CAP6938-02
Plan, Activity, and Intent Recognition

Lecture 6:

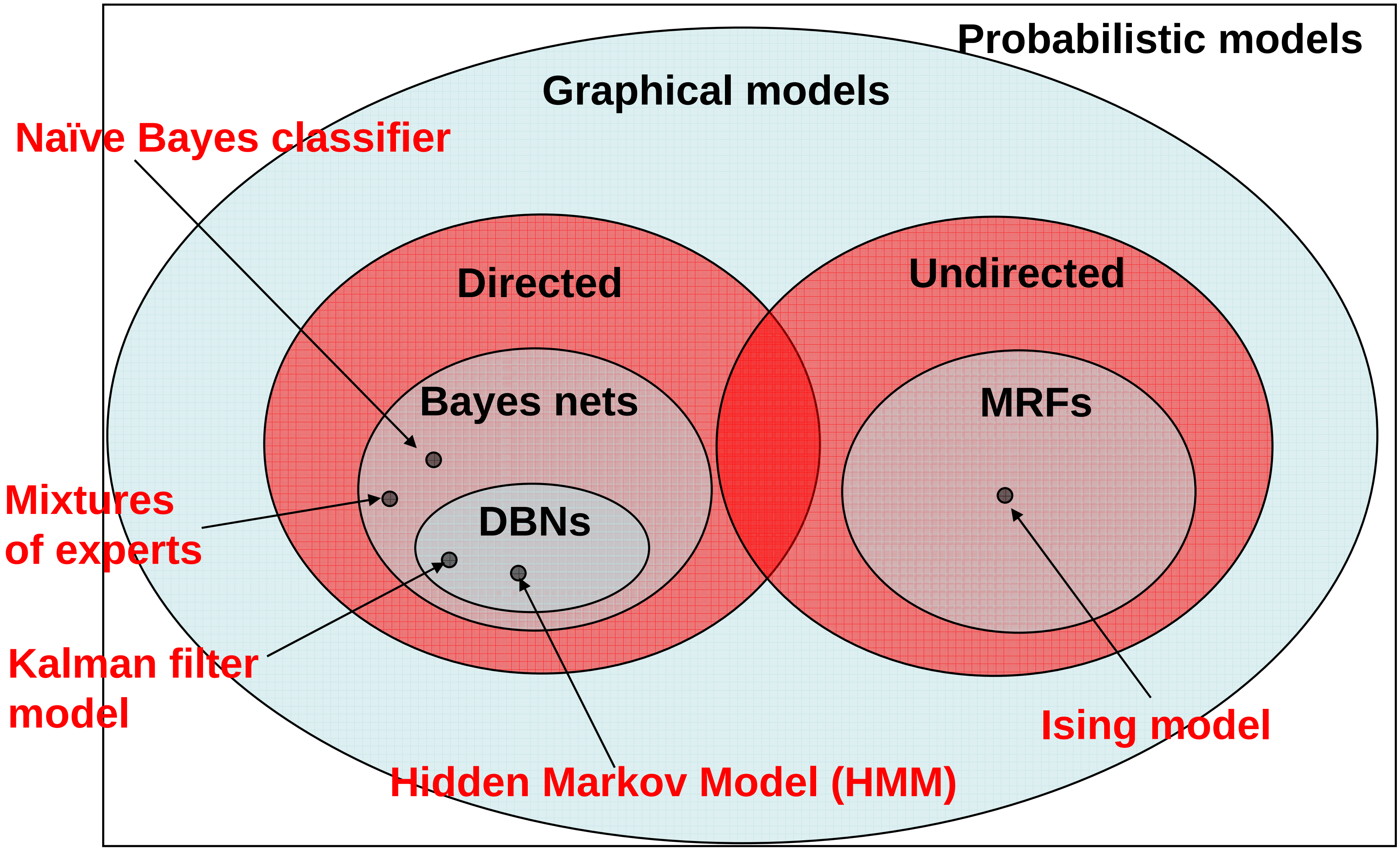
Introduction to Graphical Models (Part 1)

Instructor: Dr. Gita Sukthankar
Email: gitars@eecs.ucf.edu
Schedule: T & Th 1:30-2:45pm
Location: CL1 212
Office Hours (HEC 232):
T 3-4:30pm, Th 10-11:30am

Reminder

- Homework:
  - Start project implementation
  - Hand in 1-2 page specification of how your implementation is going to work (Sept 13)
  - Be as detailed as possible about the input and output of the system
  - Timeline:
    - This week: specification
    - Sept 20: demo your system (no writeup)
    - Sept 27: present evaluation of your system

Instances of graphical models



Graphical Model Definitions

- Representation: compactly representing joint probability distributions
- Inference: determine hidden states of a system given noisy observations
- Learning: how to estimate parameters and structure of the model
- Decision theory: how to convert belief into action

Outline (today)

Part 1: Representation

- Fundamentals
- Bayes nets
- D-separation
- Dynamic Bayes nets
  - Hidden Markov models
- Commonly used variations

Future Topics

- Part 2: Exact inference in DBNs; learning parameters from data
- Part 3: Approximate inference in DBNs
- Part 4: Undirected graphical models (Markov Random Fields)
- Part 5: Determining model structure from data (emerging research area)
- Part 6: Decision-making models (MDP, POMDP)

Applications

- Too numerous to name....
- Bayes nets:
  - User interfaces
  - Medical diagnosis
  - Fault diagnosis
- Dynamic Bayes nets:
  - Speech recognition (hidden Markov models)
  - Command line interfaces
  - Motion tracking and prediction

Bayes Net Toolbox

- Available at sourceforge: http://bnt.sourceforge.net
- Developed by Kevin Murphy
- Open-source collection of Matlab functions for inference and learning of (directed) graphical models
- Over 100,000 hits and about 30,000 downloads since May 2000
- About 43,000 lines of code (of which 8,000 are comments)

Probabilities

- Probability distribution $P(X|\xi)$
  - X is a random variable
    - Discrete
    - Continuous
  - ξ is background state of information

