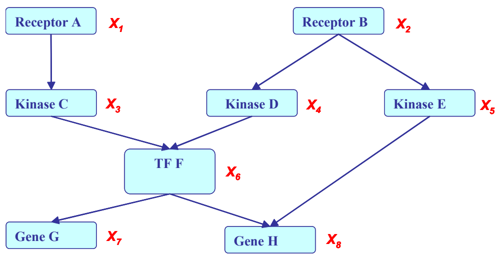
Machine Learning

10-701, Fall 2016

Graphical Models and Exact Inference

Eric Xing



Lecture 17, November 7, 2016

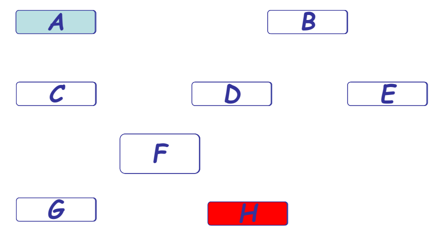
Reading: Chap. 8, C.B book

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?**



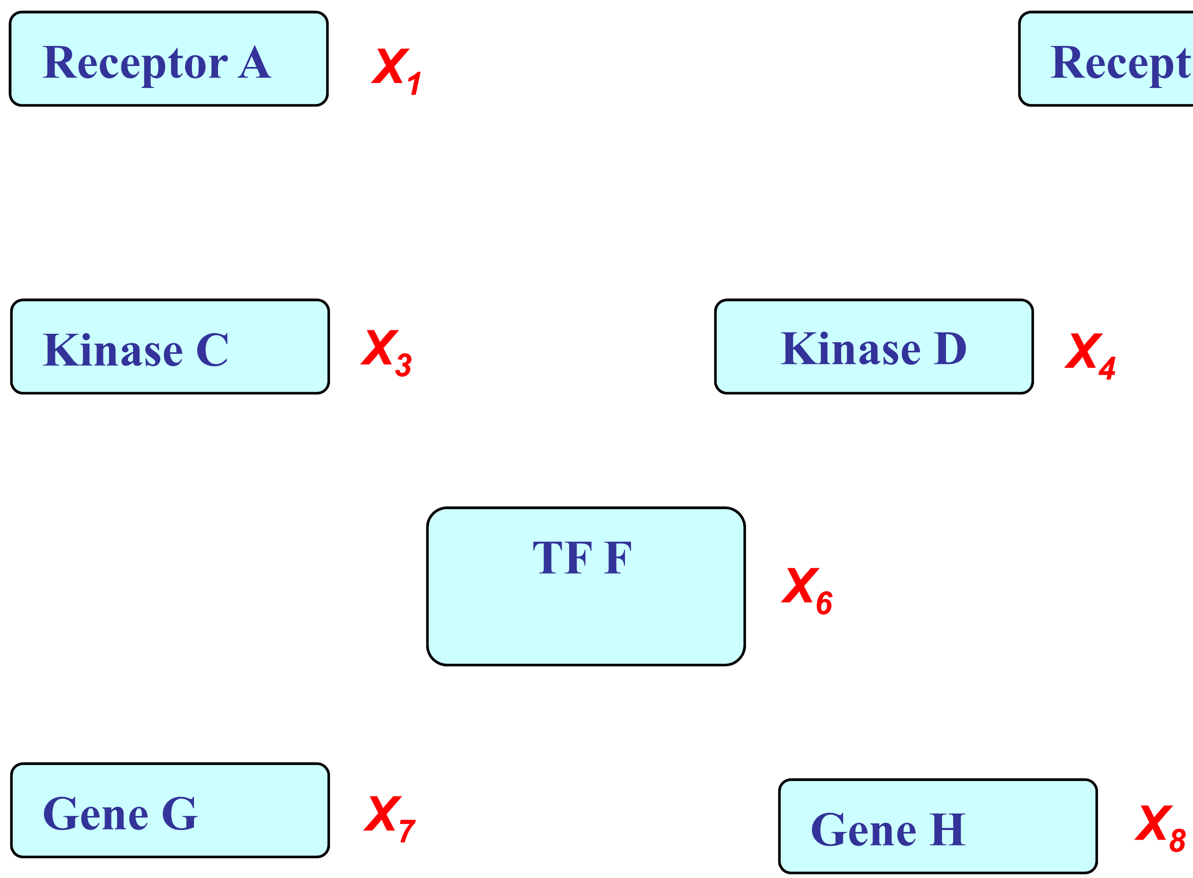
- Learning: where do we get all this probabilities?
  - Maximal-likelihood estimation? but how many data do we need?
  - Are there other est. principles?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?

- 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?

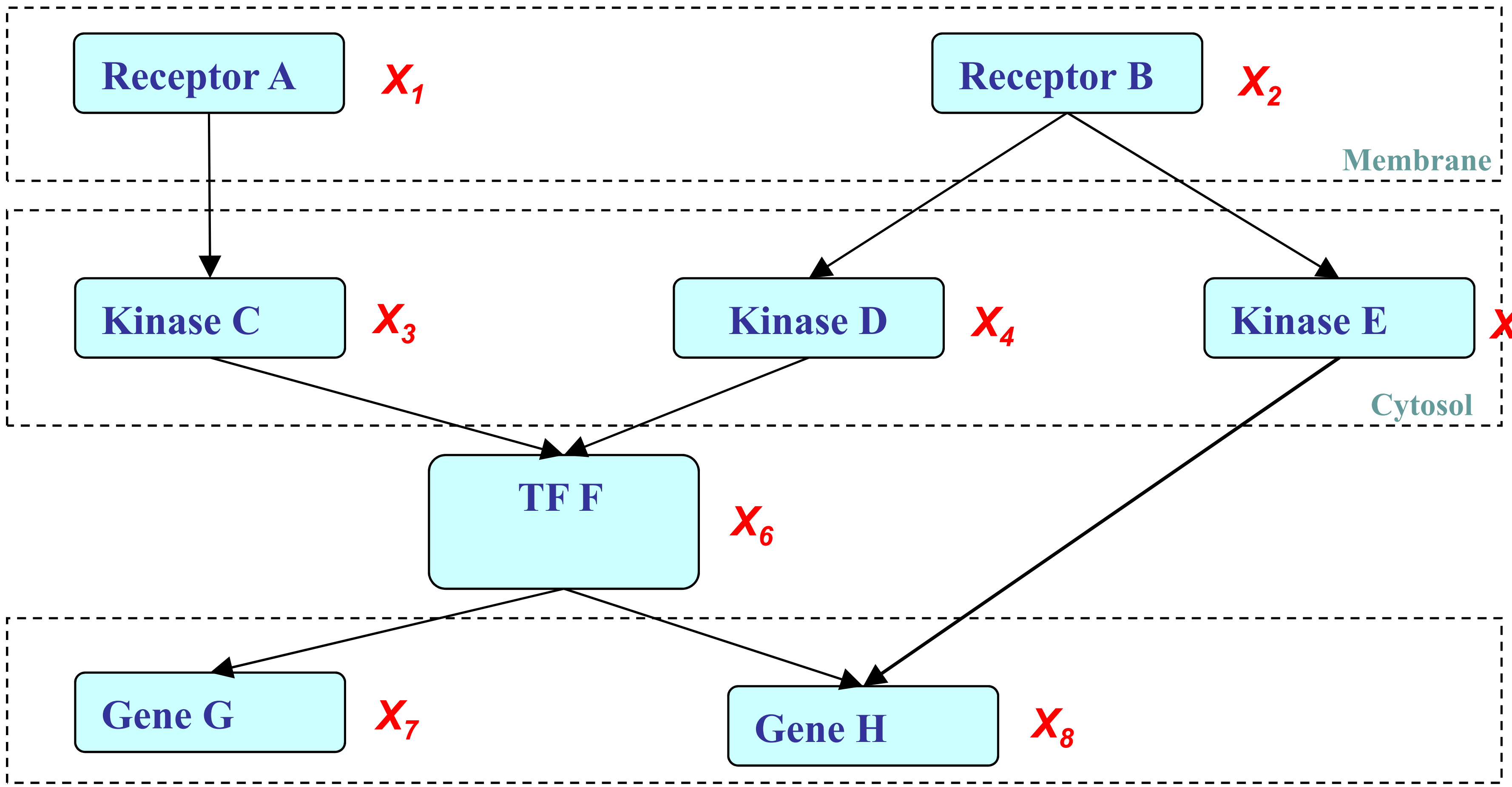
--- Multivariate Distribution in High-D Space

- A possible world for cellular signal transduction:



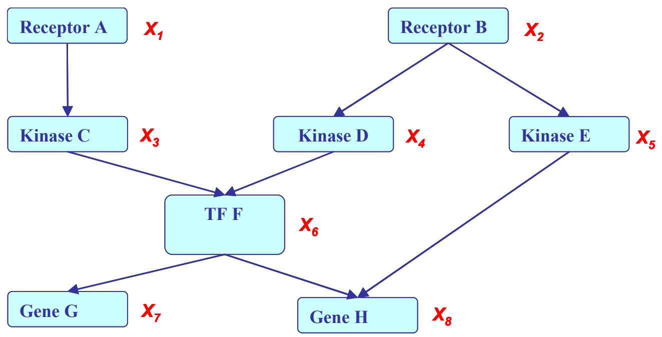
GM: Structure Simplifies Representation

- 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.,



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

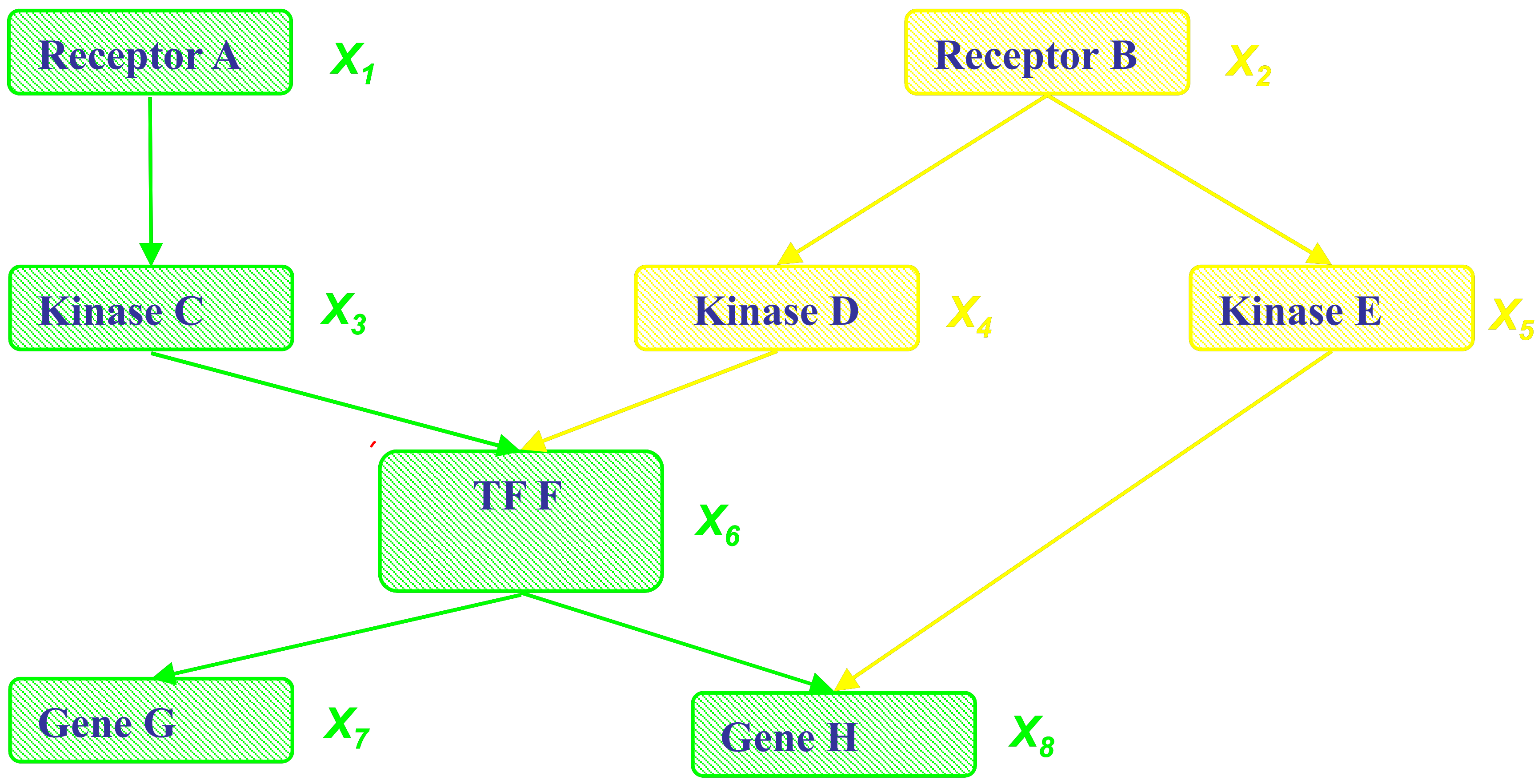
$$= P(X_1)\, P(X_2)\, P(X_3|\, X_1)\, P(X_4|\, X_2)\, P(X_5|\, X_2)\, P(X_6|\, X_3, X_4)\, P(X_7|\, X_6)\, P(X_8|\, X_5, X_6)$$

Stay tune for what are these independencies!

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

    1+1+2+2+2+4+2+4=18, a 16-fold reduction from 2^8 in representation cost !

