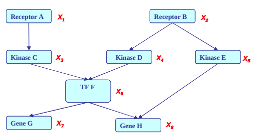
School of Computer Science
Carnegie Mellon

Introduction

Probabilistic Graphical Models (10-708)

Lecture 0, Sep 10, 2007

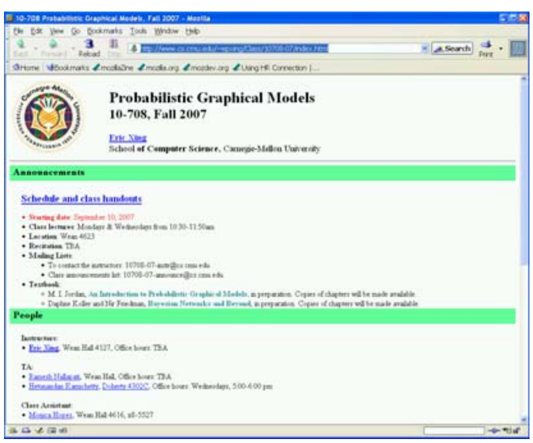


Eric Xing

Reading:

Logistics

- Class webpage:
  - http://www.cs.cmu.edu/~epxing/Class/10708-07/



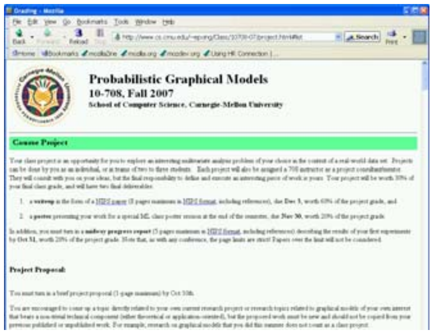
Logistics

- No formal text book, but draft chapters will be handed out in class:
  - M. I. Jordan, **An Introduction to Probabilistic Graphical Models**
  - Daphne Koller and Nir Friedman, **Structured Probabilistic Models**
- Mailing Lists:
  - To contact the instructors: 10708-07-instr@cs.cmu.edu
  - Class announcements list: 10708-07-announce@cs.cmu.edu.
- TA:
  - Hetunandan Kamichetty, Doherty 4302C, Office hours: Wednesdays, 5:00-6:00 pm
  - Dr. Ramesh Nallapati
- Class Assistant:
  - Monica Hopes, Wean Hall 4616, x8-5527

Logistics

- 4 homework assignments: 45% of grade
  - Theory exercises
  - Implementation exercises
- Final project: 30% of grade
  - Applying PGM to your research area
    - NLP, IR, Computational biology, vision, robotics …
  - Theoretical and/or algorithmic work
    - a more efficient approximate inference algorithm
    - a new sampling scheme for a non-trivial model …
- Take home final: 25% of grade
  - Theory exercises and/or analysis
- Policies …

Past projects:



- We will have a prize for the best project(s) …

- **Winner of the 2005 project:**

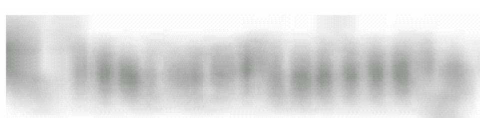
  J. Yang, Y. Liu, E. P. Xing and A. Hauptmann, Harmonium-Based Models for Semantic Video Representation and Classification , *Proceedings of The Seventh SIAM International Conference on Data Mining* **(SDM 2007)**. (**Recipient of the BEST PAPER Award**)

- **Other projects:**

  Andreas Krause, Jure Leskovec and Carlos Guestrin, **Data Association for Topic Intensity Tracking,** *23rd International Conference on Machine Learning* **(ICML 2006).**

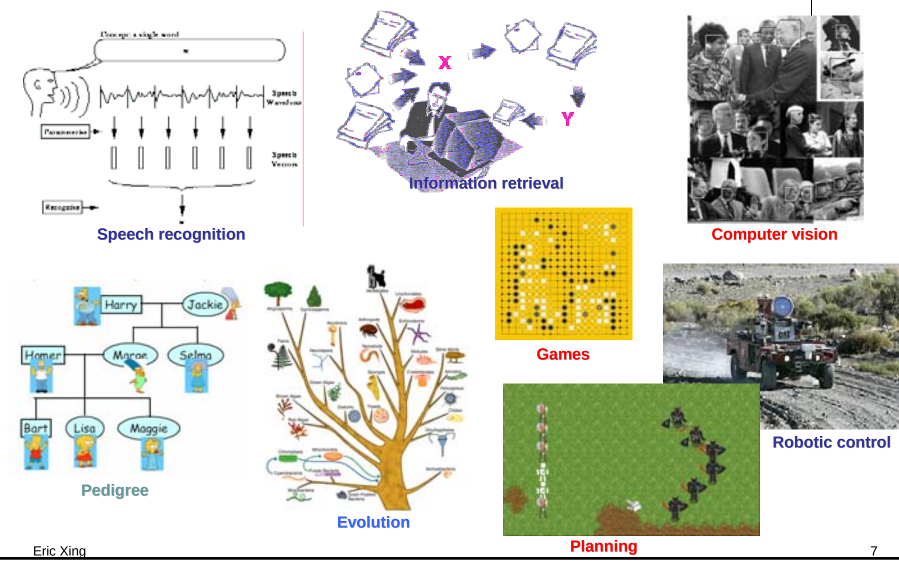
  Y. Shi, F. Guo, W. Wu and E. P. Xing, GIMscan: A New Statistical Method for Analyzing Whole-Genome Array CGH Data, *The Eleventh Annual International Conference on Research in Computational Molecular Biology* **(RECOMB 2007)**.

