in BN: pick any node as root, and direct edges away from root
    - breadth-first-search defines directions

Can we extend Chow-Liu?

- Tree augmented naïve Bayes (TAN) [Friedman et al. '97]
  - Naïve Bayes model overcounts, because correlation between features not considered
  - Same as Chow-Liu, but score edges with:

$$\widehat{I}(X_i, X_j \mid C) = \sum_{c, x_i, x_j} \widehat{P}(c, x_i, x_j) \log \frac{\widehat{P}(x_i, x_j \mid c)}{\widehat{P}(x_i \mid c)\widehat{P}(x_j \mid c)}$$

Plan for today

- (Finish) BN Structure Learning
  - Bayesian Score
  - Heuristic Search
  - Efficient tricks with decomposable scores

Bayesian score

- Bayesian view → Prior distributions:
  - Over structures
  - Over parameters of a structure

- Posterior over structures given data:

$$\log P(\mathcal{G} \mid D) \propto \log P(\mathcal{G}) + \log \int_{\theta_{\mathcal{G}}} P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}} | \mathcal{G}) d\theta_{\mathcal{G}}$$

Structure Prior

$$\log P(\mathcal{G} \mid D) \propto \log P(\mathcal{G}) + \log \int_{\theta_{\mathcal{G}}} P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}} | \mathcal{G}) d\theta_{\mathcal{G}}$$

- Common choices:
  - Uniform: P(*G*) α c
  - Sparsity prior: P(*G*) α $c^{|G|}$
  - Prior penalizing number of parameters
  - P(G) should decompose like the family score

Parameter Prior and Integrals

$$\log P(\mathcal{G} \mid D) \propto \log P(\mathcal{G}) + \log \int_{\theta_{\mathcal{G}}} P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}} | \mathcal{G}) d\theta_{\mathcal{G}}$$

- Important Result:
  - If $P(\theta_G \mid G)$ is Dirichlet, then integral has closed form!
  - And it factorizes according to families in G

$$P(D \mid G) = \prod_i \prod_{pa_i^G}$$

Dirichlet marginal likelihood for multinomial $P(X_i \mid pa_i)$

$$\frac{\Gamma(\alpha(pa_i^G))}{\Gamma(\alpha(pa_i^G) + N(pa_i^G))} \prod_{x_i} \frac{\Gamma(\alpha(x_i, pa_i^G) + N(x_i, pa_i^G))}{\Gamma(\alpha(x_i, pa_i^G))}$$

Parameter Prior and Integrals

$$\log P(\mathcal{G} \mid D) \propto \log P(\mathcal{G}) + \log \int_{\theta_{\mathcal{G}}} P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}} | \mathcal{G}) d\theta_{\mathcal{G}}$$

- How should we choose Dirichlet hyperparameters?
  - *K2 prior*: fix an α, $P(\theta_{X_i|\mathbf{Pa}X_i})$ = Dirichlet(α,…, α)
    - K2 is "inconsistent"

BDe Prior

$$\log P(\mathcal{G} \mid D) \propto \log P(\mathcal{G}) + \log \int_{\theta_{\mathcal{G}}} P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}} | \mathcal{G}) d\theta_{\mathcal{G}}$$

- BDe Prior
  - Remember that Dirichlet parameters are analogous to "fictitious samples"
  - Pick a fictitious sample size m'
  - Pick a "prior" BN
    - Usually independent (product of marginals)
  - Compute P(X_i,$\mathbf{Pa}_{X_i}$) under this prior BN

- **BDe prior**:

- Has consistency property

Chow-Liu for Bayesian score

- Edge weight $w_{X_j \rightarrow X_i}$ is advantage of adding X_j as parent for X_i

- Now have a directed graph, need directed spanning forest
  - Note that adding an edge can hurt Bayesian score – choose forest not tree
  - Maximum spanning forest algorithm works

Structure learning for general graphs

- In a tree, a node only has one parent

- **Theorem**:
  – The problem of learning a BN structure with at most d parents is NP-hard for any (fixed) $d \geq 2$

- Most structure learning approaches use heuristics
  – Exploit score decomposition
  – (Quickly) Describe two heuristics that exploit decomposition in different ways

Structure learning using local search

Starting from Chow-Liu tree

Local search,
possible moves:

Only if acyclic!!!