GM: Data Integration



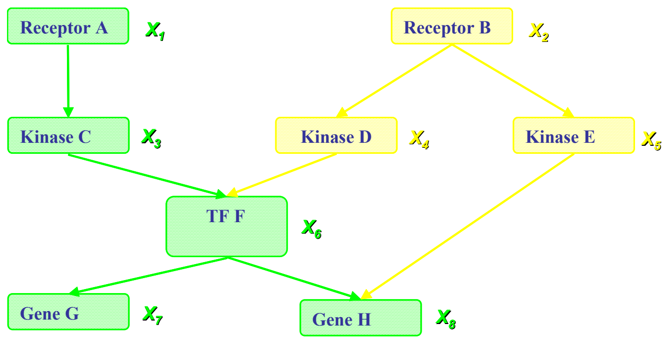
More Data Integration

- Text + Image + Network ➔ Holistic Social Media

- Genome + Proteome + Transcritome + Phenome + … ➔ PanOmic Biology

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:

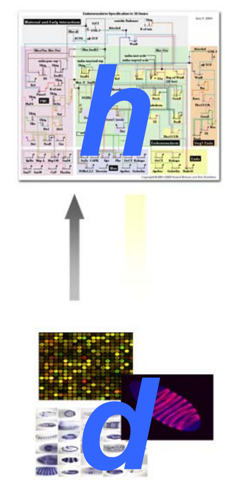
Posterior probability

Likelihood

Prior probability

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

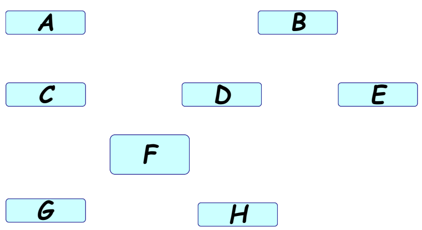
Sum over space of hypotheses



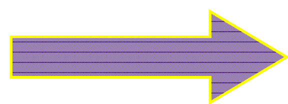
- This allows us to capture uncertainty about the model in a principled way
- But how can we specify and represent a complicated model?
  - Typically the number of genes need to be modeled are in the order of thousands!

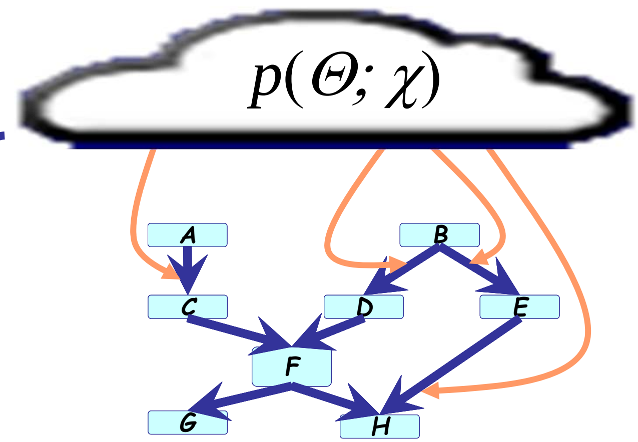
GM: MLE and Bayesian Learning

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



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





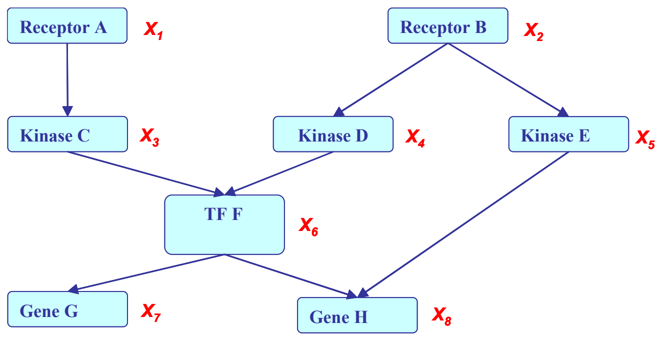
C	D	P(F | C,D)	
c	d	0.9	0.1
c	$\bar{d}$	0.2	0.8
$\bar{c}$	d	0.9	0.1
$\bar{c}$	$\bar{d}$	0.01	0.99

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

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

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.,



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

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

- 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

So What Is a PGM After All?

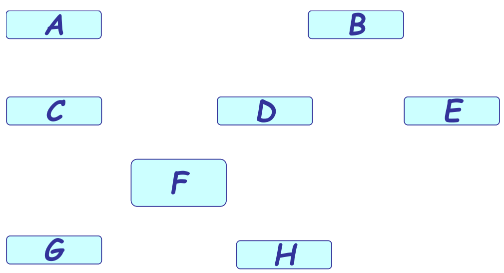
In a nutshell:

PGM = Multivariate Statistics + Structure

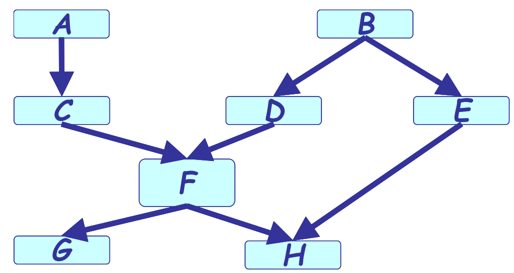
GM = Multivariate Obj. Func. + Structure

So What Is a PGM After All?

- 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***



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



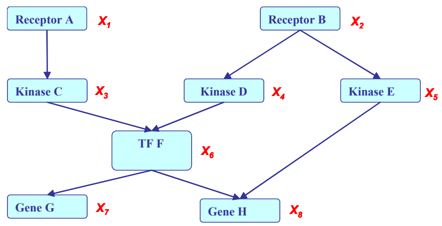
$$P(X_{1:8}) = P(X_1)P(X_2)P(X_3 \mid X_1X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$
$$P(X_6 \mid X_3, X_4)P(X_7 \mid X_6)P(X_8 \mid X_5, X_6)$$

- 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

Two types of GMs

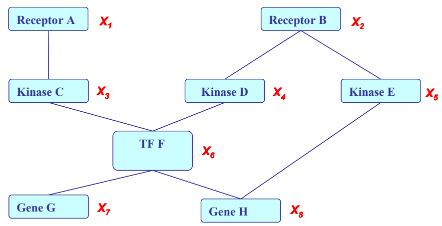
- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):

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



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

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



Towards structural specification of probability distribution

- Separation properties in the graph imply independence properties about the associated variables
- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

- **The Equivalence Theorem**

  For a graph G,

  Let $\mathcal{D}_1$ denote the family of all distributions that satisfy I(G),

  Let $\mathcal{D}_2$ denote the family of all distributions that factor according to G,

  Then $\mathcal{D}_1 \equiv \mathcal{D}_2$.

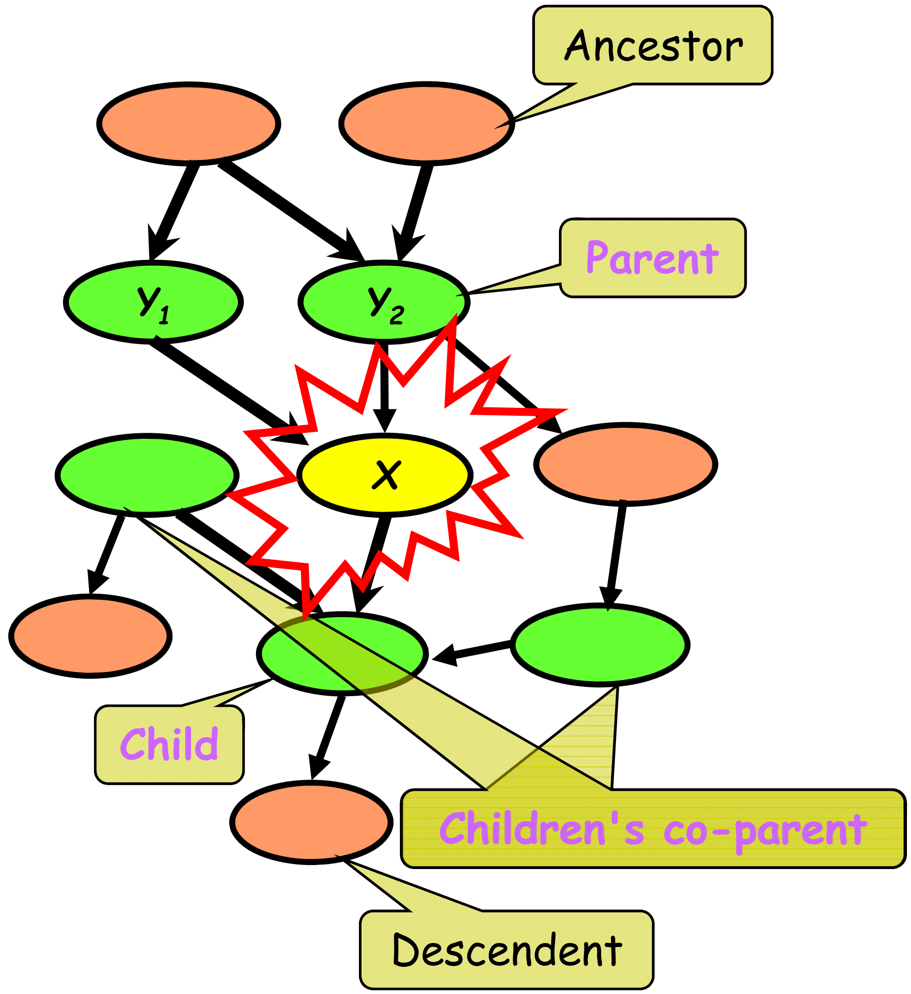
Bayesian Networks

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



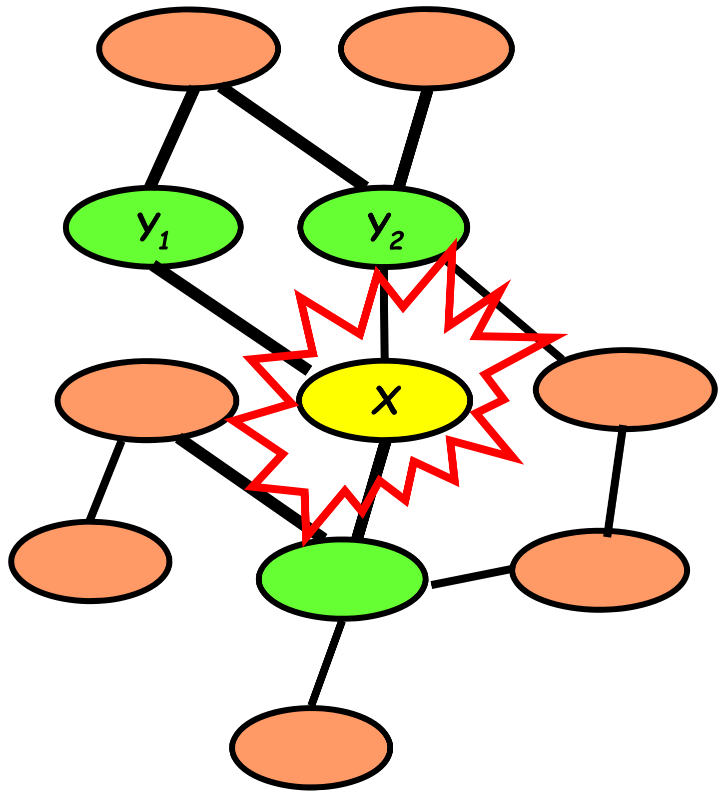
Markov Random Fields

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.

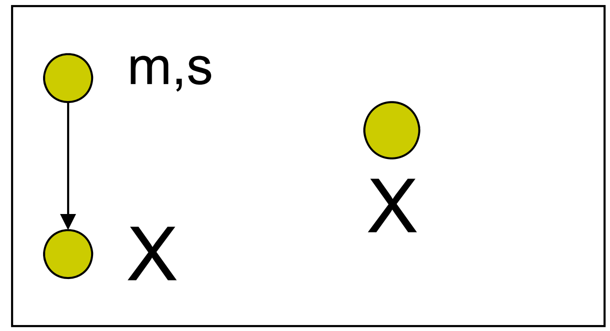
- Give **correlations** between variables, but no explicit way to generate samples



