Journal of Machine Learning Research 00 (2013) 1-32 Submitted 07/12; Published 00/00

Experiment Selection for Causal Discovery

Antti Hyttinen ANTTI.HYTTINEN@HELSINKI.FI
Helsinki Institute for Information Technology
Department of Computer Science
University of Helsinki
Helsinki, Finland

Frederick Eberhardt FDE@CALTECH.EDU
Philosophy, Division of the Humanities and Social Sciences
California Institute of Technology
Pasadena, California, USA

Patrik O. Hoyer PATRIK.HOYER@HELSINKI.FI
Helsinki Institute for Information Technology
Department of Computer Science
University of Helsinki
Helsinki, Finland

Editor:

Abstract

Randomized controlled experiments are often described as the most reliable tool available to scientists for discovering causal relationships among quantities of interest. However, it is often unclear how many and which different experiments are needed to identify the full (possibly cyclic) causal structure among some given (possibly causally insufficient) set of variables. Recent results in the causal discovery literature have explored various identifiability criteria that depend on the assumptions one is able to make about the underlying causal process, but these criteria are not directly constructive for selecting the optimal set of experiments. Fortunately, many of the needed constructions already exist in the combinatorics literature, albeit under terminology which is unfamiliar to most of the causal discovery community. In this paper we translate the theoretical results and apply them to the concrete problem of experiment selection. For a variety of settings we give explicit constructions of the optimal set of experiments and adapt some of the general combinatorics results to answer questions relating to the problem of experiment selection.

Keywords: Causality, Randomized experiments, Experiment selection, Separating systems, Completely separating systems, Cut-coverings.

1. Introduction

In a variety of scientific fields one of the main goals is to understand the causal relationships among some variables of interest. In most cases empirical data is key to discovering and verifying such relationships. While there has been much work on inference principles and algorithms for discovering causal relationships from purely 'passive observational' data, largely based on the seminal work of Spirtes et al. (1993) and Pearl (2000), randomized controlled experiments (Fisher, 1935) still often constitute the tool of choice for inferring

causal relationships. In the more recent literature on causal discovery, randomized experiments, in combination with novel inference principles, play an increasingly prominent role (Cooper and Yoo, 1999; Tong and Koller, 2001; Murphy, 2001; Eaton and Murphy, 2007; Meganck et al., 2005; Eberhardt et al., 2005; He and Geng, 2008; Hyttinen et al., 2010, 2011; Claassen and Heskes, 2010). Thus, given a set of assumptions one is willing to make in a test setting, questions concerning the optimal choices of the manipulations arise. What sequence of experiments identifies the underlying causal structure most efficiently? Or, given some background knowledge, how can one select an experiment that maximizes (in some to be defined sense) the insight one can expect to gain?

In our work (Eberhardt, 2007; Eberhardt et al., 2010; Hyttinen et al., 2011, 2012a,b) we found that many of these questions concerning the optimal selection of experiments are equivalent to graph-theoretic or combinatoric problems for which, in several cases, there exist solutions in the mathematics literature. Generally these solutions are couched in a terminology that is neither common in the literature on causal discovery, nor obvious for its connections to the problems in causal discovery. The present article is intended to bridge this terminological gap, both to indicate which problems of experiment selection already have formal solutions, and to provide explicit procedures for the construction of optimal sets of experiments.[1] It gives rise to new problems that (to our knowledge) are still open and may benefit from the exchange of research in causal discovery on the one hand, and the field of combinatorics and graph theory on the other.

2. Causal Models, Experiments, and Identifiability

We consider causal models which represent the relevant causal structure by a directed graph $G = (\mathcal{V}, \mathcal{D})$, where $\mathcal{V}$ is the set of variables under consideration, and $\mathcal{D} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of directed edges among the variables. A directed edge from $x_i \in \mathcal{V}$ to $x_j \in \mathcal{V}$ represents a direct causal influence of x_i on x_j, with respect to the full set of variables in $\mathcal{V}$ (Spirtes et al., 1993; Pearl, 2000). In addition to the graph G, a fully specified causal model also needs to describe the causal processes that determine the value of each variable given its direct causes. Typically this is achieved either by using conditional probability distributions (in the 'causal Bayes nets' framework) or stochastic functional relationships (for 'structural equation models' (SEMs)).

In addition to describing the system in its 'natural' or 'passive observational' state, a causal model also gives a precise definition of how the system behaves under manipulations. Specifically, consider an intervention that sets (that is, forces) a given variable $x_i \in \mathcal{V}$ to some value randomly chosen by the experimenter. Such a "surgical" intervention corresponds to deleting all arcs pointing *into* x_i (leaving all outgoing arcs, and any other arcs in the model, unaffected), and disregarding the specific process by which x_i normally acquires its value.[2] The resulting graph is known as the 'manipulated' graph corresponding to this intervention. If, in an experiment, the values of *several* variables are set by the experi-

1. The different constructions and bounds presented in the paper are implemented in a code package at: `http://www.cs.helsinki.fi/u/ajhyttin/exp/`
2. In Spirtes et al. (1993) this is how manipulations are defined, which are then combined with the Markov assumption to yield the Manipulation Theorem that specifies formally the relation between the passive observational and a manipulated model. In Pearl (2000) the relation is formally specified using the *do*-operator in combination with the modularity assumption. For our purposes, either connection suffices.

menter, any arcs into any of those variables are deleted. In this way, a causal model provides a concrete prediction for the behaviour of the system under any experimental conditions.

The problem of causal discovery is to infer (to the fullest extent possible) the underlying causal model, from sample data generated by the model. The data can come either from a passive observational setting (no manipulations performed by the researcher) or from one or more randomized experiments, each of which (repeatedly) sets some subset of the variables to values determined purely by chance, while simultaneously measuring the remaining variables. We define an experiment $\mathcal{E} = (\mathcal{J},\mathcal{U})$ as a partition of the variable set $\mathcal{V}$ into two mutually exclusive and collectively exhaustive sets $\mathcal{J}$ and $\mathcal{U}$, where $\mathcal{J} \subseteq \mathcal{V}$ represents the variables that are intervened on (randomized) in experiment $\mathcal{E}$, and $\mathcal{U} = \mathcal{V} \setminus \mathcal{J}$ represents the remaining variables, all of which are passively observed in this experiment. We will not consider the specific distributions employed to randomize the intervened variables, except to require that the distribution is positive over all combinations of values of the intervened variables. Note that the identifiability results mentioned below apply when the variables simultaneously intervened on in one experiment are randomized independently of one another.[3]

The extent to which the underlying causal model can be inferred then depends not only on the amount of data available (number of samples) but fundamentally also on the details of what experiments are available and what assumptions on the underlying model one can safely make. In what follows, we only consider model *identifiability*[4], that is, we disregard sample size and only examine the settings under which models can be learned in the large sample limit. The identifiability results we consider build on causal discovery procedures that make one or more of the following standard assumptions on the underlying model:

acyclicity The graph G is often assumed to be *acyclic*, that is, there exists no directed path from a node back to itself. This assumption is useful for causal discovery because finding that x causes y allows us to deduce that y does not cause x.

causal sufficiency In many cases only a subset of the variables involved in the underlying data generating process are measured. Even if some variables are unobserved, a causal model is said to be *causally sufficient* if there are no unobserved common causes of the observed variables. Unobserved common causes are typically troublesome because they bring about a dependence between two observed variables that is not due to any actual causal process among the observed variables.

faithfulness Many causal discovery procedures use independence tests as a primary tool for inferring the structure of the underlying graph. Such inferences are correct in the limit if the distribution generated by the model is *faithful* to the graph structure, that is, all independencies in distribution are consequences of the graph structure rather than the specific parameter values defining the quantitative relationships between the variables. Under faithfulness, perturbing the parameters defining the quantitative relationships will not break any of the observed independencies between the variables in the distribution.

3. For linear cyclic models this assumption can be relaxed; see Lemma 5 in Hyttinen et al. (2012b).
4. In contrast, Shpitser and Pearl (2006) and Bareinboim and Pearl (2012) consider the identifiability of single causal effects presuming the knowledge of the true causal structure.

parametric form Some discovery methods rely on the quantitative causal relations between the variables being restricted to a particular (simple or smooth) *parametric form*. The most common such assumption is linearity: the value of each variable is given by a linear sum of the values of its parents plus a stochastic error term.

It is well known that even when one makes *all* of the above assumptions (using only linearity as a parametric restriction), the true causal structure is in general underdetermined given only passive observational data, but can be identified using experiments. We can ask more generally: Under what combination of assumptions and conditions on a set of K experiments $\{\mathcal{E}_1, \ldots, \mathcal{E}_K\}$ is the underlying causal structure identified?

If a total of n observed variables is considered, it should come as no surprise that a set of $K = n$ randomized experiments, each of which intervenes on all but one of the variables, is in general sufficient to uniquely identify the graph G that represents the causal structure among the n variables. In each such experiment we can test which of the $n-1$ other variables are direct causes of the one non-intervened variable. A natural question is whether the full identification of G can be achieved with other sets of experiments, under various combinations of the above assumptions. In particular, we can ask whether identification can be reached with fewer than n experiments, or with experiments that only involve simultaneous interventions on many fewer than $n-1$ variables in each experiment.

Figure 1 provides an example with three variables. Suppose the true causal structure is the chain $x \leftarrow y \leftarrow z$, as shown in graph (i). Assuming only faithfulness *or* linearity, the three experiments intervening on two variables each are sufficient to uniquely identify the true causal graph. In particular, when intervening on both x and y, as illustrated in graph (ii), both of the edges in the true model are cut, and z is independent of both of the intervened variables, indicating that the edges $x \rightarrow z$ and $y \rightarrow z$ are both absent. On the other hand, when intervening on x and z, as shown in graph (iii), a dependence is found between y and z, indicating that $y \leftarrow z$ must be present, while the independence between x and y rules out the edge $x \rightarrow y$. Similar considerations apply when intervening on y and z, illustrated in graph (iv). Together, all potential edges in the model are established as either present or absent, and hence the full causal structure is identified. Note that this inference is possible without assuming causal sufficiency or acyclicity.

Assuming causal sufficiency, acyclicity and faithfulness, fewer experiments are needed: If one had started with an experiment only intervening on x, then a second single intervention experiment on y would be sufficient for unique identifiability. This is because the first experiment, illustrated in graph (v), rules out the edges $x \rightarrow y$ and $x \rightarrow z$, but also establishes, due to the statistical dependence, that y and z are connected by an edge (whose orientation is not yet known). Intervening on y next, as shown in graph (vi), establishes the edge $x \leftarrow y$, and the absence of the edge $y \rightarrow z$, allowing us to conclude (using (v)) that $y \leftarrow z$ must be present in the true graph. Finally, x and z are independent in this second experiment, ruling out the edge $x \leftarrow z$. Thus, the true causal structure is identified. If one had been *lucky* to start with an intervention on z then it turns out that one could have identified the true causal graph in this single experiment. But that single experiment would, of course, have been insufficient if in fact the causal chain had been oriented in the opposite direction.

