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

Frederick Eberhardt
Clark Glymour[1]
Richard Scheines

Carnegie Mellon University

Abstract

By combining experimental interventions with search procedures for graphical causal models we show that under familiar assumptions, with perfect data, N - 1 experiments suffice to determine the causal relations among N>2 variables when each experiment randomizes at most one variable. We show the same bound holds for adaptive learners, but does not hold for N > 4 when each experiment can simultaneously randomize more than one variable. This bound provides a type of ideal for the measure of success of heuristic approaches in active learning methods of causal discovery, which currently use less informative measures.

Three Methods and Their Limitations

Consider situations in which the aim of inquiry is to determine the causal structure of a kind of system with many variables, for example the gene regulation network of a species in a particular environment. The aim in other words is to determine for each pair X, Y of variables in a set of variables, S, whether X *directly* causes Y (or vice-versa), with respect to the remaining variables in S, i.e., for some assignment of values V to all the remaining variables in S, if we were to intervene to hold those variables fixed at values V while randomizing X, Y would covary with X, or vice versa. Such a system of causal relations can be represented by a directed graph, in which the variables are nodes or vertices of the graph, and X → Y indicates that X is a direct cause of Y. If there are no feedback relations among the variables, the graph is acyclic. We are concerned with the most efficient way to determine the complete structure of such a directed acyclic graph, under some simplifying assumptions.

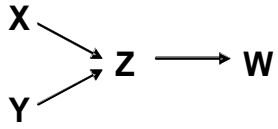
Suppose that, before collecting data, nothing is known that will provide positive or negative evidence about the influence of any of the variables on any of the others. There are several ways to obtain data and to make inferences:

[1] Second affiliation: Florida Institute for Human and Machine Cognition

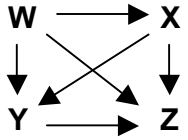
1. Conduct a study in which all variables are passively observed, and use the inferred associations or correlations among the variables to learn as much as possible about the causal relations among the variables.
2. Conduct an experiment in which one variable is assigned values randomly (randomized) and use the inferred associations or correlations among the variables to learn as much as possible about the causal relations.
3. Do (2) while intervening to hold some other variable or variables constant.

Procedure 1. is characteristic of non-experimental social science, and it has also been proposed and pursued for discovering the structure of gene regulation networks (Spirtes, et. al, 2001). Consistent algorithms for causal inferences from such data have been developed in computer science over the last 15 years Under weak assumptions about the data generating process, specifically the *Causal Markov Assumption*, which says that the direct causes of a variable screen it off from variables that are not its effects, and the *Faithfulness Assumption*, which says that all of the conditional independence relations are consequences of the *Causal Markov Assumption* applied to the directed graph representing the causal relations. Consistent search algorithms are available based on conditional independence facts - the PC-Algorithm, for example (Spirtes, et al., 2000) - and other consistent procedures are available based on assignments of prior probabilities and computation of posterior probabilities from the data (Meek, 1996; Chickering, 2002). We will appeal to facts about such procedures in what follows, but the details of the algorithms need not concern us.

There are, however, strong limitations on what can be learned from data that satisfy these assumptions, even supplemented with other, ideal simplifications. Thus suppose we have available the true joint probability distribution on the variables, and there are no unrecorded common causes of the variables (we say the variable set is *causally sufficient*), and there are no feedback relations among the variables. Under these assumptions, the algorithms can determine from the observed associations whether it is true that X and Y are adjacent, i.e., *whether* X *directly causes* Y *or* Y *directly causes* X, for all variables X, Y, but only in certain cases can the direction of causation be determined. For example, if the true structure is

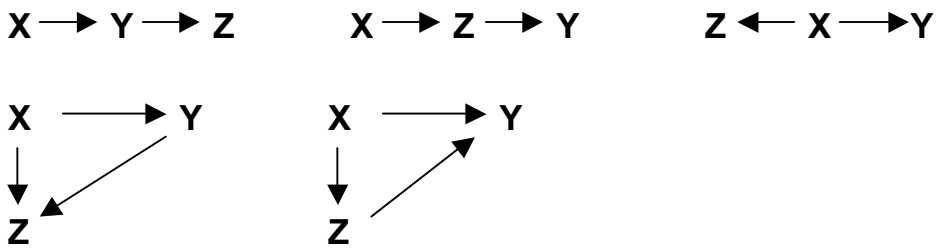


various search algorithms will find this structure uniquely. But if the true structure is



then no consistent[2] search algorithm can determine more than that all pairs of variables are adjacent in the true graph. That is, nothing about the direction of causation can be determined from a joint distribution generated by this structure.