Discrete Random Variables

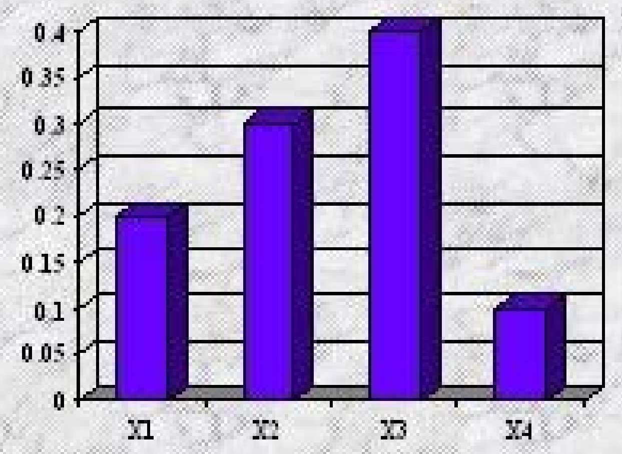
- Finite set of possible outcomes

$$X \in \{x_1, x_2, x_3, \ldots, x_n\}$$

$$P(x_i) \geq 0$$

$$\sum_{i=1}^{n} P(x_i) = 1$$

X binary: $P(x) + P(\bar{x}) = 1$



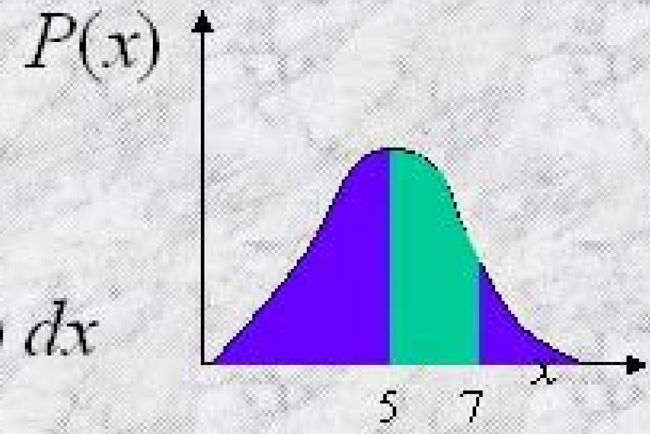
Continuous Random Variable

- Probability distribution (density function) over continuous values

$$X \in [0,10] \qquad P(x) \geq 0$$

$$\int_{0}^{10} P(x)\,dx = 1$$

$$P(5 \leq x \leq 7) = \int_{5}^{7} P(x)\,dx$$



More Probabilities

- Conditional

$$P(x \mid y) \equiv P(X = x \mid Y = y)$$

  - Probability that $X=x$ given we know that $Y=y$
- Joint

$$P(x, y) \equiv P(X = x \wedge Y = y)$$

  - Probability that both $X=x$ and Y=y

Rules of Probability

- Product Rule

$$P(X,Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

- Marginalization

$$P(Y) = \sum_{i=1}^{n} P(Y, x_i)$$

X binary: $P(Y) = P(Y, x) + P(Y, \bar{x})$

Bayes Rule

$$P(H,E) = P(H \mid E)P(E) = P(E \mid H)P(H)$$

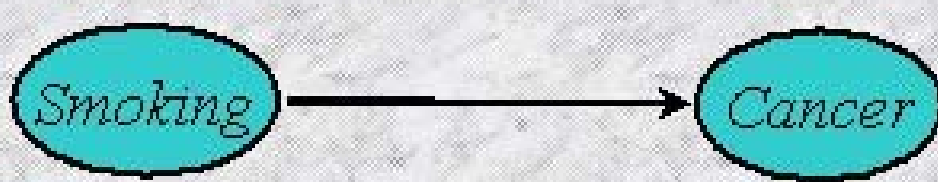
$$\boxed{P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}}$$

$$P(h \mid e) = \frac{P(e \mid h)P(h)}{P(e,h) + P(e,\bar{h})}$$

$$= \frac{P(e \mid h)P(h)}{P(e \mid h)P(h) + P(e \mid \bar{h})P(\bar{h})}$$

Bayesian Networks

$S \in \{no, light, heavy\}$



$P(S=no)$	0.80
$P(S=light)$	0.15
$P(S=heavy)$	0.05

$C \in \{none, benign, malignant\}$

Smoking=	*no*	*light*	*heavy*
$P(C=none)$	0.96	0.88	0.60
$P(C=benign)$	0.03	0.08	0.25
$P(C=malig)$	0.01	0.04	0.15

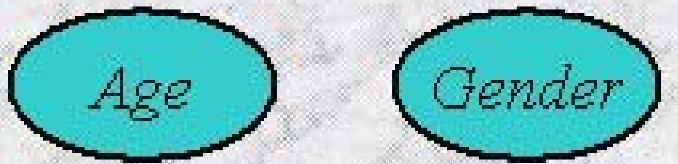
Marginalization

$S\Downarrow$ $C\Rightarrow$	*none*	*benign*	*malig*	total
no	0.768	0.024	0.008	.80
light	0.132	0.012	0.006	.15
heavy	0.035	0.010	0.005	.05
total	0.935	0.046	0.019	

P(Smoke)

P(Cancer)

Independence



Age and *Gender* are independent.