The example illustrates the sensitivity of the identifiability results to the model space assumptions. However, recent research has shown that, in several different settings (de-

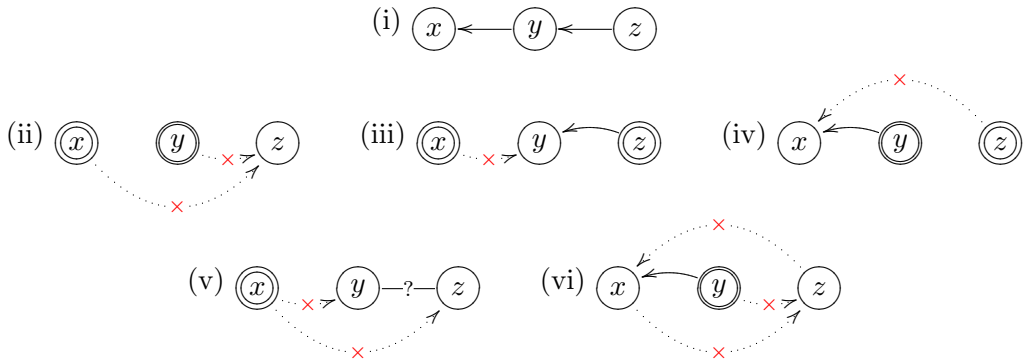

Figure 1: Graph (i) shows the true data generating structure. Graphs (ii-vi) show the possible inferences about the causal structure that can be made from experiments intervening on two variables simultaneously (ii-iv), or intervening on a single variable (v-vi). Variables that are circled twice are intervened on in the corresponding experiment. Edges determined to be present are solid, edges determined to be absent are dotted and crossed out (×). The solid line with a question mark denotes an edge determined to be present but whose orientation is unknown. See the text for a description of the background assumptions that support these inferences.

scribed explicitly below), identification hinges on the set of experiments satisfying some relatively simple conditions. Specifically, consider the following conditions:

Definition 1 (Unordered Pair Condition) *A set of experiments* $\{\mathcal{E}_1, \ldots, \mathcal{E}_K\}$ *satisfies the unordered pair condition for an unordered pair of variables* $\{x_i, x_j\} \subseteq \mathcal{V}$ *whenever there is an experiment* $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ *in* $\{\mathcal{E}_1, \ldots, \mathcal{E}_K\}$ *such that* $x_i \in \mathcal{J}_k$ *(*x_i *is intervened on) and* $x_j \in \mathcal{U}_k$ *(*x_j *is passively observed), or* $x_j \in \mathcal{J}_k$ *(*x_j *is intervened on) and* $x_i \in \mathcal{U}_k$ *(*x_i *is passively observed).*

Definition 2 (Ordered Pair Condition) *A set of experiments* $\{\mathcal{E}_1, \ldots, \mathcal{E}_K\}$ *satisfies the ordered pair condition for an ordered pair of variables* $(x_i, x_j) \in \mathcal{V} \times \mathcal{V}$ *(with* $x_i \neq x_j$*) whenever there is an experiment* $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ *in* $\{\mathcal{E}_1, \ldots, \mathcal{E}_K\}$ *such that* $x_i \in \mathcal{J}_k$ *(*x_i *is intervened on) and* $x_j \in \mathcal{U}_k$ *(*x_j *is passively observed).*

Definition 3 (Covariance Condition) *A set of experiments* $\{\mathcal{E}_1, \ldots, \mathcal{E}_K\}$ *satisfies the covariance condition for an unordered pair of variables* $\{x_i, x_j\} \subseteq \mathcal{V}$ *whenever there is an experiment* $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ *in* $\{\mathcal{E}_1, \ldots, \mathcal{E}_K\}$ *such that* $x_i \in \mathcal{U}_k$ *and* $x_j \in \mathcal{U}_k$*, that is, both variables are passively observed.*

The above conditions have been shown to underlie the following identifiability results: Assuming causal sufficiency, acyclicity and faithfulness, a set of experiments uniquely identifies the causal structure of a causal Bayes net if and only if for any two variables $x_i, x_j \in \mathcal{V}$ one of the following is true: (i) the ordered pair condition holds for the ordered pairs (x_i, x_j) and (x_j, x_i), or (ii) the unordered pair condition and the covariance condition hold for the unordered pair $\{x_i, x_j\}$ (Eberhardt, 2007). Note that the 'only if' part is a worst-case result: For any set of experiments that does not satisfy the above requirement, there exists

a causal graph such that the structure is not identified with these experiments. Since a single passive observation – a so-called *null*-experiment, as $\mathcal{J} = \emptyset$ – satisfies the covariance condition for all pairs of variables, under the stated assumptions the main challenge is to find experiments that satisfy the *unordered pair condition* for every pair of variables.

Without causal sufficiency, acyclicity *or* faithfulness, but assuming a linear data generating model, a set of experiments uniquely identifies[5] the causal structure among the observed variables of a linear SEM if and only if it satisfies the *ordered pair condition* for all ordered pairs of variables (Eberhardt et al., 2010; Hyttinen et al., 2010, 2012a,b). Similar identifiability results can be obtained for (acyclic) causal models with binary variables by assuming a noisy-OR parameterization for the local conditional probability distributions (Hyttinen et al., 2011).

Such identifiability results immediately give rise to the following questions of optimal experiment selection:

- What is the least number of experiments that satisfy the above conditions?
- Can we give procedures to construct such sets of experiments?

The above questions can be raised similarly given additional context, such as the following:

- The number of variables that can be subject to an intervention simultaneously is limited in some way.
- Background knowledge about the underlying causal structure is available.

Naturally, there are other possible scenarios, but we focus on these, since we are aware of their counterparts in the combinatorics literature. To avoid having to repeatedly state the relevant search space assumptions, we will present the remainder of this article in terms of the satisfaction of the (unordered and ordered) pair conditions, which provide the basis for the identifiability results just cited.

3. Correspondence to Separating Systems and Cut-coverings

The satisfaction of the pair conditions introduced in the previous section is closely related to two problems in combinatorics: Finding *(completely) separating systems*, and finding *(directed) cut-coverings*. Throughout, to simplify notation and emphasize the connections, we will overload symbols to the extent that there is a correspondence to the problem of experiment selection for causal discovery.

Definition 4 (Separating System) *A* separating system $\mathcal{C} = \{\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_K\}$ *is a set of subsets of an n-set* $\mathcal{V}$ *with the property that given any two distinct elements* $x_i, x_j \in \mathcal{V}$*, there exists a* $\mathcal{J}_k \in \mathcal{C}$ *such that* $x_i \in \mathcal{J}_k \wedge x_j \notin \mathcal{J}_k$ or $x_i \notin \mathcal{J}_k \wedge x_j \in \mathcal{J}_k$*.*

Definition 5 (Completely Separating System) *A* completely separating system $\mathcal{C} = \{\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_K\}$ *is a set of subsets of an n-set* $\mathcal{V}$ *with the property that given any two distinct elements* $x_i, x_j \in \mathcal{V}$*, there exist* $\mathcal{J}_k, \mathcal{J}_{k'} \in \mathcal{C}$ *such that* $x_i \in \mathcal{J}_k \wedge x_j \notin \mathcal{J}_k$ and $x_i \notin \mathcal{J}_{k'} \wedge x_j \in \mathcal{J}_{k'}$*.*

5. For simplicity, we focus in this article on identifying the causal structure among the observed variables, even though some of the cited results also permit the identification of confounding.

As can be easily verified, a set of experiments $\{\mathcal{E}_1, \ldots, \mathcal{E}_K\}$ that satisfies the *unordered* pair condition for all pairs over n variables in $\mathcal{V}$ directly corresponds to a separating system over the variable set, while a set of experiments that satisfies the *ordered* pair condition for all ordered variable pairs corresponds to a *completely* separating system over the variables.[6]

A related but more general problem is that of finding cut-coverings. First, we need to define cuts and directed cuts: A *cut* $\mathcal{E}_k$ corresponds to a partition of a set of vertices $\mathcal{V}$ of an undirected graph $H = (\mathcal{V}, \mathcal{P})$ into two sets $\mathcal{J}$ and $\mathcal{U}$. Any edge $p \in \mathcal{P}$ connecting an $x_i \in \mathcal{J}$ to an $x_u \in \mathcal{U}$ is said to be *in the cut* $\mathcal{E}_k$. For a directed graph F, a *directed cut* is a cut where only the edges *from* vertices in $\mathcal{J}$ *to* vertices in $\mathcal{U}$ are in the cut, while edges in the opposite direction are not. We are thus ready to define a *cut-covering* and a *directed cut-covering*:

Definition 6 (Cut-covering) *A cut-covering for an undirected graph $H = (\mathcal{V}, \mathcal{P})$ is a set of cuts $\{\mathcal{E}_1, \ldots, \mathcal{E}_K\}$ such that each edge $p \in \mathcal{P}$ of H is in some cut $\mathcal{E}_k$.*

Definition 7 (Directed Cut-covering) *A directed cut-covering for a directed graph $F = (\mathcal{V}, \mathcal{Q})$ is a set of directed cuts $\{\mathcal{E}_1, \ldots, \mathcal{E}_K\}$ such that each directed edge $q \in \mathcal{Q}$ of F is in some directed cut $\mathcal{E}_k$.*

The correspondence of finding cut-coverings to the problem of experiment selection is now immediate: In the case of searching for a set of experiments that satisfies the ordered pair condition for all ordered pairs of variables in $\mathcal{V}$, let the graph $F = (\mathcal{V}, \mathcal{Q})$ be a *complete* directed graph over the vertex set $\mathcal{V}$ where each ordered pair of variables is connected by a directed edge. Each directed edge represents an ordered pair condition for a pair of vertices (x_i, x_j) that needs to be satisfied by the set of experiments. Finding such a set of experiments is then equivalent to finding a directed cut-covering for F, where each experiment corresponds to a directed cut. An analogous correspondence holds for the unordered pair condition with a complete undirected graph H. We discuss the generalization and interpretation of the problem when H or F are not complete graphs in Section 6.

As our overloading of symbols suggests, most aspects of the experiment selection have direct counterparts in the cut-covering representation. However, the edges representing direct causes in a causal graph G *do not* correspond to the edges representing the satisfaction of an ordered pair condition in the directed graph F. That is, F and G in general do not share the same edge structure: G is the graph of the underlying causal model to be identified, while F represents the set of ordered pairs that are not yet satisfied. Moreover, there is a difference in how the causal graph G is changed in light of an experiment $\mathcal{E} = (\mathcal{J}, \mathcal{U})$, and how the ordered pair graph F is changed in light of the (corresponding) cut $\mathcal{E}$. The experiment $\mathcal{E}$ results in the manipulated causal graph G', which is the same as the original causal graph G except that the edges *into* the variables in $\mathcal{J}$ are removed. The corresponding *cut* $\mathcal{E}$, however, cuts the edges of the ordered pair condition graph F that are *outgoing* from variables in $\mathcal{J}$ (and simultaneously *into* variables in $\mathcal{U}$). This may seem unintuitive, but in fact these two representations of $\mathcal{E}$ (the experiment and the cut) illustrate two aspects of what an experiment achieves: It manipulates the underlying causal graph G (by breaking incoming edges on variables in $\mathcal{J}$), and it satisfies the ordered pair condition for all ordered

6. *Separating systems* and *completely separating systems* are sometimes also referred to by the terms *weakly separating systems* and *strongly separating systems*, respectively.

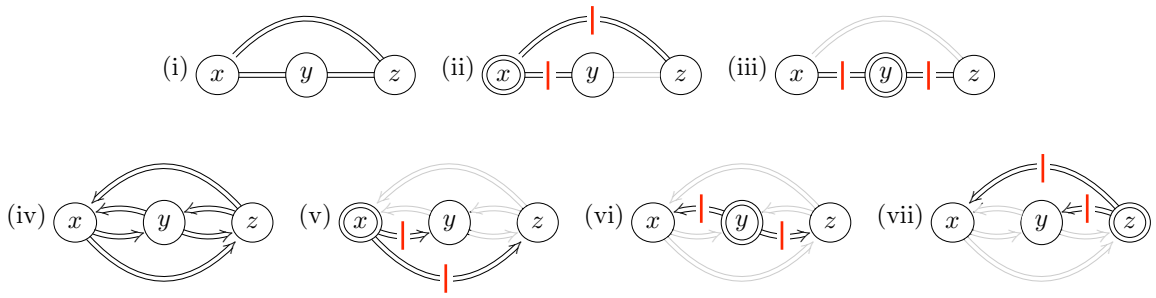

Figure 2: *Top*: Satisfaction of the unordered pair condition for all pairs of variables in a three variable model. Undirected double-lined edges indicate pairs for which the unordered pair condition needs to be satisfied (graph (i)). The cuts (|) indicate which pairs are satisfied by the experiments intervening on x (graph (ii)) and y (graph (iii)). *Bottom*: Satisfication of the ordered pair condition: Directed double-lined edges indicate ordered pairs for which the ordered pair condition needs to be satisfied (graph (iv)). Graphs (v-vii) show which pairs are in the directed cuts (are satisfied) by the respective single-intervention experiments. See text for details.

pairs (x_j, x_u) with $x_j \in \mathcal{J}$ and $x_u \in \mathcal{U}$ by determining whether x_j has a causal effect on x_u. Similarly in the unordered case, the causal graph G (or its skeleton) must not be confused with the undirected graph H representing the unordered pairs that are not yet satisfied.