For several reasons, experimental intervention, in particular on a potential cause is a preferred method of estimating causal relations, and it is the standard method described in many methodology texts. Randomization guards against the possibility that there is an unrecorded common cause of the manipulated variable and other variables. But even when we assume there are no such unrecorded confounding variables, randomization is informative about some of the directions of causal relations. By randomizing X and observing which other variables in a set covary with X, we can determine which variables are influenced by X, but the associations of other variables with X do not themselves determine which of those variables are influenced directly and which indirectly. So randomizing X does not, for example, distinguish among the following structures:



That is why it is sometimes recommended that we manipulate the system to keep some variables constant while we randomize others, as in method 3. If S = {X, Y, Z} we might randomize X while intervening to hold Z constant and see if Y covaries with X. If so, X is a direct cause of Y with respect to Z. If not, we may have to intervene to hold Z constant at other values and see if X and Y covary. If X and Y never covary for any fixed value of Z, then X is not a direct cause of Y. This procedure therefore has two difficulties that render it infeasible for large sets of variables: for each pair, X, Y, we must experimentally manipulate all remaining variables in S to hold them constant while varying X or varying Y (or both) and in the worst case we must do such a distinct experiment for every consistent assignment of values to the remaining variables. For continuous variables this is of course impossible, but even if each variable has only M values, in the worst case[3] we require at least

(N choose 2) $M^{(N-2)}$

different experiments to determine the entire structure. Suppose we have measured the messenger RNA (mRNA) expression levels of 10 genes and divide the expression levels into high, medium and low values. We would require in the worst case at least 295,245 experiments.

---

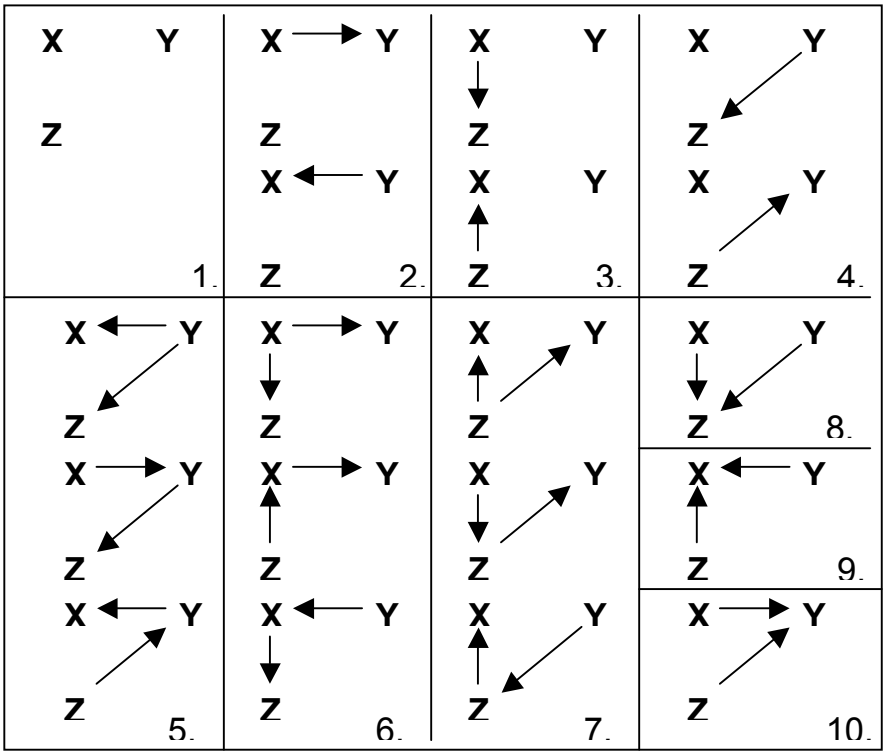
[2] By a consistent search algorithm we mean an algorithm that provably recovers as much about the graphical structure as is determined by the joint probability distribution, for all graphs and probability distributions satisfying the assumptions specified.

