

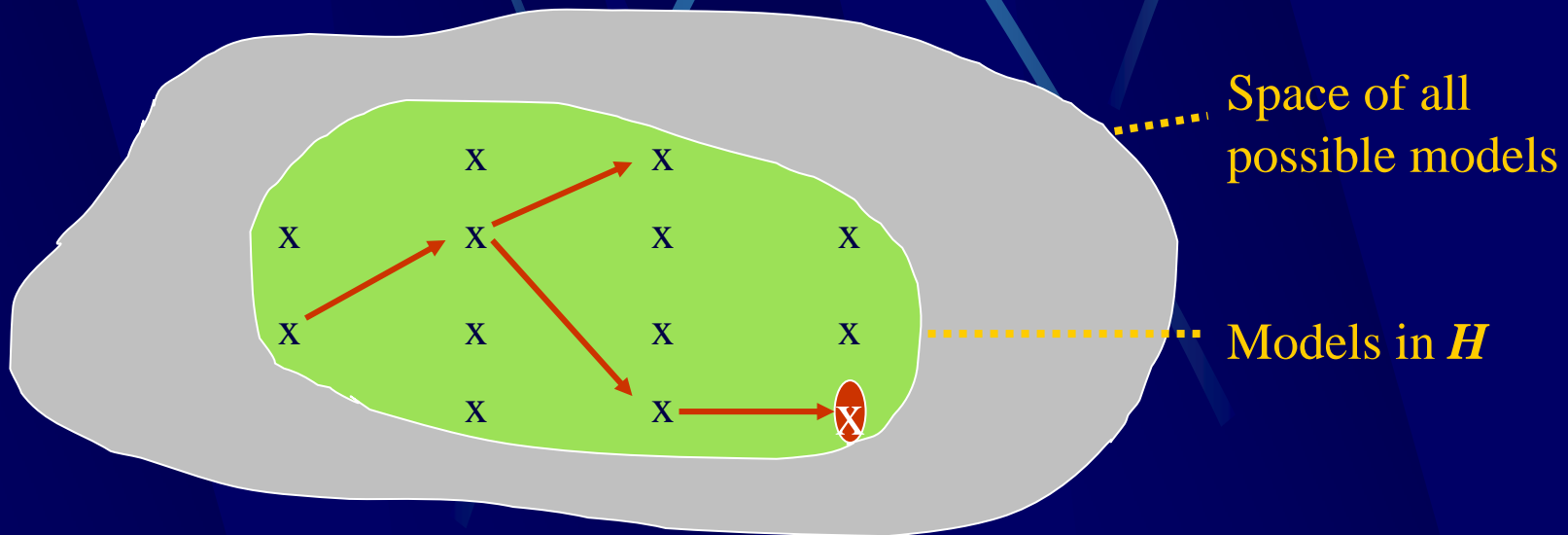
Statistical Genomics: Making Sense Of All the Data

Bayesian Networks

**C.F. Aliferis M.D., Ph.D.
May 22nd, Vanderbilt University**

Understanding Inductive Machine Learning

- Inductive Machine Learning algorithms can be designed and analysed using the following framework:
 - A language L in which we express models. The set of all possible models expressible in L constitutes our hypothesis space H
 - A scoring metric M tells us how good is a particular model
 - A search procedure S helps us identify the best model in H



Bayesian Networks: Overview

- A Note On Terminology
- Brief Historical Perspective
- The Bayesian Network Model and Its Uses
- Learning BNs
- Reference & Resources

Bayesian Networks: A Note On Terminology

- Bayesian Networks (or “Nets”): generic name
- Belief Networks: subjective probability-based, non-causal
- Causal Probabilistic Networks: frequentist probability-based, causal

Bayesian Networks: A Note On Terminology

- Various other names for special model classes:
 - Influence Diagrams (Howard and Mathesson): incorporate decision and utility nodes. Used for decision analyses
 - Dynamic Bayesian Networks (Dagum et al.): temporal semantics. Used as alternatives to multivariate time series models and dynamic control
 - Markov Decision Processes (Dean et al.): for decision policy formulation in temporally-evolving domains
 - Modifiable Temporal Belief Networks (Aliferis et al.): for well-structured and very large problem models that involve time and causation and cannot be stored explicitly

Bayesian Networks: Historical Perspective

- Naïve Bayesian Model (mutually exclusive diseases, findings independent given diseases) predominant model for medical decision support systems in the 60's and early 70's because it requires linear number of parameters and computational steps (to total findings and diseases)
- Theorem 1 (Minsky, Peot): Naïve Bayes *heuristic usefulness* (expected classification performance) over all domains gets exponentially worse as number of variables increases
- Theorem 2 (see Mitchell): Full Bayesian classifier=perfect classifier
- However FBC impractical and serves as analytical tool only

