

N-1 Experiments Suffice to Determine the Causal Relations Among N variables

Frederick Eberhardt, Clark
Glymour and Richard Scheines

Carnegie Mellon University



Coauthors



Richard Scheines



Clark Glymour





Causal Discovery

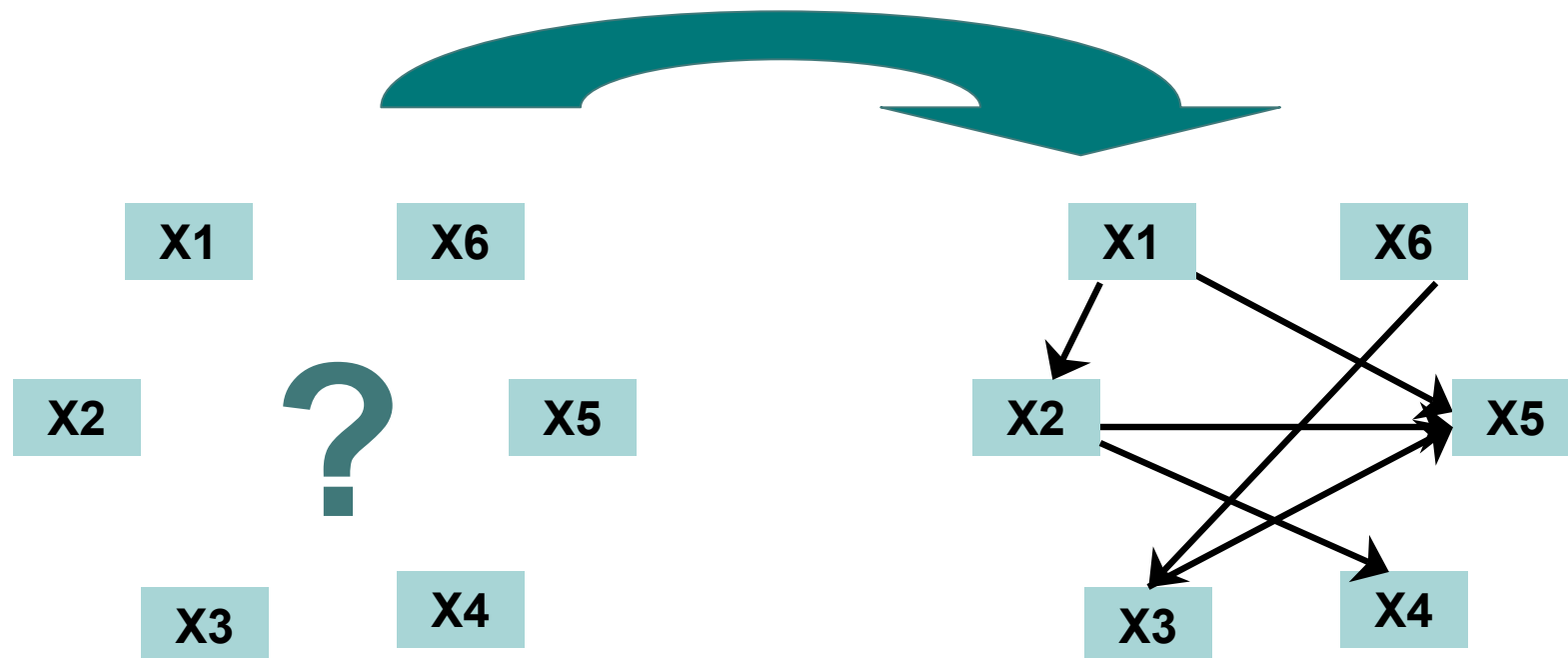
- Problem: Given a set of N variables, how many *experiments* are in the worst case necessary to determine all the causal relations?
- For any pair of variables X and Y in a set of variables S , does X directly cause Y ?



Causal Bayes Nets

- Causal Graphs $G = (V, E)$
- Requirements:
 -  Causal Markov & Faithfulness
 -  Ideal Interventions
 - Determine the distribution of their targets
 - Interventions make variables independent of their causes (breaks incoming edges)

The Problem in terms of Causal Bayes Nets



How many *experiments* are in the worst case necessary to discover the causal graph?