What is this?



- Classical AI and ML research ignored this phenomena
- The Problem (an example):
  - you want to catch a flight at 10:00am from Pitt to SF, can I make it if I leave at 7am and take a 28X at CMU?
    - partial observability (road state, other drivers' plans, etc.)
    - noisy sensors (radio traffic reports)
    - uncertainty in action outcomes (flat tire, etc.)
    - immense complexity of modeling and predicting traffic
- Reasoning under **uncertainty**!

A universal task …

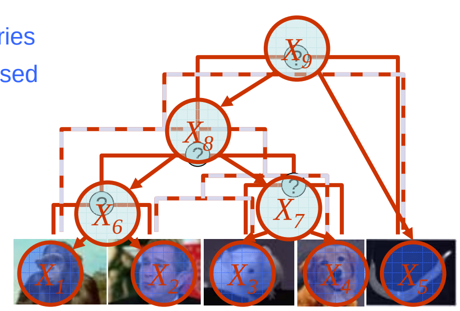


The Fundamental Questions

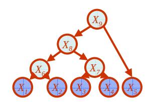
- Representation
  - How to capture/model uncertainties in possible worlds?
  - How to encode our domain knowledge/assumptions/constraints?
- Inference
  - How do I answers questions/queries according to my model and/or based given data?

    e.g.: $P(X_i \mid \mathcal{D})$
- Learning
  - What model is "right" for my data?

    e.g.: $\mathcal{M} = \arg\max_{\mathcal{M} \in \mathrm{M}} F(\mathcal{D}; \mathcal{M})$



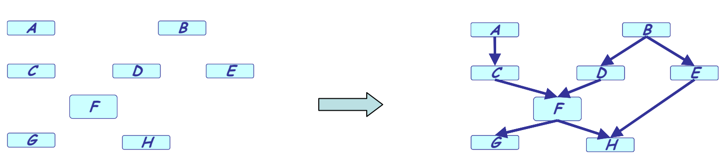
Graphical Models



- Graphical models are a marriage between graph theory and probability theory
- One of the most exciting developments in machine learning (knowledge representation, AI, EE, Stats,…) in the last two decades…
- Some advantages of the graphical model point of view
  - Inference and learning are treated together
  - Supervised and unsupervised learning are merged seamlessly
  - Missing data handled nicely
  - A focus on conditional independence and computational issues
  - Interpretability (if desired)
- Are having significant impact in science, engineering and beyond!

What is a Graphical Model?

- The informal blurb:
  - It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with structured semantics



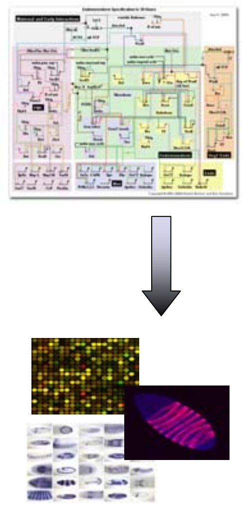
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$P(X_{18}) = P(X_1)P(X_2)P(X_3 \mid X_1X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$
$$P(X_6 \mid X_3, X_4)P(X_7 \mid X_6)P(X_8 \mid X_5, X_6)$$

- A more formal description:
  - It refers to a family of distributions on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a graph that connects these variables