$$P(A,G) = P(G)P(A)$$

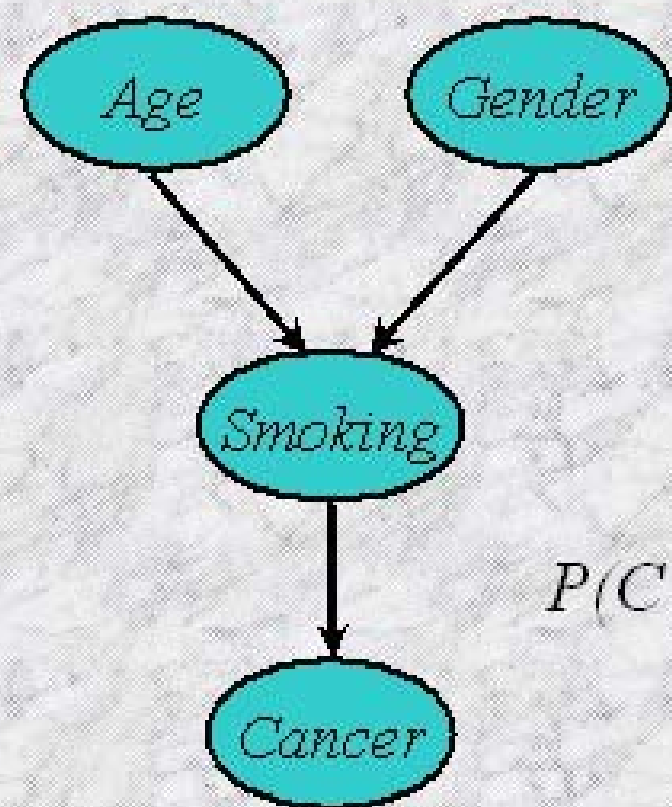
$$P(A|G) = P(A) \quad A \perp G$$
$$P(G|A) = P(G) \quad G \perp A$$

$$P(A,G) = P(G|A)\,P(A) = P(G)P(A)$$
$$P(A,G) = P(A|G)\,P(G) = P(A)P(G)$$

Independence in directed graphs

- A path between A and B is blocked if there is a node C such that:
  - The path has converging arrows at C and none of C or its descendants are given.
  - The path does not have converging arrows at C and C is given.
- If all paths between them are blocked, then A and B are independent. This kind of separation is called d-separation.

Conditional Independence



Cancer is independent of *Age* and *Gender* given *Smoking*.

$$P(C|A,G,S) = P(C|S) \quad C \perp A,G \mid S$$

General Product (Chain) Rule for Bayesian Networks

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \boldsymbol{Pa_i})$$

$$\boldsymbol{Pa_i} = parents(X_i)$$

Why are Bayes nets useful?

- Graph structure supports:
  - Modular representation of knowledge
  - Local, distributed algorithms for inference and learning
  - Intuitive (possibly causal) interpretation

- Factored representation may have exponentially fewer parameters than full joint $P(X_1,\ldots,X_n)$ =>
  - lower sample complexity (less data for learning)
  - lower time complexity (less time for inference)

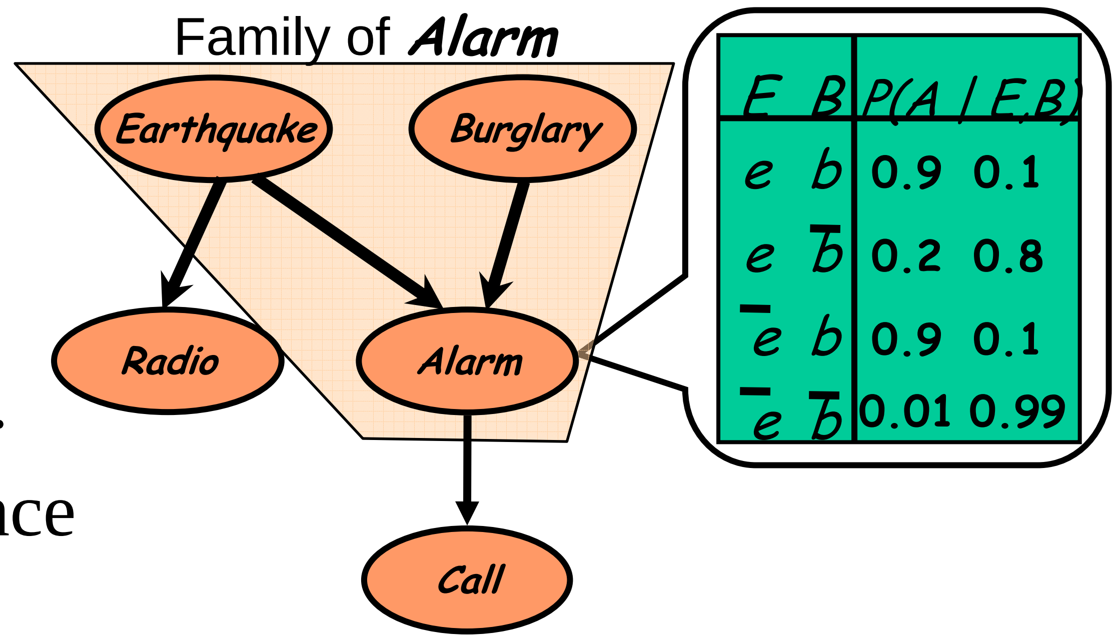
What is a Bayes (belief) net?

Compact representation of joint probability distributions via conditional independence

Qualitative part:

Directed acyclic graph (DAG)

- Nodes - random vars.
- Edges - direct influence



E	B	P(A | E,B)	
e	b	0.9	0.1
e	$\bar{b}$	0.2	0.8
$\bar{e}$	b	0.9	0.1
$\bar{e}$	$\bar{b}$	0.01	0.99

