



Causal Discovery Methods Using Causal Probabilistic Networks

MEDINFO 2004,
T02: Machine Learning Methods for Decision Support and Discovery
Constantin F. Aliferis & Ioannis Tsamardinos
Discovery Systems Laboratory
Department of Biomedical Informatics
Vanderbilt University

Desire for Causal Knowledge

■ **Diagnosis**

- Knowing that “people with cancer often have yellow-stained fingers and feel fatigue”, diagnose lung cancer

■ **Prevention**

- Need to know that “Smoking causes lung cancer” to reduce the risk of cancer

■ **Treatment**

- Knowing that “the presence of protein *X* causes cancer, inactivate protein *X*, using medicine *Y* that causes *X* to be inactive”

Causal Knowledge NOT required

Causal Knowledge required

Importance of Causal Discovery Today

- What SNP combination causes what disease
- How genes and proteins are organized in complex causal regulatory networks
- How behaviour causes disease
- How genotype causes differences in response to treatment
- How the environment modifies or even supersedes the normal causal function of genes

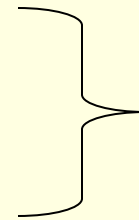
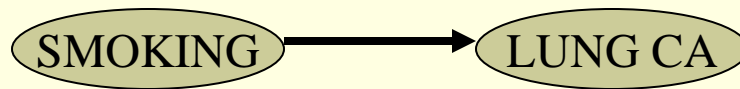
What is Causality?

- Thousands of years old problem, still debated
- Operational Informal Definition:
 - Assume the existence of a mechanism M capable of setting values for a variable A . We say that A can be manipulated by M to take the desired values.
 - Variable A causes variable B , *if*: in a hypothetical randomized controlled experiment in which A is randomly manipulated via M (i.e., all possible values a_i of A are randomly assigned to A via M) we would observe in the sample limit that $P(B=b | A=a_i) \neq P(B=b | A=a_j)$ for some $i \neq j$.
- Definition is stochastic
- Problems: self-referencing, ignores time-dependence, variables that need to be co-manipulated, etc.

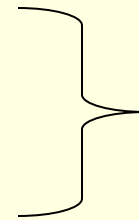
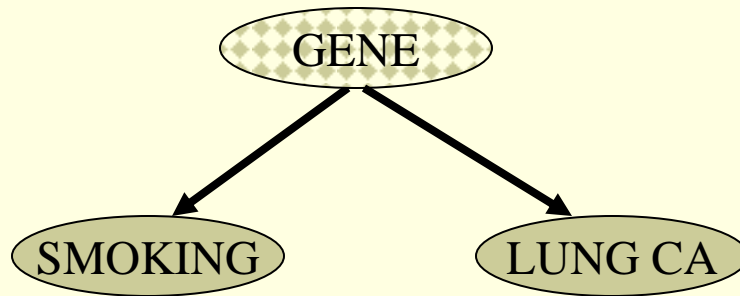
Causation and Association

- What is the relationship between the two?
- If A causes B, are A and B always associated?
- If A is associated with B are they always causes or effects of each other? (directly?, indirectly?, conditionally, unconditionally?)

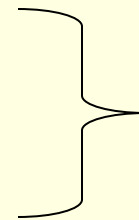
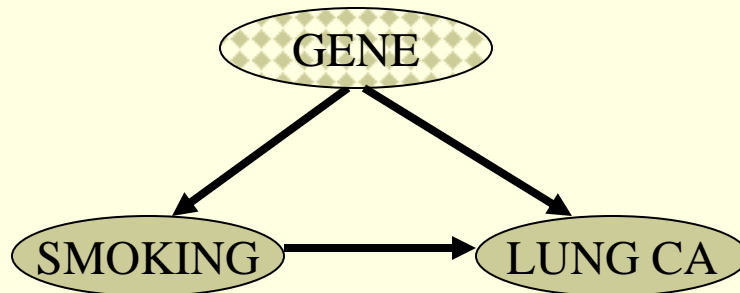
Statistical Indistinguishability