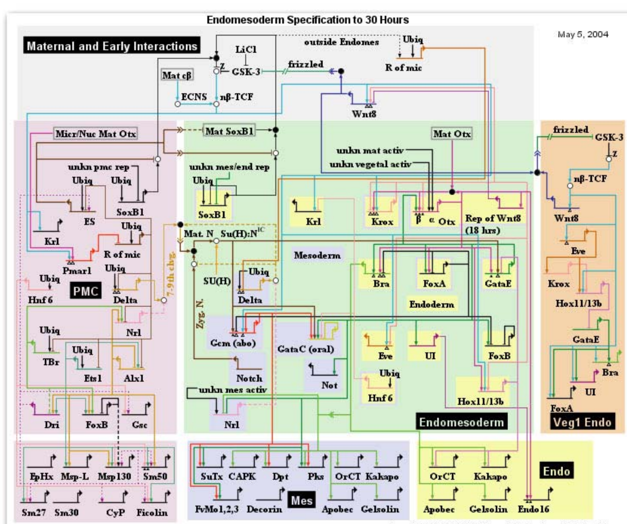
Statistical Inference

probabilistic generative model



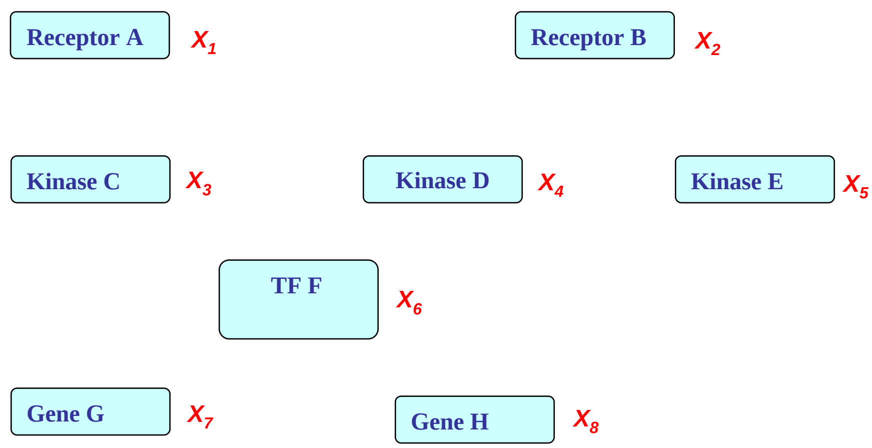
gene expression profiles

Statistical Inference



Multivariate Distribution in High-D Space

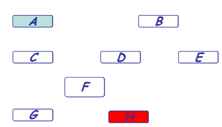
- A possible world for cellular signal transduction:



Recap of Basic Prob. Concepts

- Representation: what is the joint probability dist. on multiple variables?

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8,)$$

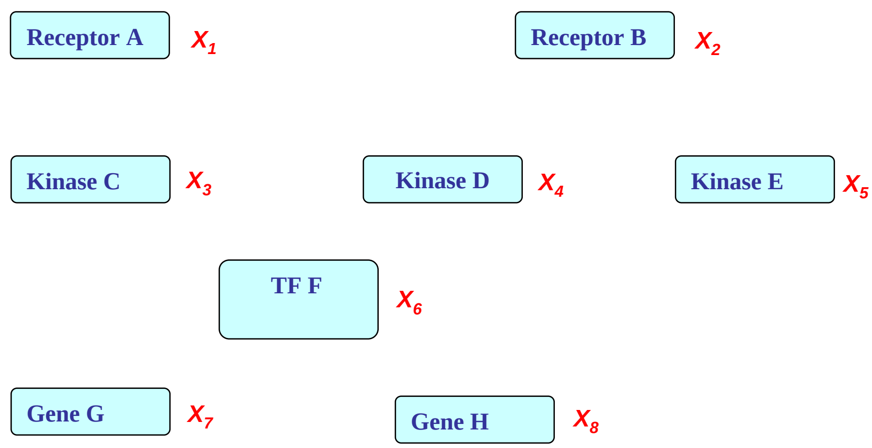


  - How many state configurations in total? --- 2^8
  - Are they all needed to be represented?
  - **Do we get any scientific/medical insight?**
- Learning: where do we get all this probabilities?
  - Maximal-likelihood estimation? but how many data do we need?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?
- Inference: If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?
  - Computing $p(H|A)$ would require summing over all 2^6 configurations of the unobserved variables

What is a Graphical Model?