- Add edge
- Delete edge
- Invert edge

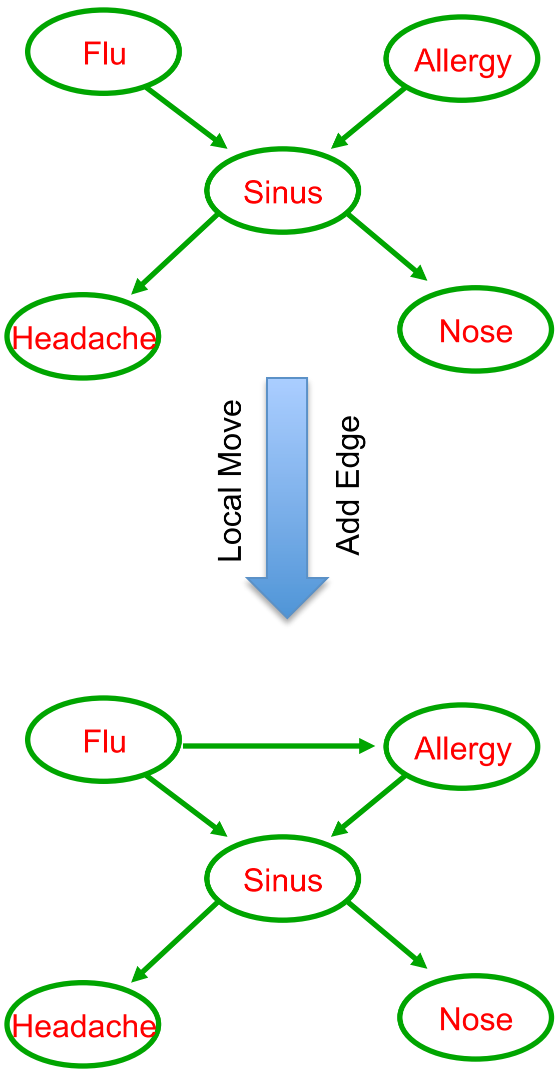
Select using favorite score

Structure learning using local search

- Problems:
  - Local maximum
  - Plateau

- Strategies
  - Random restart
  - Tabu list

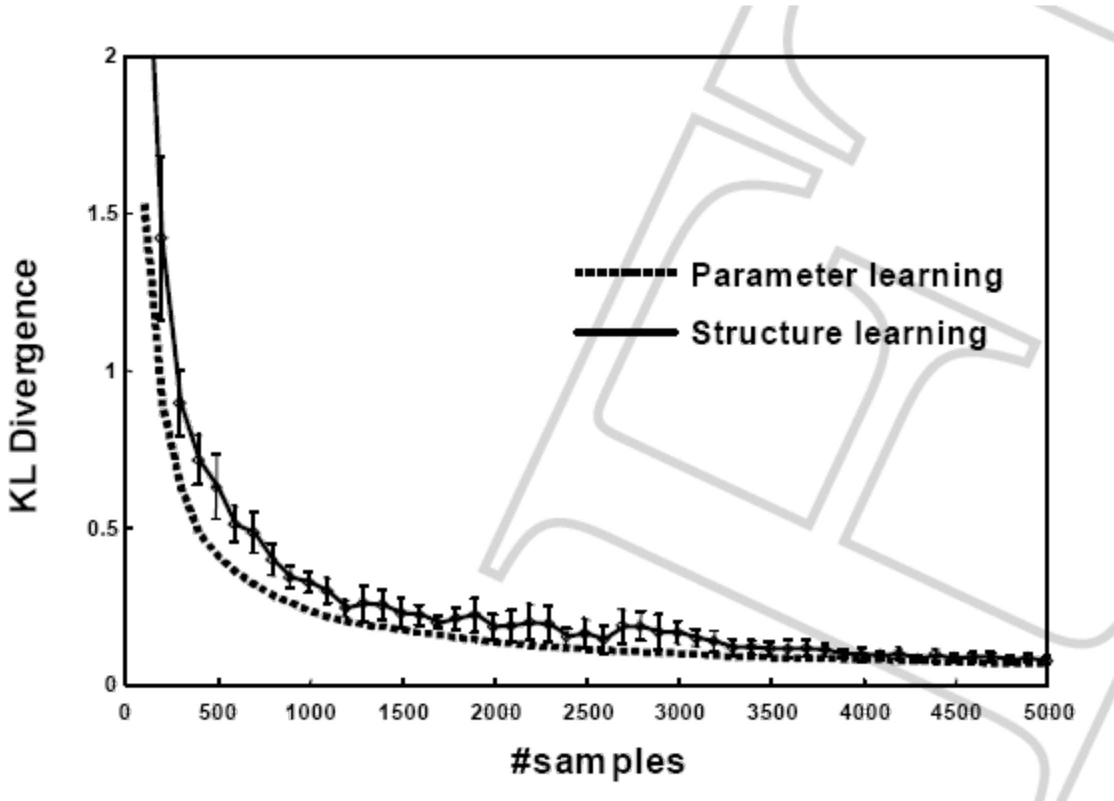
Exploit score decomposition in local search