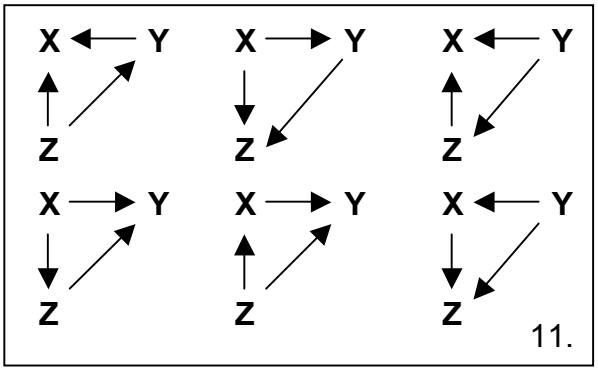
[3] The worst case occurs with a complete graph and our randomizations are on effects before causes.

Various modifications of the control procedure might improve these worst case results, and for many probability distributions over the possible causal structures the expected case number of experiments would presumably be much better. But we propose a principled result: By combining procedure 1 with procedure 2, under the assumptions so far listed, for N > 2, in the worst case, the complete causal structure on N variables can be determined with N - 1 experiments, counting the null experiment of passive observation (procedure 1) as one experiment, if conducted. Further, this is the best possible result when at most one variable is randomized in each experiment.

The Idea

Consider the case of N = 3 variables. There are 25 directed acyclic graphs on 3 vertices. In figure X we show the graphs sorted into sub-classes that are indistinguishable without experimental intervention.

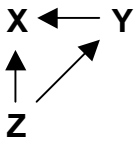




Given the joint distribution prior to an intervention, a consistent search algorithm will return the equivalence class of the graph - that is, the disjunction of all graphs in the box to which the true graph belongs. An experimental intervention, say one that randomizes X, provides extra information: the fact that the distribution of X has been randomly assigned by an external intervention tells us that in the resulting experimental data none of the remaining variables influence X. We can use that as prior information in applying a consistent search algorithm to the experimental data. If X is manipulated, the resulting joint distribution on X,Y and Z will give us, through such search procedures, information about whether X is a direct cause of Y or Z, or an indirect cause of one or the other. Thus suppose we randomize X and we find the following in the experimental distribution: Y and Z covary with X, and Y and Z are independent conditional on X. Then we know that X causes Y and Z, which tells us that the true graph is in box 6 or in box 11, and further, we know Y does not cause Z directly, and Z does not cause Y directly, because they are independent conditional on all values of X. (The Markov and Faithfulness assumptions imply that when Y and Z are independent conditional on X, there is no direct causal relation between Y and Z.) The top graph in box 6 must therefore be the true graph. By combining search procedures (in this case used informally) with experimentation, we have determined the truth with a single experiment. (We were lucky: if we had begun by randomizing Y or Z, two experiments would have been required.) When we randomize X and follow up with a consistent search procedure, which requires no additional experimentation, all of the direct connections between the remaining variables can be estimated. Only the directions of some of the edges remain unknown. Those directions can clearly be determined by randomizing each of the remaining variables.

In some cases, we lose something when we experiment. If when X is randomized, X and Y do not covary, we know that X does not cause Y, but we do not know whether Y causes X or neither causes the other, because our manipulation of X has destroyed any possible influence of Y on X. Thus in the single structure in box 9, if we randomize X, and Y and Z do not covary with X, every structure in which X is not a direct or indirect cause of Y or Z, and Y is not a cause of Z and Z is not a cause of Y, is consistent with our data. There are four such graphs. Subsequent experimental manipulation of Y will tell us the answer, of course.