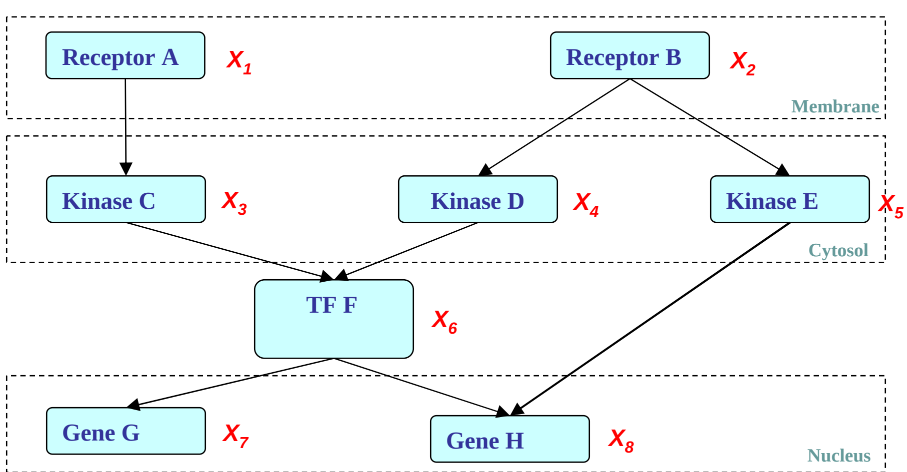
--- example from a signal transduction pathway

- A possible world for cellular signal transduction:



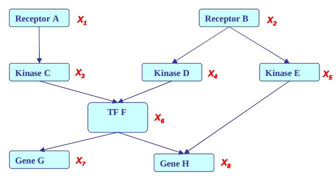
GM: Structure Simplifies Representation

- Dependencies among variables



Probabilistic Graphical Models

- If X_i's are **conditionally independent** (as described by a **PGM**), the joint can be factored to a product of simpler terms, e.g.,



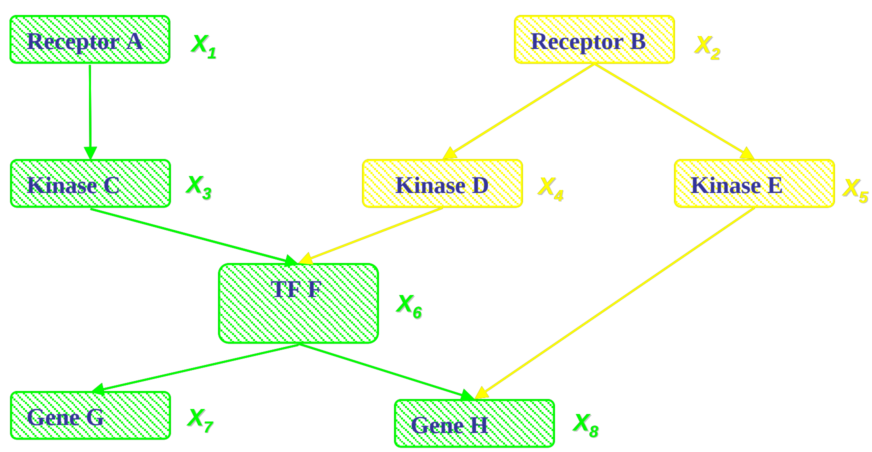
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3|\, X_1)\, P(X_4|\, X_2)\, P(X_5|\, X_2)\, P(X_6|\, X_3, X_4)\, P(X_7|\, X_6)\, P(X_8|\, X_5, X_6)$$

Stay tune for what are these independencies!

- Why we may favor a PGM?
  - Incorporation of domain knowledge and causal (logical) structures

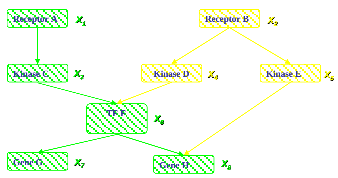
    2+2+4+4+4+8+4+8=36, an 8-fold reduction from 2^8 in representation cost !

GM: Data Integration



Probabilistic Graphical Models

- If X_i's are **conditionally independent** (as described by a **PGM**), the joint can be factored to a product of simpler terms, e.g.,



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_2)\, P(X_4|\, X_2)\, P(X_5|\, X_2)\, P(X_1)\, P(X_3|\, X_1)\, P(X_6|\, X_3, X_4)\, P(X_7|\, X_6)\, P(X_8|\, X_5, X_6)$$

- Why we may favor a PGM?
  - Incorporation of domain knowledge and causal (logical) structures
    2+2+4+4+4+8+4+8=36, an 8-fold reduction from 2^8 in representation cost !
  - Modular combination of heterogeneous parts – data fusion

Rational Statistical Inference

The Bayes Theorem:

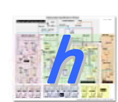
Posterior probability

Likelihood

Prior probability

$$p(h\,|\,d) = \frac{p(d\,|\,h)\,p(h)}{\sum_{h' \in H} p(d\,|\,h')\,p(h')}$$

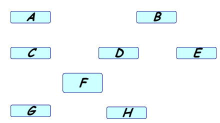
Sum over space of hypotheses



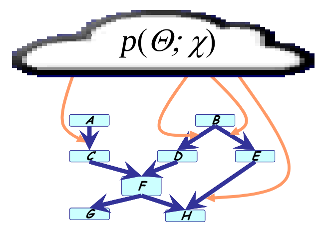
- This allows us to capture uncertainty about the model in a principled way
- But how can we specify and represent a complicated model?
  - Typically the number of genes need to be modeled are in the order of thousands!