S1

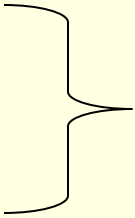
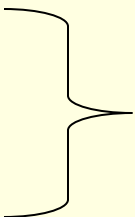
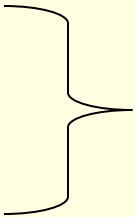
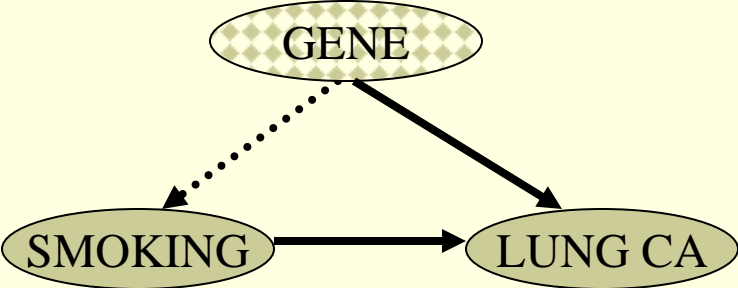
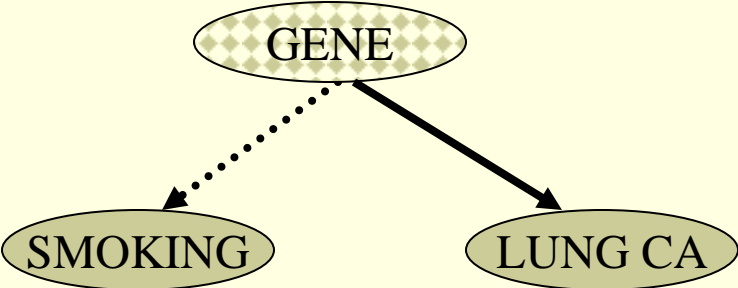
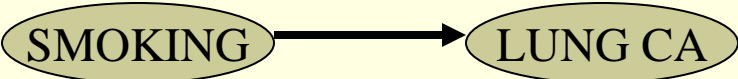


S2



S3

RANDOMIZED CONTROLLED TRIALS



S1

S2

S3



Association is still retained even after manipulating Smoking

RCTs Are *not* always feasible!

- Unethical (smoking)
- Costly/Time consuming (gene manipulation, epidemiology)
- Impossible (astronomy)
- Extremely large number

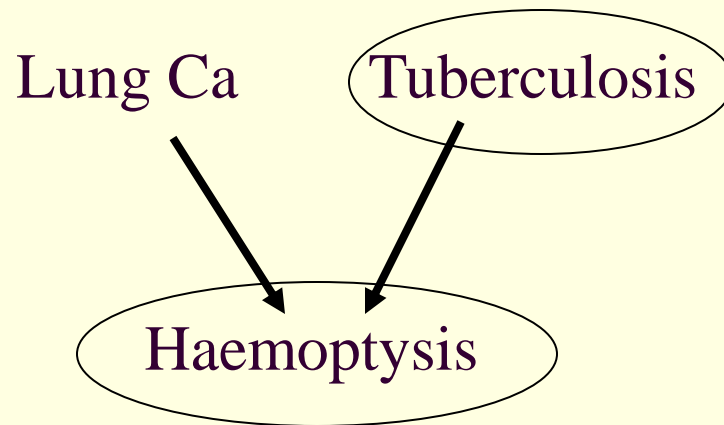
Large-Scale Causal Discovery without RCTs?

- Heuristics to the rescue...
- What is a heuristic?
 - When the heuristic condition holds, *most probably* a causal association holds

Pitfalls of Causal Heuristics

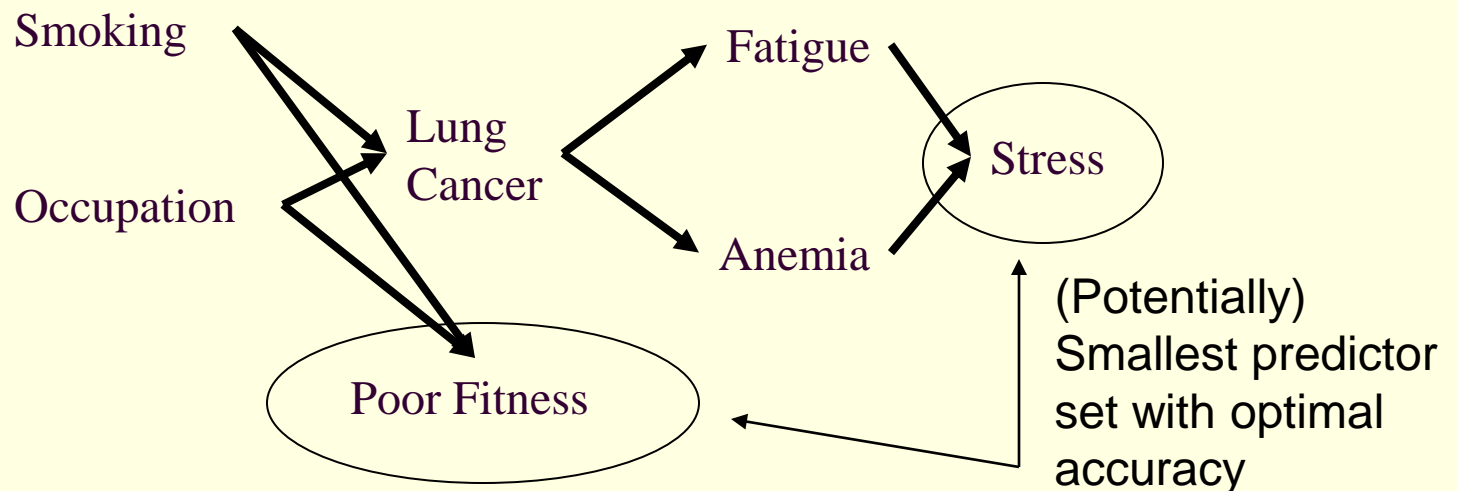
'If A is a robust and strong predictor of T then A is likely a cause of T '