Figure 2 illustrates for a three variable model (such as that in Figure 1) how the satisfaction of the (un)ordered pair condition for all pairs of variables is guaranteed using cut-coverings. Graph (i) gives the complete *undirected* graph H over $\{x, y, z\}$ illustrating the three unordered pairs for which the *unordered* pair condition needs to be satisfied. Graphs (ii) and (iii) show for which pairs the unordered pair condition is satisfied by a single intervention experiment on x (or y, respectively), that is, which pairs are in the cut (|). The pairs that remain unsatisfied by each experiment, respectively, are shown in gray for easier legibility. Together these experiments constitute a cut-covering for H. Similarly, graph (iv) gives the complete *directed* graph F over the three variables, illustrating the six ordered pairs of variables for which the *ordered* pair condition needs to be satisfied. Graphs (v-vii) show for which pairs the ordered pair condition is satisfied by a single intervention experiment on x (or y or z, respectively), that is which pairs are in the directed cut (|), while the others are again shown in gray. As can be seen, in the ordered case all three experiments are needed to provide a directed cut-covering for F.

The correspondence between the problem of finding experiments that satisfy the pair conditions on the one hand and finding separating systems or cut-coverings on the other, allows us to tap into the results in combinatorics to inform the selection of experiments in the causal discovery problem.

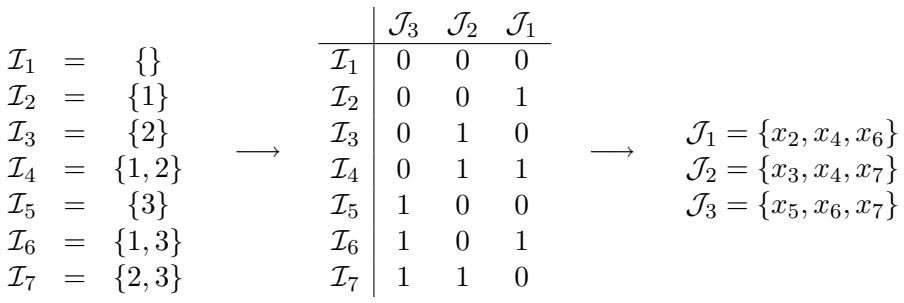

Figure 3: An illustration of the relationship between intervention sets and index sets, giving the construction of a minimal set of experiments satisfying the *unordered* pair condition for all pairs of variables in a 7 variable model. The index sets $\mathcal{I}_1, \ldots, \mathcal{I}_n$ (left) are chosen as distinct subsets of the set of experiment indexes $\{1, 2, 3\}$, and each row of the binary matrix (middle) marks the corresponding experiments. The intervention sets $\mathcal{J}_1, \ldots, \mathcal{J}_K$ (right) are then obtained by reading off the columns of this matrix. Note that one additional variable could still be added (intervened on in all three experiments) while still satisfying the unordered pair condition for all variable pairs. For nine variables, however, a minimum of four experiments would be needed.

4. Minimal Sets of Experiments

We now turn to the concrete problem of constructing sets of experiments which satisfy the pair condition for all pairs of variables, while simultaneously requiring as few experiments as possible. We divide the results concerning the unordered and the ordered pair condition into Sections 4.1 and 4.2, respectively. In what follows, we always start by presenting the explicit construction of the intervention sets, and subsequently give the available bounds on the number of experiments. The constructions are also available as special cases of the algorithms presented in Section 5, and implemented in our associated code package. Throughout, $i = 1, \ldots, n$ indexes the variables in the variable set $\mathcal{V}$, while $k = 1, \ldots, K$ indexes the experiments in the construction.

Many of the constructions of intervention sets $\mathcal{J}_1, \ldots, \mathcal{J}_K$ will be examined using so-called *index sets* $\mathcal{I}_1, \ldots, \mathcal{I}_n$, where

$$\mathcal{I}_i \;=\; \{k \,|\, x_i \in \mathcal{J}_k\}.$$

That is, the i:th index set simply lists the indexes of the intervention sets that include the variable x_i. Clearly, K experiments (or separating sets) over n variables can be defined either in terms of the intervention sets $\mathcal{J}_1, \ldots, \mathcal{J}_K$ or equivalently in terms of the index sets $\mathcal{I}_1, \ldots, \mathcal{I}_n$. See Figure 3 for an illustration.

4.1 Satisfying the Unordered Pair Condition

The earliest results (that we are aware of) relevant to finding minimal sets of experiments satisfying the unordered pair condition are given in Rényi (1961)[7] in the terminology of separating systems. He found that a separating system $\mathcal{C} = \{\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_K\}$ over $\mathcal{V}$ can be obtained by assigning distinct binary numbers to each variable in $\mathcal{V}$. That is, the strategy is to choose distinct index sets for all variables in $\mathcal{V}$. This is supported by the following Lemma:

Lemma 8 (Index sets must be distinct) *Intervention sets $\{\mathcal{J}_1, \ldots, \mathcal{J}_K\}$ satisfy the unordered pair condition for all unordered pairs of variables if and only if the corresponding index sets $\mathcal{I}_1, \ldots, \mathcal{I}_n$ are distinct.*

Proof Assume that the index sets are distinct. Since the index sets $\mathcal{I}_i$ and $\mathcal{I}_j$ of any two variables are distinct, there must exist an index k such that either $k \in \mathcal{I}_i$ and $k \notin \mathcal{I}_j$, or $k \notin \mathcal{I}_i$ and $k \in \mathcal{I}_j$. Thus, the experiment $\mathcal{E}_k$ satisfies the unordered pair condition for the pair $\{x_i, x_j\}$.

Next assume that the unordered pair condition is satisfied for all pairs. Take two arbitrary index sets $\mathcal{I}_i$ and $\mathcal{I}_j$. Since the unordered pair condition is satisfied for the pair $\{x_i, x_j\}$, there is an experiment $\mathcal{E}_k$ where x_i is intervened on and x_j is not, or x_j is intervened and x_i is not. Either way, $\mathcal{I}_i \neq \mathcal{I}_j$. ■

Again, Figure 3 is used to illustrate this concept.

Rényi (1961, p. 76) notes that the smallest separating system for a set of n variables has size

$$c(n) \quad = \quad \lceil \log_2(n) \rceil. \tag{1}$$

This is clear from the previous lemma: K experiments only allow for up to 2^K distinct index sets. Equivalent results and procedures are derived, only in the terminology of finding minimal cut-coverings for complete graphs, by Loulou (1992, p. 303). In the terminology of causal discovery, Eberhardt (2007, Theorem 3.3.4) requires $\lfloor \log_2(n) \rfloor + 1$ experiments to guarantee identifiability of a causal model.[8] For graphs over three variables, the result is obvious given the graph H (for the unordered pair condition) in Figure 2 (top, left): For a cut-covering of the three undirected edges, $2 = \lceil \log_2(3) \rceil$ cuts are necessary and sufficient. The two cuts in graphs (ii) and (iii) in Figure 2 corresponding to the experiments in the last row of Figure 1 are an example. Figure 4 shows the number of experiments needed to satisfy the unordered pair condition for all variable pairs for models of up to 5,000 variables.

4.2 Satisfying the Ordered Pair Condition

When Dickson (1969, p. 192) coined the term "completely separating systems", he also showed that as the number $n = |\mathcal{V}|$ of elements tends to infinity, the size of a minimal

7. Rényi (1961) also examines the probability of finding separating systems when subsets of $\mathcal{V}$ are selected randomly. These results apply to causal discovery when experiments are selected at random.
8. The one additional experiment sometimes required in this case derives from the need to satisfy the ordered pair condition, or unordered pair condition *and* the covariance condition, for each pair of variables, as discussed in Section 2. The covariance condition can be trivially satisfied with a single passive observational data set (a so-called *null*-experiment).

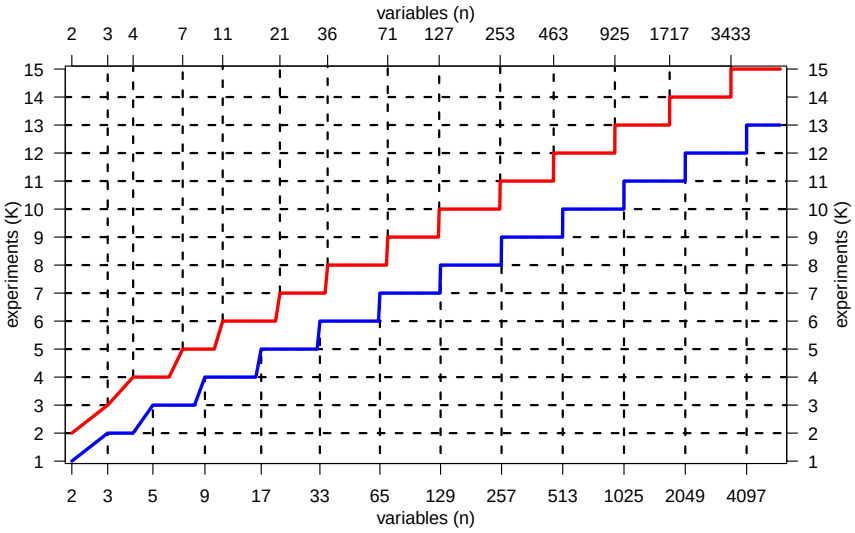

Figure 4: Sufficient and necessary number of experiments needed to satisfy the unordered (blue, lower solid line) and the ordered (red, upper solid line) pair condition for models of different sizes (in log-scale). The number of variables in the models are only ticked on the x-axis when an additional experiment is needed. For example, for a model with 100 variables, 7 experiments are needed to satisfy the unordered pair condition, while 9 experiments are needed to satisfy the ordered pair condition.

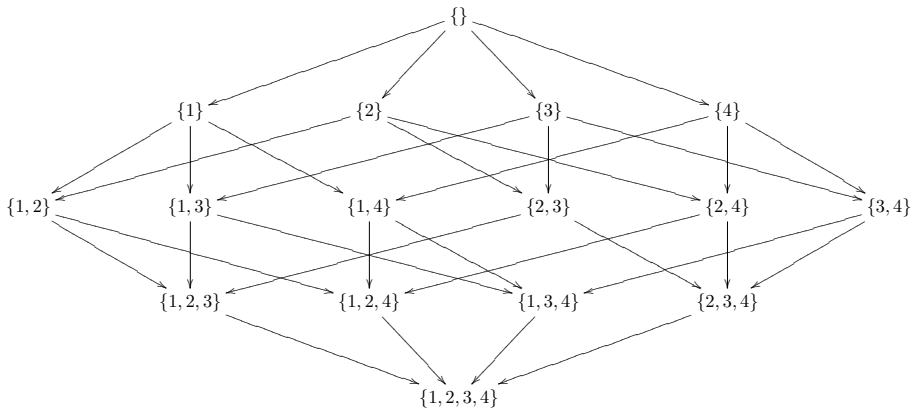

Figure 5: Subset lattice of subsets of $\{1, 2, 3, 4\}$. A directed path from set $\mathcal{S}_i$ to set $\mathcal{S}_j$ exists if and only if $\mathcal{S}_i \subset \mathcal{S}_j$. The largest antichain is the family of sets that are not connected by any directed paths, which in this case is formed by all the sets of size 2, that is, all the sets on the middle 'row' of the graph.

completely separating system approaches the size of a (standard) separating system, that is, $\log_2(n)$. However, Dickson did not derive the exact minimal size. Shortly after, Spencer (1970) recognized the connection between completely separating systems and antichains in the subset lattice, as defined below. See Figure 5 for an illustration.

Definition 9 (Antichain) *An antichain (also known as a Sperner system) $\{\mathcal{S}_i\}$ over a set $\mathcal{S}$ is a family of subsets of $\mathcal{S}$ such that $\forall i, j : \mathcal{S}_i \not\subseteq \mathcal{S}_j$ and $\mathcal{S}_i \not\supseteq \mathcal{S}_j$.*

The connection between antichains and completely separating systems (that is, satisfying the ordered pair condition) is then the following:

Lemma 10 (Index sets form an antichain) *The intervention sets $\{\mathcal{J}_1, \ldots, \mathcal{J}_K\}$ satisfy the ordered pair condition for all ordered pairs if and only if the corresponding index sets $\{\mathcal{I}_1, \ldots, \mathcal{I}_n\}$ form an antichain over $\{1, \ldots, K\}$.*

Proof Assume that the index sets form an antichain. Consider an arbitrary ordered pair (x_i, x_j). Since index sets form an antichain we have that $\mathcal{I}_i \not\subseteq \mathcal{I}_j$ and there must be experiment $\mathcal{E}_k$ such that $k \in \mathcal{I}_i$ and $k \notin \mathcal{I}_j$. This experiment satisfies the ordered pair condition for the ordered pair (x_i, x_j).