GMs are your old friends

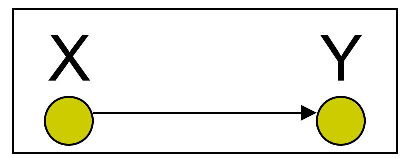
Density estimation

Parametric and nonparametric methods



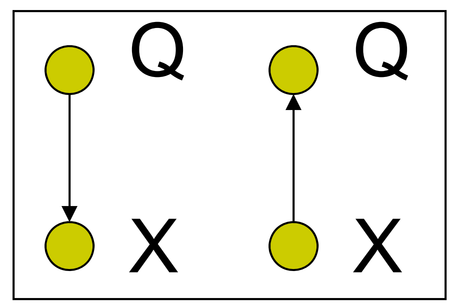
Regression

Linear, conditional mixture, nonparametric



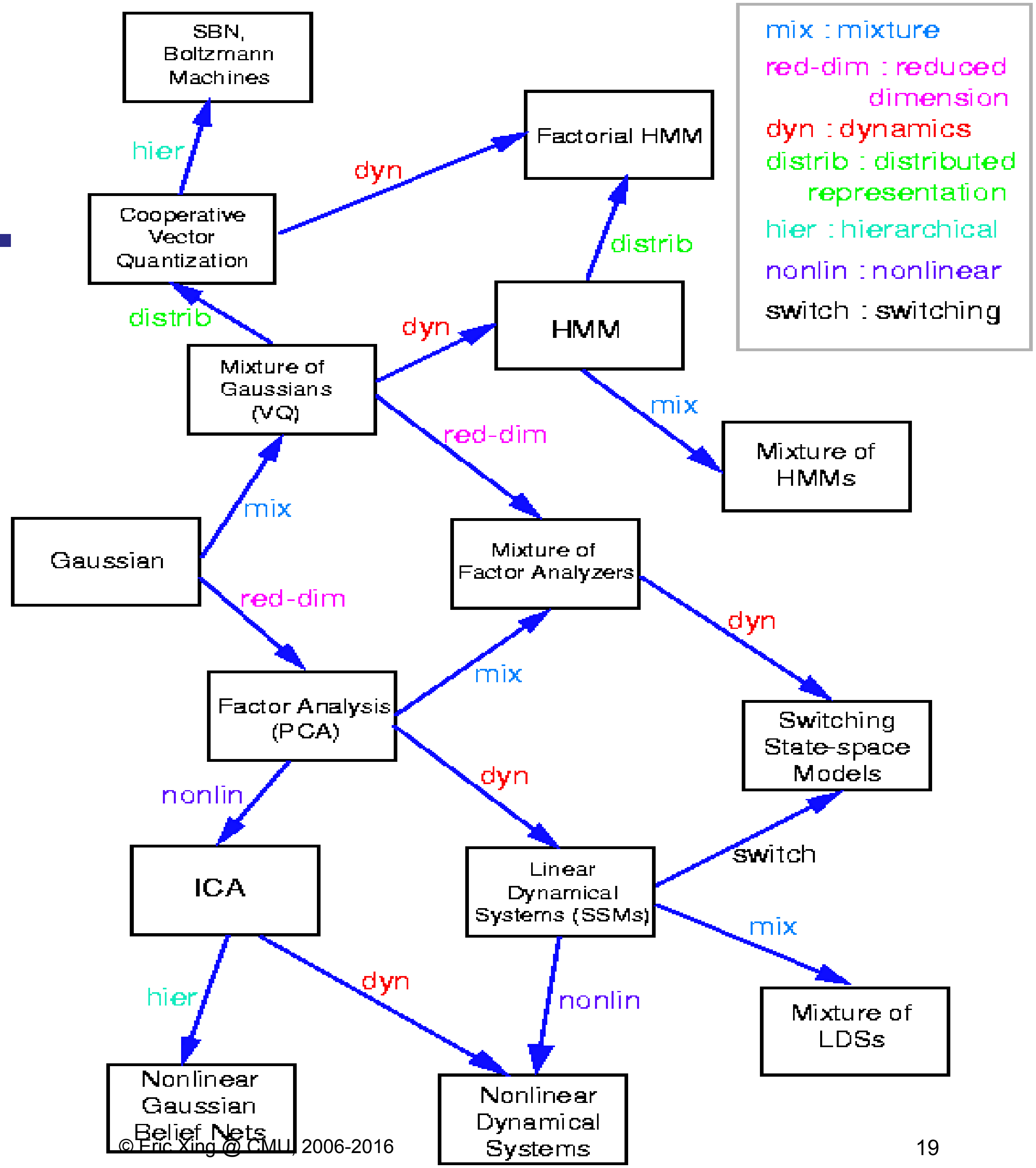
Classification

Generative and discriminative approach



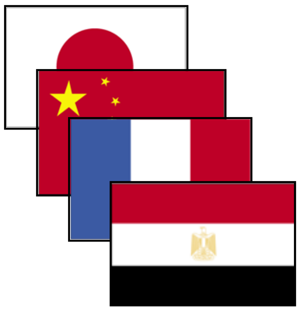
Clustering

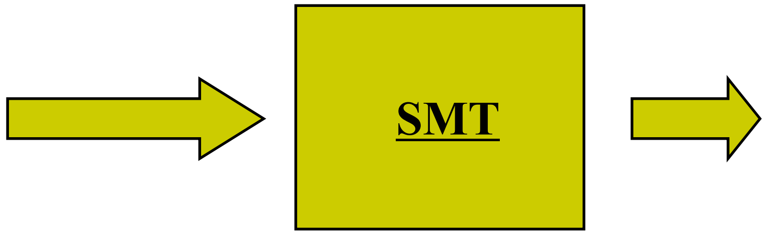
An (incomplete) genealogy of graphical models

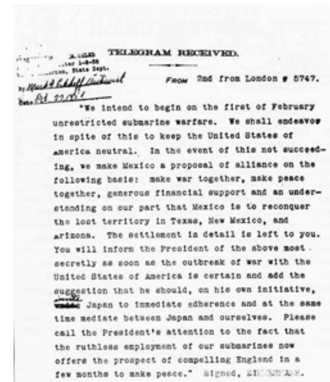


(Picture by Zoubin Ghahramani and Sam Roweis)

Fancier GMs: machine translation

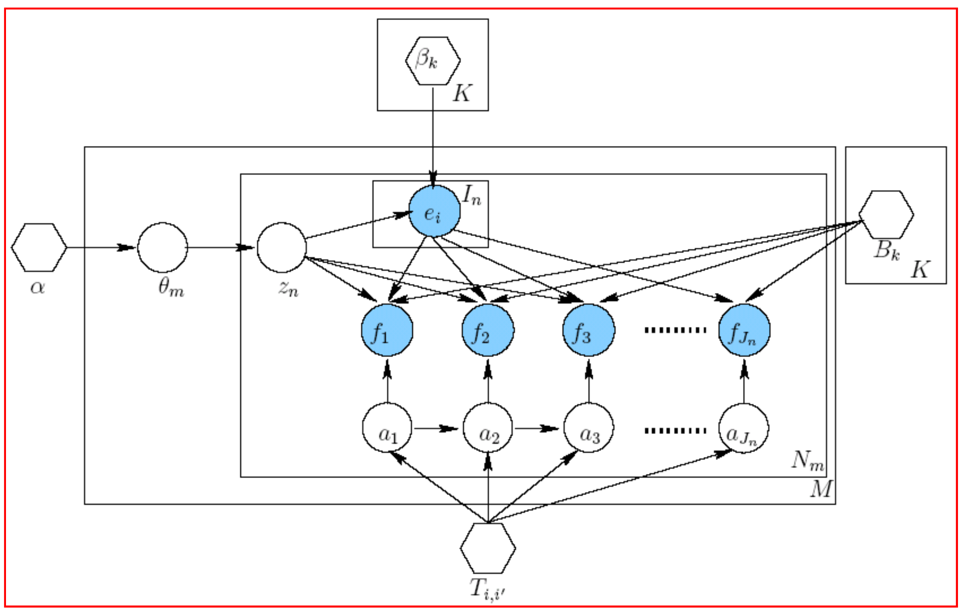





TELEGRAM RECEIVED.

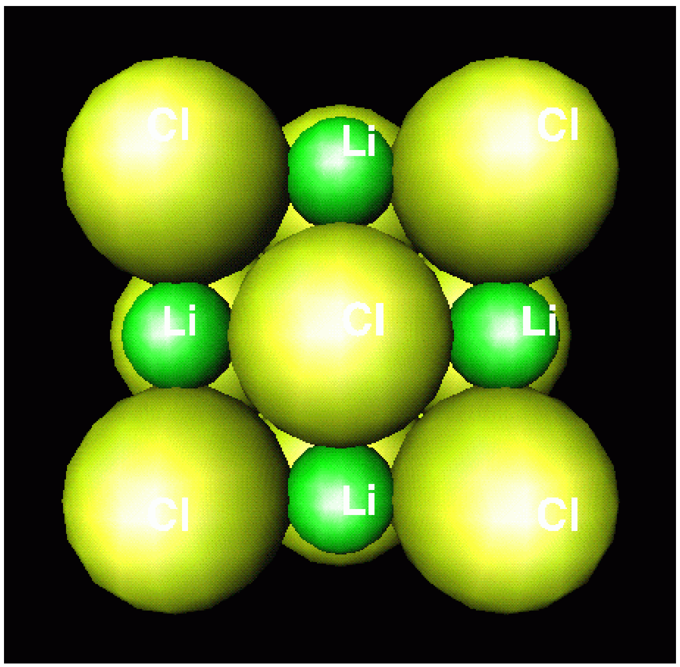
FROM 2nd from London # 5747.

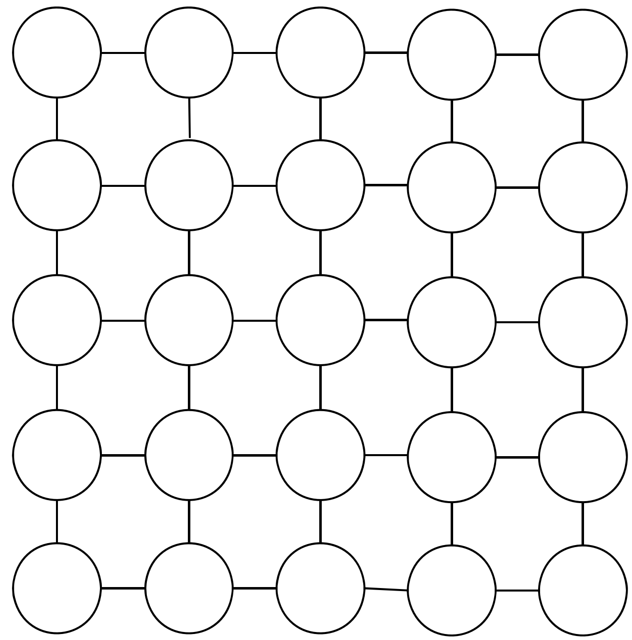
"We intend to begin on the first of February unrestricted submarine warfare. We shall endeavor in spite of this to keep the United States of America neutral. In the event of this not succeeding, we make Mexico a proposal of alliance on the following basis: make war together, make peace together, generous financial support and an understanding on our part that Mexico is to reconquer the lost territory in Texas, New Mexico, and Arizona. The settlement in detail is left to you. You will inform the President of the above most secretly as soon as the outbreak of war with the United States of America is certain and add the suggestion that he should, on his own initiative, [illegible] Japan to immediate adherence and at the same time mediate between Japan and ourselves. Please call the President's attention to the fact that the ruthless employment of our submarines now offers the prospect of compelling England in a few months to make peace." Signed, [illegible]



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

Fancier GMs: solid state physics





Ising/Potts model

A Generative Scheme for model design

Why graphical models

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

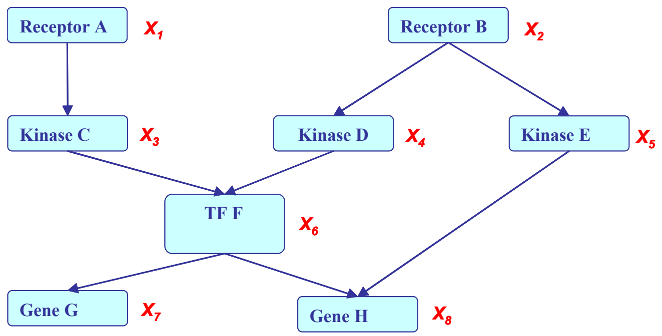
- Origins:
  - Wright 1920's
  - Independently developed by Spiegelhalter and Lauritzen in statistics and Pearl in computer science in the late 1980's

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

Bayesian Network: Factorization Theorem



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

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

- **Theorem:**

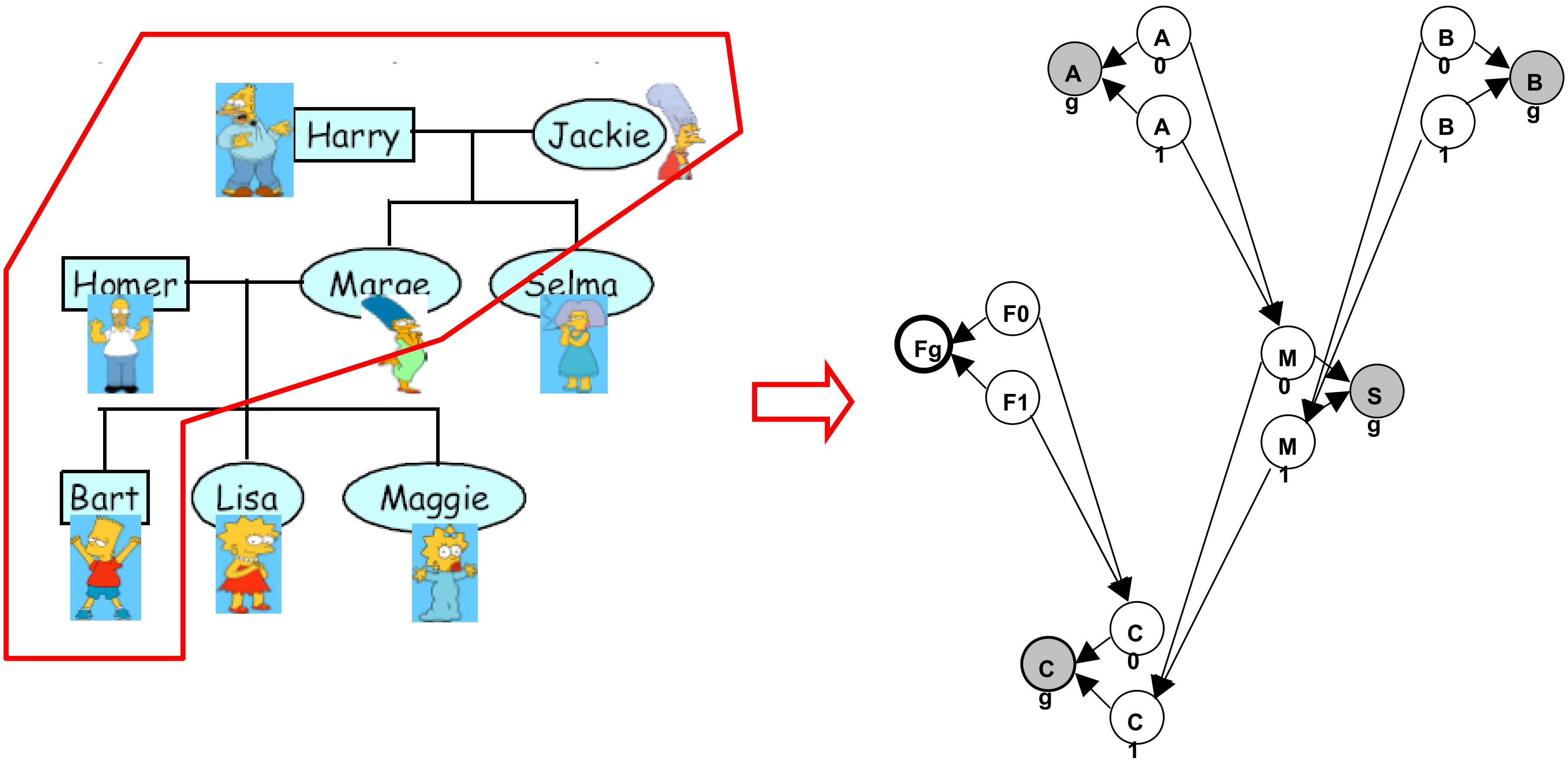
  Given a DAG, The most general form of the probability distribution that is consistent with the (probabilistic independence properties encoded in the) graph factors according to "node given its parents":

$$P(\mathbf{X}) = \prod_i P(X_i | \mathbf{X}_{\pi_i})$$

  where $\mathbf{X}_{\pi_i}$ is the set of parents of xi. d is the number of nodes (variables) in the graph.