- Add edge and delete edge:
  - Only rescore one family!

- Reverse edge
  - Rescore only two families

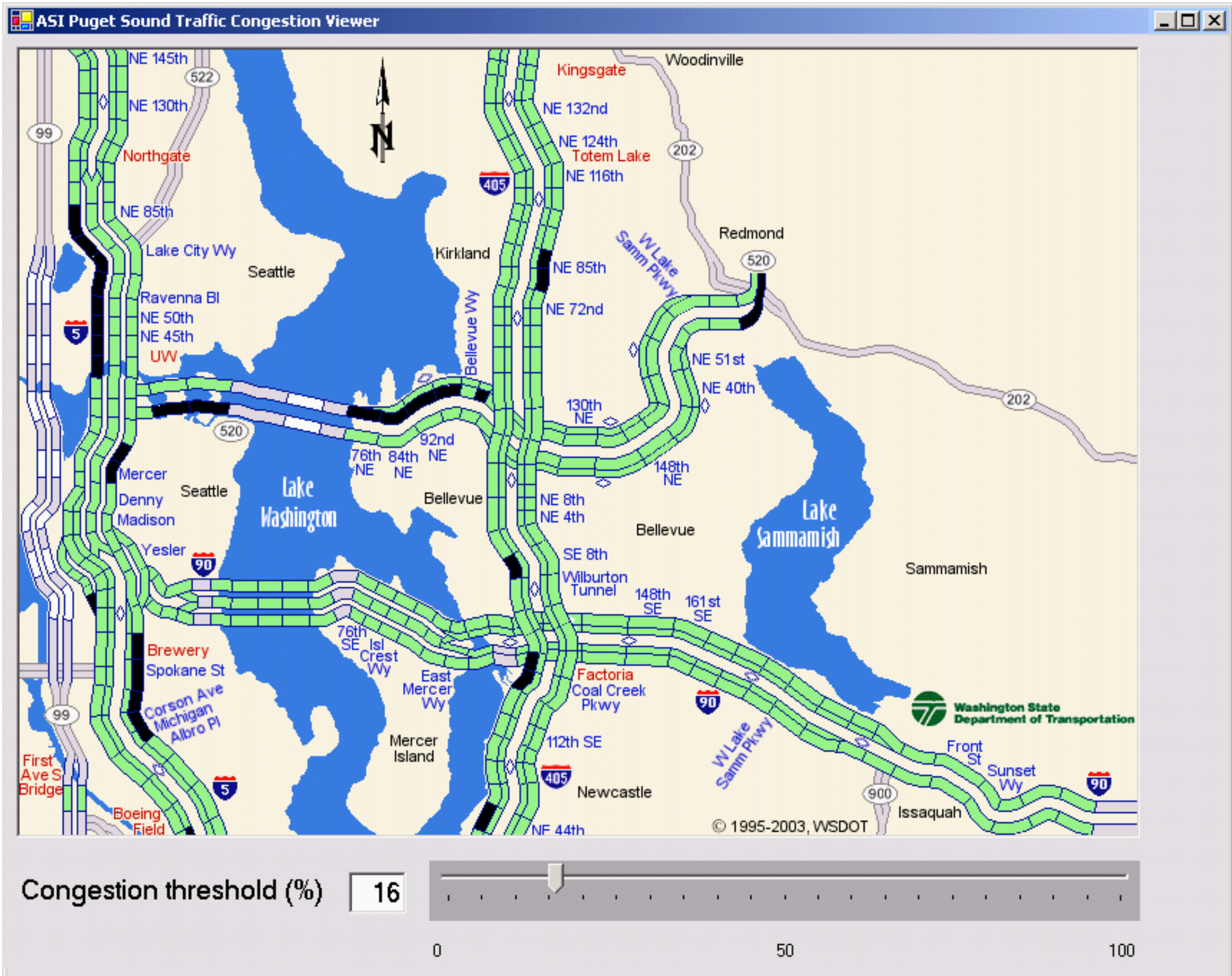