- Example: Feature selection
- Example: Predictive Rules
- In the example: Tuberculosis is a strong predictor of Lung Cancer (when Haemoptysis is included as a predictor), but not causally associated with Lung Cancer



Pitfalls of Causal Heuristics

- ‘The closer A and T are in a causal sense, the stronger their correlation’ (localizes causality as well)
- Poor Fitness may be a stronger predictor (univariately) to lung cancer than either Occupation or Smoking individually; yet the latter predictors are causally “closer” to Lung Cancer



The Problem with Causal Discovery

- Causal heuristics are unreliable
- Causation is difficult to define
- RCTs are not always doable
- Major “causal knowledge” does not have RCT backing!

Formal Computational Causal Discovery from Observational Data

- Formal algorithms exist!
- Most are based on a graphical-probabilistic language called “Causal Probabilistic Networks (a.k.a. “Causal Bayesian Networks”)
- Well-characterized properties of
 - What types of causal relations they can learn
 - Under which conditions
 - What kind of errors they may make

Types of Causal Discovery Questions

- What will be the effect of a manipulation to the system
- Is A causing B , B causing A , or neither?
- Is A causing B directly (no other observed variables interfere)?
- What is the smallest set of variables for optimally effective manipulation of A ?
- Can we infer the presence of hidden confounder factors/variables?

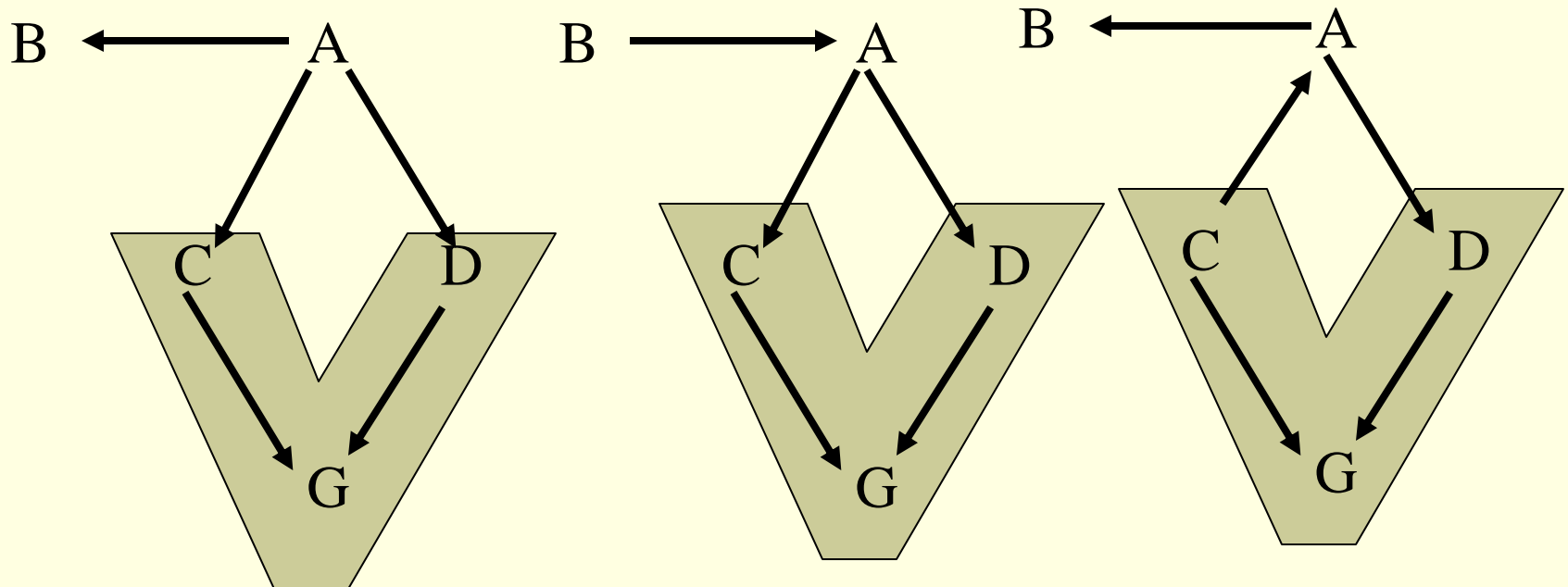
A Formal Language for Representing Causality

- Bayesian Networks
- Edges: probabilistic dependence
- Markov Condition: A node N is independent from non-descendants given its parents
- Probabilistic reasoning

- Causal Bayesian Networks
- Edges represent direct causal effects
- Causal Markov Condition: A node N is independent from non-descendants given its direct causes
- Probabilistic reasoning + causal inferences