Algorithms using passive observational Data

- PC-Algorithm (SGS, 2000)

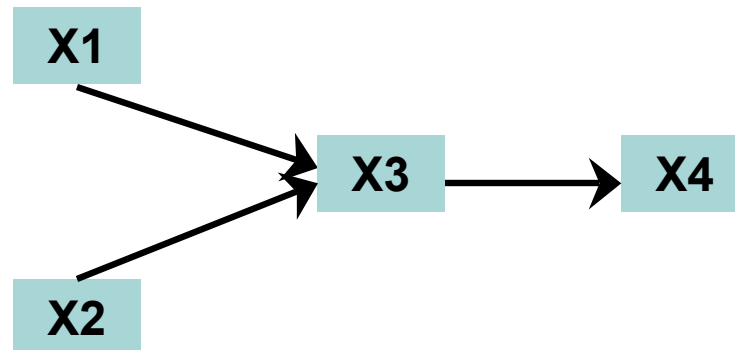
Input: (conditional) independence relations from the data

Output: Markov Equivalence class of graphs satisfying these independence relations

(adjacencies and unshielded colliders)

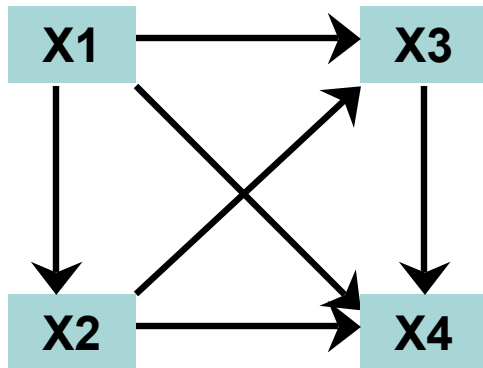
● ● ● | PC-Algorithm

- Qualitative approach to causal discovery
- Consistent search procedure
- Extended to include latent variables (FCI-Algorithm)

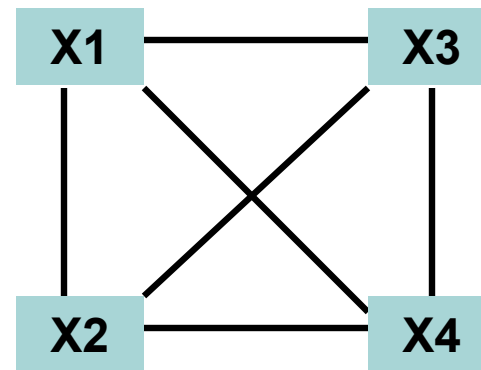


- ● ● | Limitations of Algorithms based on Passive Observation

Truth



PC-Algorithm Output



Problem: Even with perfect data it is impossible to do any better than the Markov Equivalence class, hence in dense graphs we can say little about the direction of the causal relation.




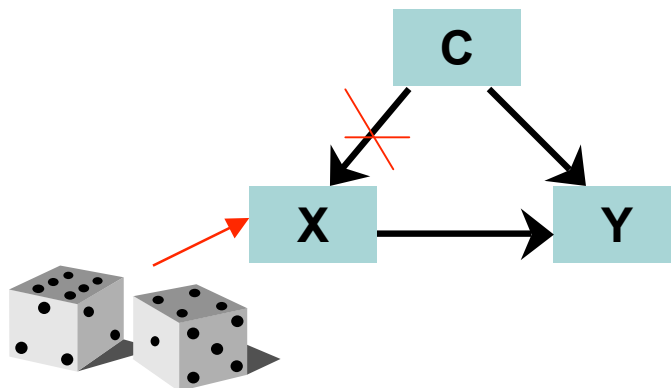
Intervention

- Randomization

- Advantages

-  Removes all confounding by breaking all the incoming edges into the intervened variable

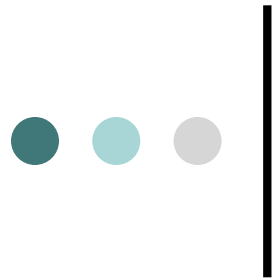
-  Provides reference distribution for further statistical analysis