Example



Alarm network

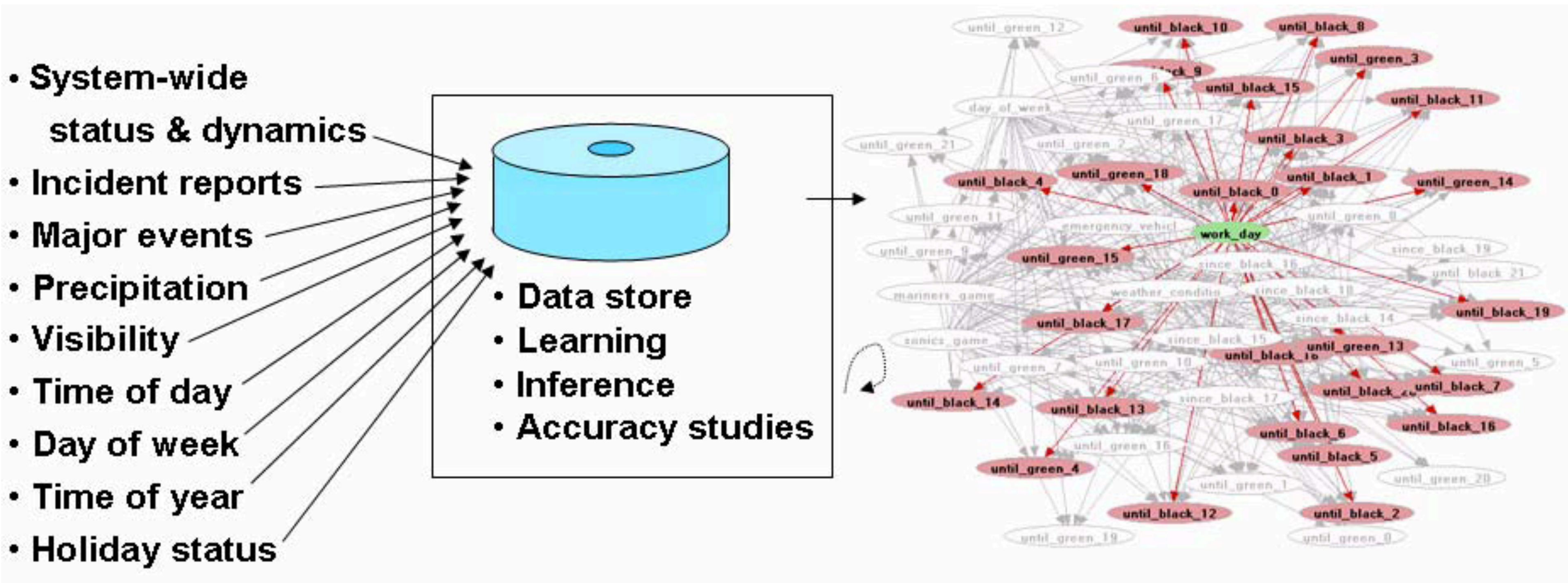
Example

- JamBayes [Horvitz et al UAI05]