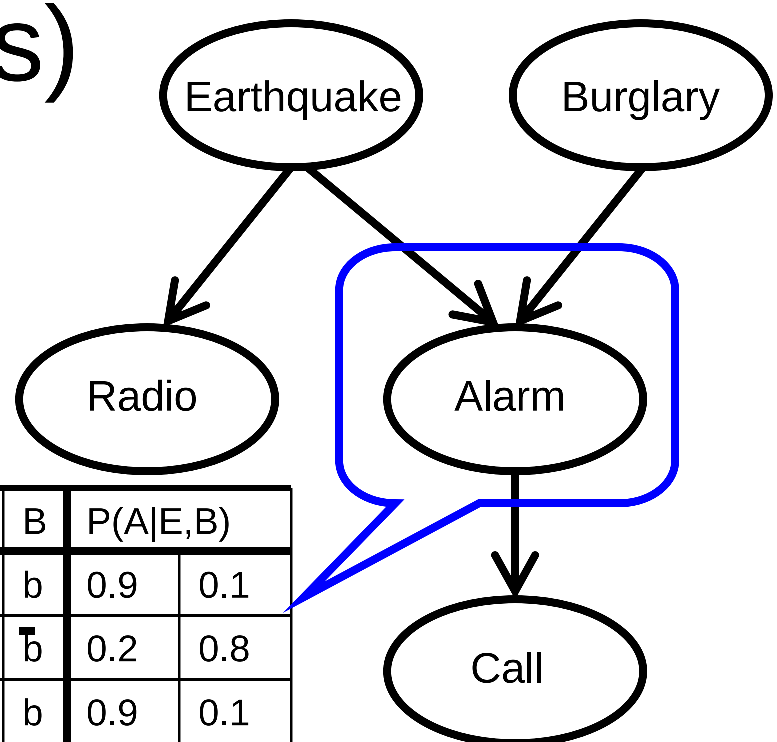
Together:
Define a unique distribution in a factored form

Quantitative part:
Set of conditional probability distributions

$$P(B,E,A,C,R) = P(B)P(E)P(A \mid B,E)P(R \mid E)P(C \mid A)$$

Figure from N. Friedman

Conditional probability distributions (CPDs)



- Each node specifies a distribution over its values given its parents values $P(X_i \mid X_{Pa_i})$
- Full table needs 2^5-1=31 parameters, BN needs 10

E	B	P(A|E,B)	
e	b	0.9	0.1
e	$\bar{b}$	0.2	0.8
$\bar{e}$	b	0.9	0.1
$\bar{e}$	$\bar{b}$	0.01	0.99

$$\underbrace{P(B,E,A,R,C)}_{31}$$
$$= \underbrace{P(B)}_{1}\underbrace{P(E|B)}_{2}\underbrace{P(A|B,E)}_{4}\underbrace{P(R|A,B,E)}_{8}\underbrace{P(C|R,A,B,E)}_{16}$$
$$= \underbrace{P(B)}_{1}\underbrace{P(E)}_{1}\underbrace{P(A|B,E)}_{4}\underbrace{P(R|E)}_{2}\underbrace{P(C|A)}_{2}$$

Pearl, 1988

Inference

- **Posterior probabilities**
  – Probability of any event given any evidence
- *Most likely explanation*
  – *Scenario that explains evidence*
- *Rational decision making*
  – *Maximize expected utility*
  – *Value of Information*
- *Effect of intervention*
  – *Causal analysis*