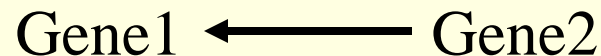
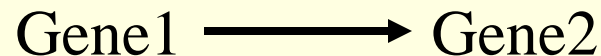
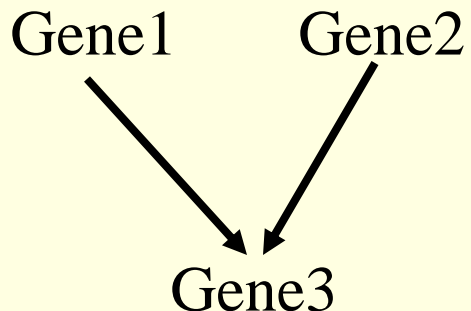
Causal Bayesian Networks

- There may be many (non-causal) BNs that capture the same distribution.
- All such BNs have the same edges (ignoring direction) same v-structures
- Statistically equivalent



Causal Bayesian Networks

- If there is a (faithful) Causal Bayesian Network that captures the data generation process, it has to have the same edges and same v-structures as any (faithful) Bayesian Network that is induced by the data.
 - We can infer what the direct causal relations are
 - We can infer some of the directions of the edges

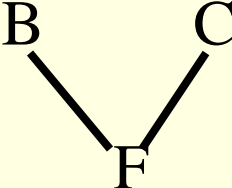


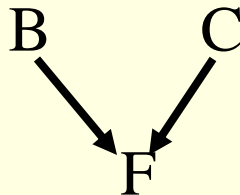
Faithfulness

- When d -separation \Leftrightarrow independence
- Intuitively, an open path between A and B means there is association between them in the data
- Previous discussion holds for faithful BNs only
- Faithful BN is a very large class of BNs

Learning Bayesian Networks: Constraint-Based Approach

- An edge $X - Y$ (of unknown direction) exists, if and only if for all sets of nodes S , $\text{Dep}(X, Y / S)$ (allows discovery of the edges)
- Test all subsets. If $\text{Dep}(X, Y | s)$ holds, add the edge, otherwise do not.

■ If structure  and for every set S that

contains F , $\text{Dep}(X, Y / S)$, then 

Learning Bayesian Networks: Constraint-Based Approach

- Tests of conditional dependences and independencies from the data
- Estimation using G^2 statistic, conditional mutual-information, etc.
- Infer structure and orientation from results of tests
- Based on the assumption these tests are accurate
- The larger the number of nodes in the conditioning set, the more samples are required to estimate the dependence, $\text{Ind}(A,B|C,D,E)$ more sample than $\text{Ind}(A,B|C,D)$
- For relatively sparse networks, we can d -separate two nodes conditioned on a couple of variables (sample requirements in the low hundreds)

Learning Bayesian Networks: Search-and-Score

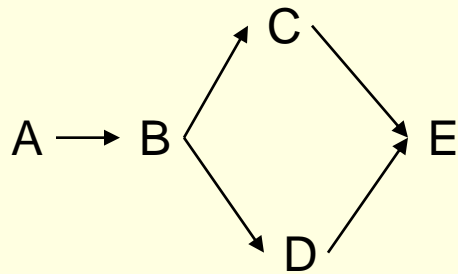
- Score each possible structure
- Bayesian score: $P(\text{Structure} \mid \text{Data})$
- Search in the space of all possible BNs structures to find the one that maximizes score.
- Search space too large. Greedy or local search is typical.
- Greedy search: add, delete, or reverse the edge that increases the score the most.

The PC algorithm (Spirtes, Glymour, Scheines 1993)

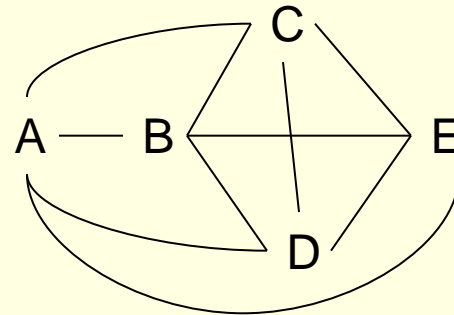
- Phase I: Edge detection
 - Start with a fully connected undirected network
 - For each subset of variables of size $n=0, 1, \dots$
 - For each remaining edge $A - B$
 - If there is a subset S of variables still connected to A or B of size n such $Ind(A; B | S)$, remove edge $A - B$
- Phase II: Edge orientation
 - For every possible V -structure $A - B - C$ with $A - C$ missing
 - If $Dep(A, C | B)$, orient $A \rightarrow B \leftarrow C$
 - While no more orientations possible
 - If $A \rightarrow B - C$ and $A - C$ missing, orient it as $A \rightarrow B \rightarrow C$
 - If there is a path $A \rightarrow \dots \rightarrow B$ orient the edge $A - B$ as $A \rightarrow B$