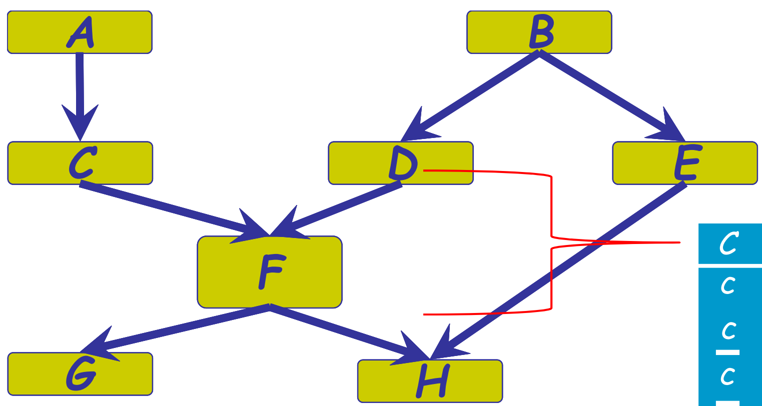
Example: a pedigree of people

- Genetic Pedigree



Specification of a BN

- There are two components to any GM:
  - the *qualitative* specification
  - the *quantitative* specification

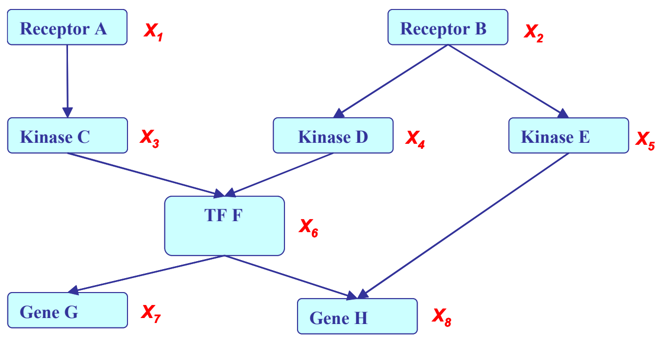


C	D	$P(F \mid C,D)$	
c	d	0.9	0.1
c	$\bar{d}$	0.2	0.8
$\bar{c}$	d	0.9	0.1
$\bar{c}$	$\bar{d}$	0.01	0.99

Qualitative Specification

- Where does the qualitative specification come from?
  - Prior knowledge of causal relationships
  - Prior knowledge of modular relationships
  - Assessment from experts
  - Learning from data
  - We simply link a certain architecture (e.g. a layered graph)
  - …

Bayesian Network: Factorization Theorem



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

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

- **Theorem:**

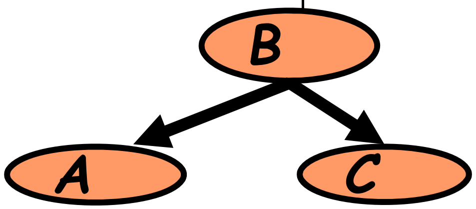
  Given a DAG, The most general form of the probability distribution that is consistent with the (probabilistic independence properties encoded in the) graph factors according to "node given its parents":

$$P(\mathbf{X}) = \prod_i P(X_i \mid \mathbf{X}_{\pi_i})$$

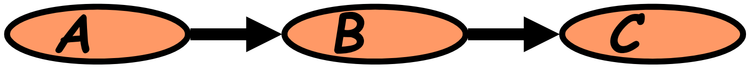
  where $\mathbf{X}_{\pi_i}$ is the set of parents of xi. d is the number of nodes (variables) in the graph.

Local Structures & Independencies

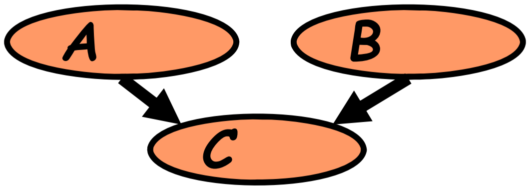
- Common parent
  - Fixing B decouples A and C
    "given the level of gene B, the levels of A and C are independent"



- Cascade
  - Knowing B decouples A and C
    "given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"

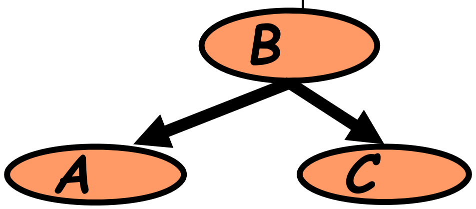


- V-structure
  - Knowing C couples A and B
    because A can "explain away" B w.r.t. C
    "If A correlates to C, then chance for B to also correlate to B will decrease"



- The language is compact, the concepts are rich!

A simple justification

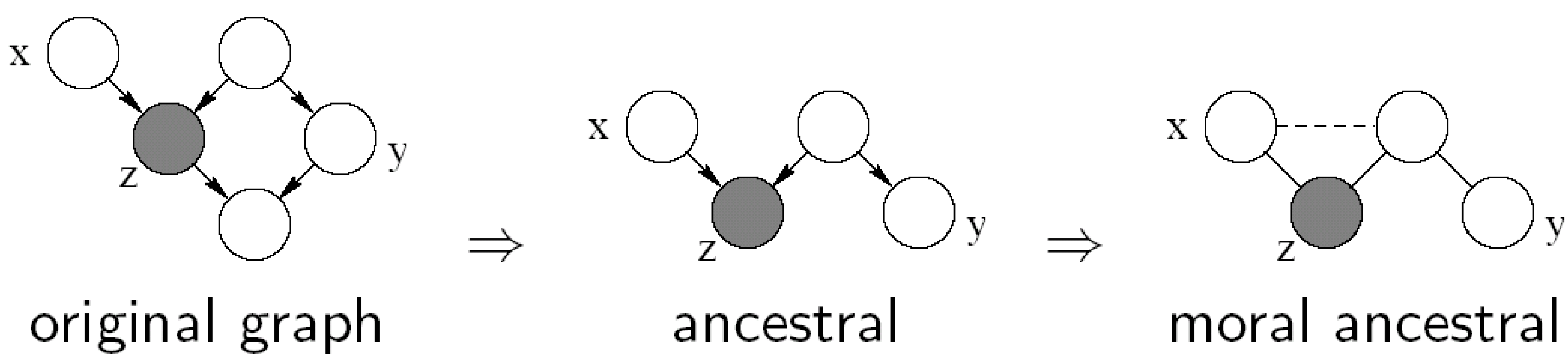


Graph separation criterion

- D-separation criterion for Bayesian networks (D for Directed edges):

  Definition: variables x and y are *D-separated* (conditionally independent) given z if they are separated in the *moralized* ancestral graph

- Example:



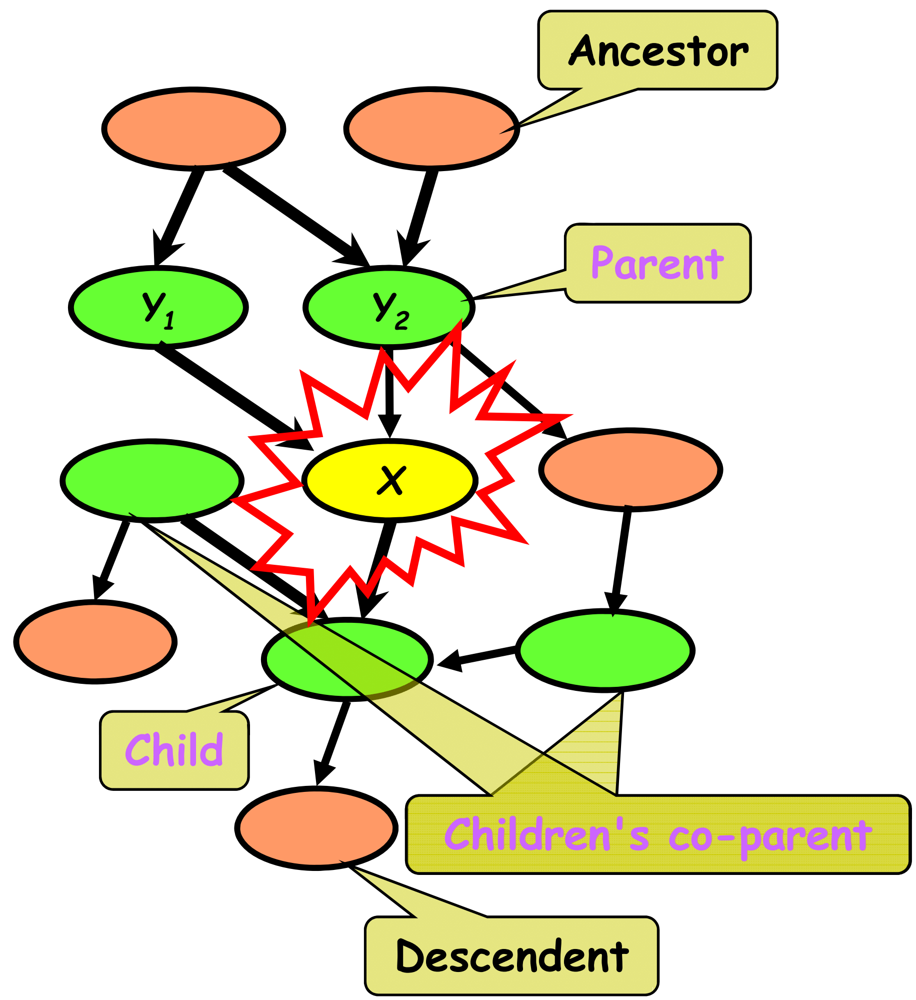
Local Markov properties of DAGs

Structure: *DAG*

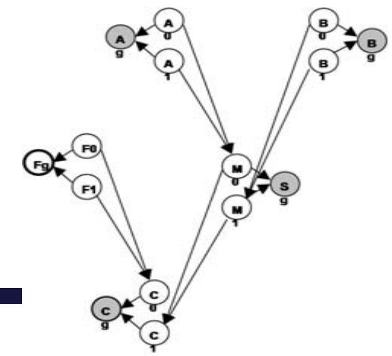
- Meaning: a node is conditionally independent of every other node in the network outside its Markov blanket

- Local conditional distributions (CPD) and the DAG completely determine the joint dist.