So, with N variables, N experiments clearly suffice to determine the structure uniquely. In fact, because of the assumption of acyclicity and because the associations among the variables not randomized in an experiment are still informative about the adjacencies among these variables, N - 1 experiments always suffice when $N > 2$. To illustrate, consider the first graph in box 11 and suppose we make the worst choices for the sequence of variables to randomize: first X, then Y, then Z.



Randomizing X we find only that X does not cause Y or Z, and that Y and Z are adjacent. Randomizing Y, we find that Y does cause X but does not cause Z, and one X and Z are adjacent. We reason as follows: Z must be a direct cause of X, because we know they are adjacent but we now know X is not a cause of Z. Similarly, Z must be a direct cause of Y because they are adjacent and Y does not cause Z. Y must be a direct cause of X, because Y is a cause of X and Y is not a cause of Z (so there cannot be a pathway $Y \rightarrow Z \rightarrow X$). We have found the true graph, and only 2 experiments were required. We show in the appendix that the same result, that at most N-1 experiments are required, is true for all $N > 2$.

The result does not hold for $N = 2$, where there are only 3 possible structures: no edges, $X \rightarrow Y$, and $X \leftarrow Y$. Suppose $X \rightarrow Y$. Suppose we randomize nothing, merely observe non-experimental values. If we find X, Y are associated, then a second experiment is required to determine the direction of the effect. Suppose instead, we begin by randomizing X. If we find X, Y are not associated, a second experiment is required to determine whether Y causes X.

The proof of the bound has three perhaps surprising corollaries. (1) Any procedure that includes passive observation in which no variables are randomized exceeds the lower bound for some cases, when the passive observation is counted as an experiment. (2) Controlling for variables by experimentally fixing their values is never an advantage. (3) "Adaptive" search procedures (Murphy, 1998; Tong and Koller, 2001) choose the most "informative" next experiment given the results of previous experiments. That is, they choose the next experiment that maximizes the expected information to be obtained. We also show that no adaptive procedure can do better than the N-1 lower bound on the number of experiments required to identify the structure in the worst case. Implementations of adaptive search procedures generally have to make simplifying assumptions in order to make updating of the distribution over all possible graphs computationally tractable or must use greedy heuristics to select the next intervention that is deemed most informative with respect to the underlying causal structure given the evidence from the previous experiments. The success of such a heuristic is commonly measured by comparing it to a strategy that randomly chooses the next experiment no matter what the evidence is so far. Other comparisons are to uniform sampling or to passive observation - the latter obviously only provides limited directional information. These comparisons indicate whether the heuristic is achieving anything at all but give little insight into how well the strategy compares with an ideal procedure. The bound we provide provides such an ideal, at least in the case when only passive observational and single intervention experiments are considered.

Discussion

A variety of theoretical issues remain. Expected complexities can differ considerably from worst case complexities, and we have not investigated the expected number of experiments required for various probability distributions on graphs. When the variable

set is not known to be causally sufficient, which is the typical scientific situation, there is a consistent search procedure, the FCI Algorithm (Spirtes et al., 1993; 2000), which unsurprisingly returns more limited information about the structure. When there are unrecorded common causes, some structures cannot be distinguished by independence and conditional independence relations alone, but can be distinguished by attention to changes in the covariation of variables in different experiments. In general, the number of experiments required is larger than N-1 but we have no bound to report. Further, we do not know by how much the N - 1 bound can be improved by experiments in which two or more variables are randomized simultaneously. However, for N>4, multiple intervention experiments do reduce the total number of experiments required to identify the causal structure even in the worst case, although it may initially seem that information on the potential edges between intervened upon vertices is lost. For example, in the case of five vertices, three such multiple simultaneous randomization experiments suffice even in the worst case; in the case of six vertices, four experiments will do.

The N-1 bound can be considered proportional to a minimum cost of inquiry when all experiments, including passive observation, have the same cost. Costs may differ when one experiment simultaneously randomizes several variables. In practice there is a cost to sample size as well, which can result in complicated trade-offs between cost and the confidence one has in the results.