Trace Example of the PC

True Graph



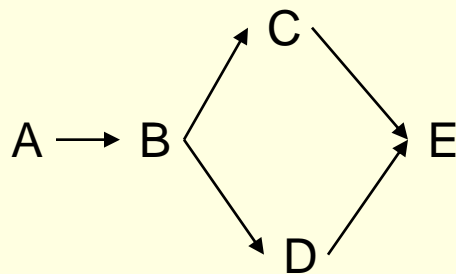
Current candidate graph



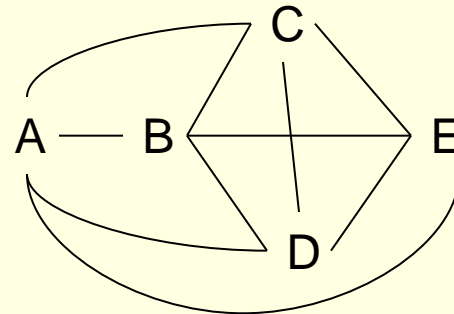
Start with a fully connected undirected network

Trace Example of the PC

True Graph



Current candidate graph



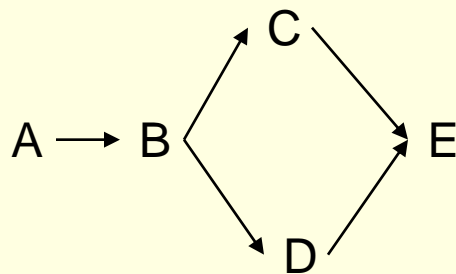
For subsets of size 0

- For each remaining edge $A - B$
 - If there is a subset S of variables still connected to A or B of size n such $Ind(A; B | S)$, remove edge $A - B$

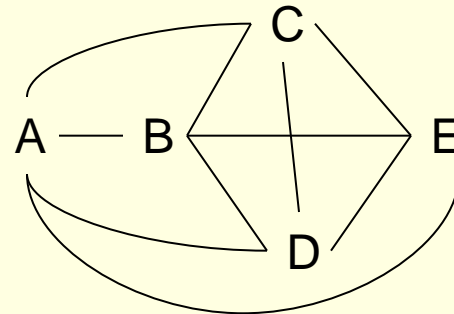
No independencies discovered

Trace Example of the PC

True Graph



Current candidate graph



For subsets of size 1

- For each remaining edge $A - B$
 - If there is a subset S of variables still connected to A or B of size n such $Ind(A; B | S)$, remove edge $A - B$

$Ind(A, C | B)$

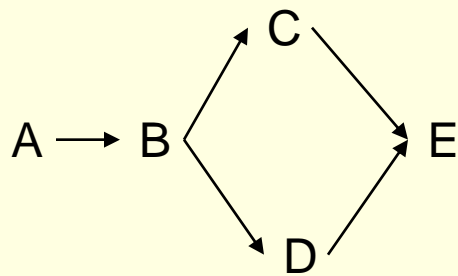
$Ind(A, E | B)$

$Ind(A, D | B)$

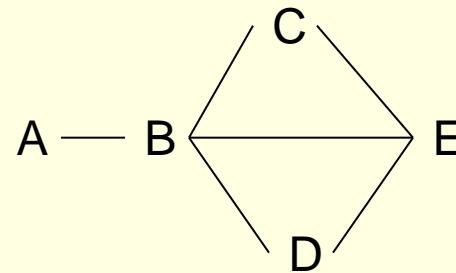
$Ind(C, D | B)$

Trace Example of the PC

True Graph



Current candidate graph



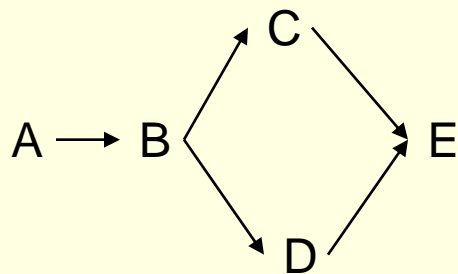
For subsets of size 2

- For each remaining edge $A - B$
 - If there is a subset S of variables still connected to A or B of size n such $Ind(A; B | S)$, remove edge $A - B$

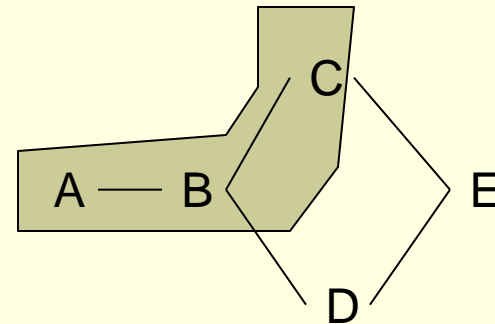
$Ind(B, E | C, D)$

Trace Example of the PC

True Graph



Current candidate graph

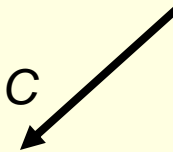


Phase II: Edge orientation

• For every possible V-structure: $A - B - C$ with $A - C$ missing

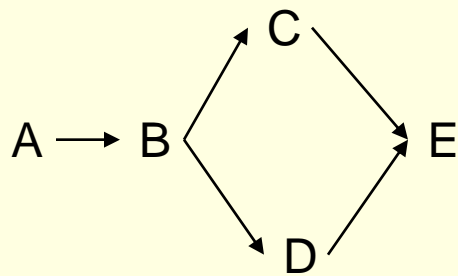
- If $Dep(A, C|B)$, orient $A \rightarrow B \leftarrow C$