Bayesian Networks: Historical Perspective

- In the late 70's and up to mid-80's this led to: Production Systems (i.e., rule-based systems, that is simplifications of first-order logic). The most influential version of PSs (Shortliffe, Buchanan) handled uncertainty through a modular account of subjective belief (*the Certainty Factor Calculus*)
- Theorem 3 (Heckerman): The CFC is inconsistent with probability theory unless rule-space search graph is a tree. Consequently, forward and backward reasoning cannot be combined in a CFC PS and still produce valid results

Bayesian Networks: Historical Perspective

- That led to research (late 80s) in Bayesian Networks which can vary expressiveness between the full dependency (or even the full Bayesian classifier) and the Naïve Bayes model (Pearl, Cooper)

Variables
Conditionally
Independent Given
Categories &
Categories Mutually
Exclusive



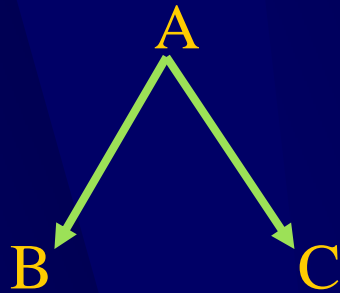
Variables
Conditionally
Dependent

Bayesian Networks: Historical Perspective

- In the early 90's researchers developed the first algorithms for learning BNs from data (Herskovits, Cooper, Heckerman)
- In the mid 90's researchers (Spirtes, Glymour, Sheines, Pearl, Verma) discovered methods to learn CPNs from observational data(!)
- Overall BNs is the brain child of computer scientists, medical informaticians, artificial intelligence researchers, and industrial engineers and is considered to be the representation language of choice for most biomedical Decision Support Systems today

Bayesian Networks: The Bayesian Network Model and Its Uses

- BN=Graph (Variables (nodes), dependencies (arcs)) + Joint Probability Distribution + Markov Property
- Graph has to be DAG (directed acyclic) in the standard BN model



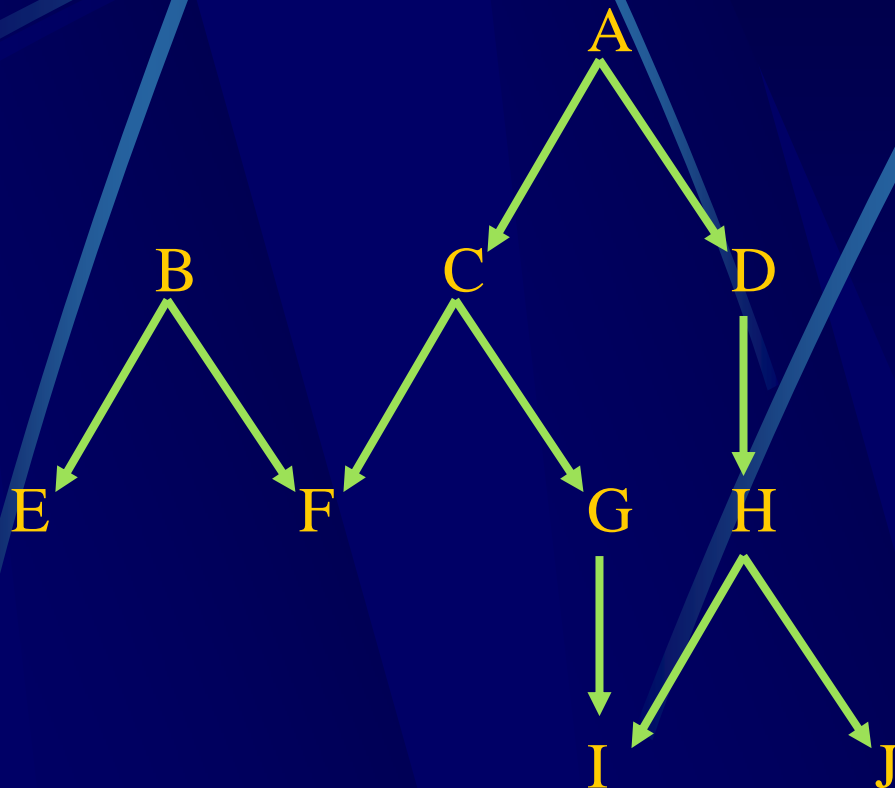
JPD

$P(A+, B+, C+) = 0.006$
$P(A+, B+, C-) = 0.014$
$P(A+, B-, C+) = 0.054$
$P(A+, B-, C-) = 0.126$
$P(A-, B+, C+) = 0.240$
$P(A-, B+, C-) = 0.160$
$P(A-, B-, C+) = 0.240$
$P(A-, B-, C-) = 0.160$