GM: MLE and Bayesian Learning

- Probabilistic statements of Θ is conditioned on the values of the observed variables $\mathbf{A}_{obs}$ and prior $p(\ |\chi)$



$$\mathbf{A} = \begin{array}{l} (A,B,C,D,E,\ldots)=(T,F,F,T,F,\ldots) \\ (A,B,C,D,E,\ldots)=(T,F,T,T,F,\ldots) \\ \ldots\ldots.. \\ (A,B,C,D,E,\ldots)=(F,T,T,T,F,\ldots) \end{array}$$



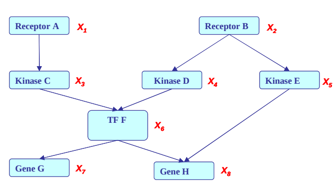
C	D	$P(F \mid C,D)$	
c	d	0.9	0.1
c	$\bar{d}$	0.2	0.8
$\bar{c}$	d	0.9	0.1
$\bar{c}$	$\bar{d}$	0.01	0.99

$$\boxed{\Theta_{Bayes} = \int \Theta\, p(\Theta \mid \mathbf{A}, \chi)\, d\Theta}$$

$$\underbrace{p(\Theta \mid \mathbf{A};\chi)}_{\text{posterior}} \propto \underbrace{p(\mathbf{A} \mid \Theta)}_{\text{likelihood}}\, \underbrace{p(\Theta;\chi)}_{\text{prior}}$$

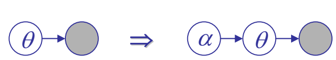
Probabilistic Graphical Models

- If X_i's are **conditionally independent** (as described by a **PGM**), the joint can be factored to a product of simpler terms, e.g.,



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3 \mid X_1)\, P(X_4 \mid X_2)\, P(X_5 \mid X_2)\, P(X_6 \mid X_3, X_4)\, P(X_7 \mid X_6)\, P(X_8 \mid X_5, X_6)$$

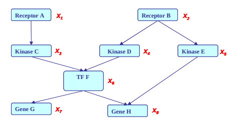
- Why we may favor a PGM?
  - Incorporation of domain knowledge and causal (logical) structures
    - 2+2+4+4+4+8+4+8=36, an 8-fold reduction from 2^8 in representation cost !
  - Modular combination of heterogeneous parts – data fusion
  - Bayesian Philosophy
    - Knowledge meets data



Two types of GMs

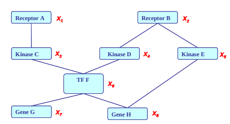
- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3|\, X_1)\, P(X_4|\, X_2)\, P(X_5|\, X_2)\, P(X_6|\, X_3, X_4)\, P(X_7|\, X_6)\, P(X_8|\, X_5, X_6)$$



- Undirected edges simply give correlations between variables (**Markov Random Field** or **Undirected Graphical model**):

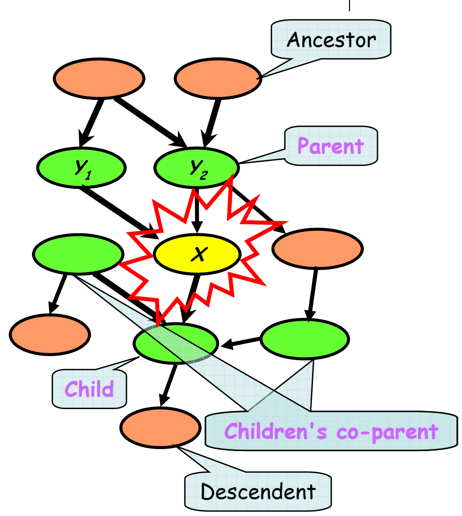
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= 1/Z \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2) + E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}$$



Bayesian Networks

Structure: ***DAG***

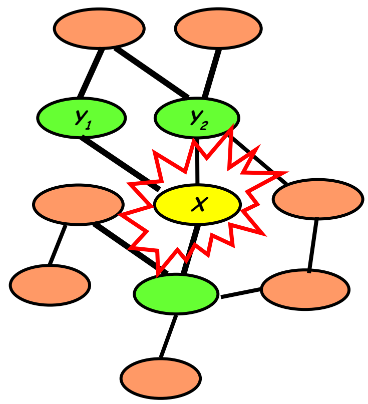
- Meaning: a node is **conditionally independent** of every other node in the network outside its **Markov blanket**
- Local conditional distributions (**CPD**) and the **DAG** completely determine the **joint** dist.
- Give **causality** relationships, and facilitate a **generative** process