Condition does not hold

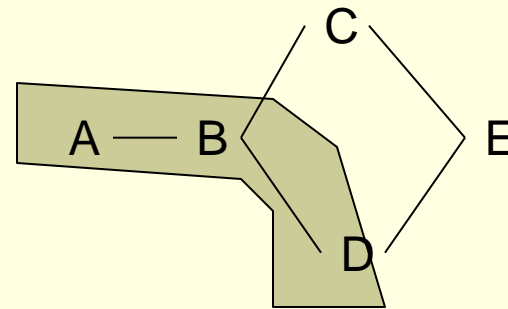


Trace Example of the PC

True Graph



Current candidate graph

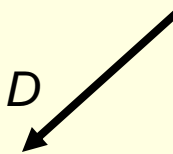


Phase II: Edge orientation

• For every possible V -structure: $A - B - D$ with $A - D$ missing

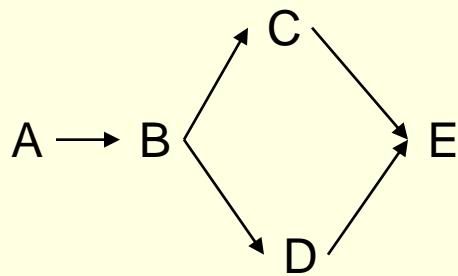
- If $Dep(A, D | B)$, orient $A \rightarrow B \leftarrow D$

Condition does not hold

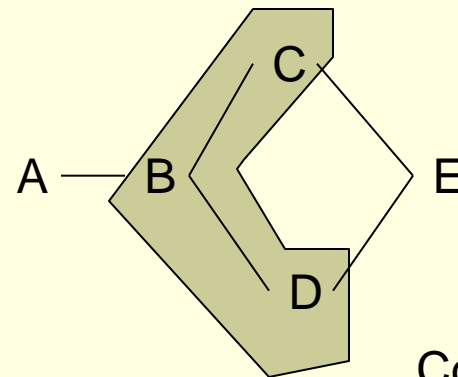


Trace Example of the PC

True Graph



Current candidate graph

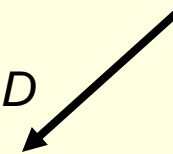


Phase II: Edge orientation

• For every possible V -structure $C - B - D$ with $C - D$ missing

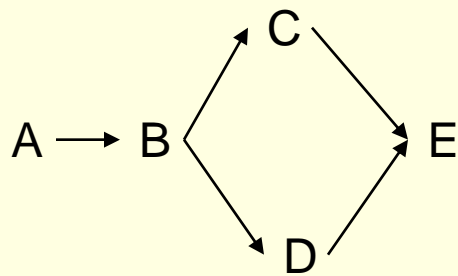
- If $Dep(C, D | B)$, orient $C \rightarrow B \leftarrow D$

Condition does not hold

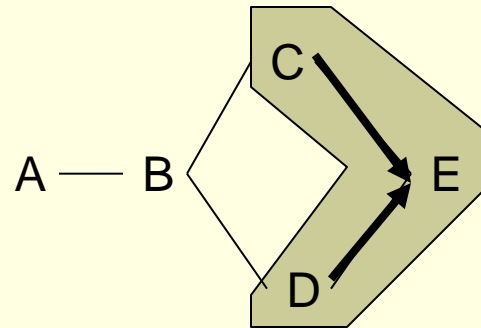


Trace Example of the PC

True Graph



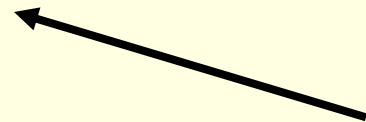
Current candidate graph



Phase II: Edge orientation

- For every possible V -structure $C - E - D$ with $C - D$ missing

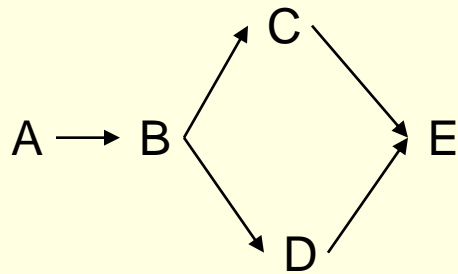
- If $Dep(C, D|E)$, orient $C \rightarrow E \leftarrow D$



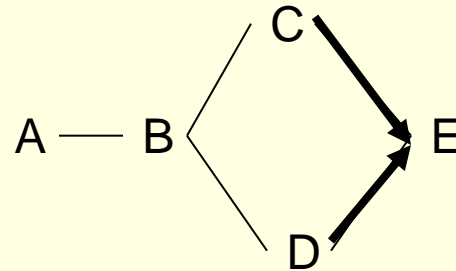
Condition holds

Trace Example of the PC

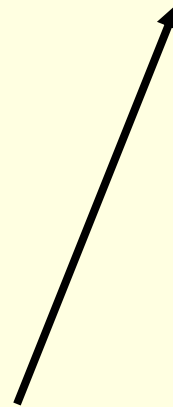
True Graph



Current candidate graph



Final output!