Definition of Experiment (1)

- An ***experiment*** is either
 - ☞ a passive observation, or
 - ☞ randomized assignment of *one* variable
- **Randomization:** The distribution of the intervened variable is determined by a known probability distribution that is *independent* of any other variable
- Returns the independence facts true of the resulting manipulated population



Differences to Literature on Experimental Design

○ Experiments here

- Determine variable for intervention (if any)
- Optimization involves the choice of intervention set

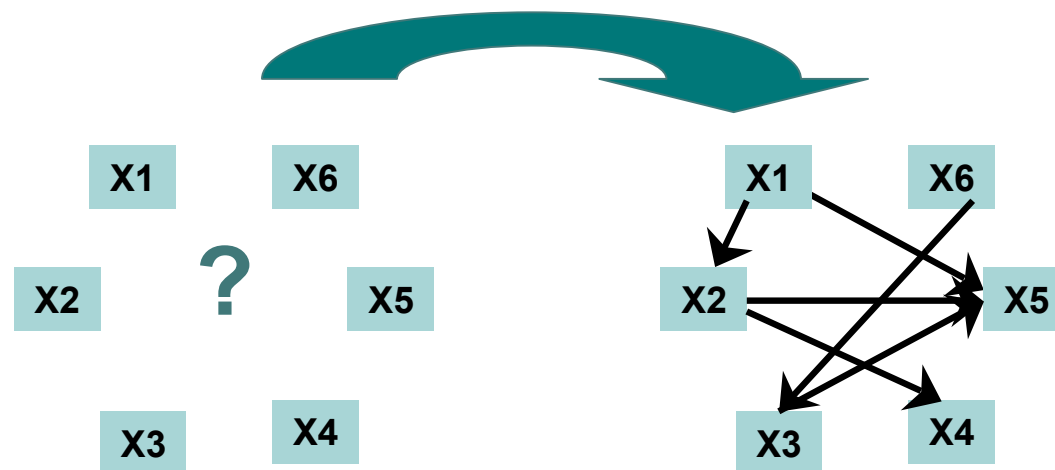
○ Experimental Design

- Treatment and effect variables are predetermined
- Optimization is on the most efficient value assignment to treatment variables



Main Problem

- How many experiments are in the worst case necessary to determine the causal structure among the variables?
- Each experiment is chosen given the knowledge gained from the previous experiments.