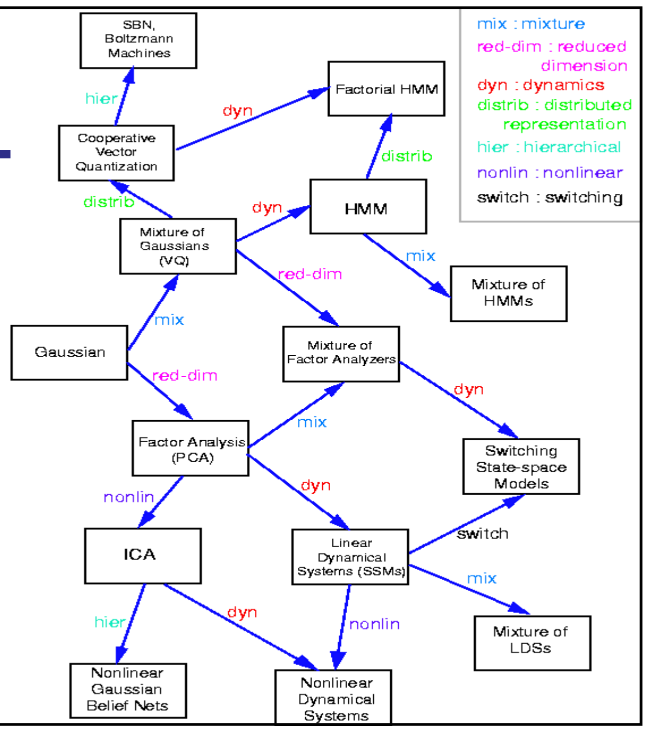
Markov Random Fields

Structure: ***undirected graph***

- Meaning: a node is **conditionally independent** of every other node in the network given its **Directed neighbors**
- Local contingency functions (**potentials**) and the **cliques** in the graph completely determine the **joint** dist.
- Give **correlations** between variables, but no explicit way to generate samples

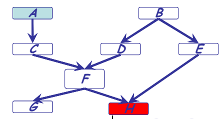


An (incomplete) genealogy of graphical models



(Picture by Zoubin Ghahramani and Sam Roweis)

Probabilistic Inference



- Computing statistical queries regarding the network, e.g.:
  - Is node X independent on node Y given nodes Z,W ?
  - What is the probability of X=true if (Y=false and Z=true)?
  - What is the joint distribution of (X,Y) if Z=false?
  - What is the likelihood of some full assignment?
  - What is the most likely assignment of values to all or a subset the nodes of the network?

- General purpose algorithms exist to fully automate such computation
  - Computational cost depends on the topology of the network
  - Exact inference:
    - The junction tree algorithm
  - Approximate inference;
    - Loopy belief propagation, variational inference, Monte Carlo sampling

A few myths about graphical models

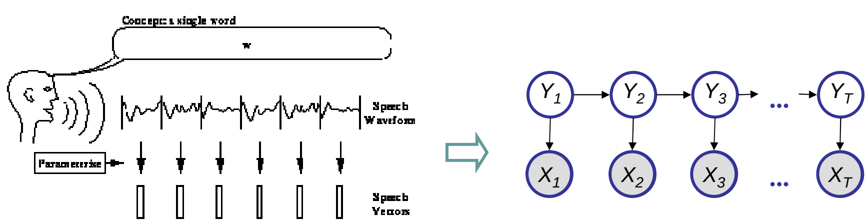
- They require a localist semantics for the nodes √
- They require a causal semantics for the edges ×
- They are necessarily Bayesian ×
- They are intractable √

Application of GMs

- Machine Learning
- Computational statistics

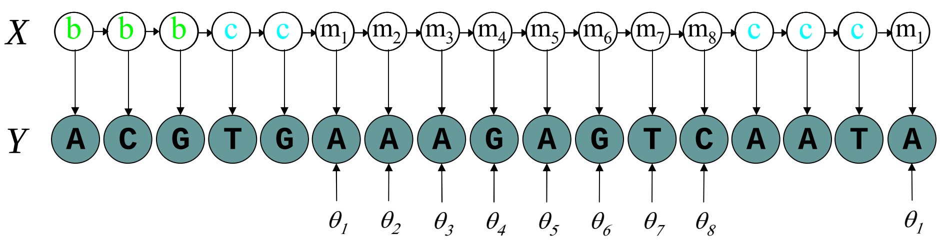
- Computer vision and graphics
- Natural language processing
- Informational retrieval
- Robotic control
- Decision making under uncertainty
- Error-control codes
- Computational biology
- Genetics and medical diagnosis/prognosis
- Finance and economics
- Etc.