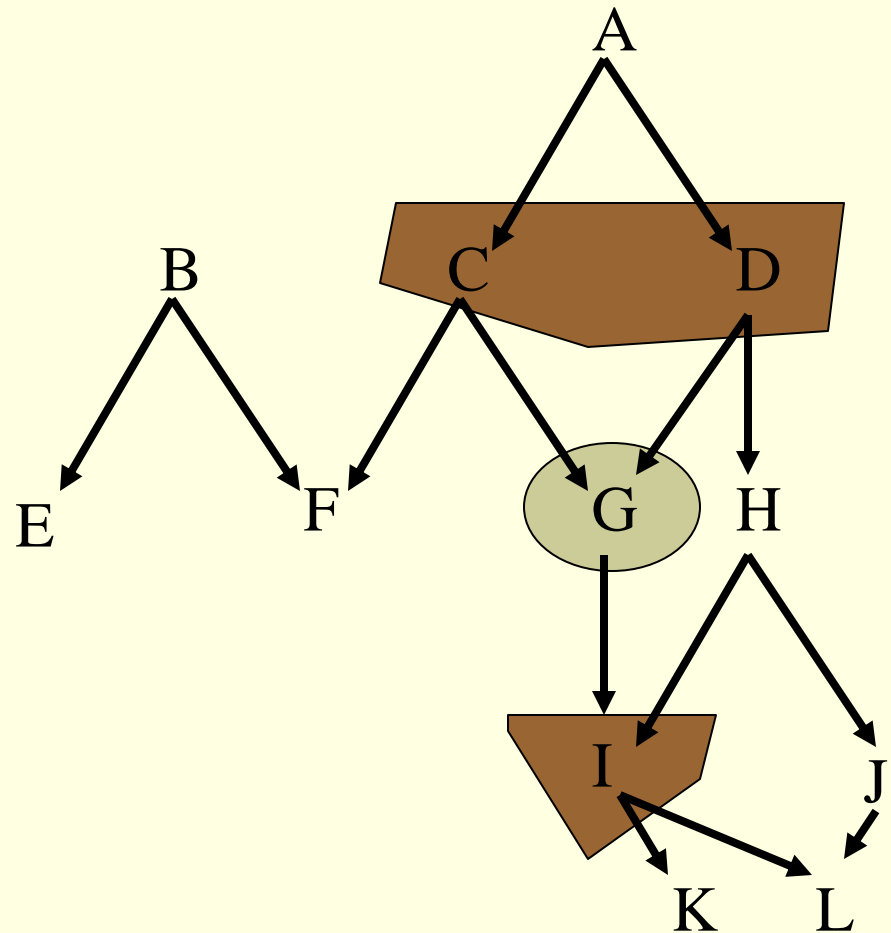


Min-Max Hill Climbing algorithm

- Brown, Tsamardinos, Aliferis, MedInfo 2004
- Based on the same ideas as PC and uses tests of conditional independence
- Uses different search strategy to identify interesting independence relations
- Similar quality results as PC but scales up to tens of thousands of variables (PC can only handle a couple of hundred variables)