Next assume that the ordered pair condition is satisfied for all ordered pairs. Take two arbitrary index sets $\mathcal{I}_i$ and $\mathcal{I}_j$. Since the ordered pair condition is satisfied for the pair (x_i, x_j), there is an experiment where x_i is intervened on and x_j is not, thus $\mathcal{I}_i \not\subseteq \mathcal{I}_j$. Symmetrically, $\mathcal{I}_j \not\subseteq \mathcal{I}_i$. Thus the index sets form an antichain. ■

Earlier, Sperner (1928) had already proven the following theorem on the maximum possible size of an antichain.

Theorem 11 (Sperner's Theorem) *The largest antichain over $\{1, \ldots, K\}$ is formed by the subsets of constant size $\lfloor K/2 \rfloor$ and thus has size $\binom{K}{\lfloor K/2 \rfloor}$.*

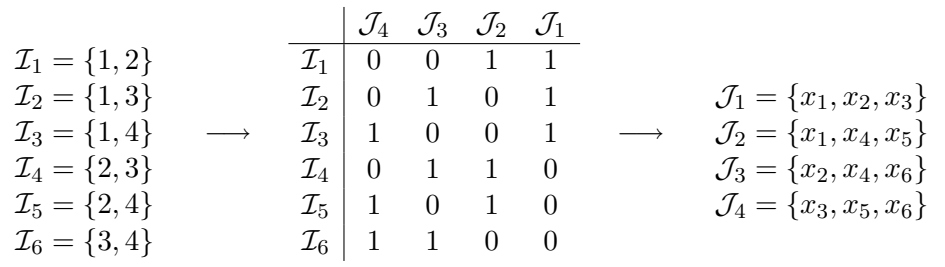

Figure 6: Designing the intervention sets of experiments satisfying the *ordered* pair condition for all ordered pairs of variables in a $n = 6$ variable model with $K = 4$ experiments. Select the index sets $\mathcal{I}_1, \ldots, \mathcal{I}_n$ as an antichain over $\{1, \ldots, K\}$ and translate the index sets into intervention sets $\mathcal{J}_1, \ldots, \mathcal{J}_K$.

Thus, the minimal completely separating system over a set of size n can always be constructed by selecting the corresponding index sets as any distinct $\lfloor K/2 \rfloor$-size subsets. See Figure 6 for an illustration. Using this rationale, Spencer (1970) notes that the cardinality $c(n)$ of a minimal completely separating system for n elements is given by

$$c(n) \quad = \quad \min\{K : \binom{K}{\lfloor K/2 \rfloor} \geq n\}, \tag{2}$$

which can be approximated using Stirling's approximation as

$$c(n) = \log_2(n) + \frac{1}{2}\log_2\log_2(n) + \frac{1}{2}\log_2(\frac{\pi}{2}) + o(1). \tag{3}$$

Equation 2 is re-proven for directed cut-coverings over complete graphs by Alon et al. (2007, Theorem 11), also using the connection to antichains. To our knowledge, tight bounds or constructions have not been previously described in the causal discovery literature on experiment selection.

For graphs over three variables, the graph F (for the ordered pair condition) in Figure 2 (bottom, left) illustrates the point: For a directed cut-covering of the six directed edges, $c(3) = 3$ directed cuts are necessary and sufficient. The three cuts shown in graphs (v-vii) of Figure 2 corresponding to the three possible single-intervention experiments are an example, but the cuts corresponding to the double-intervention experiments in graphs (ii-iv) of Figure 1 would also work. Figure 4 shows the number of experiments required for models with up to $5,000$ variables. Note that the difference between the number of experiments needed for a separating and a completely separating system over a given number of variables is only 2 or 3 experiments. Thus, in many cases the possibility of applying an inference procedure based on weaker assumptions may be worth the investigative cost of a few additional experiments to satisfy the ordered pair condition.

5. Limiting Intervention Set Size

Section 4 focused on characterizing minimal sets of experiments that guarantee identifiability, but paid no attention to the particular nature of those experiments. In some cases,

the experiments might require a simultaneous intervention on half of the variables, but of course such experiments will in many scientific contexts not be feasible. In this section we consider a generalization of the problem of finding the optimal set of experiments that can take into account additional constraints on the size of the intervention sets. We consider the following variants of the problem:

1. Given n variables and K experiments, find intervention sets $\mathcal{J}_1, \ldots, \mathcal{J}_K \subseteq \mathcal{V}$ satisfying the ordered or unordered pair condition for all variable pairs, such that the intervention sets have minimal

   (a) **average** intervention set size $\text{mean}_{k=1}^{K}|\mathcal{J}_k| = \frac{1}{K}\sum_{k=1}^{K}|\mathcal{J}_k|$ (which is equivalent to minimizing the total number of interventions), or

   (b) **maximum** intervention set size $\max_{k=1}^{K}|\mathcal{J}_k|$.

2. Given n variables and a maximum allowed intervention set size r, find the minimum number of experiments $m(n, r)$ for which there exists intervention sets $\mathcal{J}_1, \ldots, \mathcal{J}_{m(n,r)} \subseteq \mathcal{V}$ that satisfy the ordered or unordered pair condition for all variable pairs.

As will become clear from the following discussion, these problems are related. Note that, depending on the additional constraints, these problems may not have solutions (for example, for the unordered case of Problem 1(a) and 1(b), when K is smaller than the bound given in Equation 1), or they may trivially reduce to the problems of the previous sections because the additional constraints are irrelevant (for example, Problem 2 reduces for the unordered case to the problem discussed in Section 4.1 if $r \geq n/2$). As in the previous section, we separate the discussion of the results into those pertaining to the *unordered* pair condition (Section 5.1) and those pertaining to the *ordered* pair condition (Section 5.2). The algorithms presented here can also be used to construct intervention sets that satisfy the bounds discussed in Section 4.

5.1 Limiting the Intervention Set Size for the Unordered Pair Condition

We start with the simplest problem, Problem 1(a) for the *unordered* case, and give the construction of a set of experiments that achieves the smallest possible *average* intervention set size, given the number of variables n and the number of experiments K. The construction we present here is closely related to the first procedures we are aware of, given by Katona (1966). Finding a design which minimizes the average intervention set size is straightforward once one considers the problem in terms of the index sets. The sum of intervention set sizes is of course equal to the sum of index set sizes (because both represent the total number of interventions):

$$\sum_{k=1}^{K}|\mathcal{J}_k| \;=\; \sum_{i=1}^{n}|\mathcal{I}_i|. \tag{4}$$

This identity, together with Lemma 8, implies that to obtain intervention sets with minimum average size, it is sufficient to find the n smallest distinct subsets of $\{1, \ldots, K\}$ as the index sets. There are a total of $\sum_{j=0}^{p}\binom{K}{j}$ index sets $\mathcal{I}_i$ with $|\mathcal{I}_i| \leq p$. Consequently, the size l of

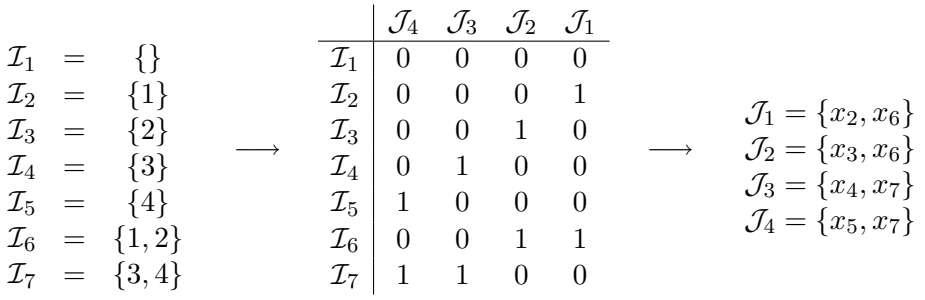

Figure 7: Intervention sets for experiments that satisfy the *unordered* pair condition for all pairs of variables in a 7 variable model such that both the maximum and average intervention set size are minimized. The index sets were chosen using Algorithm 2.

the largest required index set is the integer solution to the inequalities

$$\sum_{j=0}^{l-1} \binom{K}{j} \;<\; n \;\leq\; \sum_{j=0}^{l} \binom{K}{j}. \tag{5}$$

If there is no solution for l, the unordered pair condition cannot be satisfied with K experiments. If there is a solution, the inequalities in Equation 5 imply that when choosing the smallest index sets, we have to select *all*

$$t \;=\; \sum_{j=0}^{l-1} \binom{K}{j} \tag{6}$$

index sets of sizes 0 to $l-1$, and the remaining $n-t$ sets of size l. Since the sum of the intervention set sizes is the same as the sum of the index set sizes (Equation 4), the average intervention set size obtained is

$$\text{mean}_{k=1}^{K} |\mathcal{J}_k| \;=\; \frac{1}{K}\sum_{k=1}^{K} |\mathcal{J}_k| \;=\; \frac{1}{K}\sum_{i=1}^{n} |\mathcal{I}_i| \;=\; \frac{1}{K}\left[\sum_{j=0}^{l-1} j\binom{K}{j} + l(n-t)\right]. \tag{7}$$

For K experiments this is the minimum average intervention set size possible that satisfies the unordered pair condition for all pairs among n variables. For the case of $n = 7$ variables and $K = 4$ experiments, Figure 7 provides an example of the construction of intervention sets with minimal average size. Note that the minimum average would not have been affected if the index sets $\mathcal{I}_6$ and $\mathcal{I}_7$ had been chosen differently (but with the same size) as long as all the index sets remained distinct.

The minimum average intervention set size in Equation 7 also gives a lower bound for the lowest possible *maximum* intervention set size, given the number of experiments K and

Algorithm 1 Selects p index sets of size l, for K experiments, such that the indexes are distributed *fairly* among the index sets. The idea of this algorithm appears in a proof in Cameron (1994, accredited to D. Billington).

Fair(K,l,p)

Draw the index sets $\{\mathcal{I}_1, \ldots, \mathcal{I}_p\}$ as distinct l-size subsets of $\{1, \ldots, K\}$.

While TRUE,

Find the most frequent index M and the least frequent index m among the index sets $\{\mathcal{I}_1, \ldots, \mathcal{I}_p\}$.

If $\text{freq}(M) - \text{freq}(m) \leq 1$ then exit the loop.

Find[10] a set $\mathcal{A}$ of size $l-1$ such that $(\{M\} \cup \mathcal{A}) \in \{\mathcal{I}_1, \ldots, \mathcal{I}_p\}$ and $(\{m\} \cup \mathcal{A}) \notin \{\mathcal{I}_1, \ldots, \mathcal{I}_p\}$.

Replace the index set $(\{M\} \cup \mathcal{A})$ with $(\{m\} \cup \mathcal{A})$ in $\{\mathcal{I}_1, \ldots, \mathcal{I}_p\}$.

Return the index sets $\{\mathcal{I}_1, \ldots, \mathcal{I}_p\}$.

variables n (Problem 1(b)): At least one intervention set of size $\lceil \text{mean}_{k=1}^{K} |\mathcal{J}_k| \rceil$ or larger is needed, because otherwise the intervention sets would yield a lower average. Next, we will show that this is also an upper bound.

The size of an arbitrary intervention set $\mathcal{J}_k$ is equal to the number of index sets that contain the index k. We say that index sets are selected *fairly* when the corresponding intervention sets satisfy

$$|\mathcal{J}_k| - |\mathcal{J}_{k'}| \quad \leq \quad 1 \quad \forall k, k'. \tag{8}$$

In the construction that minimizes the average intervention set size, the index sets $\mathcal{I}_1, \ldots, \mathcal{I}_t$ constitute *all* possible subsets of $\{1, \ldots, K\}$ of size $l-1$ or less, and consequently all indexes appear equally often in these sets.[9] The remaining $n - t$ index sets can be chosen *fairly* using Algorithm 1. It finds fair index sets by simply switching the sets until the experiment indexes appear fairly. Since (i) the average intervention set size remains unchanged by this switching, (ii) the minimum average constitutes a lower bound, and (iii) the intervention set sizes differ by at most one, it follows (see Appendix A) that the lowest maximum intervention set size is given by

$$\max_{k=1}^{K} |\mathcal{J}_k| \quad = \quad \lceil \text{mean}_{k=1}^{K} |\mathcal{J}_k| \rceil. \tag{9}$$

Thus, if the construction of index sets is *fair*, then *both* the minimum average and the smallest maximum intervention set size is achieved, simultaneously solving both Problem 1(a) and 1(b). Algorithm 2 provides this complete procedure. Note that in general it is possible to select index sets such that the maximum intervention set size is not minimized, even though the average is minimal (for example, if $\mathcal{I}_7$ had been $\{1, 4\}$ in Figure 7). Figure 8 (top) shows the lowest possible average intervention set sizes. Rounding these figures up

9. This is easily seen if the index sets are represented as binary numbers, as in the center of Figure 7. It is also clear from considerations of symmetry.