Explaining away effect

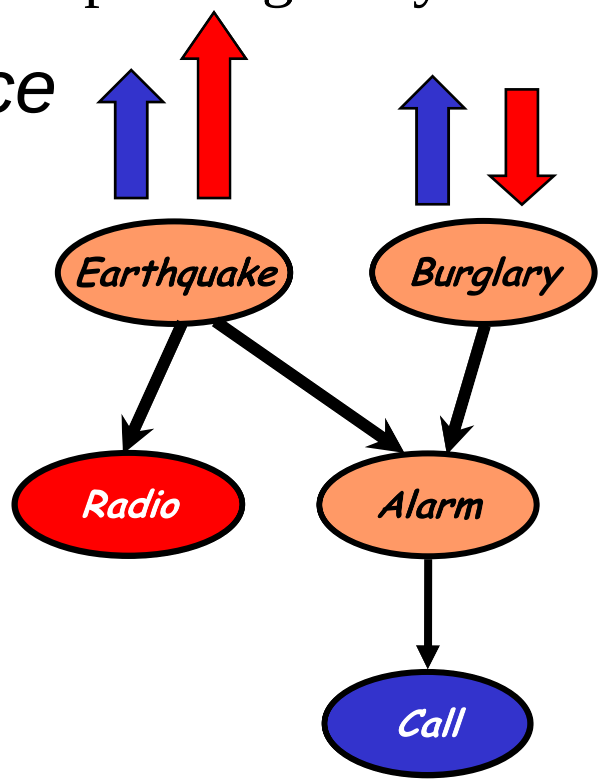


Figure from N. Friedman

A real Bayes net: Alarm

Domain: Monitoring Intensive-Care Patients

- 37 variables
- 509 parameters

…instead of 2^{37}

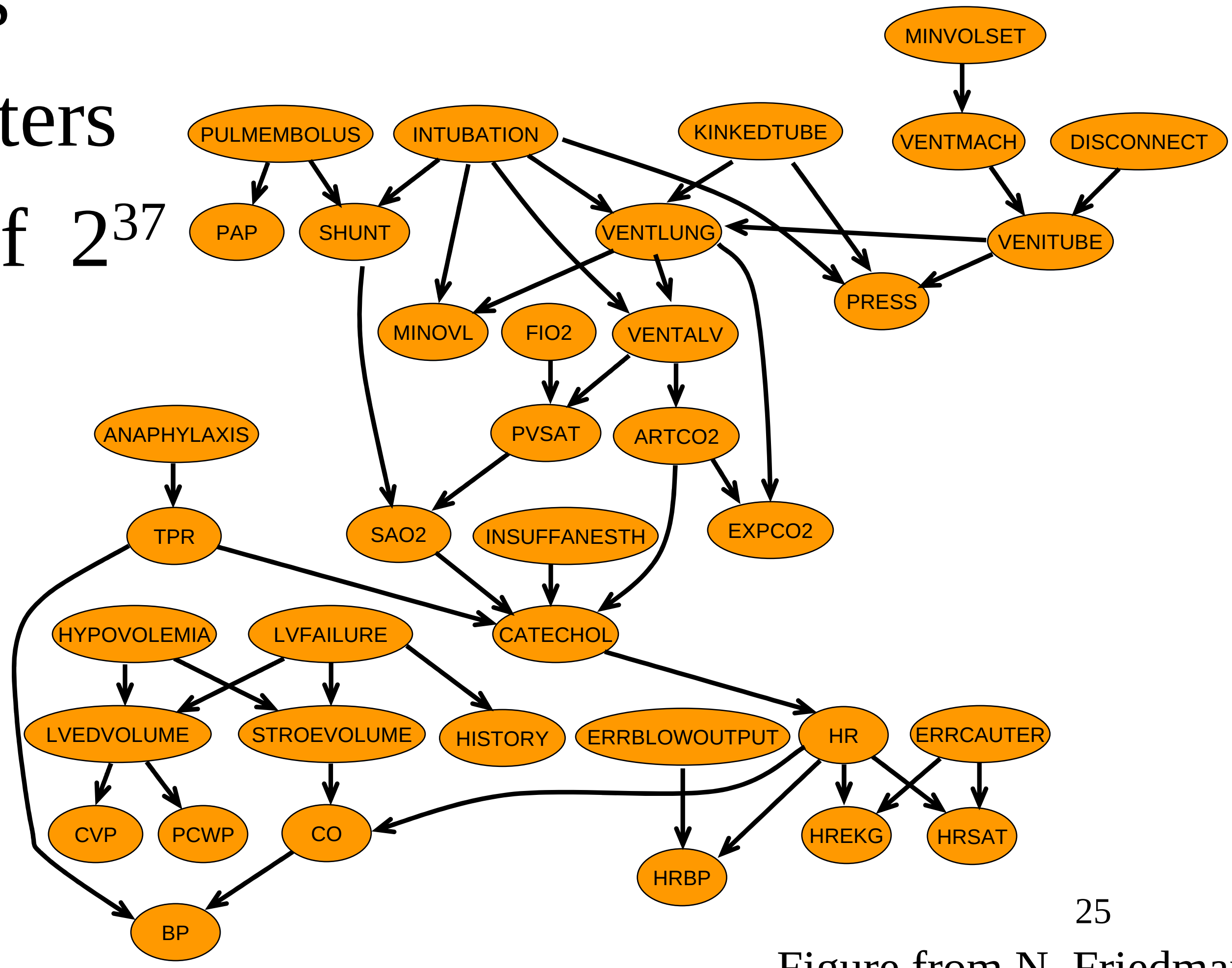


Figure from N. Friedman

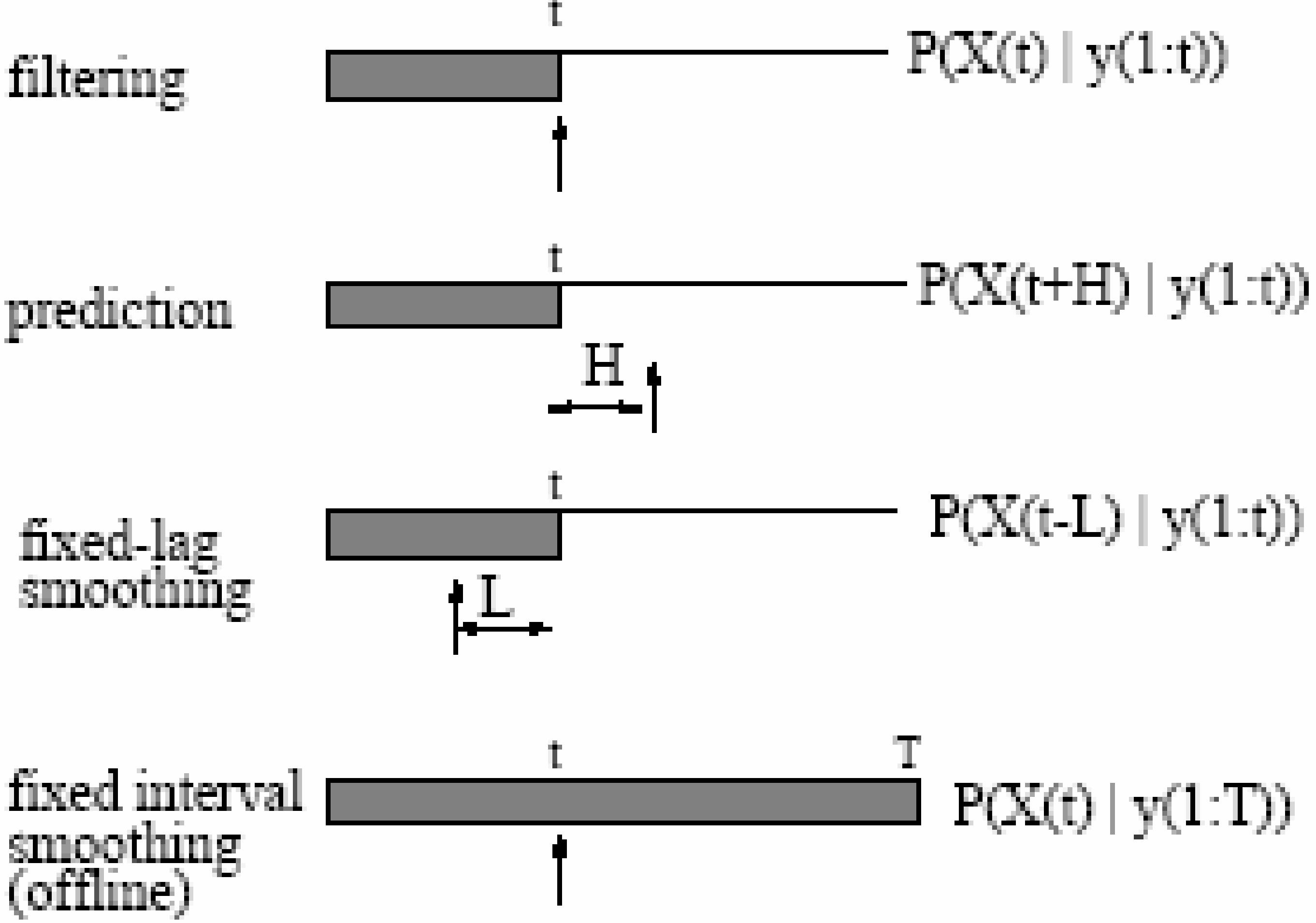
Dealing with time

- In many systems, data arrives sequentially
- Dynamic Bayes nets (DBNs) can be used to model such time-series (sequence) data
- Special cases of DBNs include
  - State-space models
  - Hidden Markov models (HMMs)