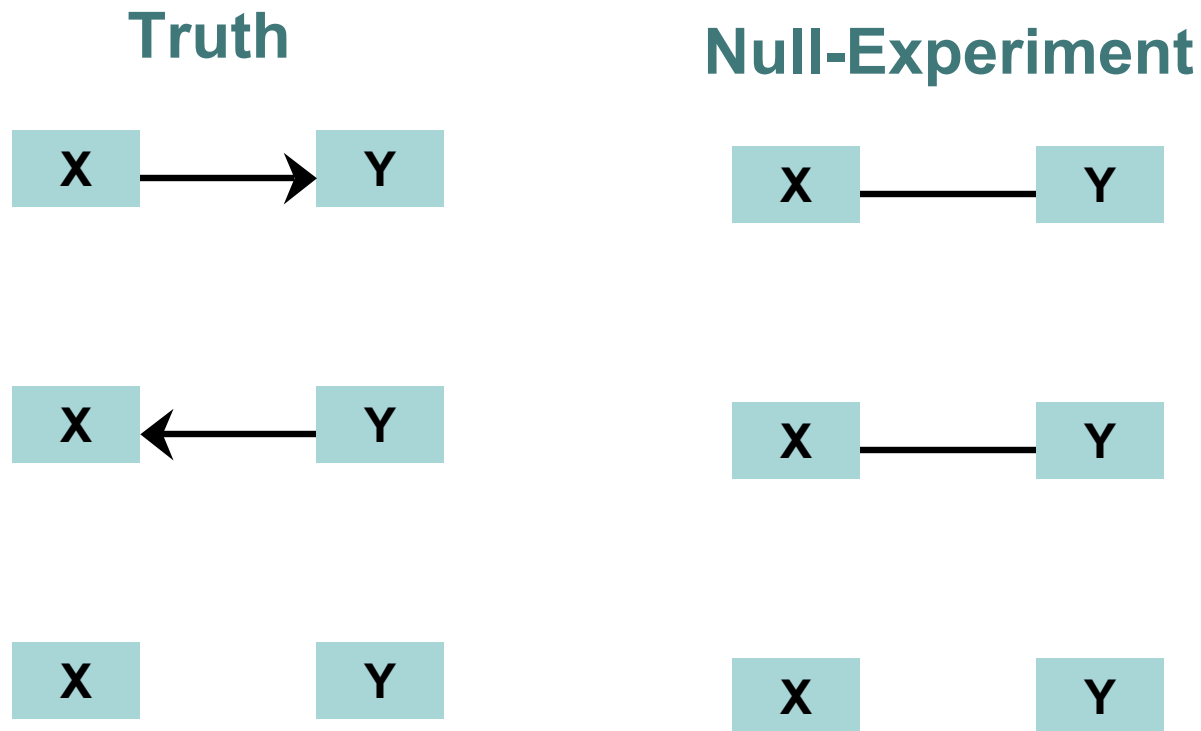




Assumptions

- No cycles \rightarrow DAGs
- No latent variables *
- Population independencies and conditional independencies
- No cost function for interventions

Passive Observation – null Experiment



- With respect to a pair of variables, the null experiments only tells us about **adjacencies**, but not about directions.

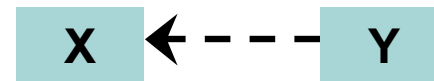


Single Intervention Experiment

Truth



Intervention on X



- With respect to a pair of variables, a single intervention tells us whether the intervened variable is the cause, but cannot distinguish between incoming edges and independency: **directional**.



Compare Experiments

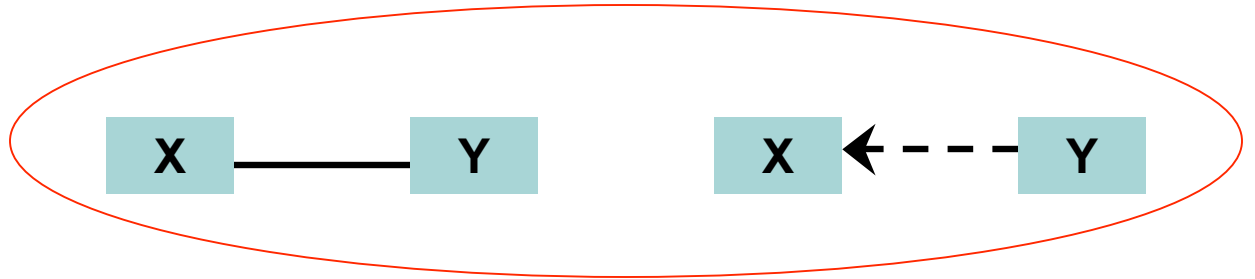
Truth



Null-Experiment



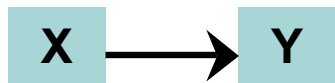
Intervention on X



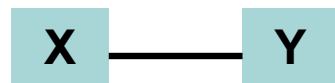


Compare Experiments

Truth

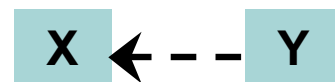
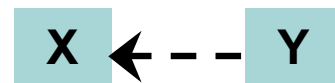


Null-Experiment



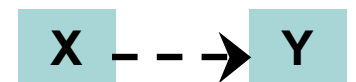
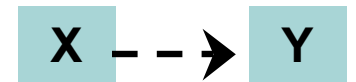
Adjacency test

Intervention on X



Directional test



Intervention on Y



Directional test



Adjacency and Directional Tests

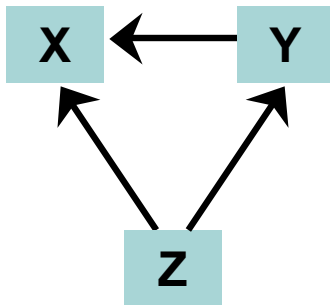
- The causal graph can be determined if each pair of variables in the graph is subject to one of the following:
 -  An adjacency test and a directional test
 -  Two opposing directional experiments