Algorithm 2 Constructs K intervention sets that satisfy the *unordered* pair condition for all pairs among n variables with a minimum average and smallest maximum intervention set size.

FairUnordered(n, K)

Determine the maximum index set size l from Equation 5, if no such l exists, then the unordered pair condition cannot be satisfied for n variables with K experiments.

Assign all subsets $\mathcal{S} \subseteq \{1, \ldots, K\}$ such that $|\mathcal{S}| \leq l-1$ to index sets $\mathcal{I}_1, \ldots, \mathcal{I}_t$.

Draw the remaining l-size index sets with: $\mathcal{I}_{t+1}, \ldots, \mathcal{I}_n \leftarrow$ Fair(K, l, $n-t$)

Return the intervention sets $\mathcal{J}_1, \ldots, \mathcal{J}_K$ corresponding to the index sets $\mathcal{I}_1, \ldots, \mathcal{I}_n$.

to the closest integer gives the lowest possible maximum intervention set sizes. All of these numbers are achieved by intervention sets constructed using Algorithm 2.

Using constructions similar to the above, Katona (1966) and Wegener (1979) were able to derive the following bounds for Problem 2, the minimum number of experiments $m(n, r)$ given an upper bound r on the intervention set sizes:

$$m(n,r) = \lceil \log_2 n \rceil, \quad \text{if } r > n/2, \tag{10}$$

$$\frac{\log_2 n}{\log_2(e \cdot n/r)} \frac{n}{r} \leq m(n,r) \leq \left\lceil \frac{\log_2 n}{\log_2 \lceil n/r \rceil} \right\rceil (\lceil n/r \rceil - 1), \quad \text{if } r \leq n/2, \tag{11}$$

where e denotes Euler's number. Equation 10 just restates Equation 1, since the constructions in Section 4.1 without a constraint on the maximum intervention set size result in intervention sets with no more than $n/2$ variables. For any practical values of n and r, the value of $m(n, r)$ can be found by simply evaluating Equations 5, 7 and 9 for different values of K (starting from the lower bound in 11), so as to find the smallest value of K for which the maximum intervention set size is smaller than or equal to r. Figure 8 (bottom) illustrates the behavior of the function $m(n, r)$ and the bounds given by Equation 11.

5.2 Limiting the Intervention Set Size for the Ordered Pair Condition

Recall from Lemma 10 that satisfaction of the *ordered* pair condition requires that the n *index sets* of a set of K experiments form an antichain over $\{1, \ldots, K\}$. Thus, no matter whether we seek to minimize the average or the maximum intervention set size, we have to ensure that the index sets form an antichain. We begin by considering the (now *ordered* versions of) Problems 1(a) and 1(b): Given n and K, we want to specify experiments minimizing either the average or maximum intervention set size.

First, we note that to obtain an antichain with n elements from K experiments, at least one of the index sets must be of cardinality l or larger, where l is chosen to satisfy

$$\binom{K}{l-1} < n \leq \binom{K}{l}. \tag{12}$$

10. We can always find such a set $\mathcal{A}$: If there were no such set $\mathcal{A}$, then for all sets $\mathcal{B}$ of size $l-1$ such that $(\{M\} \cup \mathcal{B}) \in \{I_1, \ldots, I_p\}$ we would also have that $(\{m\} \cup \mathcal{B}) \in \{I_1, \ldots, I_p\}$. But then, $\text{freq}(M) - \text{freq}(m) \leq 0$ and the algorithm would have exited already on the previous line.

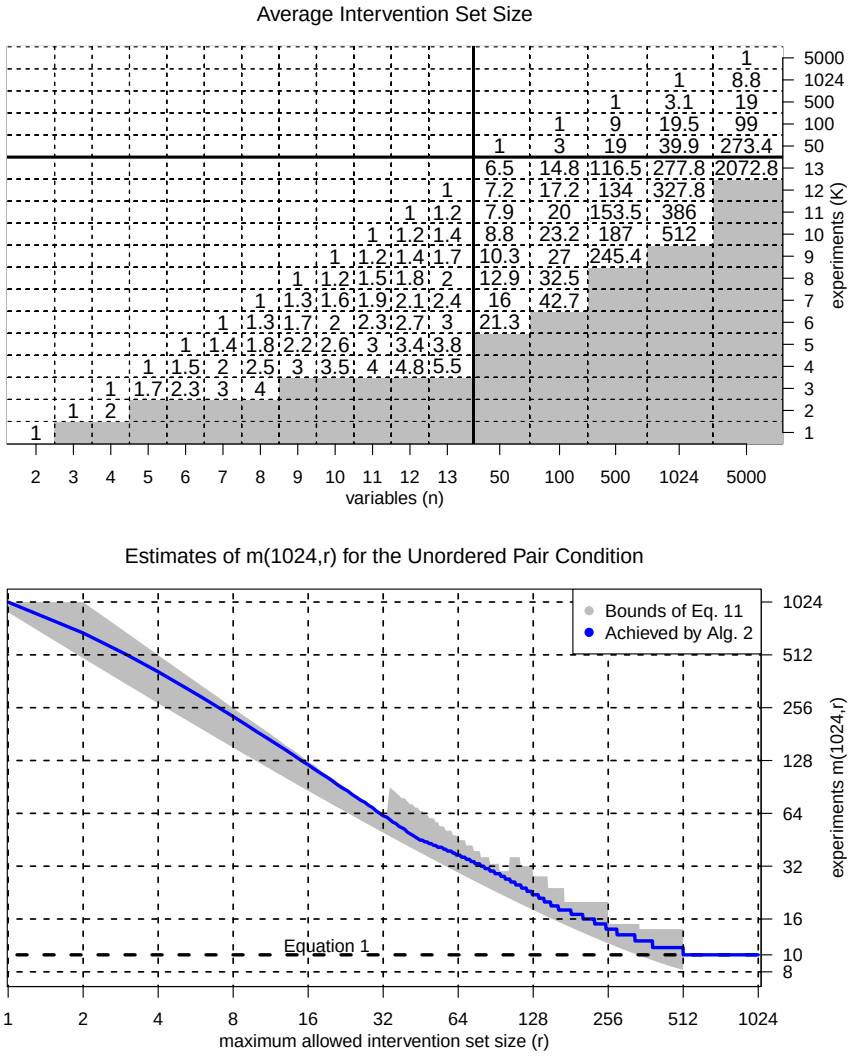

Figure 8: Satisfying the unordered pair condition while limiting the intervention set sizes. *Top:* Lowest achievable average intervention set sizes for models with n variables using K experiments. The lowest achievable maximum intervention set size is the ceiling of the average intervention set size shown in the figure. Grey areas denote an insufficient number of experiments to satisfy the unordered pair condition. Blank areas are uninteresting, since the average intervention set size can be lowered here by including irrelevant passive observational (*null*-)experiments. *Bottom:* The number of experiments needed for $n = 1024$ variables, with a limit r on the maximum allowed intervention set size.

Algorithm 3 Constructs K intervention sets that satisfy the *ordered* pair condition for all ordered pairs among n variables and approximates (and sometimes achieves) the minimum average and smallest maximum intervention size.

FairOrdered(n, K)

Determine the maximum index set size l from Equation 12, if no such l exists, then the ordered pair condition cannot be satisfied for n variables with K experiments.

Draw the l-size index sets with: $\mathcal{I}_1, \ldots, \mathcal{I}_n \leftarrow$ Fair(K, l, n)

Return the intervention sets $\mathcal{J}_1, \ldots, \mathcal{J}_K$ corresponding to the index sets $\mathcal{I}_1, \ldots, \mathcal{I}_n$.

This must be the case because it can be shown using the Lemmas in the proof of Theorem 11 (see p. 546 bottom in Sperner (1928)) that the largest antichain with sets of at most size $l-1$ has size $\binom{K}{l-1}$, which – given how l was constructed in Equation 12 – is not enough to accommodate all n index sets. On the other hand, it is equally clear that it *is* possible to obtain an antichain by selecting the n index sets to all have sizes l.

Thus, a simple approach to attempt to minimize the *maximum* intervention set size (that is, solve Problem 1 (b)) is to select the n index sets all with sizes l and use Algorithm 3, exploiting Algorithm 1, to construct a *fair* set of index sets. This construction is not fully optimal in all cases because all sets are chosen with size l while in some cases a smaller maximum intervention set size is achievable by combining index sets of different sizes. It is easily seen that Algorithm 3 will generate sets of experiments that have an average and a maximum intervention set size of

$$\text{mean}_{k=1}^{K}|\mathcal{J}_k| = \frac{1}{K}\sum_{k=1}^{K}|\mathcal{J}_k| = \frac{1}{K}\sum_{i=1}^{n}|\mathcal{I}_i| = \frac{n\cdot l}{K}, \tag{13}$$

$$\max_{k=1}^{K}|\mathcal{J}_k| = \left\lceil \text{mean}_{k=1}^{K}|\mathcal{J}_k| \right\rceil = \left\lceil \frac{n\cdot l}{K} \right\rceil. \tag{14}$$

Figure 9 (top) shows the maximum intervention set size in the output of Algorithm 3 for several values of n and K. Given some of the subsequent results it can also be shown that some of these are guaranteed to be optimal.[11] While this scheme for solving the directed case of Problem 1(b) is quite good in practice, we are not aware of any efficient scheme that is always guaranteed to minimize the maximum intervention set size.

Alternatively, one may focus on Problem 1 (a) and thus attempt to minimize the *average* intervention set size. For this problem, there exists an efficient procedure that obtains the optimum. Griggs et al. (2012) have recently provided some results in this direction. Here we apply their findings to the problem of experiment selection. To present the construction we need to start with some results pertaining to antichains.

An antichain is said to be *flat* if for any pair of sets $\mathcal{I}_i$ and $\mathcal{I}_j$ in the antichain, the cardinalities satisfy

$$|\mathcal{I}_i| - |\mathcal{I}_j| \leq 1, \quad \forall i, j. \tag{15}$$

11. But, for example, when $n = 30$ and $K = 8$, Algorithm 3 results in a maximum intervention set size of $\lceil \frac{30\cdot 3}{8} \rceil = 12$, although for this case Algorithm 4 can be used to construct suitable intervention sets with at most 11 members.

Note that *flatness* requires a selection of index sets that are *themselves* close in size, while *fairness* (8) requires a selection of index sets such that the *intervention sets* are close in size. Using this notion of flatness, Lieby (1994) originally formulated the following theorem as a conjecture:

Theorem 12 (Flat Antichain Theorem) *For every antichain there exists a flat antichain with the same size and the same average set size.*[12]

Since the sum of the index set sizes is identical to the sum of the intervention set sizes, Theorem 12 shows that whenever the ordered pair condition can be satisfied, the average intervention set size can be minimized by a set of flat index sets. From Equation 12 it is thus clear that an antichain minimizing the average set size can be selected solely from the sets of sizes $l-1$ and l. The question then becomes, how can we choose as many sets as possible of size $l-1$, thus minimizing the number of sets needed of size l, nevertheless obtaining a valid antichain?

From the Kruskal-Katona Theorem (Kruskal, 1963; Katona, 1968) it follows that an optimal solution can be obtained by choosing the *first* p index sets of size l and the *last* $n-p$ index sets of size $l-1$ from the *colexicographical order* of each of these sets (separately), defined below.