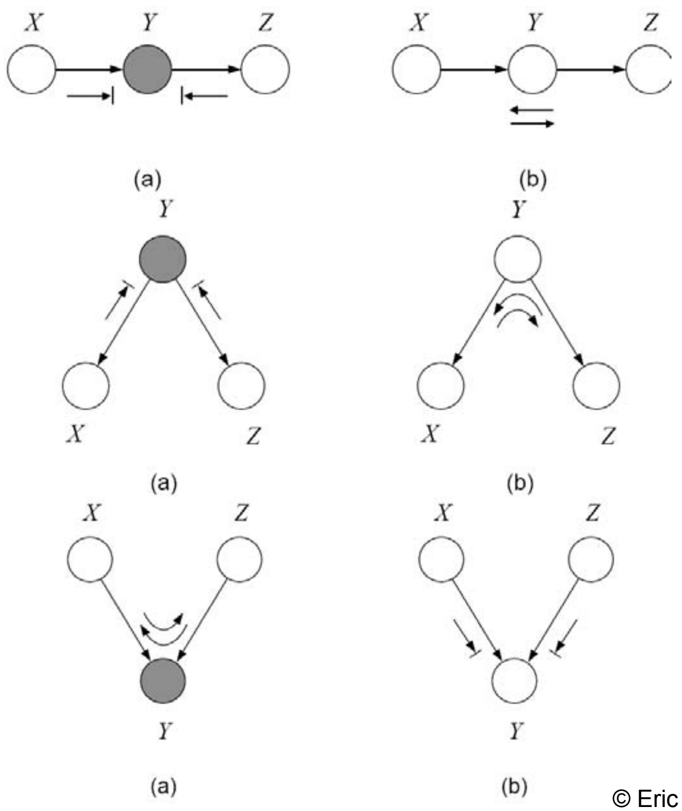
- Give causality relationships, and facilitate a generative process



Global Markov properties of DAGs



- X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "*Bayes-ball*" algorithm illustrated bellow (and plus some boundary conditions):

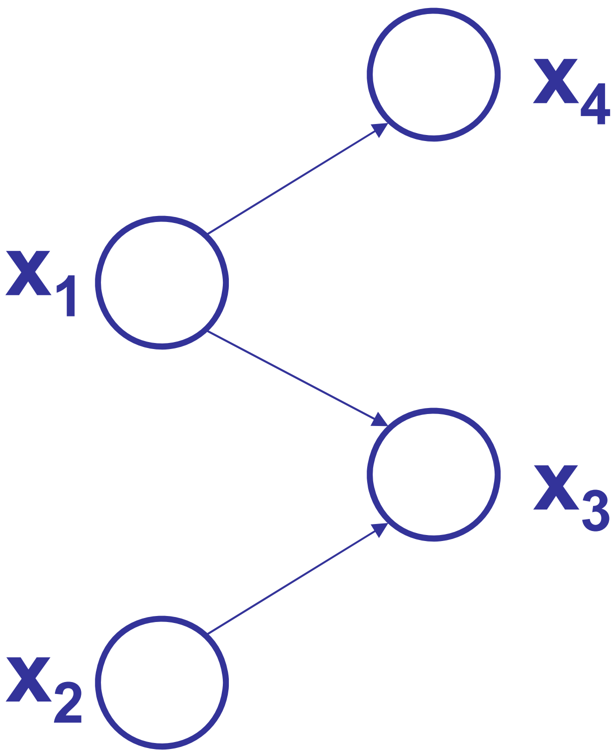


- **Defn: *I*(*G*)=all independence properties that correspond to d-separation:**

$$I(G) = \left\{ X \perp Z \middle| Y : \mathrm{dsep}_G(X; Z|Y) \right\}$$

- **D-separation is sound and complete**

Example:



- Complete the I(G) of this graph:

Essentially: A BN is a database of Pr. Independence statements among variables.

Towards quantitative specification of probability distribution

- Separation properties in the graph imply independence properties about the associated variables
- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

- **The Equivalence Theorem**

  For a graph G,

  Let $\mathcal{D}_1$ denote the family of all distributions that satisfy I(G),

  Let $\mathcal{D}_2$ denote the family of all distributions that factor according to G,

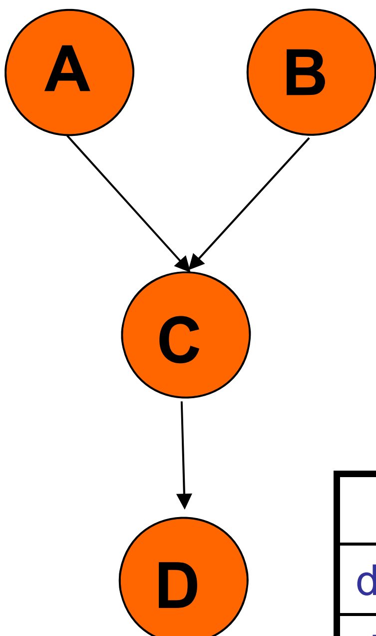
  Then $\mathcal{D}_1 \equiv \mathcal{D}_2$.

Conditional probability tables (CPTs)

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

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



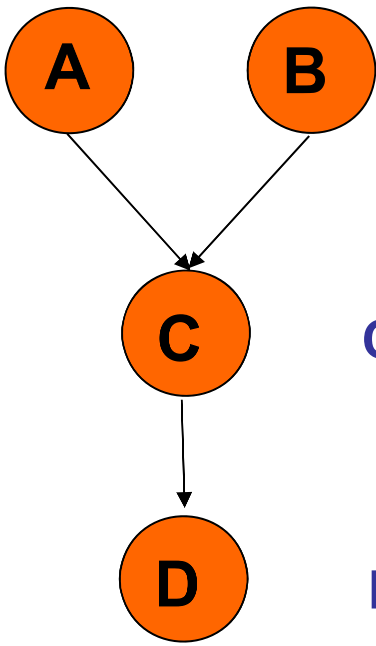
	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

	c^0	c^1
d^0	0.3	0.5
d^1	07	0.5

Conditional probability density func. (CPDs)

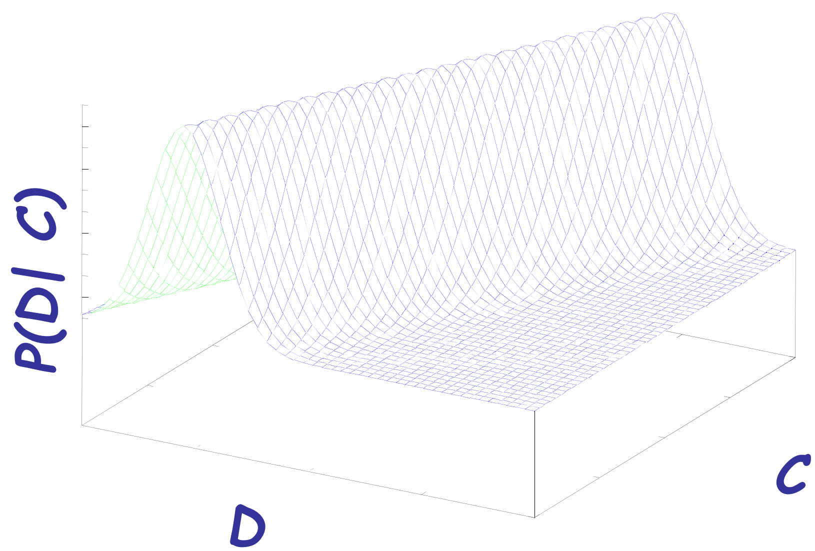
$A \sim N(\mu_a, \Sigma_a)$ $B \sim N(\mu_b, \Sigma_b)$

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

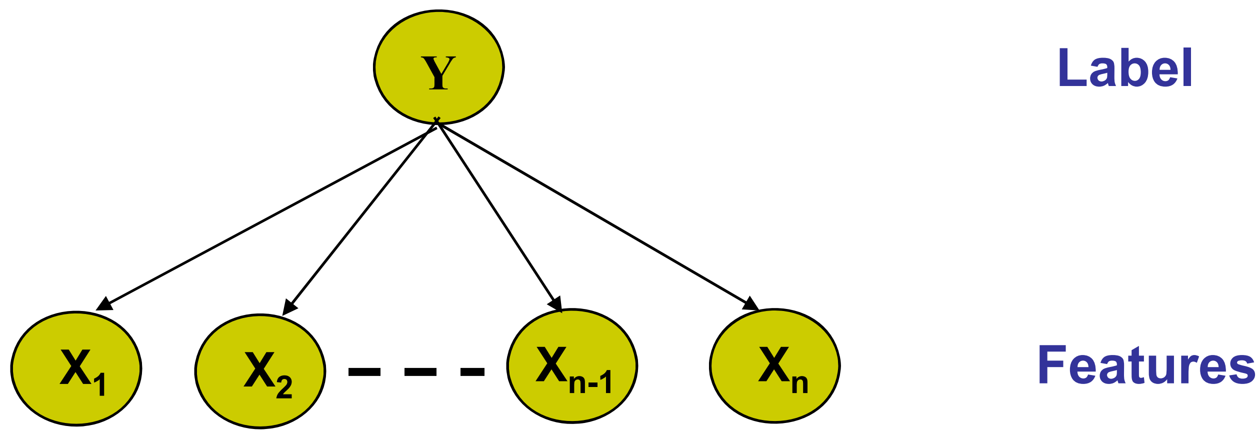


$C \sim N(A+B, \Sigma_c)$

$D \sim N(\mu_a + C, \Sigma_a)$



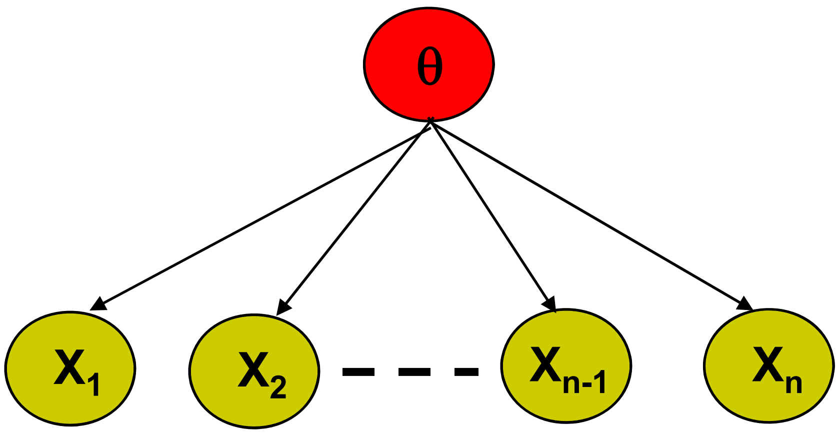
Conditional Independencies



What is this model

1. **When Y is observed?**
2. **When Y is unobserved?**

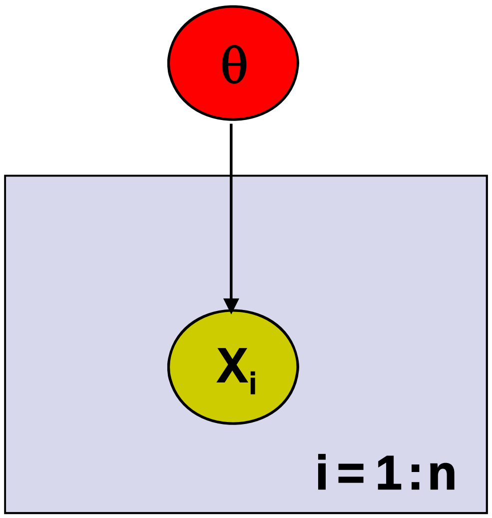
Conditionally Independent Observations



Model parameters

Data = $\{y_1, \ldots y_n\}$

"Plate" Notation



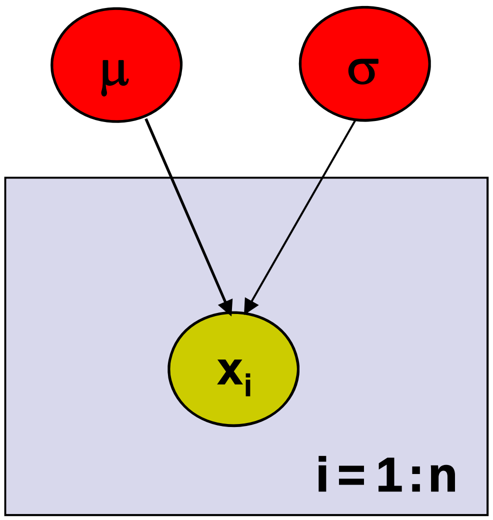
Model parameters

Data = $\{x_1,\ldots x_n\}$

Plate = rectangle in graphical model

variables within a plate are replicated
in a conditionally independent manner

Example: Gaussian Model



Generative model:

$p(x_1,\ldots x_n \mid \mu, \sigma) \quad = P \; p(x_i \mid \mu, \sigma)$

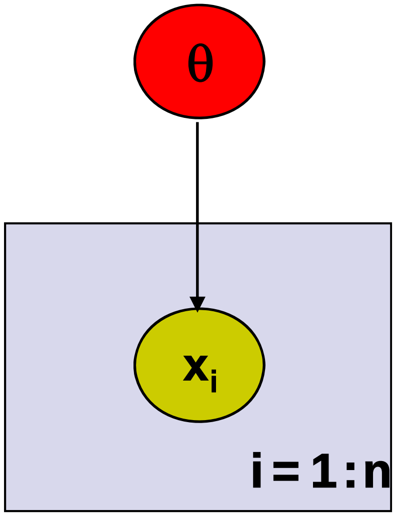
$= \;$ p(data | parameters)

$= \; p(D \mid \theta)$

where $\theta = \{\mu, \sigma\}$

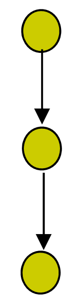
- Likelihood = p(data | parameters)
  $= p(D \mid \theta)$
  $= L(\theta)$
- Likelihood tells us how likely the observed data are conditioned on a particular setting of the parameters
  - Often easier to work with $\log L(\theta)$

Bayesian models



Summary

- Represent dependency structure with a directed acyclic graph
  - Node <-> random variable
  - Edges encode dependencies
    - Absence of edge -> conditional independence
  - Plate representation
  - A GM is a database of prob. Independence statement on variables



- The factorization theorem of the joint probability
  - Local specification → globally consistent distribution
  - Local representation for exponentially complex state-space
  - 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

- Support efficient inference and learning

Inference and Learning

- We now have compact representations of probability distributions: **BN**
- A BN *M* describes a unique probability distribution *P*
- Typical tasks:
  - Task 1: How do we answer **queries** about P?
    - We use **inference** as a name for the process of computing answers to such queries
  - Task 2: How do we estimate a **plausible model** *M* from data *D*?
    1. We use **learning** as a name for the process of obtaining point estimate of *M*.
    2. But for *Bayesian*, they seek $p(M \mid D)$, which is actually an **inference** problem.
    3. When not all variables are observable, even computing point estimate of *M* need to do **inference** to impute the *missing data*.