- Theorem 4 (Neapolitan): any JPD can be represented in BN form

Bayesian Networks: The Bayesian Network Model and Its Uses

- Markov Property: the probability distribution of any node N given its parents P is independent of any subset of the non-descendent nodes W of N



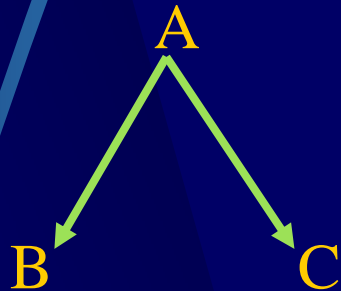
e.g., :

$$D \perp \{B, C, E, F, G \mid A\}$$

$$F \perp \{A, D, E, F, G, H, I, J \mid B, C\}$$

Bayesian Networks: The Bayesian Network Model and Its Uses

- Theorem 5 (Pearl): the Markov property enables us to decompose (factor) the joint probability distribution into a product of prior and conditional probability distributions



$$P(V) = \prod_i p(V_i | Pa(V_i))$$

The original JPD:

$P(A+, B+, C+) = 0.006$
$P(A+, B+, C-) = 0.014$
$P(A+, B-, C+) = 0.054$
$P(A+, B-, C-) = 0.126$
$P(A-, B+, C+) = 0.240$
$P(A-, B+, C-) = 0.160$
$P(A-, B-, C+) = 0.240$
$P(A-, B-, C-) = 0.160$

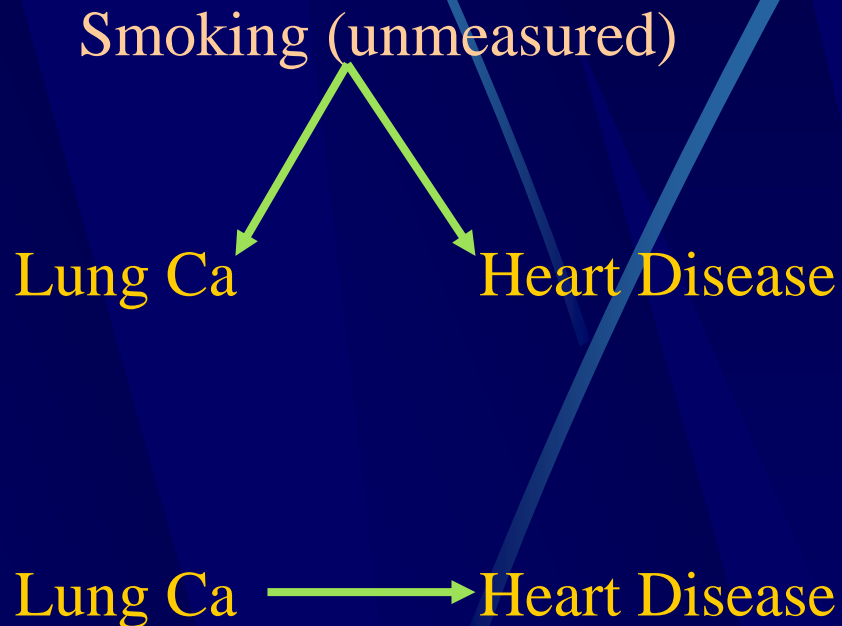
Becomes:

$P(A+) = 0.8$
$P(B+ A+) = 0.1$
$P(B+ A-) = 0.5$
$P(C+ A+) = 0.3$
$P(C+ A-) = 0.6$

**Up to
Exponential
Saving in
Number of
Parameters!**

Bayesian Networks: The Bayesian Network Model and Its Uses

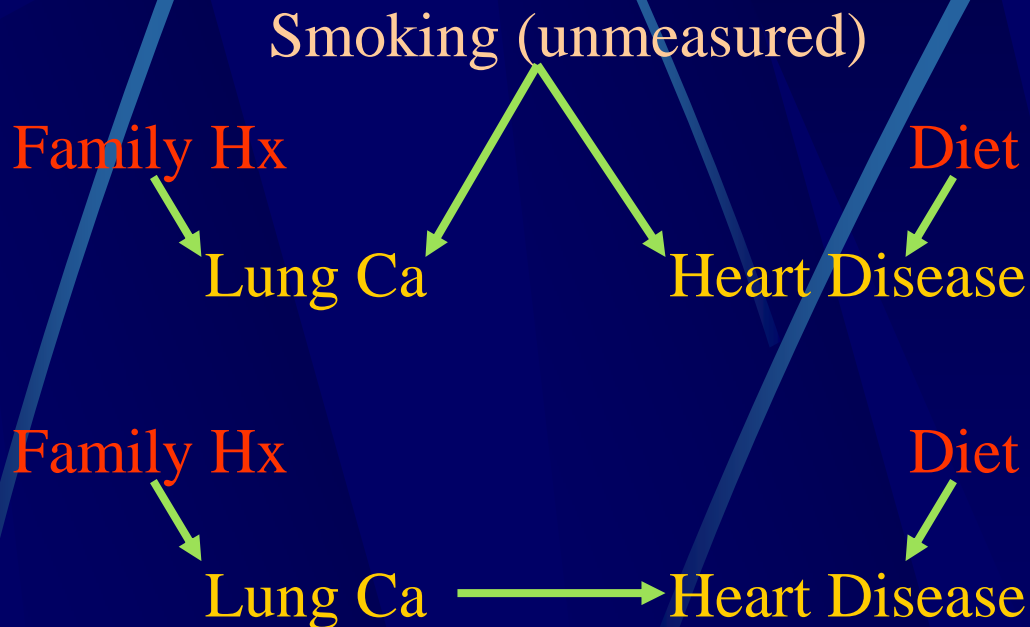
- BNs can help us learn causal relationships without doing experiments!