Aim: Allocate for each pair of variables either 1) or 2) within the minimum number of experiments

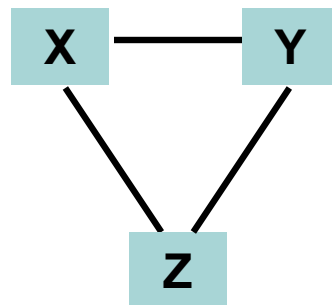


Trade-off Example

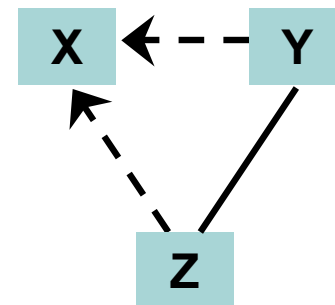
Truth



Null-Experiment



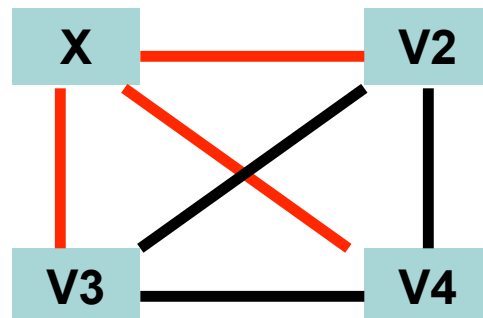
Intervention on X





Single Intervention on X

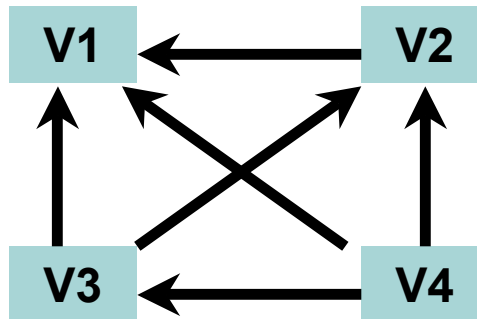
- Directional test for all pairs of variables (X , \underline{V})
- Adjacency test for all pairs of variables ($V1$, $V2$), where $V1, V2 \neq X$





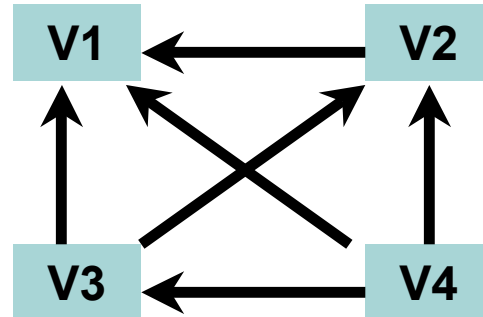
The Worst Case

- Complete graph
- Every intervention happens to be on the sink (common effect of other variables)



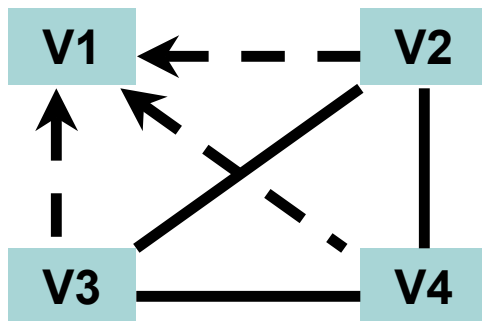


Nature hides its secrets

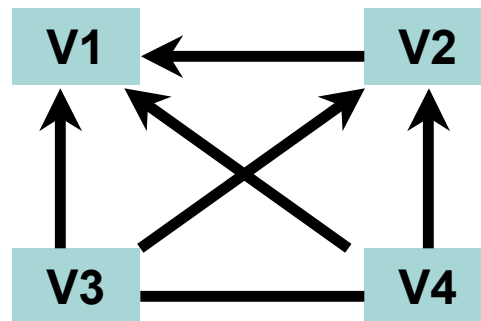


Truth
(unknown)

