Leveraging Causal Graphs for Blocking in Randomized Experiments

Abhishek K. Umrawal AUMRAWAL@PURDUE.EDU
School of Industrial Engineering, Purdue University
Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County

Editors: Mihaela van der Schaar, Dominik Janzing and Cheng Zhang

Abstract

Randomized experiments are often performed to study the causal effects of interest. *Blocking* is a technique to precisely estimate the causal effects when the experimental material is not homogeneous. It involves stratifying the available experimental material based on the covariates causing non-homogeneity and then randomizing the treatment within those strata (known as *blocks*). This eliminates the unwanted effect of the covariates on the causal effects of interest. We investigate the problem of finding a *stable* set of covariates to be used to form blocks, that minimizes the variance of the causal effect estimates. Using the underlying causal graph, we provide an efficient algorithm to obtain such a set for a general *semi-Markovian* causal model.

Keywords: Causal graphs, randomized experiments, block designs, variance reduction.

1. Introduction

1.1. Motivation

Studying the *causal* effect of some variable(s) on the other variable(s) is of common interest in social sciences, computer science, and statistics. However, a mistake that people usually make is confusing the causal effect with an *associational* effect. For instance, if high levels of bad cholesterol and the presence of heart disease are observed at the same time, it does not mean that the heart disease is caused by high levels of bad cholesterol. The question is then how do we get to know if one variable causes the other? If the answer is yes, then what is the direction (positive or negative) and what is the magnitude, of the causal effect? Fisher (1926) provided the framework of *randomized experiments* to study the causal effect, where the variable whose causal effect is to be studied (known as *treatment* or cause), is randomized over the available experimental units (like humans, rats, agricultural plots, etc.) and changes in the variable on which the causal effect is to be studied (known as *response* or effect), are recorded. A statistical comparison of values of the response with or without the treatment can therefore be done to study the existence, direction, and magnitude of the cause-effect relationship of interest.

Randomized experiments work on three basic principles, viz., *randomization*, *replication*, and *local control* (Fisher, 1926). Randomization states that the assignment of the treatment has to be random, and replication states the treatment should be given to multiple but homogeneous units, i.e., there are multiple observations of the effect variable for both with and without the treatment. Therefore, as long as the available experimental units are homogeneous (for instance, the fertility of all the agricultural plots is the same, the responsiveness of all the humans is the same for the drug, etc.), then a 'good' randomized experiment can be carried out using the first two principles, viz., randomization and replication, which gives rise to *completely randomized design* (CRD). However, when the available experimental units are not homogeneous, i.e., some attributes of experimental

© 2023 A.K. Umrawal.

units (known as *covariates*) differ from each other, then the causal effect may get influenced by the covariates causing non-homogeneity (like fertility, responsiveness, etc). The remedy to this problem is the third principle of randomized experiments, i.e., local control (also known as *blocking*), which states to stratify the available experimental units based on that covariate(s) causing non-homogeneity, and then randomize the treatment within those strata to eliminate the effect of the covariates. These strata are called *blocks*, and this third principle along with the first two gives rise to *randomized block design* (RBD). Blocking tries to control/eliminate the variability in the response attributed through the covariates and leads to a more *precise* estimation of the causal effect of interest. Precision is defined as the inverse of the variance of the estimate of the causal effect of interest.

In this paper, we focus on the problem of deciding which covariates to be used for forming the blocks while performing a randomized experiment. We consider a non-parametric setting and assume that we have access to the underlying causal structure (Pearl, 2009). We provide an efficient algorithm to obtain a *stable* set of covariates to be used to form blocks for a general semi-Markovian causal model. The term stable refers to the idea of finding and excluding the covariates that one should never use to create blocks. The notion and the method of achieving this stability are discussed in Section 3.4.