Definition 13 (Colexicographical order) *The* colexicographical order *over two sets A and B, where $|A| = |B|$, is defined by $A < B$ if and only if $\exists\, i : (A[i] < B[i]$ and $\forall j > i : A[j] = B[j])$, where $A[i]$ denotes the i:th element of the set A, when the elements of the set are arranged in numerical order.*

For example, comparing the sets $A = \{2, 3, 6\}$ and $B = \{1, 4, 6\}$ (note that they are already written in numerical order), we obtain $A < B$ because $A[2] = 3$ which is less than $B[2] = 4$, while $A[3] = B[3] = 6$. (See Figure 10 for a further illustration.)

Furthermore, the theory also allows for easily computing the smallest p (and hence largest $n-p$) for which a valid antichain is obtained: Any choice for p can be written in a unique l-cascade form

$$p \;=\; \sum_{j=1}^{l} \binom{a_j}{j}, \tag{16}$$

where the integers $a_1, \ldots, a_l$ can be computed using a simple greedy approach (for details see Jukna (2011, p. 146-8) and the code package accompanying this paper). Then, as discussed by Jukna, the number q of sets of size $l-1$ that are not subsets of the first p sets in the appropriate colexicographical order is given by

$$q \;=\; \binom{K}{l-1} - \sum_{j=1}^{l} \binom{a_j}{j-1}. \tag{17}$$

Thus, if one can pick the smallest p such that $p + q \geq n$, then it is possible to construct a flat antichain of size n that maximizes the number of index sets with size $l-1$.

12. Partial proofs follow from the work of Kleitman and Milner (1973), Lieby (1999) and Roberts (1999), and quite recently, Kisvölcsey (2006) was able to provide the full proof.

Algorithm 4 Obtains a set of K intervention sets satisfying the *ordered* pair condition for all ordered pairs among n variables that minimizes the average intervention set size.

Flat(n, K)

Determine the maximum index set size l from Equation 12, if no such l exists, the ordered pair condition cannot be satisfied for n variables with K experiments.

For p from 1 to n,

Find coefficients $a_1, \ldots, a_l$ for a cascade presentation of p in Equation 16.

Calculate the number q of available index sets of size $l-1$ by Equation 17.

If $p + q \geq n$ exit the for-loop.

Choose the index sets $\mathcal{I}_1, \ldots, \mathcal{I}_p$ as the *first* sets of size l in the colexicographical order.

Choose the index sets $\mathcal{I}_{p+1}, \ldots, \mathcal{I}_n$ as the *last* sets of size $l-1$ in the colexicographical order.

Return the intervention sets $\mathcal{J}_1, \ldots, \mathcal{J}_K$ corresponding to the index sets $\mathcal{I}_1, \ldots, \mathcal{I}_n$.

Algorithm 4 thus considers all values of p starting from 1, until Equations 16 and 17 imply that there is a flat antichain of size at least n. It then selects the *first* p index sets in the colexicographical order of sets of size l, and the *last* $n-p$ sets in the colexicographical order of sets of size $l-1$. The Kruskal-Katona Theorem ensures that the chosen $(l-1)$-sized sets will not be subsets of the chosen l-sized sets, thereby guaranteeing the antichain property (Jukna, 2011, p. 146-8). See Figure 10 for an example. Thus, Algorithm 4 returns a set of index sets that minimize the average intervention set size, solving (the directed version of) Problem 1 (a). Figure 9 (middle) shows the optimal average sizes for various values of n and K.

Trivially, the ceiling of the minimum average intervention set size for n variables in K experiments gives a lower bound on the lowest maximum intervention set size, that is, for Problem 1 (b). This allows us to determine the optimality of some of the outputs of Algorithm 3 in Figure 9 (top).

Problem 2 reverses the free parameters and asks for the minimum number of experiments $m(n, r)$ given a limit r on the maximum size of any intervention set. Cai (1984b) shows that

$$m(n, r) \quad = \quad \left\lceil \frac{2n}{r} \right\rceil, \quad \text{if} \quad 2 \leq \frac{1}{2} r^2 < n. \tag{18}$$

With input $K = \lceil 2n/r \rceil$, Algorithm 3 generates intervention sets of at most size r (see Appendix B) – this verifies that $m(n, r) \leq \lceil 2n/r \rceil$. Cai's result also gives an exact minimum number of experiments when the maximum intervention set size has to be small (see Figure 9 (bottom)). It can also be used to construct a lower bound on the maximum intervention set size when the number of experiments K is given: If $m(n, r) > K$ for some r and K, then the maximum intervention set size given n variables and K experiments must be at least $r + 1$. Again, we use this connection to determine the optimality of some of the outputs of Algorithm 3 in Figure 9 (top).

For cases when r does not satisfy the restrictions of Cai's result, Kündgen et al. (2001) use a similar construction to provide the following bounds:[13]

$$\min\{K|n \leq \binom{K}{\lceil Kr/n \rceil}\} \;\leq\; m(n,r) \;\leq\; \min\{K|n \leq \binom{K}{\lfloor Kr/n \rfloor}\}, \quad \text{if} \quad r \leq \frac{n}{2}. \tag{19}$$

Again the upper bound can be easily verified: With the upper bound K as input, Algorithm 3 will generate intervention sets of at most size r (see Appendix C). The lower bound is an application of classic results of Kleitman and Milner (1973) concerning average index set sizes. In many cases we can get an improved lower bound on $m(n,r)$ using Algorithm 4 (which optimally minimizes the average number of interventions per experiment, for given n and K): Find the smallest K such that Algorithm 4 returns intervention sets with an average size less than r. In this case we know that the minimum number of experiments given a maximum intervention set size of r must be at least K (see Figure 9 (bottom)).

Finally, note that Ramsay and Roberts (1996) and Ramsay et al. (1998) have considered the problem equivalent to finding a set of experiments where instead of a limited maximum intervention set size, the intervention sets are constrained to have *exactly* some given size. Sometimes more experiments are needed in order to satisfy this harder constraint.

6. Background Knowledge

The results and procedures of the previous two sections apply to scenarios in which there is no background knowledge concerning the possible causal relationships among the variables. In those cases, for complete identification, the chosen experiments must satisfy the (ordered or unordered, depending on the assumptions) pair condition for all pairs of variables. However, in many cases there exists domain knowledge that can assist in inferring the underlying system. For instance, it may be the case that background knowledge rules out certain causal relationships. Hyttinen et al. (2010) gave details of how, in the linear case, such prior knowledge can be integrated into the discovery procedure, reducing the set of ordered pair conditions that need to be satisfied by the experiments. As another example, under causal sufficiency, acyclicity, and faithfulness, if by prior knowledge a given edge is known to be present, or it is known that there is no causal relation between a given pair of variables, this translates directly to an unordered pair that does *not* need to be satisfied by the set of experiments. Finally, any background knowledge that is equivalent to knowledge of the outcome of some experiment can be described in terms of satisfied pair conditions. In this section, we thus consider the selection of experiments when the experiments only need to satisfy the pair condition for a given subset of all variable pairs, but acknowledge that not all background knowledge is representable in terms of satisfied pair conditions.

When the pair condition only needs to be satisfied for a subset of the variable pairs, the search problem is equivalent to that of finding a minimal cut-covering of a given graph. As described in Section 3, we represent the satisfaction of the *unordered* pair condition by a graph H over the vertices $\mathcal{V}$, where an undirected edge between a pair of variables indicates that the unordered pair condition is *not* yet satisfied for that pair (and hence needs to be satisfied by the experiments we select), while the absence of an edge indicates it is already satisfied for the pair (and does not need to be satisfied by our experiments). Similarly,

13. Roberts and Rylands (2005) give the exact values for $m(n,r)$ for $n \leq 10$ variables and all suitable r.

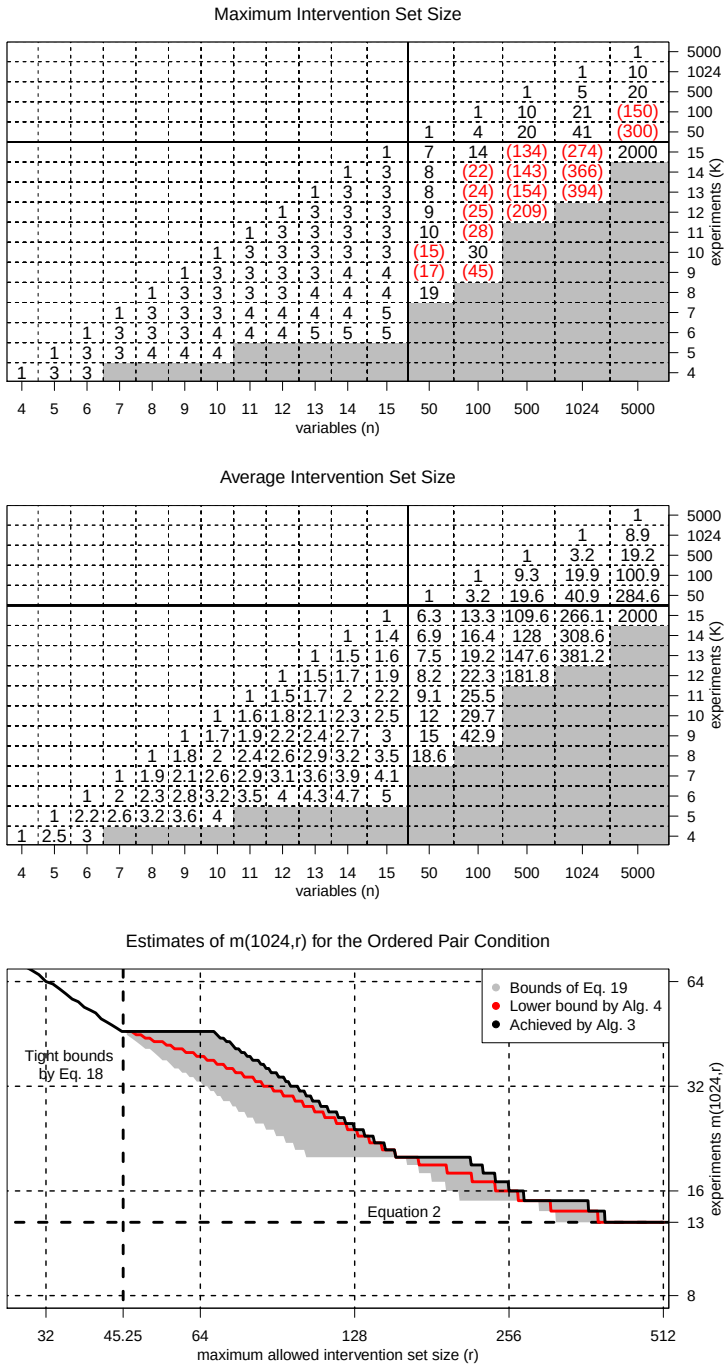

Figure 9: Satisfying the ordered pair condition while limiting the size of the intervention sets. *Top:* Maximum intervention set sizes achieved by Algorithm 3. Black numbers mark the cases where the achieved maximum intervention set size is known to be optimal, while red numbers in parentheses mark cases that are not known to be optimal. *Middle:* Average intervention set sizes achieved by Algorithm 4, all guaranteed to be optimal. *Bottom:* Number of experiments needed to satisfy the ordered pair condition for $n = 1024$ variables with a limit r on the maximum intervention set size.

$$
\begin{array}{ccccccccccc}
\mathbf{\{1,2,3\}} & < & \mathbf{\{1,2,4\}} & < & \mathbf{\{1,3,4\}} & < & \mathbf{\{2,3,4\}} & < & \mathbf{\{1,2,5\}} & < & \mathbf{\{1,3,5\}} \\
 & & & & & & & & & & \wedge \\
\mathbf{\{1,3,6\}} & > & \mathbf{\{1,2,6\}} & > & \mathbf{\{3,4,5\}} & > & \mathbf{\{2,4,5\}} & > & \mathbf{\{1,4,5\}} & > & \mathbf{\{2,3,5\}} \\
\wedge & & & & & & & & & & \\
\mathbf{\{2,3,6\}} & < & \mathbf{\{1,4,6\}} & < & \mathbf{\{2,4,6\}} & < & \boxed{\mathbf{\{3,4,6\}}} & < & \{1,5,6\} & < & \cdots
\end{array}
$$

$$
\begin{array}{ccccccccccc}
\cdots & < & \{3,6\} & < & \{4,6\} & < & \boxed{\mathbf{\{5,6\}}} & < & \mathbf{\{1,7\}} & < & \mathbf{\{2,7\}} \\
 & & & & & & & & & & \wedge \\
\mathbf{\{2,8\}} & > & \mathbf{\{1,8\}} & > & \mathbf{\{6,7\}} & > & \mathbf{\{5,7\}} & > & \mathbf{\{4,7\}} & > & \mathbf{\{3,7\}} \\
\wedge & & & & & & & & & & \\
\mathbf{\{3,8\}} & < & \mathbf{\{4,8\}} & < & \mathbf{\{5,8\}} & < & \mathbf{\{6,8\}} & < & \mathbf{\{7,8\}} & &
\end{array}
$$