Inferential Query 1: Likelihood

- Most of the queries one may ask involve **evidence**
  - Evidence $\mathbf{x}_v$ is an assignment of values to a set $\mathbf{X}_v$ of nodes in the GM over varialbe set $\mathbf{X}=\{X_1, X_2, \ldots, X_n\}$
  - Without loss of generality $\mathbf{X}_v=\{X_{k+1}, \ldots, X_n\}$,
  - Write $\mathbf{X}_H=\mathbf{X}\backslash\mathbf{X}_v$ as the set of hidden variables, $\mathbf{X}_H$ can be $\varnothing$ or $\mathbf{X}$

- Simplest query: compute probability of evidence

$$P(\mathbf{x}_v) = \sum_{\mathbf{x}_H} P(\mathbf{X}_H, , \mathbf{X}_v) = \sum_{x_1} \cdots \sum_{x_k} P(x_1, \ldots, x_k, \mathbf{x}_v)$$

  - this is often referred to as computing the **likelihood** of $\mathbf{x}_v$

Inferential Query 2: Conditional Probability

- Often we are interested in the **conditional probability distribution** of a variable given the evidence

$$P(\mathbf{X_H} \mid \mathbf{X_V} = \mathbf{x_V}) = \frac{P(\mathbf{X_H}, \mathbf{x_V})}{P(\mathbf{x_V})} = \frac{P(\mathbf{X_H}, \mathbf{x_V})}{\sum_{\mathbf{x_H}} P(\mathbf{X_H} = \mathbf{x_H}, \mathbf{x_V})}$$

  - this is the ***a posteriori* belief** in $\mathbf{X_H}$, given evidence $\mathbf{x_v}$

- We usually query a subset $\mathbf{Y}$ of all hidden variables $\mathbf{X_H}=\{\mathbf{Y},\mathbf{Z}\}$ and "don't care" about the remaining, $\mathbf{Z}$:

$$P(\mathbf{Y} \mid \mathbf{x_V}) = \sum_{\mathbf{z}} P(\mathbf{Y}, \mathbf{Z} = \mathbf{z} \mid \mathbf{x_V})$$

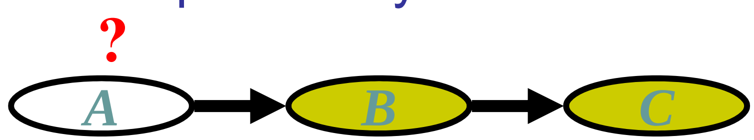
  - the process of summing out the "don't care" variables *z* is called **marginalization**, and the resulting $P(\mathbf{Y}|\mathbf{x_v})$ is called a **marginal** prob.

Applications of *a posteriori* Belief

- **Prediction**: what is the probability of an outcome given the starting condition

  - the query node is a descendent of the evidence

- **Diagnosis**: what is the probability of disease/fault given symptoms



  - the query node an ancestor of the evidence

- **Learning** under partial observation
  - fill in the unobserved values under an "EM" setting (more later)

- The directionality of information flow between variables is not restricted by the directionality of the edges in a GM
  - probabilistic inference can combine evidence form all parts of the network

Inferential Query 3: Most Probable Assignment

- In this query we want to find the **most probable joint assignment** (MPA) for ***some*** variables of interest

- Such reasoning is usually performed under some given evidence $\mathbf{x}_V$, and ignoring (the values of) other variables $\boldsymbol{Z}$:

$$\mathbf{Y}^* \mid \mathbf{x}_V = \arg\max_{\mathbf{y}} P(\mathbf{Y} \mid \mathbf{x}_V) = \arg\max_{\mathbf{y}} \sum_{\mathbf{z}} P(\mathbf{Y}, \mathbf{Z} = \mathbf{z} \mid \mathbf{x}_V)$$

  - this is the **maximum *a posteriori*** configuration of $\mathbf{Y}$.

Complexity of Inference

Thm:

Computing $P(X_H=x_H| x_v)$ in an arbitrary GM is NP-hard

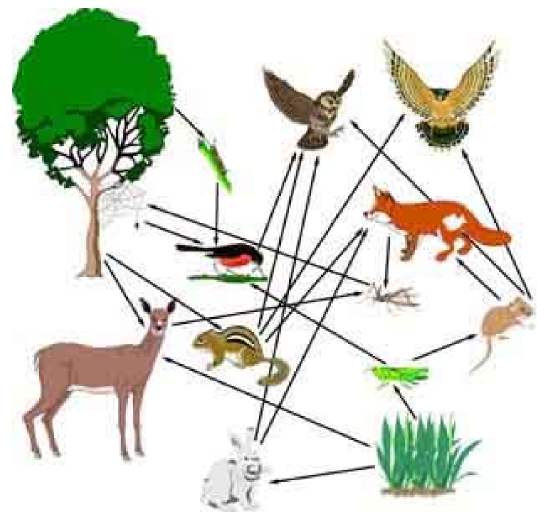
- **Hardness does not mean we cannot solve inference**
  - **It implies that we cannot find a general procedure that works efficiently for arbitrary GMs**
  - **For particular families of GMs, we can have provably efficient procedures**

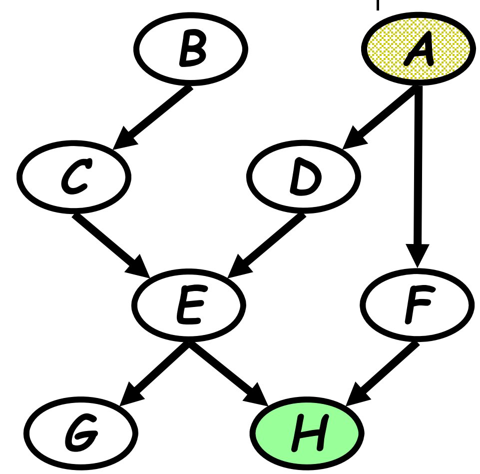
Approaches to inference

- Exact inference algorithms
  - The elimination algorithm
  - Belief propagation
  - The junction tree algorithms (but will not cover in detail here)
- Approximate inference techniques
  - Variational algorithms
  - Stochastic simulation / sampling methods
  - Markov chain Monte Carlo methods

Marginalization and Elimination

- A food web:





What is the probability that hawks are leaving given that the grass condition is poor?

Query: $P(\boldsymbol{h})$

$$P(h) = \sum_g \sum_f \sum_e \sum_d \sum_c \sum_b \sum_a P(a,b,c,d,e,f,g,h)$$

a naïve summation needs to enumerate over an exponential number of terms

- By chain decomposition, we get

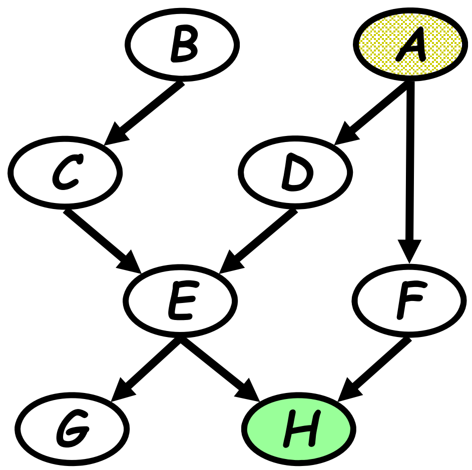
$$= \sum_g \sum_f \sum_e \sum_d \sum_c \sum_b \sum_a P(a)P(b)P(c \mid b)P(d \mid a)P(e \mid c,d)P(f \mid a)P(g \mid e)P(h \mid e,f)$$

Variable Elimination

- Query: *P(A | h)*
  - Need to eliminate: *B,C,D,E,F,G,H*
- Initial factors:

$$P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)P(g|e)P(h|e,f)$$

- Choose an elimination order: *H,G,F,E,D,C,B*

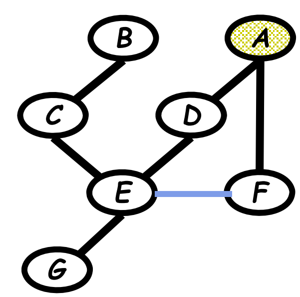


- Step 1:
  - **Conditioning** (fix the evidence node (i.e., *h*) on its observed value (i.e., $\tilde{h}$)):

$$m_h(e,f) = p(h = \tilde{h} \,|\, e,f)$$

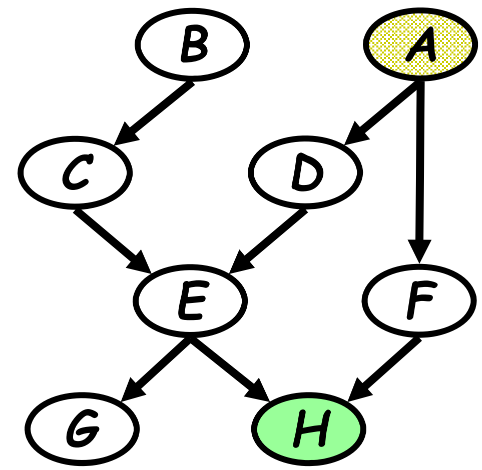
  - This step is isomorphic to a marginalization step:

$$m_h(e,f) = \sum_h p(h|e,f)\,\delta(h = \tilde{h})$$



Example: Variable Elimination

- Query: $P(B \mid h)$
  - Need to eliminate: B,C,D,E,F,G



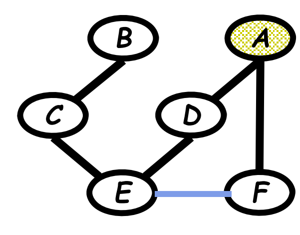
- Initial factors:

$$P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)P(g|e)P(h|e,f)$$
$$\Rightarrow P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)\underline{P(g|e)}m_h(e,f)$$

- Step 2: Eliminate G
  - compute

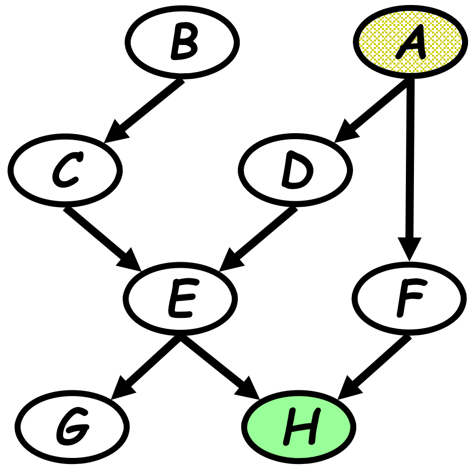
$$m_g(e)=\sum_g p(g|e)=1$$

$$\Rightarrow P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)m_g(e)m_h(e,f)$$
$$=P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)\underline{m_h(e,f)}$$



Example: Variable Elimination

- Query: $P(B\,|h)$
  - Need to eliminate: B,C,D,E,F



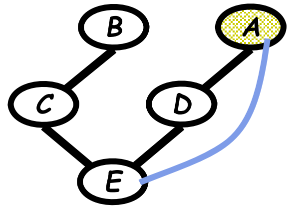
- Initial factors:

$$P(a)P(b)P(c\,|\,b)P(d\,|\,a)P(e\,|\,c,d)P(f\,|\,a)P(g\,|\,e)P(h\,|\,e,f)$$
$$\Rightarrow P(a)P(b)P(c\,|\,b)P(d\,|\,a)P(e\,|\,c,d)P(f\,|\,a)P(g\,|\,e)m_h(e,f)$$
$$\Rightarrow P(a)P(b)P(c\,|\,b)P(d\,|\,a)P(e\,|\,c,d)\underline{P(f\,|\,a)m_h(e,f)}$$

- Step 3: Eliminate F
  - compute

$$m_f(e,a) = \sum_f p(f\,|\,a)m_h(e,f)$$

$$\Rightarrow P(a)P(b)P(c\,|\,b)P(d\,|\,a)P(e\,|\,c,d)\underline{m_f(a,e)}$$

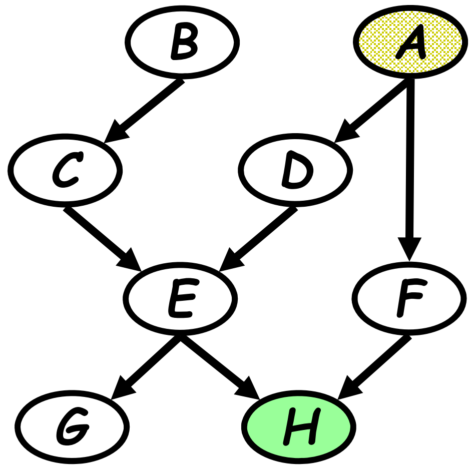


Example: Variable Elimination

- Query: $P(B \mid h)$
  - Need to eliminate: B, C, D, E

- Initial factors:

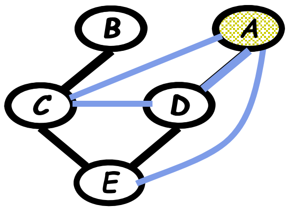
$$P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)P(g|e)P(h|e,f)$$
$$\Rightarrow P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)P(g|e)m_h(e,f)$$
$$\Rightarrow P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)m_h(e,f)$$
$$\Rightarrow P(a)P(b)P(c|b)P(d|a)\underline{P(e|c,d)m_f(a,e)}$$