1.2. Literature Review

Statistics literature (Yates, 1935; Kempthorne, 1952, 1955; Cochran and Cox, 1948) presents profound discussions on relative efficiency (ratio of the variances of causal effects) of RBD over CRD. It can be seen that in general, RBD is always more or at least as efficient as CRD. Therefore, the natural question is how do we do blocking in an intelligent manner such that maximum gain in precision can be attained? More specifically, we are interested in answering the following question. Can we decide which covariates to be used for forming blocks using the causal structure (diagram) of all the variables, viz., treatment, response, and covariates (observed and unobserved) provided the causal structure is given?

In the past, efforts have been made to address the problem of efficient adjustments in causal graphical models under static interventions (Henckel et al., 2019; Rotnitzky and Smucler, 2020) and under dynamic interventions (Smucler et al., 2021). In contrast to these, we provide an efficient algorithm to obtain the set of covariates to be used for forming blocks for a general semi-Markovian causal model. citetcinelli2020crash provided a good discussion on good and bad controls for different causal graphs and emphasized the importance of good adjustments in causal estimation. Some other partially related works include (Kallus, 2018) which discusses the problem of experimental design in terms of variance reduction and describes the functional assumptions placed on the potential outcome by blocking, (Harshaw et al., 2019) which shows that discrepancy minimization (difference in means between treatment and control covariate groups) implicitly performs ridge regression and (Li and Ding, 2019) talks about re-randomization and regression adjustment.

In this paper, we focus on deciding what covariates should form the basis of forming blocks. However, a problem that follows our problem is, how to efficiently form blocks by grouping the experimental units which are close to each other based on the chosen covariates. Moore (2012); Higgins et al. (2016) focus on the latter problem.

1.3. Contribution

In this paper, we formalize the problem of finding a *stable* set of covariates to be used to form blocks, that minimizes the variance of the causal effect estimates. We leverage the underlying causal graph and provide an efficient algorithm to obtain such a set for a general *semi-Markovian* causal model.

1.4. Organization

The rest of the paper is organized as follows. Section 2 provides some background and formulates the problem of interest. Section 3 discusses our methodology that includes motivating examples, key observations, important lemmas, main results, and the final algorithm. Section 4 provide experimental evaluations, and Section 5 concludes the paper and provides future directions.

2. Background and Problem Formulation

In this section, we provide some background and formulate the problem of interest in this paper.

Let $\mathcal{V}$ be the set of observed random variables and $\mathcal{U}$ be the set of unobserved random variables. Let $\mathcal{G}$ be a directed acyclic graph that depicts the causal relationships among the random variables in $\mathcal{U}$ and $\mathcal{V}$. Formally, we are assuming that the probability distribution of the random variables in $\mathcal{U}$ and $\mathcal{V}$ factorizes according to $\mathcal{G}$ (the causal Markov condition (Geiger and Pearl, 1990)). Furthermore, we assume that no variables other than the ones that are d-separated (Pearl, 2009) in $\mathcal{G}$ are independent, i.e. independence in the joint probability distribution of the random variables in $\mathcal{U}$ and $\mathcal{V}$ is encoded by the causal graph (the causal faithfulness condition (Meek, 1995)).

We are interested in studying the causal effect of $X \in \mathcal{V}$ (treatment) on $Y \in \mathcal{V}$ (response). Studying the causal effect of X on Y means learning the (interventional) probability distribution, $\mathbb{P}(Y|do(X))$ (Pearl, 2009).

2.1. Gain in Precision due to Blocking

Consider a simple setting where X is an indicator random variable defined as

$$X = \begin{cases} 1, & \text{if the treatment is applied,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

and Y is a real-valued random variable.

Let there be n experimental units available to study the causal effect of X on Y by performing a randomized experiment. Let $\mathcal{Z}$ denote the set of covariates that vary across the experimental units.