But Fisher says these two causal graphs are not distinguishable without doing an experiment (!?)

Bayesian Networks: The Bayesian Network Model and Its Uses

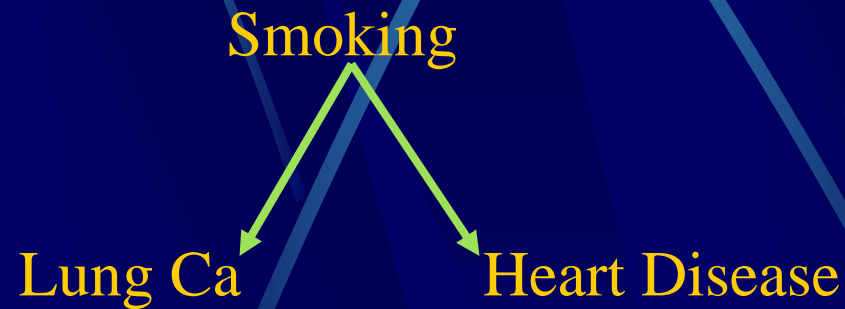
- BNs can help us learn causal relationships without doing experiments!



Fisher is right of course; however if we know a cause of each variable of interest then, in many cases, we can derive causal associations without an experiment

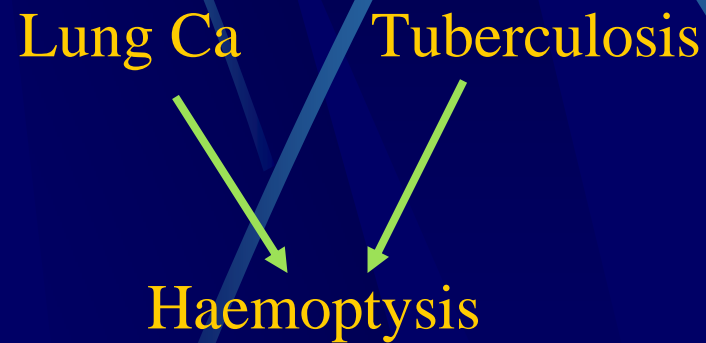
Bayesian Networks: The Bayesian Network Model and Its Uses

- The Markov property captures causality:
 - Revealing confounders



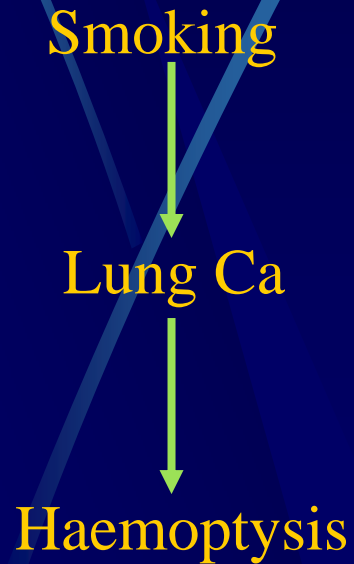
Bayesian Networks: The Bayesian Network Model and Its Uses

- The Markov property captures causality:
 - Modeling “explaining away”
 - Modeling/understanding selection bias