- Step 4: Eliminate E
  - compute

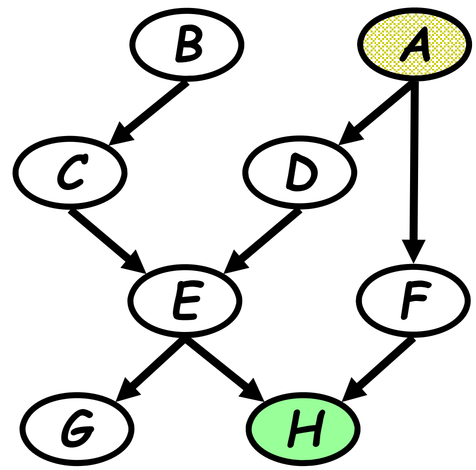
$$m_e(a,c,d) = \sum_e p(e|c,d)m_f(a,e)$$

$$\Rightarrow P(a)P(b)P(c|b)P(d|a)\underline{m_e(a,c,d)}$$



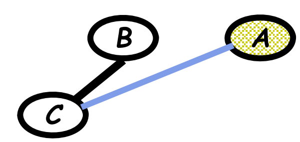
Example: Variable Elimination

- Query: $P(B\,|h)$
  - Need to eliminate: B,C,D
- Initial factors:



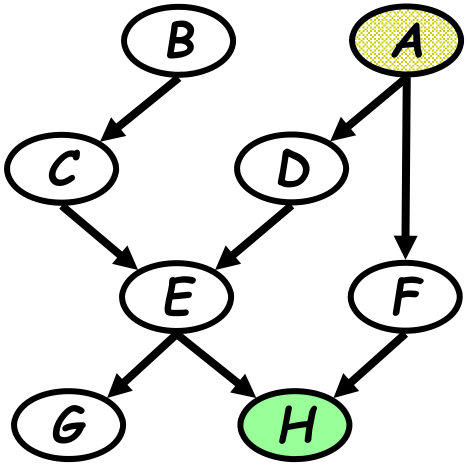
$$P(a)P(b)P(c\,|\,b)P(d\,|\,a)P(e\,|\,c,d)P(f\,|\,a)P(g\,|\,e)P(h\,|\,e,f)$$
$$\Rightarrow P(a)P(b)P(c\,|\,b)P(d\,|\,a)P(e\,|\,c,d)P(f\,|\,a)P(g\,|\,e)m_h(e,f)$$
$$\Rightarrow P(a)P(b)P(c\,|\,b)P(d\,|\,a)P(e\,|\,c,d)P(f\,|\,a)m_h(e,f)$$
$$\Rightarrow P(a)P(b)P(c\,|\,b)P(d\,|\,a)P(e\,|\,c,d)m_f(a,e)$$
$$\Rightarrow P(a)P(b)P(c\,|\,b)\underline{P(d\,|\,a)m_e(a,c,d)}$$

- Step 5: Eliminate D
  - compute $$m_d(a,c)=\sum_d p(d\,|\,a)m_e(a,c,d)$$



$$\Rightarrow P(a)P(b)P(c\,|\,d)\underline{m_d(a,c)}$$

Example: Variable Elimination



- Query: $P(B\,|h)$
  - Need to eliminate: B,C

- Initial factors:

$$P(a)P(b)P(c\,|\,d)P(d\,|\,a)P(e\,|\,c,d)P(f\,|\,a)P(g\,|\,e)P(h\,|\,e,f)$$
$$\Rightarrow P(a)P(b)P(c\,|\,d)P(d\,|\,a)P(e\,|\,c,d)P(f\,|\,a)P(g\,|\,e)m_h(e,f)$$
$$\Rightarrow P(a)P(b)P(c\,|\,d)P(d\,|\,a)P(e\,|\,c,d)P(f\,|\,a)m_h(e,f)$$
$$\Rightarrow P(a)P(b)P(c\,|\,d)P(d\,|\,a)P(e\,|\,c,d)m_f(a,e)$$
$$\Rightarrow P(a)P(b)P(c\,|\,d)P(d\,|\,a)m_e(a,c,d)$$
$$\Rightarrow P(a)P(b)\underline{P(c\,|\,d)m_d(a,c)}$$

- Step 6: Eliminate C
  - compute $$m_c(a,b)=\sum_c p(c\,|\,b)m_d(a,c)$$

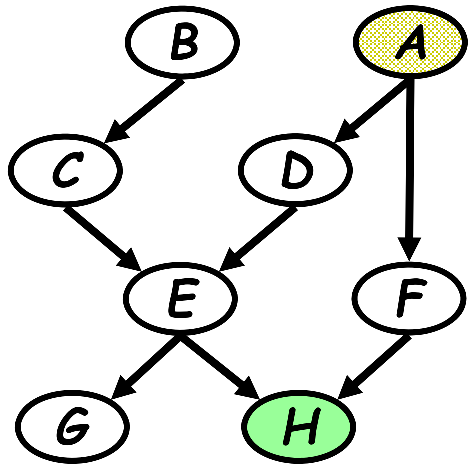
$$\Rightarrow P(a)P(b)P(c\,|\,d)\underline{m_d(a,c)}$$

Example: Variable Elimination

- Query: $P(B\,|h)$
  - Need to eliminate: B

- Initial factors:

$$P(a)P(b)P(c|d)P(d|a)P(e|c,d)P(f|a)P(g|e)P(h|e,f)$$
$$\Rightarrow P(a)P(b)P(c|d)P(d|a)P(e|c,d)P(f|a)P(g|e)m_h(e,f)$$
$$\Rightarrow P(a)P(b)P(c|d)P(d|a)P(e|c,d)P(f|a)m_h(e,f)$$
$$\Rightarrow P(a)P(b)P(c|d)P(d|a)P(e|c,d)m_f(a,e)$$
$$\Rightarrow P(a)P(b)P(c|d)P(d|a)m_e(a,c,d)$$
$$\Rightarrow P(a)P(b)P(c|d)m_d(a,c)$$
$$\Rightarrow P(a)\underline{P(b)m_c(a,b)}$$



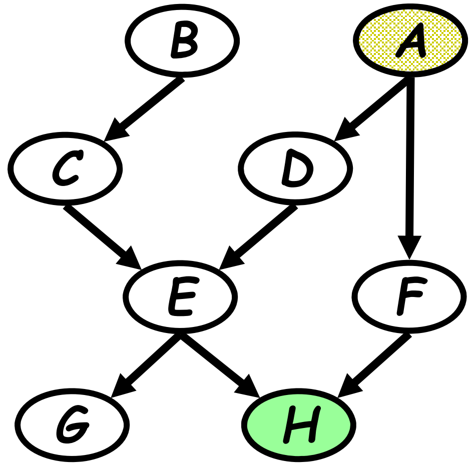
- Step 7: Eliminate B
  - compute

$$m_b(a) = \sum_b p(b)m_c(a,b)$$

$$\Rightarrow P(a)\underline{m_b(a)}$$

Example: Variable Elimination

- Query: $P(B \mid h)$
  - Need to eliminate: B



- Initial factors:

$$
\begin{aligned}
& P(a)P(b)P(c \mid d)P(d \mid a)P(e \mid c,d)P(f \mid a)P(g \mid e)P(h \mid e,f) \\
\Rightarrow\ & P(a)P(b)P(c \mid d)P(d \mid a)P(e \mid c,d)P(f \mid a)P(g \mid e)m_h(e,f) \\
\Rightarrow\ & P(a)P(b)P(c \mid d)P(d \mid a)P(e \mid c,d)P(f \mid a)m_h(e,f) \\
\Rightarrow\ & P(a)P(b)P(c \mid d)P(d \mid a)P(e \mid c,d)m_f(a,e) \\
\Rightarrow\ & P(a)P(b)P(c \mid d)P(d \mid a)m_e(a,c,d) \\
\Rightarrow\ & P(a)P(b)P(c \mid d)m_d(a,c) \\
\Rightarrow\ & P(a)P(b)m_c(a,b) \\
\Rightarrow\ & P(a)m_b(a)
\end{aligned}
$$

- Step 8: Wrap-up

$$p(a,\tilde{h}) = p(a)m_b(a)\,, \quad p(\tilde{h}) = \sum_a p(a)m_b(a)$$

$$\Rightarrow P(a \mid \tilde{h}) = \frac{p(a)m_b(a)}{\sum_a p(a)m_b(a)}$$

Complexity of variable elimination

- Suppose in one elimination step we compute

$$m_x(y_1,\ldots,y_k)=\sum_x m'_x(x,y_1,\ldots,y_k)$$

$$m'_x(x,y_1,\ldots,y_k)=\prod_{i=1}^{k} m_i(x,\mathbf{y}_{c_i})$$

This requires

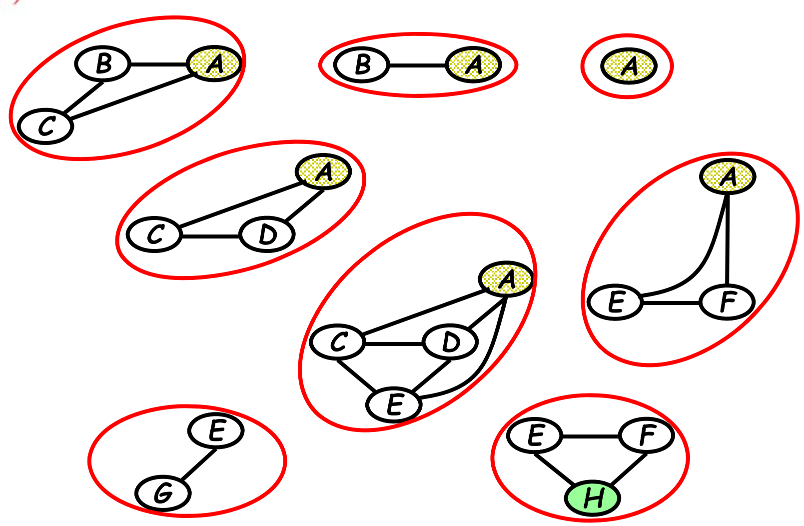
- $k \bullet |\mathrm{Val}(X)| \bullet \prod_i |\mathrm{Val}(\mathbf{Y}_{C_i})|$ multiplications
  - For each value of $x, y_1, \ldots, y_k$, we do k multiplications
- $|\mathrm{Val}(X)| \bullet \prod_i |\mathrm{Val}(\mathbf{Y}_{C_i})|$ additions
  - For each value of $y_1, \ldots, y_k$, we do $|Val(X)|$ additions

Complexity is **exponential** in number of variables in the intermediate factor

Elimination Clique

- Induced dependency during marginalization is captured in elimination cliques
  - Summation <-> elimination
  - Intermediate term <-> elimination clique

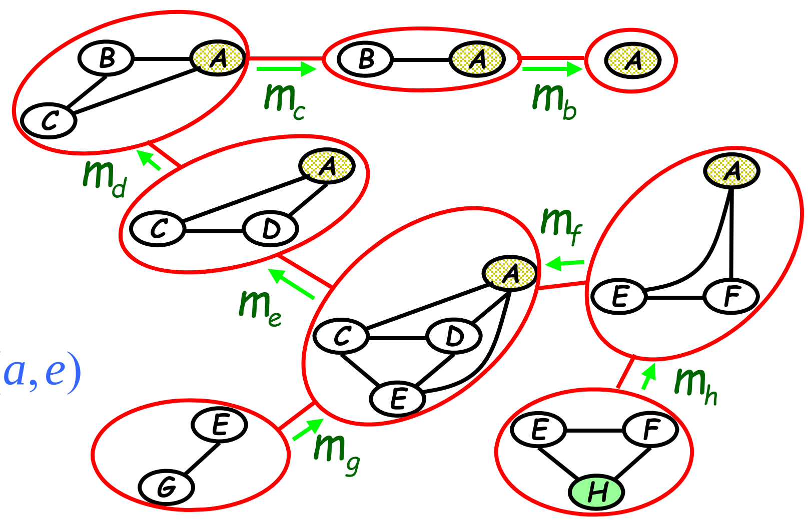
$$
\begin{aligned}
& P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)P(g|e)P(h|e,f) \\
\Rightarrow\ & P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)P(g|e)\phi_h(e,f) \\
\Rightarrow\ & P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)\phi_g(e)\phi_h(e,f) \\
\Rightarrow\ & P(a)P(b)P(c|b)P(d|a)P(e|c,d)\phi_f(a,e) \\
\Rightarrow\ & P(a)P(b)P(c|b)P(d|a)\phi_e(a,c,d) \\
\Rightarrow\ & P(a)P(b)P(c|b)\phi_d(a,c) \\
\Rightarrow\ & P(a)P(b)\phi_c(a,b) \\
\Rightarrow\ & P(a)\phi_b(a) \\
\Rightarrow\ & \phi(a)
\end{aligned}
$$



  - Can this lead to an generic inference algorithm?