We want to demonstrate the gain in precision due to blocking. Thus, we first ignore these covariates and perform a CRD as follows. We randomize X over the available experimental units, i.e., randomly assign a unit to either have the treatment ($X = 1$) or not ($X = 0$) and then observe the response. Let (X_i, Y_i) denote the treatment and response pair for the ith experimental unit. Define $I_x := \{i : X_i = x\}$ and $n_x := |I_x|$.

To demonstrate the gain in precision due to blocking, we focus on estimating $\mathbb{E}(Y|do(X))$. For simplicity, define $Y(x) := Y|do(X = x)$. For our simple setting, we have only two possibilities of that conditional expectation, viz. $\mathbb{E}(Y(1))$ and $\mathbb{E}(Y(0))$. Similar to Heckman (1992), Clements et al.

(1994), and Heckman and Smith (1995), we define the *effect of treatment* (ET) or *average treatment effect* (ATE) in the population as

$$\beta := \mathbb{E}\left(Y(1)\right) - \mathbb{E}\left(Y(0)\right). \tag{2}$$

A natural non-parametric estimator of $\mathbb{E}\left(Y(x)\right)$ is the corresponding sample average, given as

$$\bar{Y}(x) := \frac{1}{n_x} \sum_{i \in I_x} Y_i. \tag{3}$$

Therefore, we can estimate β as

$$\hat{\beta} = \bar{Y}(1) - \bar{Y}(0) \tag{4}$$

It can be shown that $\mathbb{E}(\hat{\beta}) = \beta$, i.e., $\hat{\beta}$ is *unbiased* for β. The variance of $\hat{\beta}$ is given as follows.

$$\text{Var}\left(\hat{\beta}\right) = \mathbb{E}_{\mathcal{Z}}\left(\frac{\text{Var}\left(Y(1)|\mathcal{Z}\right)}{n_{1,z}} + \frac{\text{Var}\left(Y(0)|\mathcal{Z}\right)}{n_{0,z}}\right) + \mathbb{E}_{\mathcal{Z}}(\beta(\mathcal{Z}) - \beta)^2, \tag{5}$$

where $\beta(\mathcal{Z}) := \mathbb{E}\left(Y(1)|\mathcal{Z}\right) - \mathbb{E}\left(Y(0)|\mathcal{Z}\right)$ is called the *$\mathcal{Z}$-specific causal effect* or *conditional average treatment effect* (CATE). For proofs of unbiasedness and variance of $\hat{\beta}$, refer to Appendix A.1.

We now perform an RBD as follows. We first stratify the experimental units such that units within each stratum (known as a block) are identical, i.e., the covariates in $\mathcal{Z}$ remain the same with a block. We next randomize X over the experimental units within each block and then observe the response. Let $(X_i, Y_i, \mathcal{Z}_i)$ denote the treatment, response, and covariates triplet for the ith experimental unit. Define $I_{x,z} := \{i : X_i = x, \mathcal{Z}_i = z\}$ and $n_{x,z} := |I_{x,z}|$.

Following (Petersen et al., 2006), ET defined in (2) can be re-written as

$$\begin{aligned}
\beta &= \mathbb{E}_{\mathcal{Z}}\left(\mathbb{E}\left(Y(1)|\mathcal{Z}\right) - \mathbb{E}_{\mathcal{Z}}\left(\mathbb{E}\left(Y(0)|\mathcal{Z}\right)\right.\right., && (6)\\
&= \mathbb{E}_{\mathcal{Z}}\left(\mathbb{E}\left(Y(1)|\mathcal{Z}\right) - \mathbb{E}\left(Y(0)|\mathcal{Z}\right)\right), && (7)\\
&= \sum_{\mathcal{Z}}\left(\mathbb{E}\left(Y(1)|\mathcal{Z}\right) - \mathbb{E}\left(Y(0)|\mathcal{Z}\right)\right)\mathbb{P}\left(\mathcal{Z}\right) && (8)
\end{aligned}$$