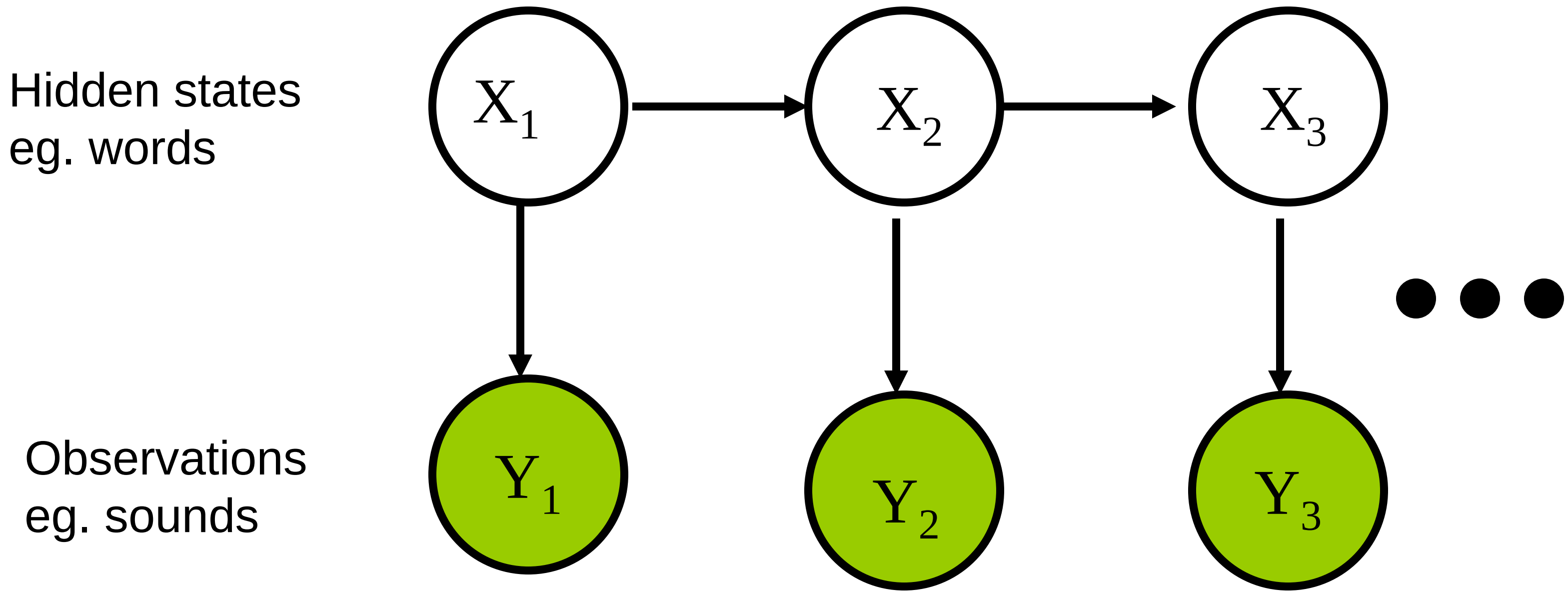
DBNs are a kind of graphical model

- In a graphical model, nodes represent random variables, and (lack of) arcs represents conditional independencies.
- Directed graphical models = Bayes nets = belief nets.
- DBNs are Bayes nets for dynamic processes.
- Informally, an arc from X_i to X_j means X_i "causes" X_j. (Graph must be acyclic!)

Kinds of inference for DBNs

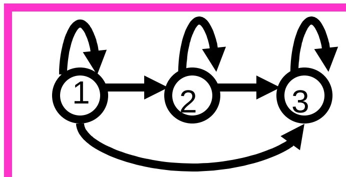


Example BN: Hidden Markov Model (HMM)