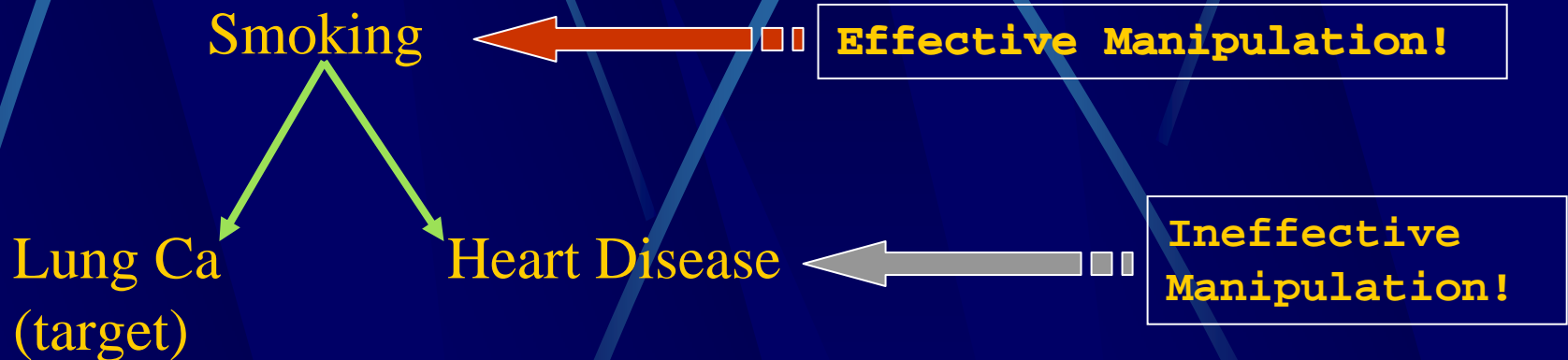
Bayesian Networks: The Bayesian Network Model and Its Uses

- The Markov property captures causality:
 - Modeling causal pathways



Bayesian Networks: The Bayesian Network Model and Its Uses

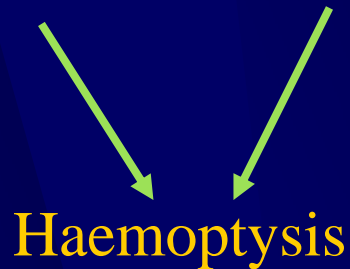
- The Markov property captures causality:
 - Manipulation in the presence of confounders



Bayesian Networks: The Bayesian Network Model and Its Uses

- The Markov property captures causality:
 - Manipulation in the presence of selection bias