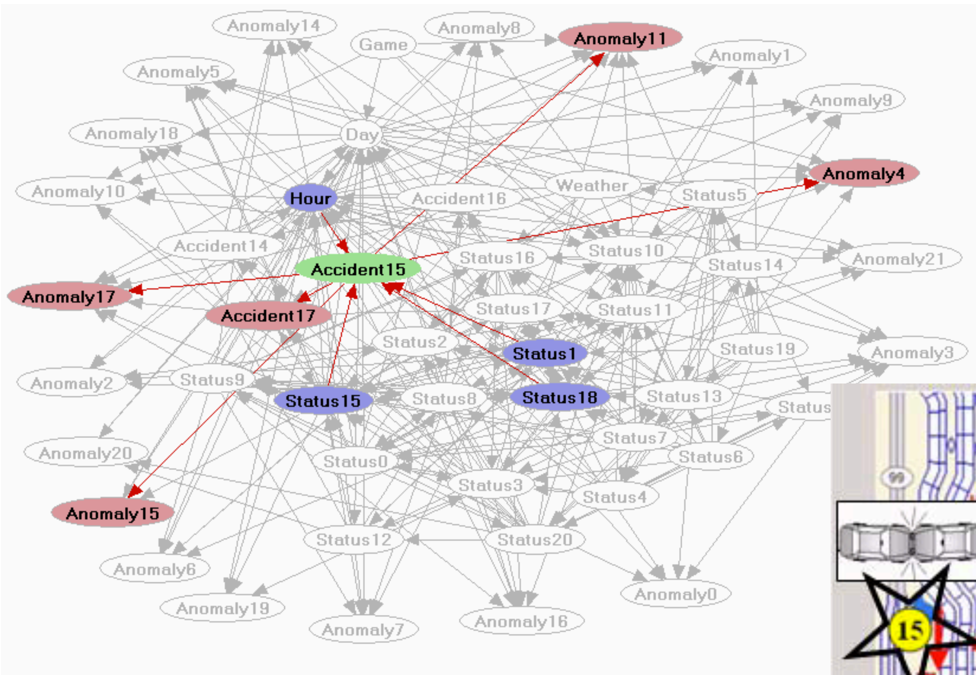


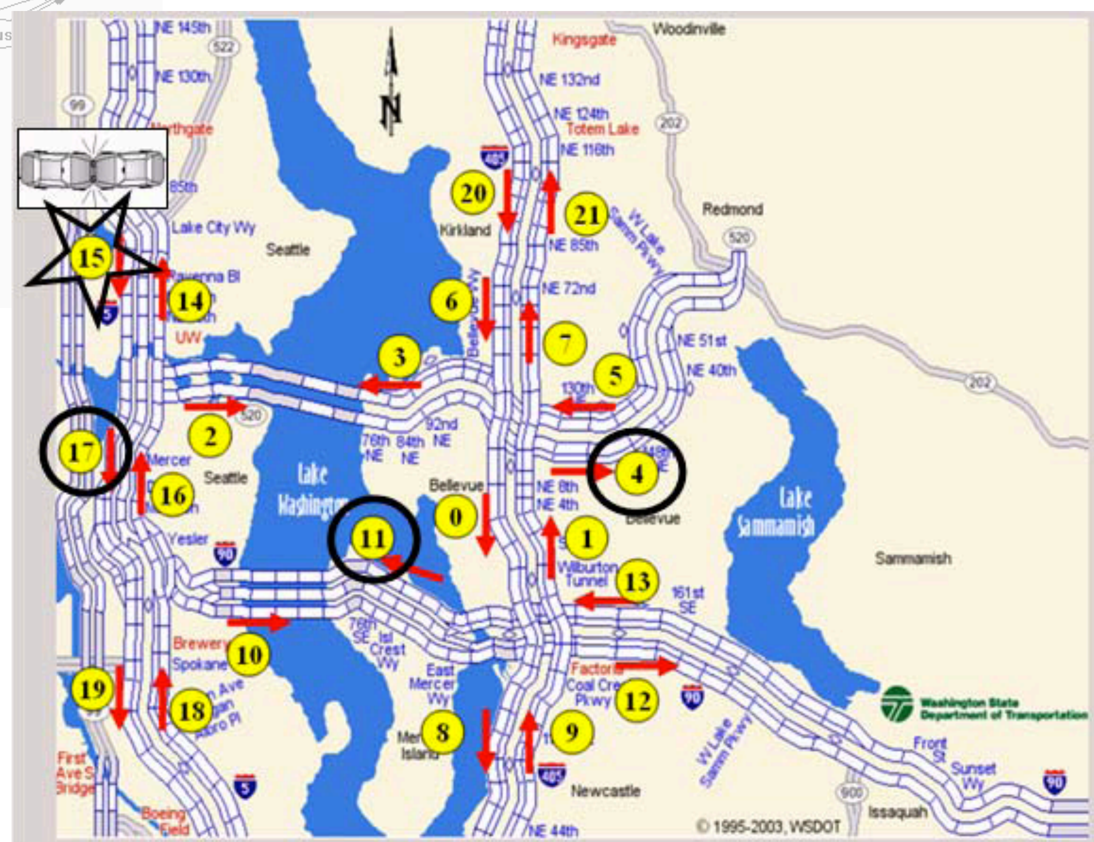
Example

- JamBayes [Horvitz et al UAI05]



Example


Anomaly14
Game
Anomaly8
Anomaly11
Anomaly1
Anomaly5
Anomaly9
Anomaly18
Day
Anomaly4
Anomaly10
Hour
Accident16
Weather
Status5
Accident14
Status16
Status10
Status14
Anomaly21
Accident15
Anomaly17
Accident17
Status17
Status11
Status2
Status1
Status19
Anomaly3
Anomaly2
Status9
Status15
Status8
Status18
Status13
Status
Status7
Anomaly20
Status0
Status6
Status3
Status4
Anomaly15
Status12
Status20
Anomaly6
Anomaly0
Anomaly19
Anomaly7
Anomaly16


N
Woodinville
Kingsgate
NE 145th
NE 130th
NE 132nd
NE 124th
Totem Lake
NE 116th
Redmond
Kirkland
Seattle
Lake City Wy
NE 85th
NE 72nd
UW
NE 51st
NE 40th
Mercer
Seattle
Lake Washington
Bellevue
NE 8th
NE 4th
Lake Sammamish
Sammamish
Yesler
Wilburton Tunnel
Brewery
Newcastle
Issaquah
Washington State Department of Transportation
© 1995-2003, WSDOT
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

Bayesian model averaging

- So far, we have selected a single structure
- But, if you are really Bayesian, must average over structures
  - Similar to averaging over parameters

$$\log P(D \mid \mathcal{G}) = \log \int_{\theta_{\mathcal{G}}} P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}} | \mathcal{G}) d\theta_{\mathcal{G}}$$

BN: Structure Learning: What you need to know

- Score-based approach
  - Log-likelihood score
    - Use θ_{MLE}
    - Information theoretic interpretation
    - Overfits! Adding edges only helps
  - Bayesian Score
    - Priors over structure and priors over parameters for a structure
    - If dirichlet closed form expression for P(D|G)
    - K2 dirichlet not enough; Need BDe for consistency
- Structure Search
  - For trees
    - Chow-Liu: max-weight spanning tree
    - Can be extended to forests and TAN
  - General graphs
    - Heuristic Search
    - Efficiency tricks due to decomposable score