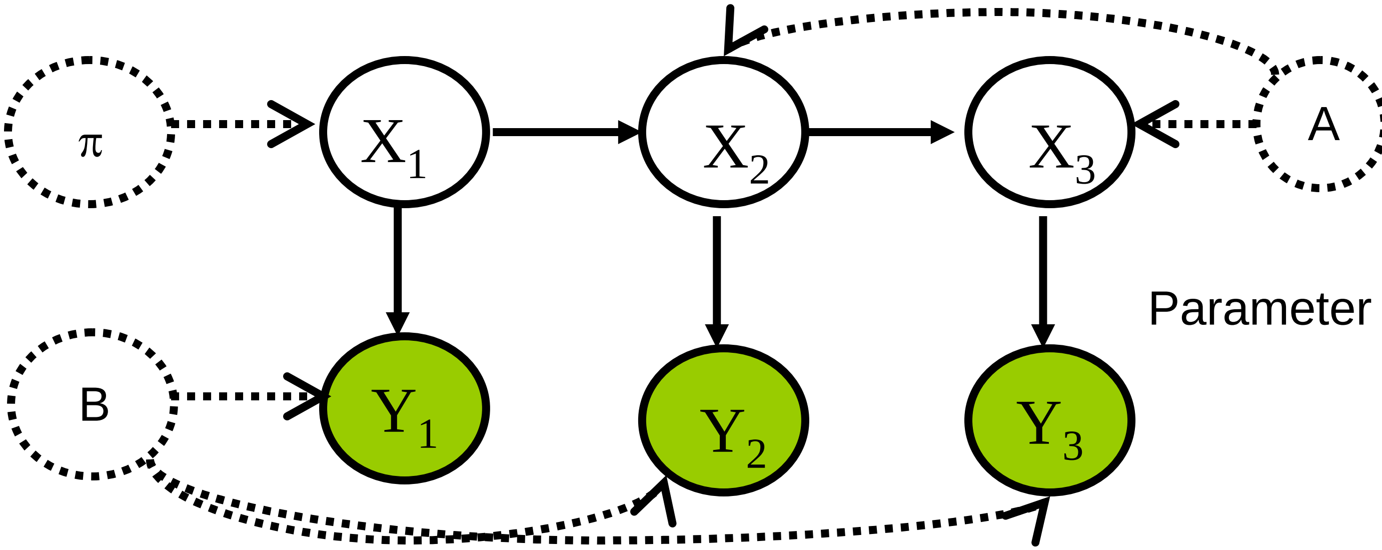


$$P(X_{1:T}, Y_{1:T}) = P(X_1)P(Y_1|X_1)$$
$$P(X_2|X_1)P(Y_2|X_2)\ldots$$

CPDs for HMMs



$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix}$$



Parameter tyeing

$$P(X_{1:T}, Y_{1:T}) = P(X_1)P(Y_1|X_1) \prod_{t=2}^{T} P(X_t|X_{t-1})P(Y_t|X_t)$$

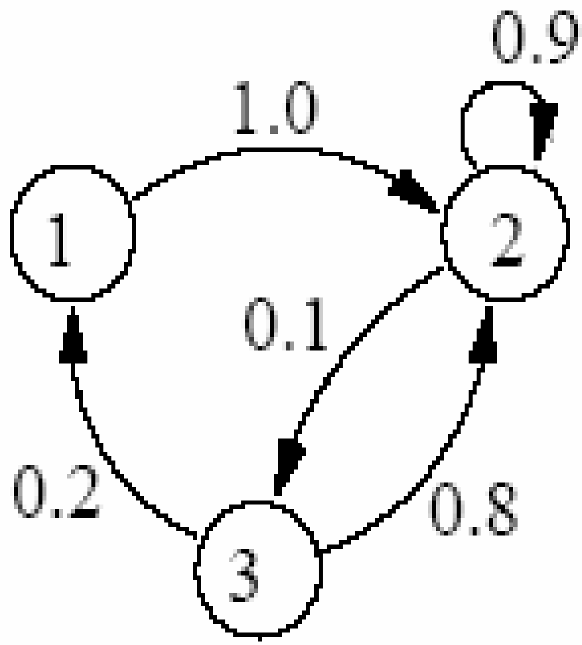
Transition matrix $P(X_t = j | X_{t-1} = i) = A(i,j)$

Observation matrix $P(Y_t = j | X_t = i) = B(i,j)$

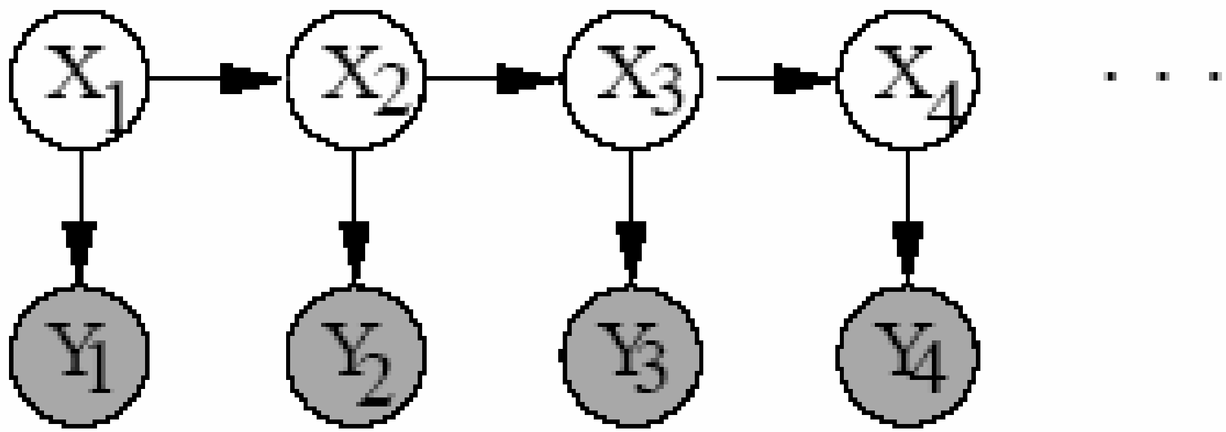
Initial state distribution $P(X_1 = i) = \pi(i)$

HMM state transition diagram

- Nodes represent states.
- There is an arrow from i to j iff $A(i,j) > 0$.



HMM represented as a DBN



- This graph encodes the assumptions

$$Y_t \perp Y_{t'} | X_t \text{ and } X_{t+1} \perp X_{t-1} | X_t \text{ (Markov)}$$

- Shaded nodes are observed, unshaded are hidden.
- Structure and parameters repeat over time.