In practice, with real data, search procedures tend to be unreliable for dense graphs. The reasons differ for different algorithms, but basically reflect the fact that in such graphs conditional probabilities must be assessed based on many conditioning variables. Each conditioning set of values corresponds to a subsample of the data, and the more variables conditioned on, the smaller the sample, and the less reliable the estimate of the conditional joint probability of two variables. Some search algorithms, such as PC and FCI, test for conditional independence relations and use the results as an oracle for graphical specification. So it would be of interest to know the effects on the worst case number of experiments required when a bound is placed on the number of variables conditioned on in the search procedure.

Acknowledgements

We are grateful to Teddy Seidenfeld, who contributed to early discussion on the work in this paper and asked several key questions that helped to focus our efforts.

The second author is supported by ONR grant N00014-04-1-0384 and NASA grants NCC2-1399 and NCC 2-1377. The third author is supported by a grant from the James S. McDonnell Foundation.

Appendix: Proofs

We assume the reader's familiarity with some fundamental ideas from directed graphical models, or Bayes nets, including the property of d-separation (Pearl, 1988) and search algorithms that exploit that property (Spirtes, et. al, 2000).

Assumptions:

We make the following assumptions in our proof of the worst case bound on the number of experiments required to identify the causal graph underlying N variables.

Faithfulness: The distribution over the variables is faithful to a directed acyclic graph on the variables in the data.
Full scale D-Separation: It is possible to condition on any subset of variables to determine d-separation relations.
Perfect Data: The data is not supposed to be of any concern. In particular, we are not concerned with weak causal links, insufficient or missing data. The data is such that we can identify the conditional independencies if there are any.
Interventions: Interventions are possible on every variable.

Definitions:

An **experiment** randomizes at most one variable and returns the joint distribution of all variables.
A **procedure** is a sequence of experiments and a structure learning algorithm applied to the results of these experiments.
A procedure is **reliable** for an N vertex problem iff for all DAGs on N vertices the procedure determines the correct graph uniquely.
A procedure is **order reliable** for an N vertex problem iff it is reliable for all non-redundant orderings of experiments.
A procedure is **adaptive** iff it chooses at each step one from among the possible subsequent experiments as a non-trivial function of the results of the previous experiments.

Claims

Proposition 1: *For N > 2, there is an order reliable procedure that in the worst case requires no more than N - 1 experiments, allowing only single interventions.*

Proof: Consider a graph with N vertices where N > 2 and let $X_1,\ldots, X_N$ specify an arbitrary ordering of these vertices. Let each experiment consist of an intervention on one variable. Perform N - 1 experiments, one intervention on each X_i where $1 \leq I \leq N\text{-}1$. By Lemma 1 below, applying the PC algorithm to the first experiment determines the

adjacencies among at least $X_2,\ldots, X_N$. The k-th experiment determines the directions of all edges adjacent to X_k : iff X_j is adjacent to X_k, then X_k is a direct cause of X_j if and only if X_j covaries with X_k when X_k is randomized (since if X_k were only an indirect cause of X_j, and since X_j and X_k are adjacent, X_j would have to be a direct cause of X_k, and there would be a cycle); otherwise, X_j is a direct cause of X_k. X_N has not been randomized, but its adjacencies with every other variable have been determined by the N-1 experiments. Suppose X_N and X_k are adjacent. Since X_k has been randomized, X_k is a cause of X_N if and only if X_N covaries with X_k when X_k is randomized. In that case, if X_k were an indirect but not a direct cause of X_N, then X_N would be a direct cause of X_k, because X_N and X_k are adjacent, and hence there would be a cycle. If X_N and X_k do not covary when X_k is randomized, then, since they are adjacent, X_N is a direct cause of X_k. If X_k and X_N are not adjacent, then this missing edge would have been identified in one of the interventions on X_j, where $j \neq k$. These are all of the cases. Q.E.D.