Figure 10: Selecting the index sets in colexicographical order for $n = 30$ and $K = 8$. Selecting 16 index sets of size 3 (up to $\{3,4,6\}$, in bold) and 14 index sets of size 2 (starting from $\{5,6\}$, in bold), gives a total of 30 index sets and achieves the lowest possible average intervention set size for the given n and K. Note that none of the selected index sets is a subset of another, thus the sets form an antichain. If we were to select only 15 index set of size 3 (up to $\{2,4,6\}$), we could still only select 14 index sets of size 2 (from $\{5,6\}$), ending up with only 29 index sets. If we were to select 17 index sets of size 3 (up to $\{1,5,6\}$), we could select 13 index sets of size 2 (from $\{1,7\}$), and find 30 index sets, but the average intervention set size would then be $1/8$ higher.

we use a *directed* graph F to represent the satisfaction of the *ordered* pair condition, in the analogous way. Essentially, the combinatorial problems discussed in the two previous sections can thus be interpreted as finding a minimal cut-covering for a *complete* directed or undirected graph, while in this section we consider the problem of finding a minimal cut-covering for an *arbitrary* directed or undirected graph.[14]

First, consider the satisfaction of the unordered pair condition for an arbitrary subset of all variable pairs. Unlike the case without background knowledge, discussed in Sections 4.1, the problem of finding the smallest set of experiments to satisfy the unordered pair condition for a subset of all pairs is known to be hard. Cai (1984a) establishes the connection to minimal graph colorings by showing (in his Theorem 5) that the smallest cardinality $c(H)$ of a cut-covering of an undirected graph H relates to its chromatic number $\chi(H)$ (the smallest number of colors required to vertex-color graph H) as

$$c(H) \quad = \quad \lceil \log_2(\chi(H)) \rceil. \tag{20}$$

The result indicates that the main constraint to reducing the number of experiments are cliques of variables for which the unordered pair condition is not satisfied. Equation 20 constitutes a generalization of the results shown in Section 4.1. Furthermore, it follows from Cai's Theorem 6 that the problem of finding a minimal set of experiments given background knowledge for arbitrary pairs is NP-hard, though constructing the appropriate experiments

14. Note that Cai (1984a) uses the terminology of 'separating systems' in relation to arbitrary graphs, but this use of the terminology does not seem to be in widespread use so we do not adopt it here.

Algorithm 5 Constructs a set of intervention sets satisfying the *unordered* pair condition for a given *arbitrary* set of unordered pairs represented by an undirected graph H over n variables. The algorithm is adapted from the proof of Theorem 5 in Cai (1984a).

BackgroundUnordered(H)

Obtain a partition $\mathrm{Col}(H)$ of the variables $\mathcal{V}$ of H into q color classes $C_1, \ldots, C_q$ (where $\chi(H) \leq q \leq n$, $\chi(H)$ is the chromatic number of H and $n = |\mathcal{V}|$), such that no adjacent vertices belong to the same class (for example, using the approximation algorithm in Welsh and Powell (1967)).

Let $K = \lceil \log_2(q) \rceil$.

Obtain "intervention sets" $\mathcal{J}'_1, \ldots, \mathcal{J}'_K$ over the color classes $C_1, \ldots, C_q$ that satisfy the unordered pair condition for all pairs of the q color classes (for example, by calling FairUnordered(q,K) (Algorithm 2)).

For each k from 1 to K,

Determine $\mathcal{J}_k = \bigcup_{C_i \in \mathcal{J}'_k} C_i$. Intervention set $\mathcal{J}_k$ consists of all variables colored with a color in $\mathcal{J}'_k$.

Return intervention sets $\mathcal{J}_1, \ldots, \mathcal{J}_K$.

given a graph coloring is very simple. Various approximation algorithms used for graph coloring could be applied, see for example Welsh and Powell (1967); Motwani and Naor (1993); Halldórsson (1993); Bussieck (1994) and Liberti et al. (2011) for proposals, bounds and simulations. Algorithm 5 calls a graph coloring method (in the code package we use the simple approximation algorithm by Welsh and Powell (1967)) and constructs intervention sets based on the graph coloring. It results in $\lceil \log_2(\chi(H) + c) \rceil$ experiments, where c is the number of colors the coloring algorithm uses in excess of the chromatic number $\chi(H)$. So it achieves Cai's optimal bound on the minimum number of experiments (Equation 20) if the coloring method uses the smallest number of colors.

In certain restricted cases the problem is easier. When the underlying model is known to be acyclic and causally sufficient, and the background knowledge derives from passive observational data or suitable previous experiments, the knowledge can be represented in terms of an (interventional) Markov equivalence class. These are sets of causally sufficient acyclic causal models that are indistinguishable given passive observational data or the experiments performed so far. Algorithm 3 in Hauser and Bühlmann (2012) constructs in *polynomial* time a minimal set of experiments that is sufficient and in the worst case necessary to identify the true causal model within such an (interventional) Markov equivalence class, assuming the skeleton of the true causal graph is identifiable given the set of experiments. We can translate this situation to our framework as follows: With such background knowledge, for complete identifiability, the *un*ordered pair condition only needs to be satisfied for the pairs of variables *adjacent* in the skeleton, for which the orientation of the edge is unknown. Thus, if we again consider the example graph (i) in Figure 11 (repeating graph (i) from Figure 1), and assume that we have a passive observational data set to determine the Markov equivalence class (graph (ii)), then the undirected graph H representing the pairs for which the unordered pair condition remains *un*satisfied is given by graph (iii). (In this case it has the same structure as the skeleton of the Markov equivalence class of the true graph, but that is not generally the case.) Given H (graph (iii)), it is obvious that a single

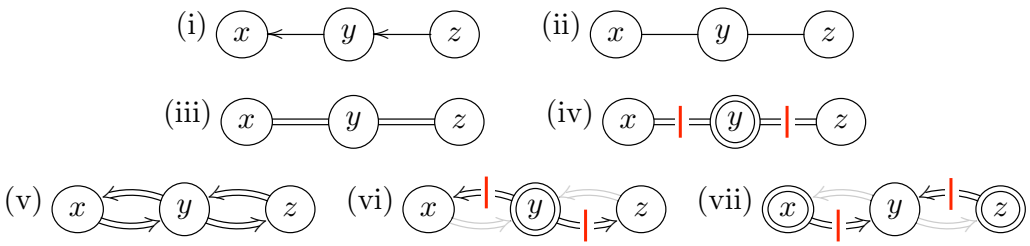

Figure 11: Graph (i) is the true causal generating model (repeating Figure 1, (i)), graph (ii) shows the corresponding passive observational Markov equivalence class (MEC). Graph (iii) illustrates the remaining pairs for which the *unordered* pair condition is not satisfied given the MEC in (ii), and graph (iv) shows that a single intervention on y resolves these pairs, that is, it provides a cut covering. Given background knowledge obtained from a passive observational data set of graph (i), graph (v) shows the ordered pairs for which the *ordered* pair condition remains unsatisfied. In this case two further experiments are required to provide a directed cut covering, one intervening on y (graph (vi)) and one intervening on x and z simultaneously (graph (vii)).

experiment $\mathcal{E} = (\mathcal{J}, \mathcal{U}) = (\{y\}, \{x, z\})$ would resolve the remaining pairs, that is, the single cut $\mathcal{E}$ is a minimal cut-covering (graph (iv)). Under the assumption of acyclicity, causal sufficiency and faithfulness, this is sufficient for identifiability of the causal structure.

Satisfaction of the *ordered* pair condition for an arbitrary subset of the ordered variable pairs is also known to be hard. Cai (1984a) shows in his Theorem 7 that the problem of determining a minimal cut-covering for an arbitrary directed graph is NP-hard. More recently, Watanabe et al. (2000) offered the following bounds on the cardinality of the minimal directed cut-covering $c(F)$

$$\log_2(\chi(F)) \quad \leq \quad c(F) \quad \leq \quad \lceil \log_2(\chi(F)) \rceil + \lceil \log_2 \lceil \log_2(\chi(F)) + 1 \rceil \rceil, \tag{21}$$

where $\chi(F)$ is the chromatic number of the directed graph F representing the unsatisfied *ordered* pair conditions.[15] These bounds constitute a generalization of the results in Section 4.2 where no background knowledge was assumed. In general, the conversion of background knowledge to pairs satisfying the *ordered* pair condition is more complicated because the equivalence classes of causal structures whose identifiability depends on the satisfaction of the ordered pair condition is less well understood (consider, for example, the equivalence classes for linear cyclic models over a causally insufficient set of variables). But for simple cases it can still be done: Given background knowledge derived from a passive observational data set over graph (i) in Figure 11, graph (v) is the directed graph F indicating the ordered pairs for which the ordered pair condition remains unsatisfied. Graphs (vi) and (vii) then show the two directed cuts that are still required to obtain a directed cut-covering of F. If we assume linearity, but not acyclicity or causal sufficiency, the two experiments

15. Strictly speaking $\chi(F)$ is the chromatic number of the undirected graph with the same set of vertices as F and in which an undirected edge exists between a pair of vertices if an only if they are adjacent in F.

corresponding to these cuts would be necessary and sufficient for identifiability given the background knowledge represented in graph (v).

7. Discussion: Related Work and Open Problems

The combinatorial results and procedures we have translated and combined in this paper update and generalize a variety of results in the causal discovery literature on the selection of experiments. For example, Eberhardt (2007, Theorem 3.3.17) only provides an upper bound on the minimum number of experiments sufficient for the satisfaction of the ordered pair condition. With Spencer's result on completely separating systems (Equation 2) we now have an exact result and the experiments that satisfy this result can be constructed using Algorithm 3. Similarly, Eberhardt (2007, Theorem 3.3.29) gave an upper bound on the minimum number of experiments sufficient to satisfy the unordered pair condition when the maximum intervention set size was restricted. The translation of the results of Katona (1966) and Wegener (1979) in Equations 10 and 11 now provide much better bounds, and Algorithm 2 can be used to construct the appropriate intervention sets.

Tong and Koller (2001) and Murphy (2001) use a greedy Bayesian procedure to select the next best single-intervention experiment, but given the computational complexity cannot solve for models with more than five variables. Meganck et al. (2005), He and Geng (2008), Eberhardt (2008) and Hauser and Bühlmann (2012), in their different ways, also try to find the best next experiment given background knowledge, typically knowledge of the (interventional) Markov equivalence class. In Section 6 we showed that the complexity of the satisfaction of the (un)ordered pair condition given background knowledge is known to be NP-hard in general, and Cai (1984a) already showed (though in the terminology of separating systems) that the minimum number of experiments required is a function of the chromatic number of the graph of unsatisfied pair conditions. Except for special cases (see, for example, Hauser and Bühlmann (2012)), graph-coloring heuristics seem to provide the best approach to problems of this type, and we have implemented one such procedure in Algorithm 5.

In addition, a variety of open problems remain: In Section 5.2 we noted that for the ordered pair condition we are not aware of a general algorithm that generates intervention sets for which the maximum intervention set size is minimized. We also do not know whether the maximum and average intervention set size can be minimized simultaneously (as we showed is possible for the unordered pair condition in Section 5.1). Nevertheless, for both cases we have shown that Algorithm 3 provides a very good approximation and can be shown to provide optimal output in many cases.

More generally, the type of background knowledge we considered in Section 6 may have to be integrated into a search procedure that is subject to constraints on the size of the intervention sets. How to compute the optimal set of experiments in such cases is an open combinatorical problem, for which we are not aware of any solutions that are not brute force searches.