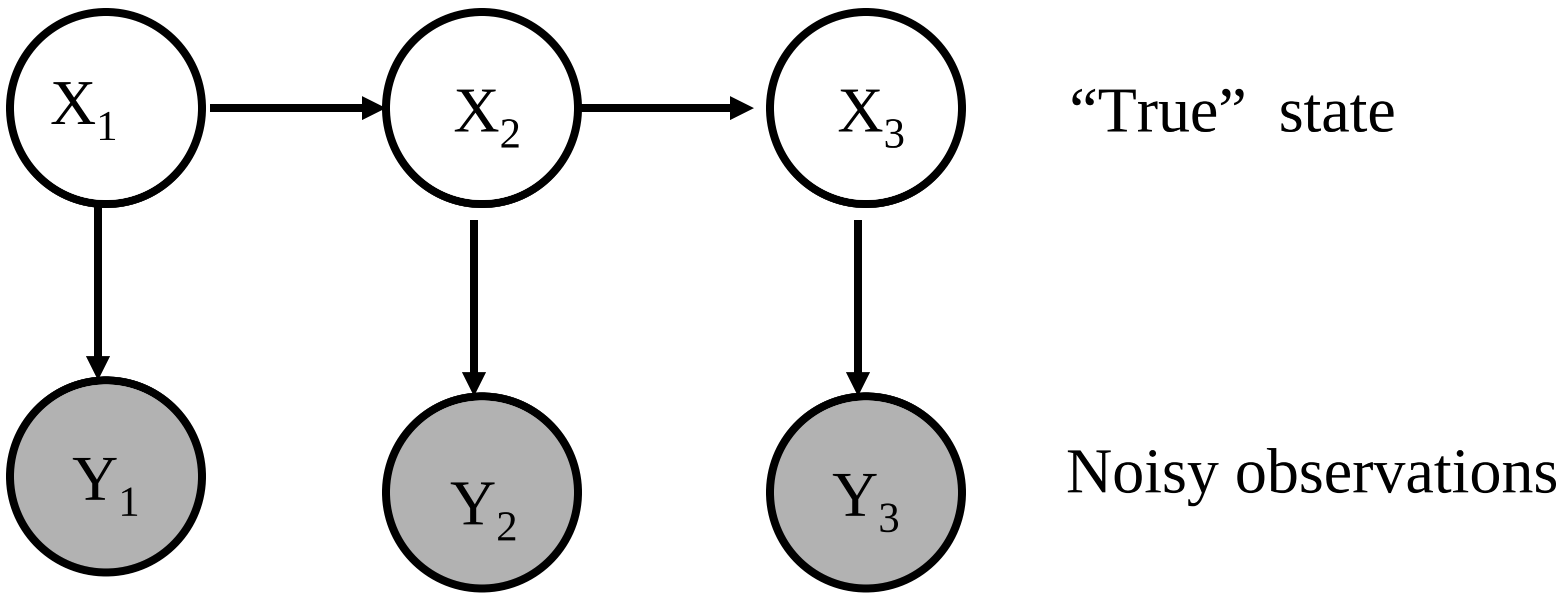
DBNs vs HMMs

- An HMM represents the state of the world using a single discrete random variable, $X_t \in \{1, \ldots, K\}$.
- A DBN represents the state of the world using a set of random variables, $X_t^{(1)}, \ldots, X_t^{(D)}$ (factored/ distributed representation).
- A DBN represents $P(X_t|X_{t-1})$ in a compact way using a parameterized graph.

⇒ A DBN may have exponentially fewer parameters than its corresponding HMM.

⇒ Inference in a DBN may be exponentially faster than in the corresponding HMM.

State-space model (SSM)/ Linear Dynamical System (LDS)



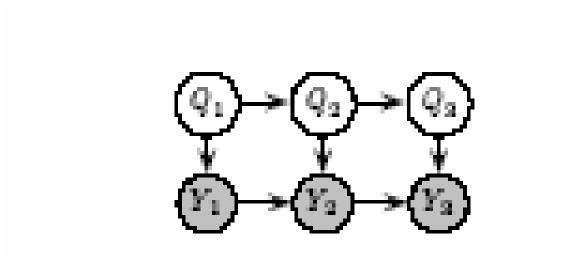
$$p(X_t|X_{t-1}) = \mathcal{N}(X_t; AX_{t-1}, Q)$$

$$p(Y_t|X_t) = \mathcal{N}(Y_t; BX_t, R)$$

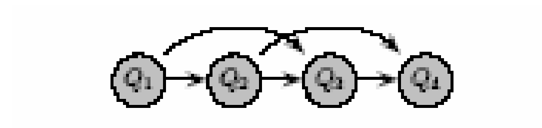
The 3 main tasks for HMMs

- Computing likelihood: $P(y_{1:t}) = \sum_i P(X_t = i, y_{1:t})$
- Viterbi decoding (most likely explanation): $\arg\max_{x_{1:t}} P(x_{1:t}|y_{1:t})$
- Learning: $\hat{\theta}_{ML} = \arg\max_\theta P(y_{1:T}|\theta)$, where $\theta = (A, B, \pi)$.
  - Learning can be done with Baum-Welch (EM).
  - Learning uses inference as a subroutine.
  - Inference (forwards-backwards) takes $O(TK^2)$ time, where K is the number of states and T is sequence length.

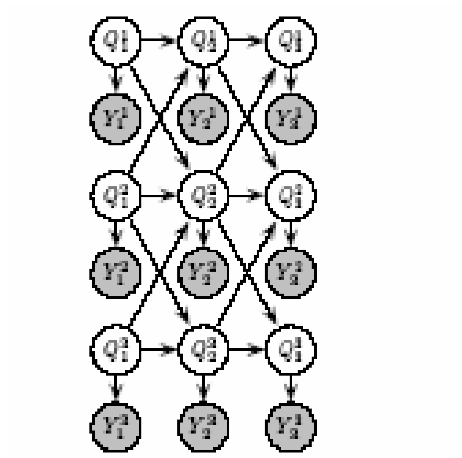
Other Types of HMMs



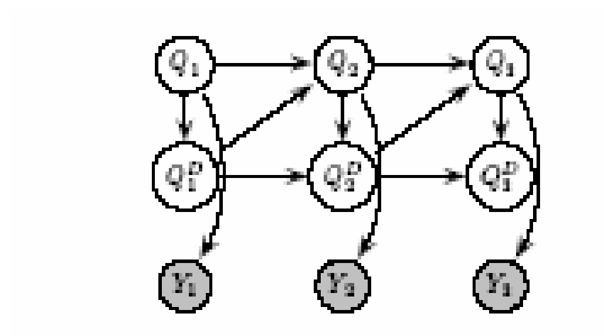
Auto-regressive HMM



Trigram models



Coupled HMMs



Hidden Semi-Markov Models

References

- J. Breese and D. Koller, Bayesian Networks and Decision-Theoretic Reasoning for Artificial Intelligence
- K. Murphy, Tutorial on DBNs
- K. Murphy, Graphical Models and BNT