Lemma 1: *If G is a causal graph over a set of variables V, and G' the manipulated graph resulting from an ideal intervention on variable X in G, then for all pairs of variables Z,Y distinct from X, Z and Y are d-separated by some $S \subseteq V$ in G if and only Z and Y are d-separated by some $S' \subseteq V$ in G'.*

Proof: G' is identical to G except that all edges into X in G do not occur in G'.

L-to-R: First assume Z and Y are d-separated by some $S \subseteq V$ in G. Then no undirected path between Z and Y in G d-connects those variables relative to S. Suppose for reductio that Z and Y are not d-separated by S in G'. Then some path between Z and Y in G' must now be active, i.e., constitutes a d-connection. The paths between Z and Y in G' are a subset of those in G. Thus some path between Z and Y that was inactive in G must now be active in G'. Thus all nodes on such a path that were inactive in G must now be active in G'. But if X was inactive on a path in G relative to S, it will still be inactive in G' relative to S. For it to be otherwise, X would either have to switch from a non-collider to a collider, which cannot happen by removing edges into X, or for X to be a collider in G with no ancestor in S but to be a collider in G' with an ancestor, which also cannot happen by removing edges into X. A similar argument applies equally to non-X nodes, so Z and Y are d-separated by S in G'.

R-to-L: Next assume that Z and Y are not d-separated by some $S \subseteq V$ in G, that is, they are d-connected by every $S \subseteq V$ in G. Then Z and Y are adjacent in G, and an intervention on X does not remove this adjacency, thus they are still adjacent in G' and thus d-connected by every $S \subseteq V$ in G'. Q.E.D.

Proposition 2: *No order reliable procedure randomizing a single variable at each step requires fewer than N -1 experiments for an N variable problem in the worst case.*

Proof: In order to show that N-1 experiments are in the worst case necessary given N variables let $X_1, \ldots, X_N$ again specify an arbitrary ordering of the N vertices. Suppose only N-2 interventions were performed in sequence, one each on X_1 to X_{N-2}. Suppose that in the true underlying causal graph X_{N-1} and X_N happen to both be (direct) causes of each

X_i, where $1 \leq i \leq N\text{-}2$, and that X_{N-1} and X_N are adjacent. It does not matter in which direction this edge is pointing, but assume, without loss of generality, that X_N is a parent of X_{N-1}. Note that in this case all of the interventions on X_1, …, X_{N-2} will indicate that there is an edge between X_N and X_{N-1}, but none will be able to direct it. Hence, an (N - 1)th experiment is required. Q.E.D.

Comment: A similar situation occurs when each X_i, where $1 \leq i \leq N - 2$, is a (direct) common cause of X_N and X_{N-1} and when, again, X_N is the parent of X_{N-1} or vice versa. Here also, none of the N-2 experiments will be able to identify the direction of the edge between X_N and X_{N-1}. It follows that N-1 experiments are sufficient and in the worst case necessary to identify the causal graph underlying N vertices. N - 1 is a tight bound for the worst case number of single intervention experiments.

The fact that the sequence of experimental interventions is arbitrary in the previous proof suggests that this result is still true for the worst case even when the choice of the next experiment is adaptive, that is, even if at each point during the sequence of experiments the "best" experiment given the evidence from the previous experiment is chosen. Although Proposition 3 follows from the previous two proofs as a corollary, the proof below emphasizes the aspect that no *adaptive* strategy will do any better in the worst case.

Proposition 3: *Every reliable adaptive procedure for which each experiment randomizes a single variable requires, in the worst case, at least N -1 experiments for an N vertex problem.*

Proof: Clearly N -1 experiments are sufficient to identify the causal graph underlying N vertices since they are sufficient for the non-adaptive case. In the following we will show that in the worst case N - 1 experiments are necessary even if an adaptive strategy is adopted for the experimental sequence. The situation can be viewed as a game between experimenter and nature: The experimenter specifies an experiment and nature returns the independence relations true of the graph, possibly modified by the experimental intervention. At each point in the game, however, nature may return the independence relations implied by the largest equivalence class of graphs that are consistent with the independence relations supplied to the experimenter in the previous experiments. The claim of proposition 3 amounts then to the claim that there always exists a strategy for nature that ensures that the experimenter requires N-1 experiments to reduce the equivalence class of graphs over the N variables to one, i.e. to identify the underlying causal structure uniquely. Consequently, no matter what the experimenter's adaptive strategy is, it will take N-1 experiments in this game.