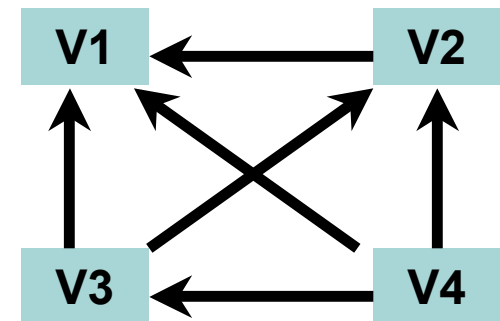
Intervention on V1



Intervention on V2



Intervention on V3



3 Experiments!!

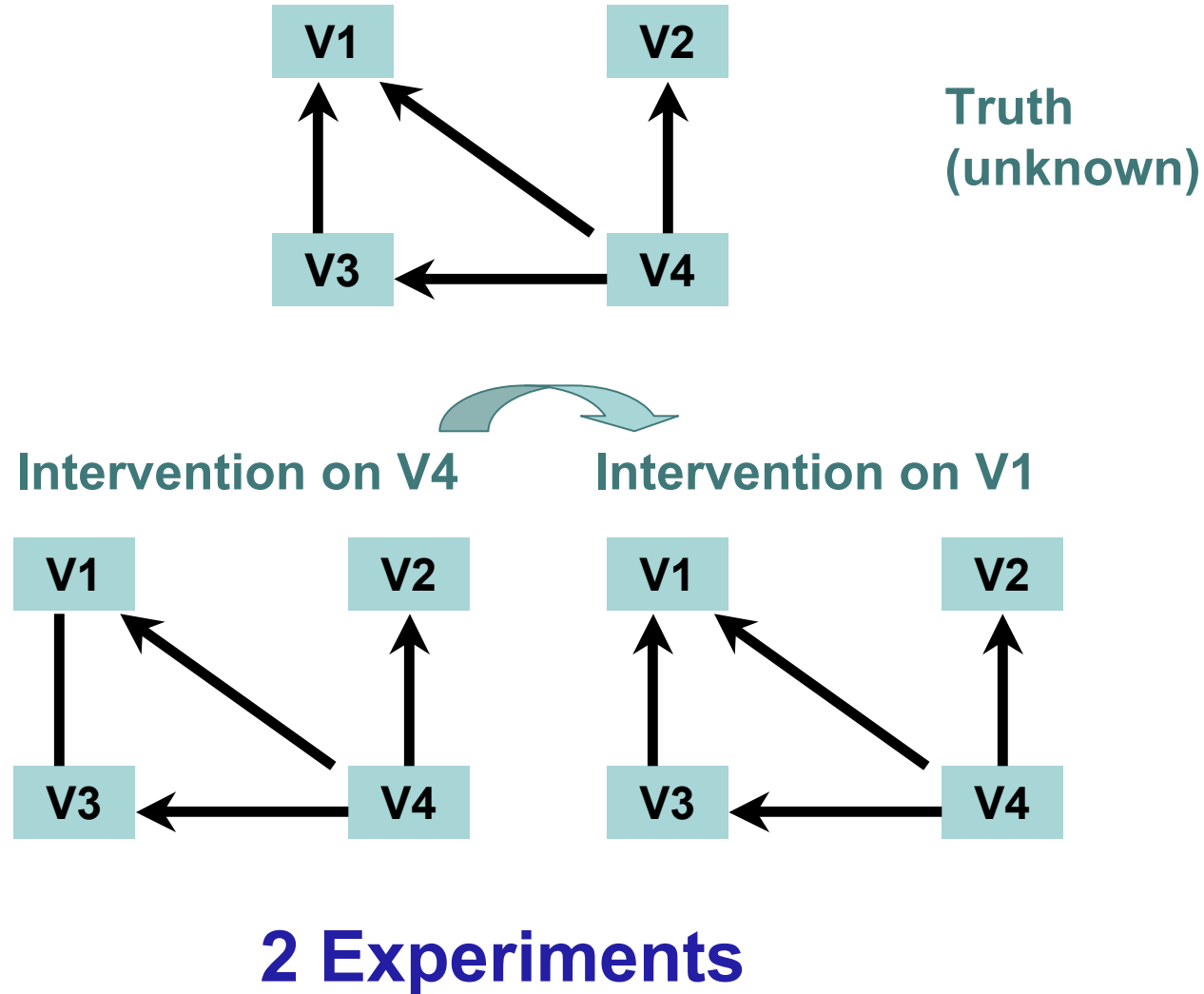


Result

- $N-1$ experiments are in the worst case necessary to determine the causal relations among $N > 2$ variables
- Corollary: $N-1$ experiments are sufficient to determine the causal structure among $N > 2$ variables



Sometimes we can do better



● ● ● | N=2 is a problem



- There is no single experiment that determines the causal graph uniquely in every case.



Can we do better?

- Can the worst case bound for discovering the causal structure be improved beyond $N-1$?

- YES!



Multiple Simultaneous Interventions

- Randomize more than one variable in each experiment
- Randomizing distribution has to make the set of intervened upon variables independent

➡ change definition of “*experiment*”



Definition of Experiment (2)

- An ***experiment*** is either
 - a passive observation, or
 - an intervention on *a set of variables*
- **Randomization:** The value of any variable in the intervention set is determined by a known probability distribution over its values *independent* of any other variable



Result for Multiple Interventions

- Theorem: $\log_2(N)+1$ experiments are sufficient and in the worst case necessary to determine the causal structure among $N>1$ variables.

Surprising, since all information about what is going on between the variables in the intervention set is lost.