From Elimination to Message Passing

- Elimination ≡ message passing on a **clique tree**

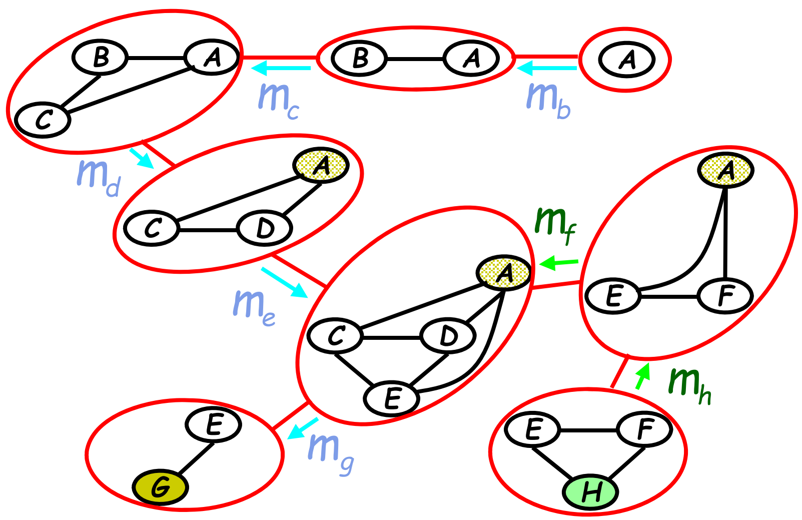


$$m_e(a,c,d) = \sum_e p(e \mid c,d) m_g(e) m_f(a,e)$$

- Messages can be reused

From Elimination to Message Passing

- Elimination ≡ message passing on a **clique tree**
  - **Another query ...**



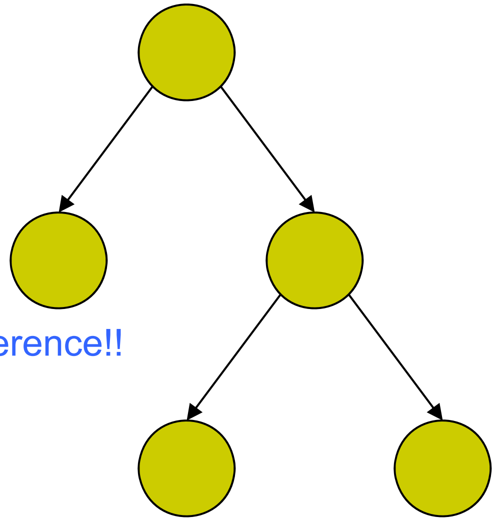
- Messages m_f and m_h are reused, others need to be recomputed

From elimination to message passing

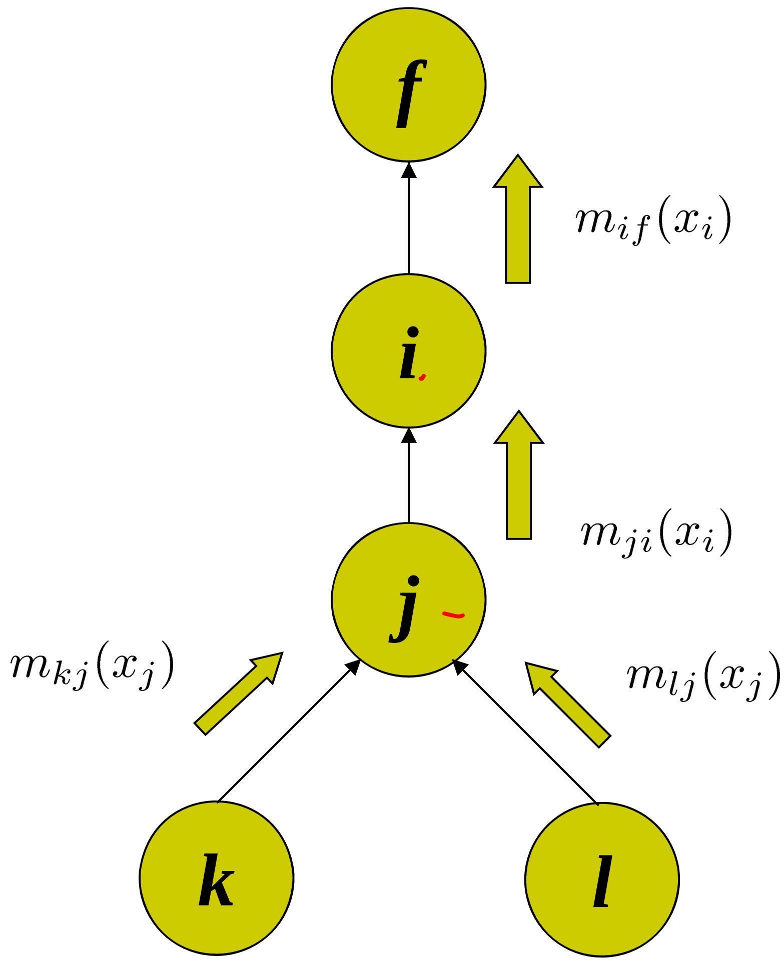
- Recall ELIMINATION algorithm:
  - Choose an ordering $\mathcal{Z}$ in which query node f is the final node
  - Place all potentials on an active list
  - Eliminate node i by removing all potentials containing i, take sum/product over x_i.
  - Place the resultant factor back on the list

- For a TREE graph:
  - Choose query node f as the root of the tree
  - View tree as a directed tree with edges pointing towards from f
  - Elimination ordering based on depth-first traversal
  - Elimination of each node can be considered as message-passing (or Belief Propagation) directly along tree branches, rather than on some transformed graphs

  → thus, we can use the tree itself as a data-structure to do general inference!!



Message passing for trees



Let $m_{ij}(x_i)$ denote the factor resulting from eliminating variables from bellow up to ***i***, which is a function of x_i:

$$m_{ji}(x_i) = \sum_{x_j} \left(\psi(x_j)\psi(x_i, x_j) \prod_{k \in N(j)\setminus i} m_{kj}(x_j) \right)$$

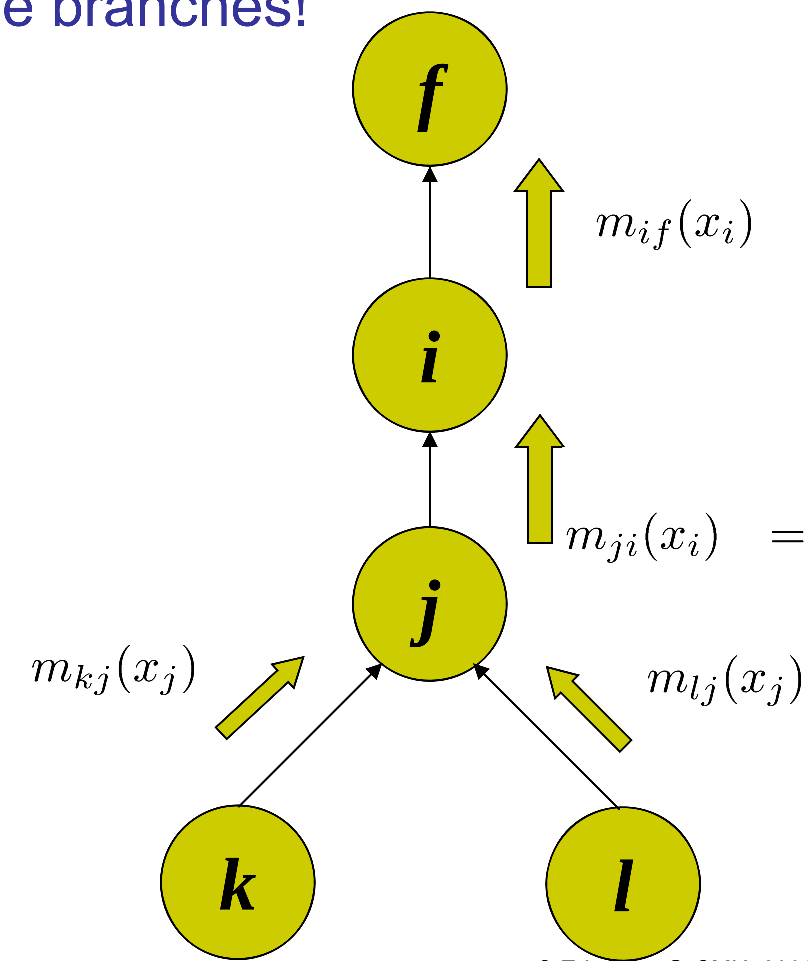
This is reminiscent of a ***message*** sent from ***j*** to ***i***.

$$m_{ji}(x_i) = \sum_{x_j} \left(\psi(x_j)\psi(x_i, x_j) \prod_{k \in N(j)\setminus i} m_{kj}(x_j) \right)$$

$$p(x_f) \propto \psi(x_f) \prod_{e \in N(f)} m_{ef}(x_f)$$

$m_{ij}(x_i)$ represents a "belief" of x_i from x_j!

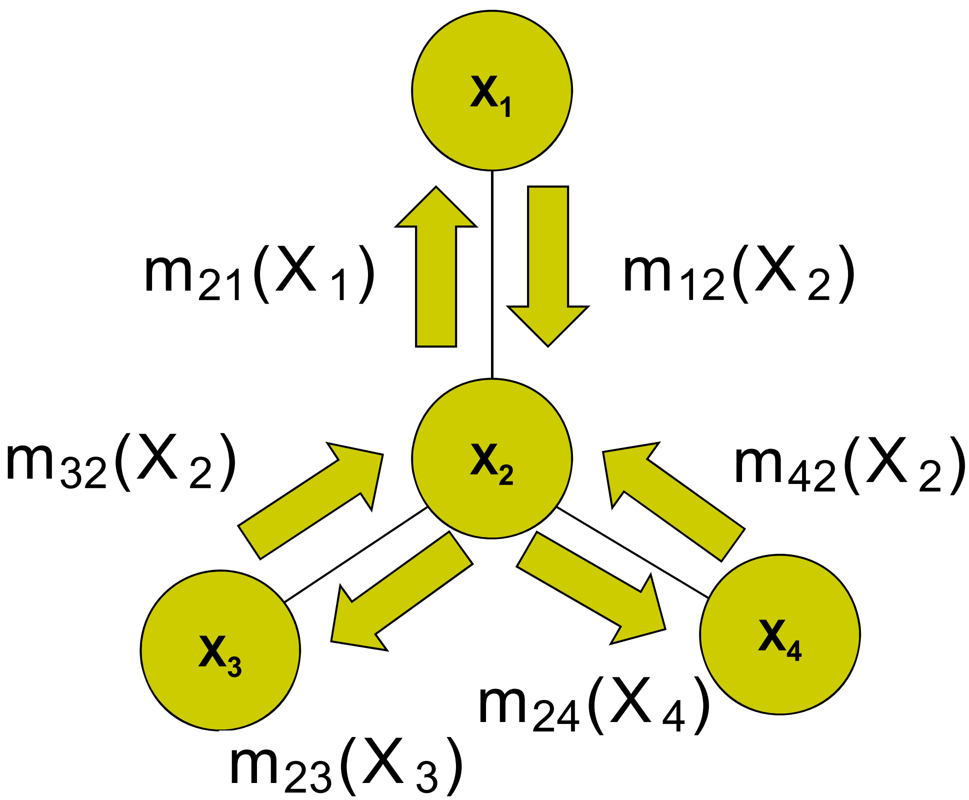
- Elimination on trees is equivalent to message passing along tree branches!



$$m_{ji}(x_i) = \sum_{x_j} \left(\psi(x_j)\psi(x_i, x_j) \prod_{k \in N(j) \setminus i} m_{kj}(x_j) \right)$$

The message passing protocol:

- A two-pass algorithm:



Belief Propagation (SP-algorithm): Sequential implementation

$\textsc{Sum-Product}(\mathcal{T}, E)$
  $\textsc{Evidence}(E)$
  $f = \textsc{ChooseRoot}(\mathcal{V})$
  for $e \in \mathcal{N}(f)$
    $\textsc{Collect}(f, e)$
  for $e \in \mathcal{N}(f)$
    $\textsc{Distribute}(f, e)$
  for $i \in \mathcal{V}$
    $\textsc{ComputeMarginal}(i)$

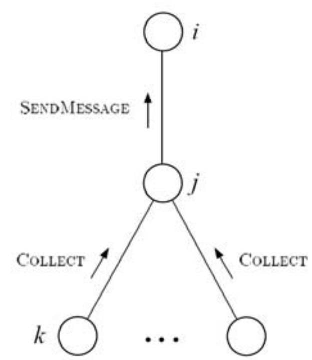
$\textsc{Evidence}(E)$
  for $i \in E$
    $\psi^E(x_i) = \psi(x_i)\delta(x_i, \bar{x}_i)$
  for $i \notin E$
    $\psi^E(x_i) = \psi(x_i)$

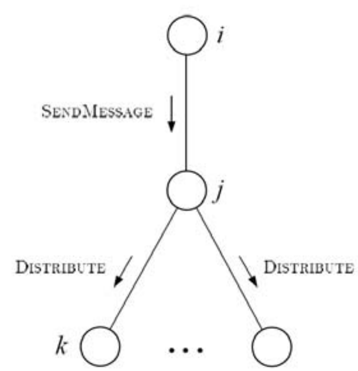
$\textsc{Collect}(i, j)$
  for $k \in \mathcal{N}(j)\backslash i$
    $\textsc{Collect}(j, k)$
  $\textsc{SendMessage}(j, i)$

$\textsc{Distribute}(i, j)$
  $\textsc{SendMessage}(i, j)$
  for $k \in \mathcal{N}(j)\backslash i$
    $\textsc{Distribute}(j, k)$

$\textsc{SendMessage}(j, i)$
$$m_{ji}(x_i) = \sum_{x_j}\left(\psi^E(x_j)\psi(x_i, x_j) \prod_{k \in \mathcal{N}(j)\backslash i} m_{kj}(x_j)\right)$$

$\textsc{ComputeMarginal}(i)$
$$p(x_i) \propto \psi^E(x_i) \prod_{j \in \mathcal{N}(i)} m_{ji}(x_i)$$





Inference on general GM

- Now, what if the GM is not a tree-like graph?

- Can we still directly run message
  message-passing protocol along its edges?

- For non-trees, we do not have the guarantee that message-passing will be consistent!

- Then what?
  - Construct a graph data-structure from P that has a tree structure, and run message-passing on it!

→ Junction tree algorithm

Summary

- The simple Eliminate algorithm captures the key algorithmic Operation underlying probabilistic inference:
  --- That of taking a sum over product of potential functions

- The computational complexity of the Eliminate algorithm can be reduced to purely graph-theoretic considerations.

- This graph interpretation will also provide hints about how to design improved inference algorithms

- What can we say about the overall computational complexity of the algorithm? In particular, how can we control the "size" of the summands that appear in the sequence of summation operation.