Nature's strategy is as follows: Let V_1, V_2,… be the sequence of variables the experimenter intervenes upon. When the experimenter intervenes upon variable V_i, nature maintains the equivalence class of graphs that satisfy the following conditions: The class contains *all* the graphs that have complete subgraphs among the non-intervened variables, i.e. V_{i+1},…,V_N and V_k is a direct cause of V_j , where $1 \leq k \leq i$ and $1 \leq j \leq k$ with

$j \neq k$. In other words, whichever variable the experimenter intervenes upon, it is the sink of all the variables that have not yet been intervened upon, that is V_N is a (direct) parent of all other vertices, V_{N-1} is a parent of $V_1,\ldots, V_{N-2}$ etc. Then V_2 is a parent of V_1 only and V_1 is a child of all other vertices. Note, that the equivalence class of graphs nature maintains is the set of graphs which are all isomorphic to each other among the variables not intervened upon.

Now consider the adaptive strategy of the experimenter trying to identify the graph. At each stage in the game, she has no information about the directions of the edges among the non-intervened variables. So, in particular, after N-2 experiments, she has no information on the direction of the edge between V_{N-1} and V_N. Hence an (N-1)th experiment is required. It follows that even with an adaptive strategy, N-1 experiments are in the worst case necessary to identify the causal graph among N variables. Q.E.D.

Other Types of Experiments

In the previous two proofs an experiment was always assumed to consist of an intervention on one particular variable. However, it might be thought that other types of experiments, such as passive observations or interventions on more than one variable might improve the worst case result of N-1 experiments. While it is true that multiple interventions (randomizing more than one variable at a time) can shorten the experimental sequence, this is not the case for passive observational studies. We call a passive observational experiment a null-experiment.

The above proofs indicate that the worst case always occurs for particular complete graphs. If one were to run a null-experiment at any point in the experiment sequence when the underlying graph is complete - the most likely time would probably be at the beginning - then one would realize that one is confronted with a complete graph. However, this information (and more) is obtained anyway from two sequential experiments, each consisting of an intervention on a particular variable. The null-experiment paired with any other experiment cannot generate more information about the graph than two single intervention experiments, since a single intervention experiment also identifies all adjacencies except for those into the intervened variable. But a second intervention on a different variable would identify these interventions, too. So the only advantage of the null-experiment is in the case where only one experiment is run. The above proofs only apply to graphs of three or more variables, which certainly cannot always be identified by one experiment alone. In fact, even for two variables, two experiments are needed in the worst case (see discussion in main body of the paper).

References

D.M. Chickering (2002). Learning Equivalence Classes of Bayesian-Network Structures. *Journal of Machine Learning Research*, 2:445-498.

C. Meek, (1996). Ph.D Thesis, Department of Philosophy, Carnegie Mellon University

K.P. Murphy, (2001). Active Learning of Causal Bayes Net Structure, Technical Report, Department of Computer Science, U.C. Berkeley.

J. Pearl, (1988). *Probabilistic Reasoning in Intelligent Systems*, San Mateo, CA. Morgan Kaufmann.

J. Pearl, (2000). *Causality*, Oxford University Press.

P. Spirtes, C.Glymour and R. Scheines, (1993). *Causation, Prediction and Search*, Springer Lecture Notes in Statistics, 1993; 2nd edition, MIT Press, 2000.

P. Spirtes, C. Glymour, R. Scheines, S. Kauffman, W. Aimalie, and F. Wimberly, (2001). Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data, in *Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems and Technology*, Duke University, March.

S. Tong and D. Koller, (2001). Active Learning for Structure in Bayesian Networks, *Proceedings of the International Joint Conference on Artificial Intelligence*.