Naturally, there are also further generalizations of the problem settings we considered. For example, it will often not be possible to perform all desired experiments. Some experiments will be more expensive than others or certain combinations of variables may not be manipulable simultaneously. Given a restricted set of experiments, how to select the

smallest subset that preserves the discriminatory power of the full set, is known as the "test collection problem". There is a large literature on variants of this problem. Halldórsson et al. (2001) describe the connection to finding minimal cut-coverings and analyze the complexity of the problem. They show that even good approximation algorithms are hard to obtain for this problem, but Moret and Shapiro (1985) analyze and test a variety of heuristics and show that in a real world setting there is reason for optimism that an (almost) minimal set of experiments can still be found relatively easily.

8. Conclusion

We have summarized and presented combinatorial results for the optimal selection of experiments when the goal is to learn the causal structure of a system. Most results were originally derived for so-called *(completely) separating systems* or *minimal cut-coverings*. We used these results to specify the minimum number of experiments necessary and sufficient for identifiability when there is no background knowledge (Section 4), when there are limitations on the size of the intervention sets (Section 5), and when background knowledge is available (Section 6). Where possible, we presented algorithms that actually construct the experiments that satisfy (or closely approximate) the specified bounds and constraints, and we indicated where extant heuristics can be applied. We hope that the constructive way of presenting the results and how to obtain them, may also provide useful guidelines on which experiments to conduct in settings with assumptions not considered here.

For the combinatorics community, we have provided a novel area of application. We have also given a unifying, more easily understandable framework for the set constructions (which are otherwise often hidden in the proofs of the reported bounds), along with clear examples and computer code. This should help in understanding and comparing the different bounds and constructions. Perhaps this compilation also helps in identifying where new theoretical findings would also have a practical value. We hope that this note provides a translation aid and helps to produce a more congenial flow of research problems and results between the fields of study.

Acknowledgements

The authors would like to thank M. Koivisto and P. Kaski for helpful discussions and three anonymous reviewers for their helpful comments that improved the article. A.H. & P.O.H. were supported by the Academy of Finland. F.E. was supported by a grant from the James S. McDonnell Foundation on 'Experimental Planning and the Unification of Causal Knowledge'.

Appendix A. Proof of Equation 9

Marking the largest intervention set by $\mathcal{J}_h$ we have that

$$\begin{aligned} |\mathcal{J}_h| - \text{mean}_{k=1}^{K}|\mathcal{J}_k| &= \frac{1}{K}\sum_{k=1}^{K}(|\mathcal{J}_h| - |\mathcal{J}_k|) \quad || \text{ Eq. } 8, h \in \{1,\ldots,K\} \\ &\leq \frac{K-1}{K} < 1. \end{aligned}$$

Since the maximum intervention set size is an integer which is lower bounded by the average intervention set size, yet less than one above it, Equation 9 follows directly.

Appendix B. Upper Bound of Cai (1984b)

We verify the upper bound in Equation 18: The minimum number of experiments $m(n, r)$, given a limit r on the maximum intervention set size, has an upper bound of $\lceil 2n/r \rceil$, when $n > \frac{1}{2}r^2$. Consider that Algorithm 3 is run with input n and $K = \lceil 2n/r \rceil$. The first step is to find l that satisfies Equation 12. The upper bound in Equation 12 is satisfied when $l = 2$ under the assumption that $n > \frac{1}{2}r^2$:

$$\begin{aligned}
\binom{K}{l} &= \frac{K(K-1)}{2} \quad ||K \geq \frac{2n}{r} \\
&\geq \frac{2n(K-1)}{2r} \quad ||K \geq \frac{2n}{r} > \frac{2\frac{1}{2}r^2}{r} = r \Rightarrow K - 1 \geq r \\
&\geq \frac{2nr}{2r} = n.
\end{aligned}$$

Thus, the used index set size l will be at most two, as there exists an antichain of n index sets of constant size $l \leq 2$ over $K = \lceil 2n/r \rceil$ experiments. Then, Algorithm 3 will produce intervention sets with average size (Equation 13) bounded by r:

$$\begin{aligned}
\text{mean}_{k=1}^{K}|\mathcal{J}_k| &= \frac{n \cdot l}{K} \quad ||\frac{1}{K} \leq \frac{r}{2n} \\
&\leq \frac{nlr}{2n} \quad ||l \leq 2 \\
&\leq \frac{2nr}{2n} = r.
\end{aligned}$$

Because the index sets are chosen fairly, the maximum intervention set size (Equation 14) is also bounded by integer r:

$$\max_{k=1}^{K}|\mathcal{J}_k| = \lceil \text{mean}_{k=1}^{K}|\mathcal{J}_k| \rceil \leq r.$$

Appendix C. Upper Bound of Kündgen et al. (2001)

We verify the upper bound in Equation 19: The minimum number of experiments $m(n, r)$, given a limit r on the maximum intervention set size, has an upper bound of $\min\{K' | n \leq \binom{K'}{\lfloor K'r/n \rfloor}\}$, when $r \leq n/2$. Consider that Algorithm 3 is run with input n and $K = \min\{K' | n \leq \binom{K'}{\lfloor K'r/n \rfloor}\}$. The first step is to find l that satisfies Equation 12. The upper bound in Equation 12 is satisfied when $l = \lfloor Kr/n \rfloor$, simply by the definition of K. Thus, the used index set size l will be at most $\lfloor Kr/n \rfloor$, as there exists an antichain of n index sets of constant size $l \leq Kr/n$ over K experiments. Then, Algorithm 3 will produce intervention sets with average size (Equation 13) bounded by r:

$$\begin{aligned}
\text{mean}_{k=1}^{K}|\mathcal{J}_k| &= \frac{n \cdot l}{K} \quad ||l \leq \frac{Kr}{n} \\
&\leq \frac{nKr}{Kn} = r.
\end{aligned}$$

Because the index sets are chosen fairly, the maximum intervention set size (Equation 14) is also bounded by integer r:

$$\max_{k=1}^{K}|\mathcal{J}_k| \quad = \quad \lceil \text{mean}_{k=1}^{K}|\mathcal{J}_k| \rceil \quad \leq \quad r.$$

References

N. Alon, B. Bollobás, A. Gyárfás, J. Lehel, and A. Scott. Maximum directed cuts in acyclic digraphs. *Journal of Graph Theory*, 55(1):1–13, 2007.

E. Bareinboim and J. Pearl. In *Proceedings of the Twenty-Eight Conference on Uncertainty in Artificial Intelligence*, pages 113–120, 2012.

M. Bussieck. The minimal cut cover of a graph. Technical Report TR-94-02, Pennsylvania State University, 1994.

M.-C. Cai. On separating systems of graphs. *Discrete Mathematics*, 49(1):15 – 20, 1984a.

M.-C. Cai. On a problem of Katona on minimal completely separating systems with restrictions. *Discrete Mathematics*, 48(1):121 – 123, 1984b.

P. J. Cameron. *Combinatorics: Topics, Techniques, Algorithms*. Cambridge University Press, 1994.

T. Claassen and T. Heskes. Causal discovery in multiple models from different experiments. In *NIPS '10*, 2010.

G. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *UAI '99*, pages 116–125, 1999.

T. J. Dickson. On a problem concerning separating systems of a finite set. *Journal of Combinatorial Theory*, 7(3):191 – 196, 1969.

D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *AISTATS '07*, 2007.

F. Eberhardt. *Causation and Intervention*. PhD thesis, Carnegie Mellon, 2007.

F. Eberhardt. Almost optimal intervention sets for causal discovery. In *UAI '08*, 2008.

F. Eberhardt, C. Glymour, and R. Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *UAI '05*, pages 178–184, 2005.

F. Eberhardt, P. O. Hoyer, and R. Scheines. Combining experiments to discover linear cyclic models with latent variables. In *AISTATS '10*, 2010.

R. A. Fisher. *The design of experiments*. Hafner, 1935.

J. R. Griggs, Sven Hartman, Uwe Leck, and I. T. Roberts. Full and maximal squashed flat antichains of minimum weight. Technical report, (in preparation), 2012.

B. Halldórsson, M. Halldórsson, and R. Ravi. On the approximability of the minimum test collection problem. In Friedhelm auf der Heide, editor, *Algorithms ESA 2001*, volume 2161 of *Lecture Notes in Computer Science*, pages 158–169. Springer, 2001.

M. Halldórsson. A still better performance guarantee for approximate graph coloring. *Information Processing Letters*, 45(1):19 – 23, 1993.

A. Hauser and P. Bühlmann. Two optimal strategies for active learning of causal models from interventions. In *Proceedings of the The 6th European Workshop on Probabilistic Graphical Models*, 2012.

Y.-B. He and Z. Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9:2523–2547, 2008.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Causal discovery for linear cyclic models with latent variables. In *PGM '10*, 2010.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Noisy-or models with latent confounding. In *UAI '11*, 2011.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Causal discovery of linear cyclic models from multiple experimental data sets with overlapping variables. In *UAI '12*, 2012a.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13:3387–3439, 2012b.

S. Jukna. *Extremal Combinatorics*. Springer, 2011.

G. Katona. On separating systems of a finite set. *Journal of Combinatorial Theory*, 1(2): 174 – 194, 1966.

G. Katona. A theorem of finite sets. In P. Erdos and G. Katona, editors, *Theory of graphs*. Akademiai Kiado and Academic Press, 1968.

Á. Kisvölcsey. Flattening antichains. *Combinatorica*, 26:65–82, 2006.

D.J. Kleitman and E.C. Milner. On the average size of the sets in a Sperner family. *Discrete Mathematics*, 6(2):141 – 147, 1973.

J. B. Kruskal. The number of simplices in a complex. In R. Bellman, editor, *Mathematical Optimization Techniques*. University of California Press, 1963.

A. Kündgen, D. Mubayi, and P. Tetali. Minimal completely separating systems of k-sets. *Journal of Combinatorial Theory, Series A*, 93(1):192 – 198, 2001.

L. Liberti, L. Alfandari, and M.-C. Plateau. Edge cover by connected bipartite subgraphs. *Annals of Operations Research*, 188:307–329, 2011.

P. Lieby. The separation problem. Honours thesis, Northern Territory University, 1994.

P. Lieby. *Extremal Problems in Finite Sets*. PhD thesis, Northern Territory University, 1999.

R. Loulou. Minimal cut cover of a graph with an application to the testing of electronic boards. *Operations Research Letters*, 12(5):301 – 305, 1992.

S. Meganck, B. Manderick, and P. Leray. A decision theoretic approach to learning Bayesian networks. Technical report, Vrije Universiteit Brussels, 2005.

B. M. E. Moret and H. D. Shapiro. On minimizing a set of tests. *SIAM Journal on Scientific and Statistical Computing*, 6(4):983–1003, 1985.

R. Motwani and J. Naor. On exact and approximate cut covers of graphs. Technical report, Stanford University, 1993.

K. P. Murphy. Active learning of causal Bayes net structure. Technical report, U.C. Berkeley, 2001.

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

C. Ramsay and I. T. Roberts. Minimal completely separating systems of sets. *Australasian Journal of Combinatorics*, 13:129–150, 1996.

C. Ramsay, I. T. Roberts, and F. Ruskey. Completely separating systems of k-sets. *Discrete Mathematics*, 183:265–275, 1998.

A. Rényi. On random generating elements of a finite Boolean algebra. *Acta Sci. Math. (Szeged)*, 22:75–81, 1961.

I. T. Roberts. The flat antichain conjecture for small average set size. Technical report, Northern Territory University, 1999.

I. T. Roberts and Leanne J. Rylands. Minimal (n) and (n,h,k) completely separating systems. *Australasian Journal of Combinatorics*, 33:57 – 66, 2005.

I. Shpitser and J. Pearl. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444, 2006.

J. Spencer. Minimal completely separating systems. *Journal of Combinatorial Theory*, 8 (4):446 – 447, 1970.

E. Sperner. Ein Satz über Untermengen einer endlichen Menge. *Mathematische Zeitschrift*, 27:544–548, 1928.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, second edition in 2000, 1 edition, 1993.

S. Tong and D. Koller. Active learning for structure in Bayesian networks. In *UAI '01*, pages 863–869, 2001.

K. Watanabe, M. Sengoku, H. Tamura, K. Nakano, and S. Shinoda. A scheduling problem in multihop networks. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E83-A(6):1222 – 1227, 2000.

I. Wegener. On separating systems whose elements are sets of at most k elements. *Discrete Mathematics*, 28(2):219 – 222, 1979.

D. J. A. Welsh and M. B. Powell. An upper bound for the chromatic number of a graph and its application to timetabling problems. *The Computer Journal*, 10(1):85–86, 1967.