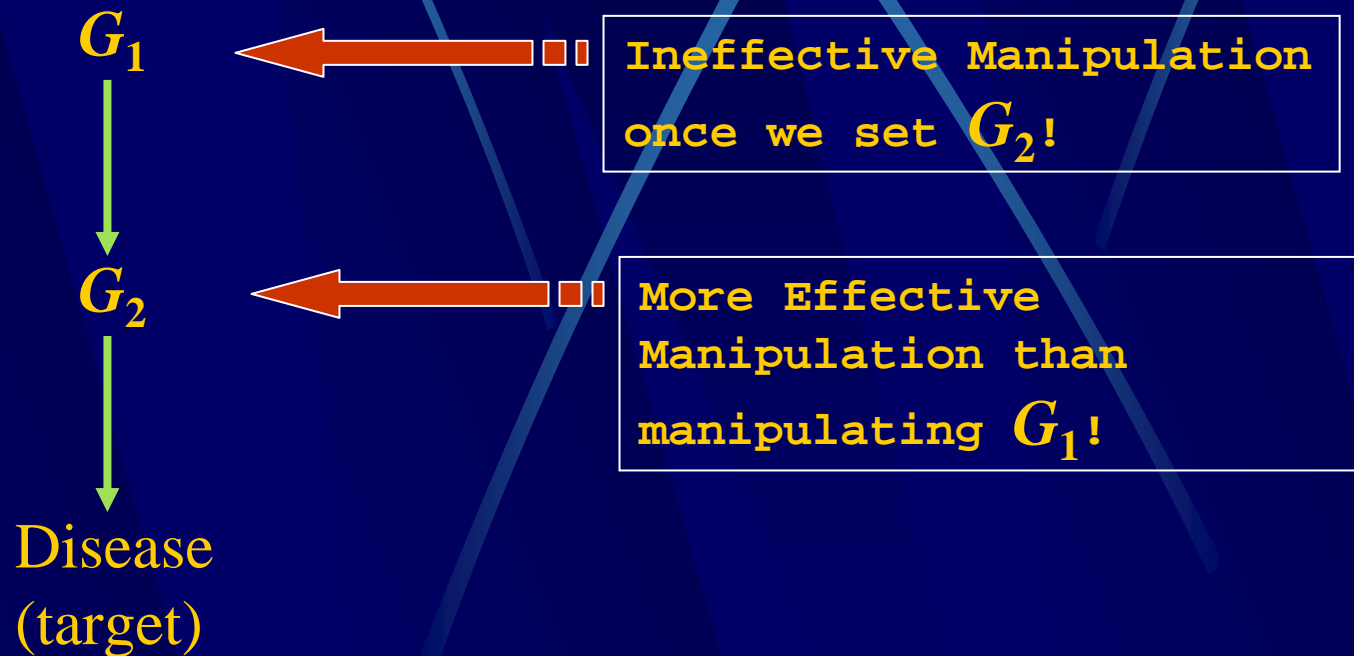
Lung Ca (target) Tuberculosis



**Ineffective
Manipulation!**

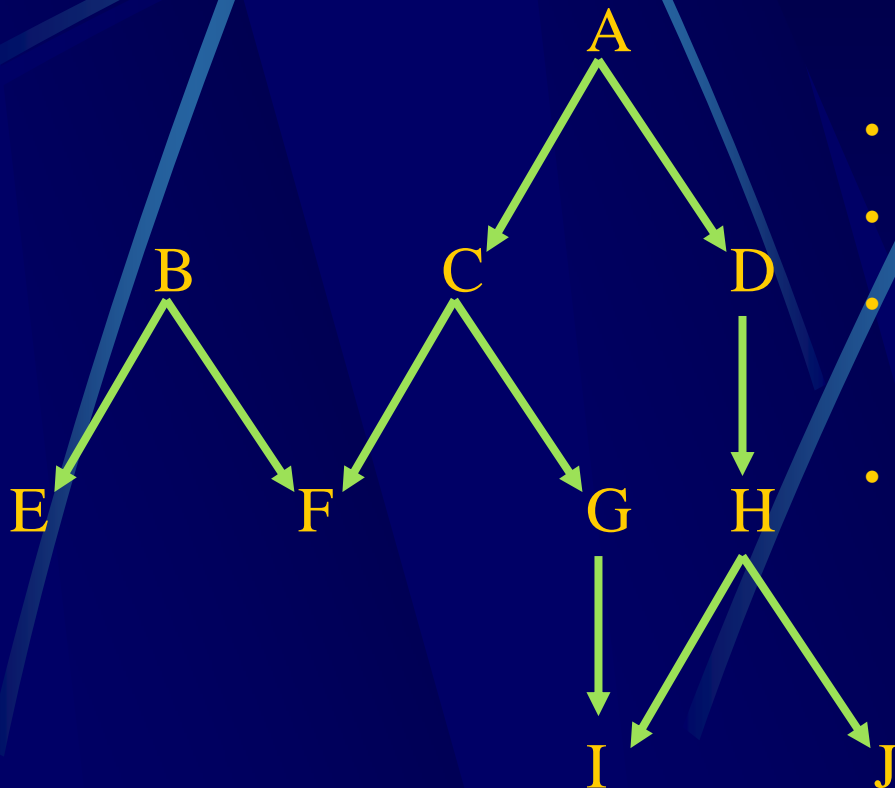
Bayesian Networks: The Bayesian Network Model and Its Uses

- The Markov property captures causality:
 - Identifying targets for manipulation in causal chains



Bayesian Networks: The Bayesian Network Model and Its Uses

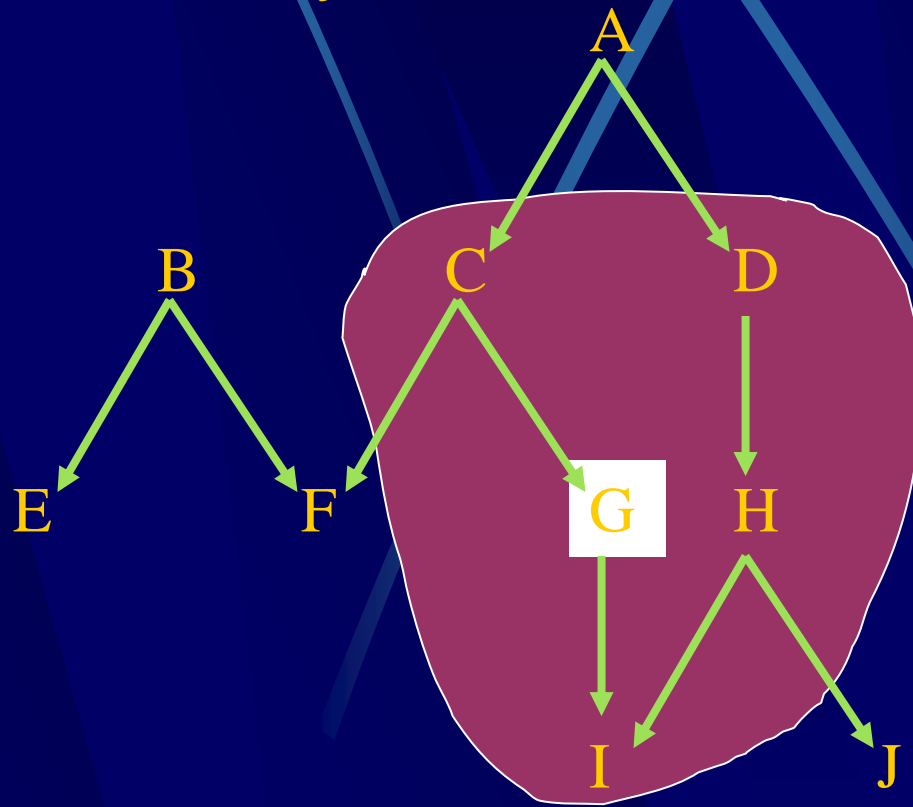
- Once we have a BN model of some domain we can ask questions:



- Forward: $P(D+, I- | A+) = ?$
- Backward: $P(A+ | C+, D+) = ?$
- Forward & Backward:
 $P(D+, C- | I+, E+) = ?$
- Arbitrary abstraction/Arbitrary predictors/predicted variables

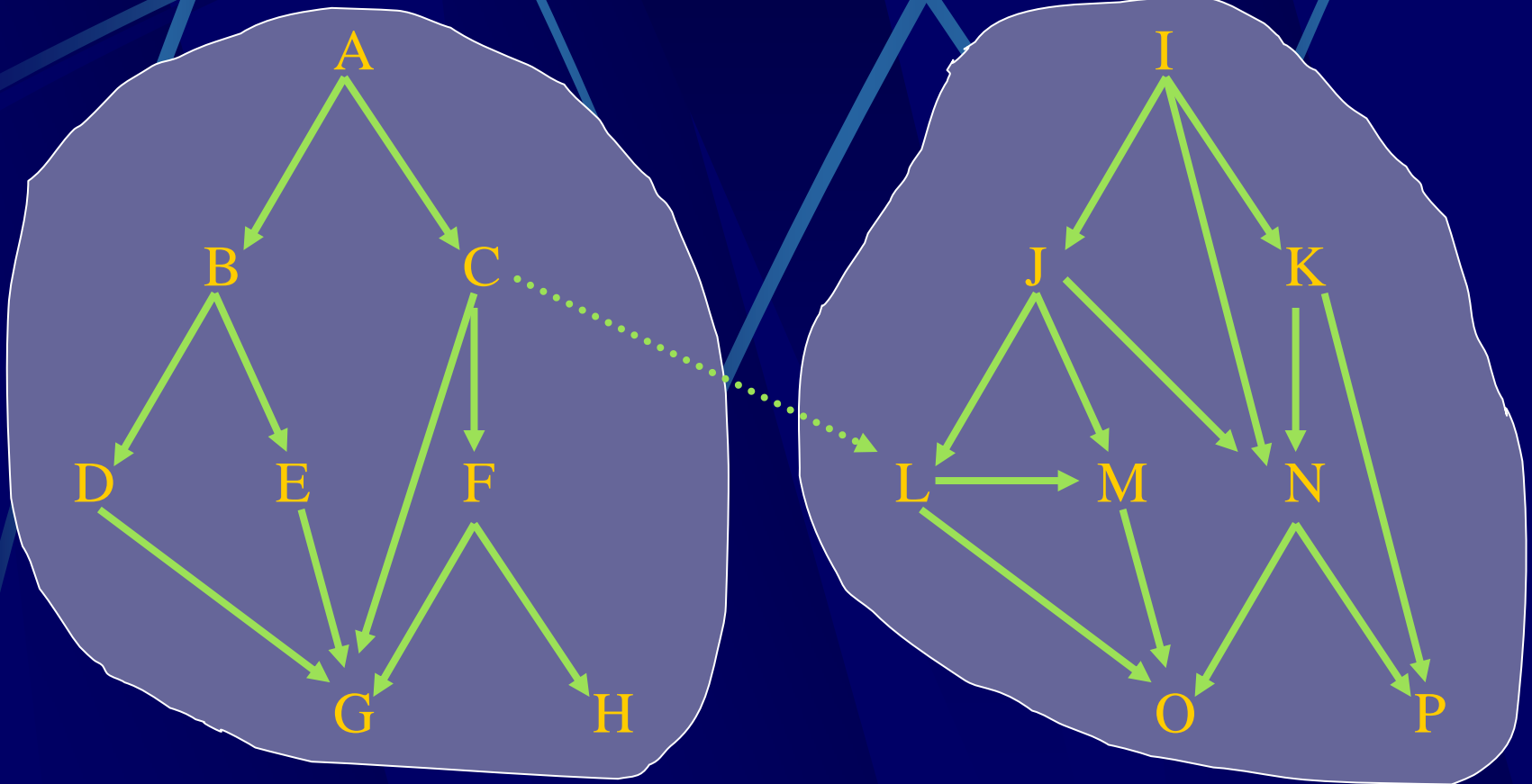
Bayesian Networks: The Bayesian Network Model and Its Uses

- The Markov property tells us which variables are important to predict a variable (Markov Blanket), thus providing a **principled way to reduce variable dimensionality**



Bayesian Networks: The Bayesian Network Model and Its Uses

- BNs can serve as sound (i.e., non-heuristic) **alternatives to associative (i.e., non-similarity-based) clustering**



Bayesian Networks: Practical Considerations

- Problem: very big networks (as in genomic datasets)
 - How good is learning with BNs?
 - “Sparse candidate” algorithm
 - Learn partial models
 - Reduce number of variables
 - Divide and conquer

Bayesian Networks: How Good Is learning?