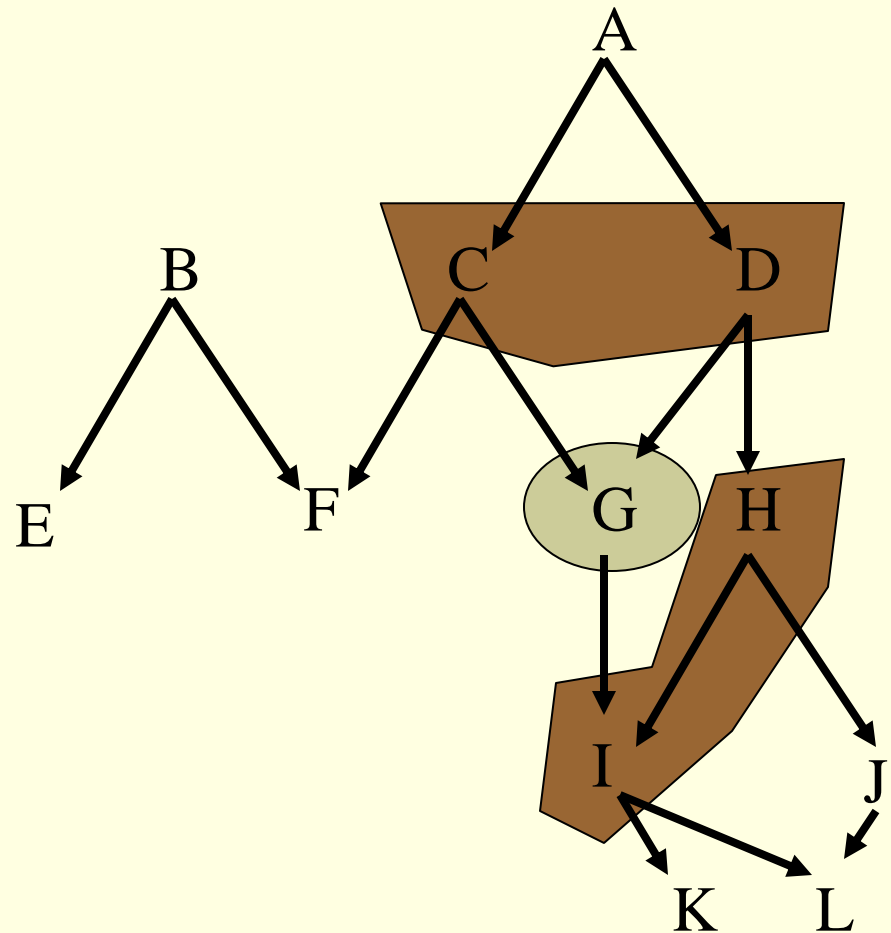
Local Causal Discovery

- Max-Min Parents and Children: returns the parents and children of a target variable
- Tsamardinos, Aliferis, Statnikov KDD 2003
- Scales-up to tens of thousands of variables



Local Causal Discovery

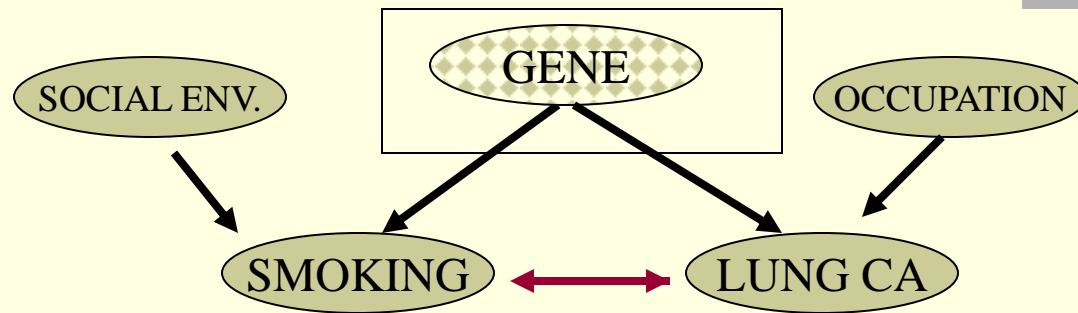
- Max-Min Markov Blanket: returns the parents and children of a target variable
- Scales-up to tens of thousands of variables
- HITON (Aliferis, Tsamardinos, Statnikov AMIA 2003) close variant: different heuristic+wrapping with a classifier to optimize for variable selection tasks



Local Causal Discovery- Different Flavor

- Mani&Cooper 2000, 2001, Silverstein, Brin, Motwani, Ullman
- Rule 1: A, B, C pairwise dependent, $Ind(A, C|B)$, A has no causes within the observed variables (e.g. temperature in a gene expression experiment), then
 - $A \rightarrow \dots \rightarrow B \rightarrow \dots \rightarrow C$
- Rule 2: $Dep(A, B|\emptyset)$, $Dep(A, C|\emptyset)$, $Ind(B, C|\emptyset)$, $Dep(B, C|A)$, then
 - $B \rightarrow \dots \rightarrow A \leftarrow \dots \leftarrow C$
- Discovers a coarser causal model (ancestor relations and indirect causality)

FCI – Causal Discovery with Hidden Confounders



- $\text{Ind}(\text{SE}, \text{LC} | \emptyset)$
- $\text{Dep}(\text{SE}, \text{LC} | \text{SM})$
- $\text{Ind}(\text{SM}, \text{OC} | \emptyset)$
- $\text{Dep}(\text{SM}, \text{OC} | \text{LC})$
- The only consistent model with all tests is one that has a hidden confounder

Other Causal Discovery Algorithms

- Large body of work in Bayesian (or other) search and score methods; still similar set of assumptions (Neapolitan 2003)
- Learning with linear Structural Equation Models in systems in static equilibria (allows feedback loops) (Richardson, Spirtes 1999)
- Learning in the presence of selection bias (Cooper 1995)
- Learning from mixtures of experimental and observational data (Cooper, Yoo, 1999)

Conclusions

- It is possible to perform causal discovery from observational data without Randomized Controlled Trials!
- Heuristic methods are typically used instead of formal causal discovery methods; their properties and their relative efficacy are unknown
- Causal discovery algorithms also make assumptions but have well-characterized properties
- There is a plethora of different algorithms with different properties and assumptions for causal discovery
- There is still plenty of work to be done

Suggested Further Reading

- Neapolitan, Learning Bayesian Networks, Prentice Hall, 2003
- Ed. Glymour, Cooper, Computation, Causation, and Discovery, AAAI Press/MIT Press, 1999
- Ed. Jordan, Learning in Graphical Models, MIT Press, 1998
- Spirtes, Glymour, Scheines, Causation, Prediction, Search, MIT Press 2000