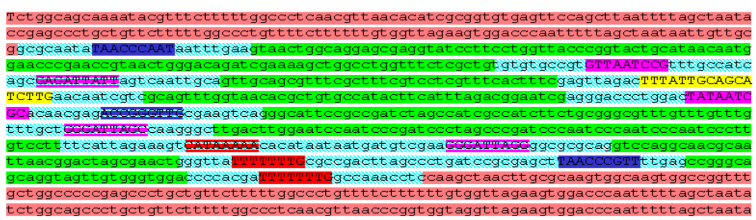
Speech recognition



Hidden Markov Model

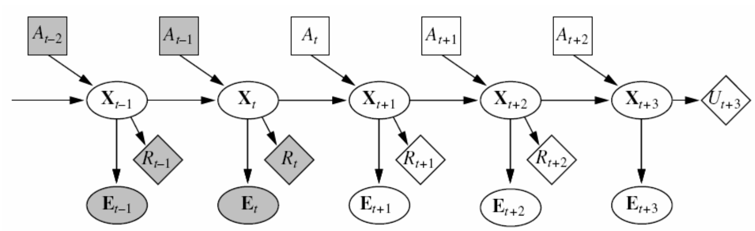
Segmentation and Pattern Recog.
(in Bio, Vision, NLP)



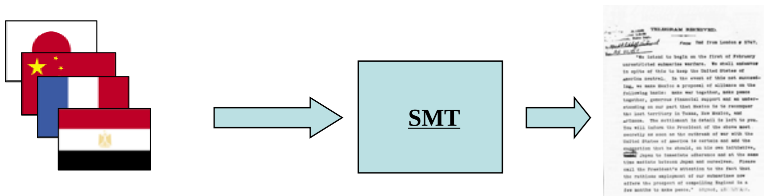


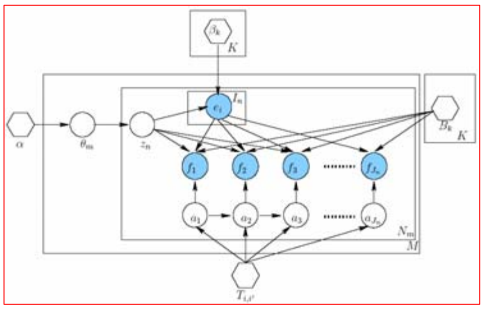
Reinforcement learning

- Partially observed Markov decision processes (POMDP)



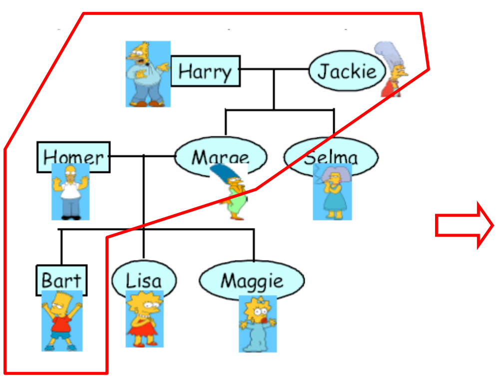
Machine translation

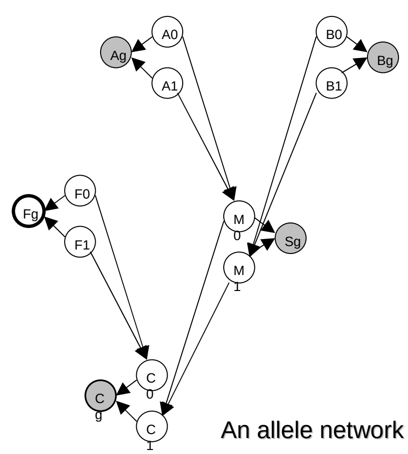




The HM-BiTAM model (B. Zhao and E.P Xing, ACL 2006)