- In discovering causal structure (Aliferis and Cooper, simulated data):
 - K2 algorithm discovers >70% of arcs 94% of the time
 - 94% of the time K2 does not add more than 10% superfluous arcs
 - Mean correctly identified arcs=94%
 - Mean superfluous arcs=4.7%
- In predicting outcomes & reducing number of predictors (Cooper, Aliferis et al., M. Fine pneumonia PORT data): K2 and Tetrad algorithms almost as good as best algorithm for domain, but requiring 6 instead of >200 variables

Bayesian Networks: Sparse Candidate Algorithm

Repeat

Select candidate parents C_i for each variable X_i

Set new best NW B to be G_n s.t. G_n maximizes a Bayesian score $\text{Score}(G|D)$ where G is a member of class of BNs for which: $\text{Pa}_G(X_i) \subseteq \text{Pa}_{B_{\text{prev}}}(X_i) \forall X_i$

Restriction Step

Maximization Step

Until Convergence

Return B

Bayesian Networks: Sparse Candidate Algorithm

- SCA proceeds by selecting up to k candidate parents for each variable on the basis of pair-wise association
- Then search is performed for a best network within the space defined by the union of all potential parents identified in the previous step
- The procedure is iterated by feeding the parents in the currently best network to the restriction step
- Theorem 6 (Friedman) : SCA monotonically improves the quality of the examined networks
- Convergence criterion: no gain in score, and maximum number of cycles with no improvement in score

Bayesian Networks: Learning Partial Models

- Partial model: feature (Friedman et al.)
- Examples:
 - Order relation (is X an ascendant or descendent of Y ?)
 - Markov Blanket membership (Is A in the MB of B ?)
- We want:

$$P(f(G|D)) = \sum_G (f(G) * p(G|D))$$

- And we approximate it by:

$$\text{Conf}(f) = \frac{1}{m} * \sum_{i=1}^m f(G_i)$$

Bayesian Networks: Reference

- Simple Bayes weakness:
 - M. Peot, Proc. Proc. UAI 96
 - M. Minsky, Transactions of IRE, 49:8-30, 1961
- Simple Bayes application:
 - H. Warner et al. Annals of NYAS, 115:2-16, 1964
 - F. de Dombal et al. BMJ, 1:376-380, 1972
- Full Bayesian Classifier:
 - T. Mitchell, Machine Learning, McGraw Hill, 1997
- Bayesian Networks as a knowledge representation:
 - J. Pearl, Probabilistic Reasoning in Expert Systems, Morgan Kaufmann, 1988
- Certainty Factor/PSs weaknesses:
 - D. Heckerman et al., Proc. UAI 86

Bayesian Networks: Reference

- Causal discovery using BNs:
 - P. Spirtes et al. , Causation, Prediction and Search, MIT Press 2000
 - C. Glymour, G. Cooper, Computation, Causation and Discovery, AAAI Press/MIT Press, 1999
 - C. Aliferis, G. Cooper, Proc. UAI 94
- Textbooks on BNs:
 - R. Neapolitan, Probabilistic Reasoning in Expert Systems, John Wiley, 1990
 - F. Jensen, An Introduction to Bayesian Networks, UCL Press, 1996
 - E. Castillo, et al. Expert Systems and Probabilistic Network Models, Springer 1997
- Learning BNs:
 - G. Cooper et al. Machine Learning 9:309-347, 1992
 - E. Herskovits, Report No. STAN-CS-91-1367 (Thesis)
 - D. Heckerman, Technical report Msr TR-95-06, 1995
 - J. Pearl, Causality, Gambridge University Press, 2001
 - N. Friedman et al. J Comput Biol, 7(3/4):601-620, 2000, and Proc. UAI 99
- Comparison to other learning algorithms:
 - G. Cooper, C. Aliferis et al. Artificial Intelligence in Medicine, 9:107-138, 1997

Bayesian Networks: For More...

**Medical Artificial Intelligence I:
Decision Support Systems and Machine
Learning For Biomedicine (BMI 330)**

SPRING 2001-02