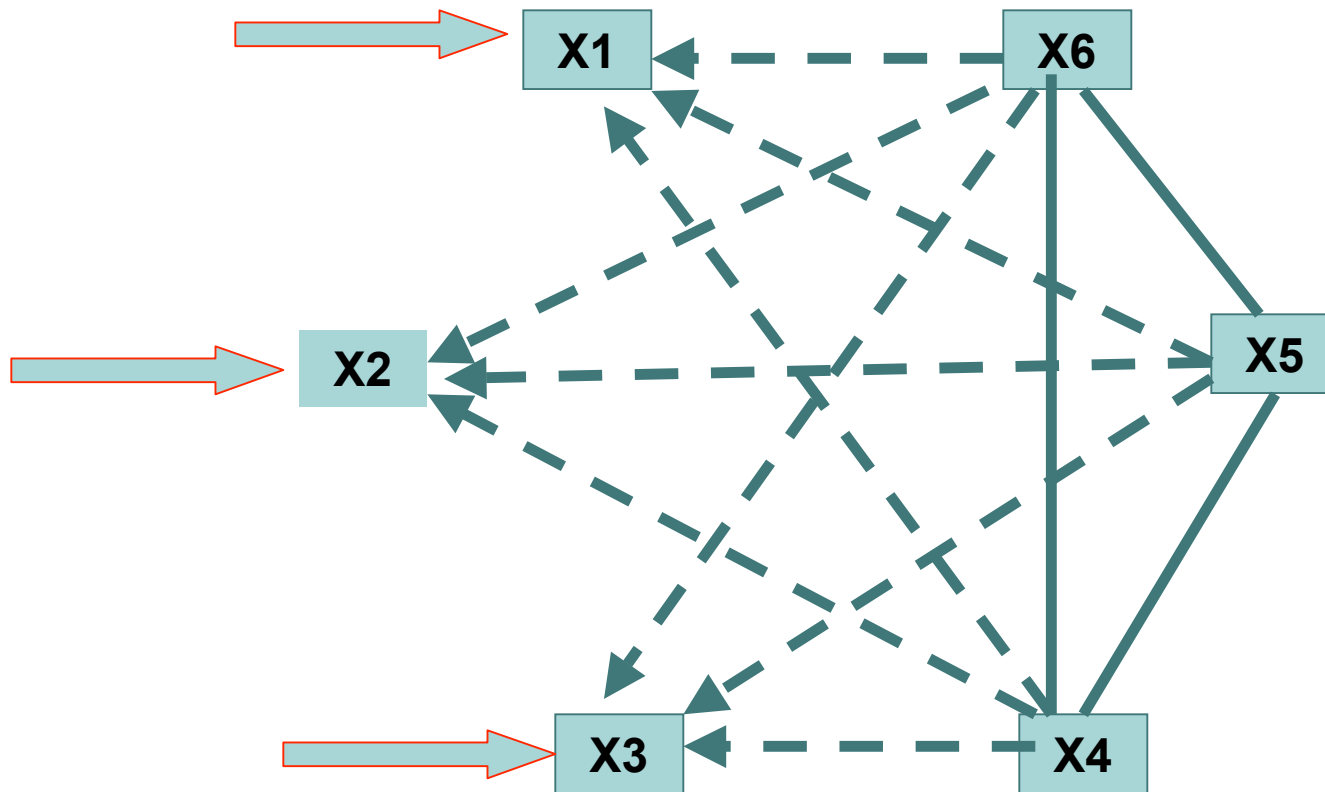


Idea

- Given an experiment with an intervention set of size k , then the number of pairs of variables subject to a directional test is $(N-k)k$, which is maximized at $N/2$.
- Hence, intervene on $N/2$ variables each time.

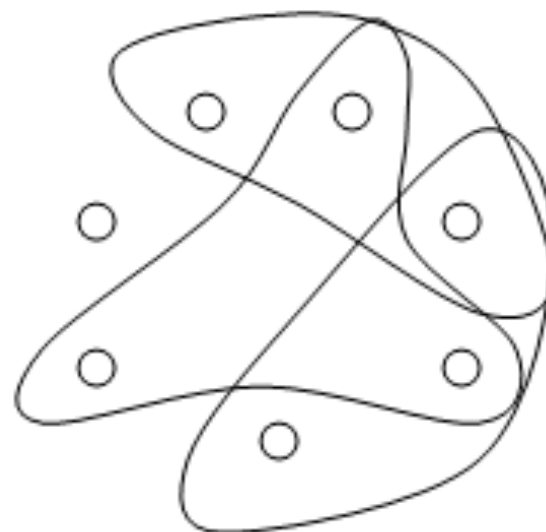
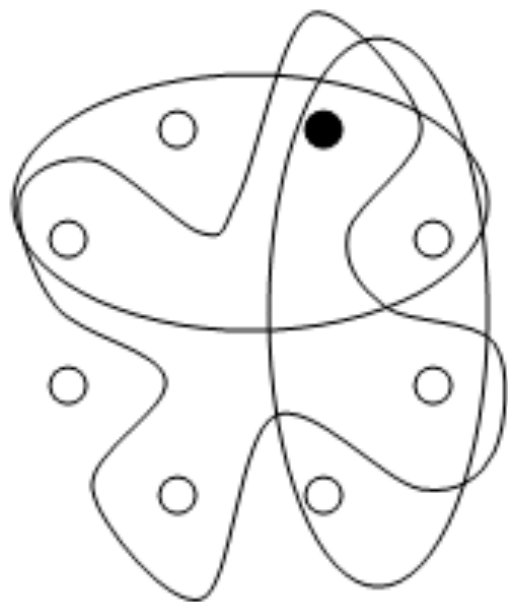


Intervention on $N/2$ variables





Intervention Sets





Results

- Worst case bound for the number of experiments required to determine the causal structure under the given assumptions
- Specify a sequence of experiments that finds the causal structure within the $N-1$ or $\log(N)$ experiments



What the results are NOT

- They do not specify the *optimal* strategy to find the causal structure in every case.
- They do not say anything about the expected number of experiments required, although...



Relaxing the Assumptions

- Latent Variables
- Imperfect Data: adding the statistical problems to the graph theoretic problem
- Specifying the optimal strategy for choosing the next experiment
- Cost functions on interventions and limitations of conditional independence



The Problems

- Counting Problems
- Computational Problems
- Statistical Problems



Frederick Eberhardt, CMU
fde@cmu.edu
www.andrew.cmu.edu/~fde