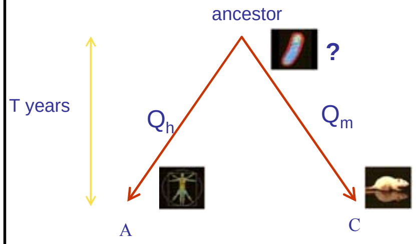
Genetic pedigree

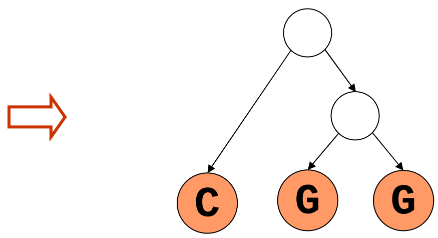




An allele network

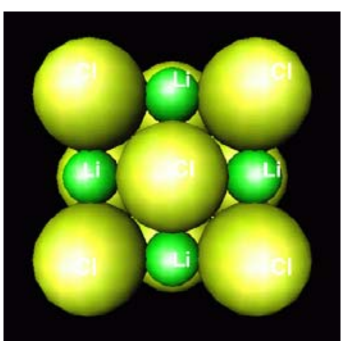
Evolution

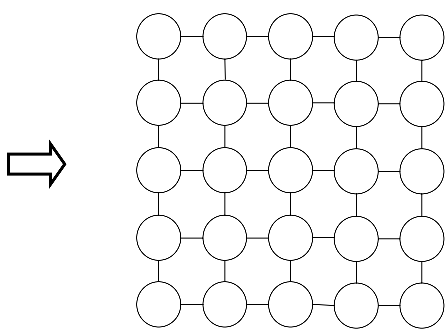




Tree Model

Solid State Physics

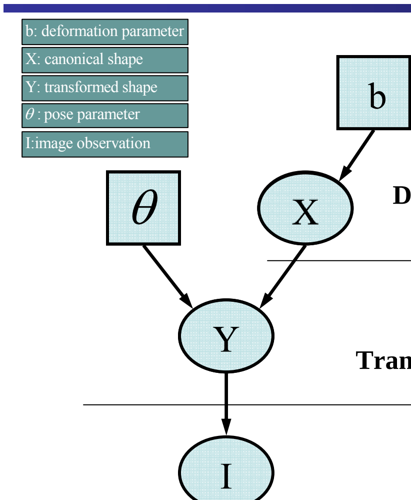




Ising/Potts model

Computer Vision

A Generative GM



P(X|b)
Deformation

- Unimodal
- Multimodal
- Example-based

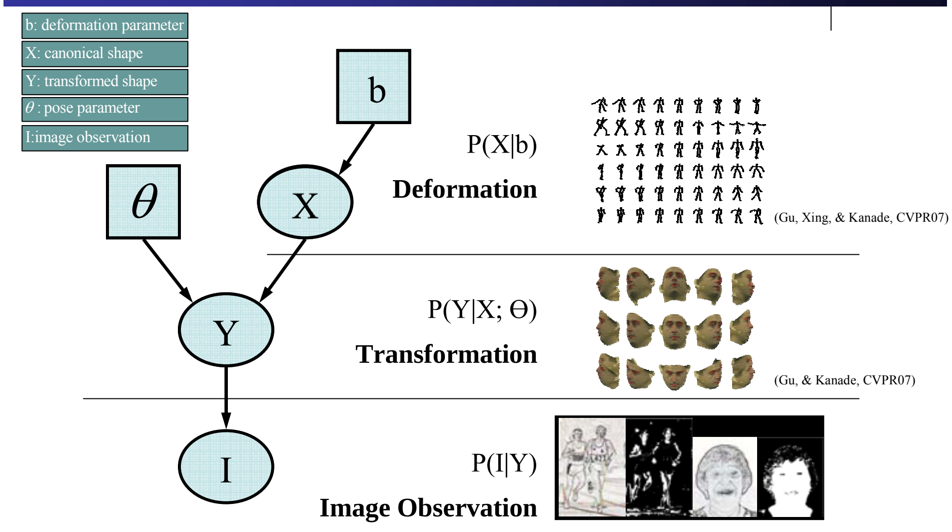
P(Y|X; Θ)
Transformation

- Rigid
- Perspective

P(I|Y)
Image Observation

- Boundary/Regional
- Searching Path/Region

A Generative GM



Why graphical models

- A language for communication
- A language for computation
- A language for development

- Origins:
  - Wright 1920's
  - Independently developed by Spiegelhalter and Lauritzen in statistics and Pearl in computer science in the late 1980's

Why graphical models

- **Probability theory** provides the **glue** whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data.
- The **graph theoretic** side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms.
- **Many of the classical multivariate probabilistic systems** studied in fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics **are special cases of the general graphical model formalism**
- The graphical model framework provides a way to view all of these systems as instances of a **common underlying formalism**.

--- M. Jordan

Plan for the Class

- Fundamentals of Graphical Models:
  - Bayesian Network and Markov Random Fields
  - Continuous and Hybrid models, exponential family, GLIM
  - Basic representation, inference, and learning
- Case studies: Popular Bayesian networks and MRF
  - Multivariate Gaussian Models
  - Temporal models
  - Trees models
  - Intractable popular BNs and MRFs: e.g., Dynamic Bayesian networks, Bayesian admixture models (LDA)
- Approximate inference
  - Monte Carlo algorithms
  - Vatiational methods
- Advanced topics
  - Learning in structured input-output space
  - Nonparametric Bayesian model
- Applications

Notation

- Variable, value and index
- Random variable
- Random vector
- Random matrix
- Parameters