A natural non-parametric estimator of $\mathbb{E}\left(Y(x)|\mathcal{Z}\right)$ is the corresponding sample average, given as

$$\bar{Y}(x)|\mathcal{Z} := \frac{1}{n_{x,z}} \sum_{i \in I_{x,z}} Y_i \tag{9}$$

Define $n_z := n_{1,z} + n_{0,z}$. $\mathbb{P}\left(\mathcal{Z}\right)$ can be estimated as

$$\hat{\mathbb{P}}\left(\mathcal{Z}\right) := \frac{n_z}{n} \tag{10}$$

Therefore, we can estimate β as

$$\hat{\beta}_{\mathcal{Z}} = \frac{1}{n} \sum_{\mathcal{Z}} n_z \left(\bar{Y}(1)|\mathcal{Z} - \bar{Y}(0)|\mathcal{Z}\right) \tag{11}$$

It can be shown that $\mathbb{E}(\hat{\beta}_{\mathcal{Z}}) = \beta$, i.e., $\hat{\beta}_{\mathcal{Z}}$ is *unbiased* for β. The variance of $\hat{\beta}_{\mathcal{Z}}$ is given as follows.

$$\text{Var}\left(\hat{\beta}_{\mathcal{Z}}\right) = \mathbb{E}_{\mathcal{Z}}\left(\frac{\text{Var}\left(Y(1)|\mathcal{Z}\right)}{n_{1,z}} + \frac{\text{Var}\left(Y(0)|\mathcal{Z}\right)}{n_{0,z}}\right) + \mathbb{E}_{\mathcal{Z}}\left(\sum_{\mathcal{Z}} \beta(\mathcal{Z})\hat{\mathbb{P}}\left(\mathcal{Z}\right) - \beta\right)^2, \tag{12}$$

For proofs of unbiasedness and variance of $\hat{\beta}_{\mathcal{Z}}$, refer to Appendix A.2.

As $\hat{\mathbb{P}}(\mathcal{Z})$ is as an unbiased estimate of $\mathbb{P}(\mathcal{Z})$, $\sum_{\mathcal{Z}} \beta(\mathcal{Z})\hat{\mathbb{P}}\left(\mathcal{Z}\right) = \mathbb{E}_{\mathcal{Z}}(\beta(\mathcal{Z})) = \beta$. Hence, the second term on the right in (12) is zero. By comparing (5) and (12), we observe that $\text{Var}\left(\hat{\beta}_{\mathcal{Z}}\right) \leq \text{Var}\left(\hat{\beta}\right)$, i.e., blocking improves the precision of the estimate of the population average causal effect.

Note 1 Blocking is guaranteed to reduce the variance of the estimates of the causal effects under the condition the experimental units are homogeneous within each block and heterogeneous across different blocks. If the blocks are formed using some covariates that do not ensure this then blocking may, in fact, increase the variance. Hence, in practice, blocking involves a trade-off as it partitions the experimental units and thus reduces the sample size for estimating the within-block causal effect which increases the variance but may reduce the inherent variability of the response within each block which can reduce the variance.

2.2. Challenges with Blocking

We observed that in experimental studies under non-homogeneity of the experimental units, blocking improves the precision of causal effect estimation. However, the practical difficulty with blocking is that depending on the number of covariates and the number of distinct values (denoted as $v(\cdot)$) of each covariate, the number of blocks can be very large. For covariates in the set $\mathcal{Z}$, we need to form $\prod_{Z \in \mathcal{Z}} v(Z)$ different blocks. For example, if we want to study the effect of a drug on curing some heart disease where the subjects under consideration have the following attributes (which can potentially affect the effectiveness of the drug). Gender: Male and Female, i.e., $v(\text{Gender}) = 2$, Age: <25, 25-45, 45-65, >65, i.e., $v(\text{Age}) = 4$, Weight: Underweight, Normal, Overweight, Obese I, Obese II, i.e., $v(\text{Weight}) = 5$, Blood Pressure: Normal, Prehypertension, Hypertension I, Hypertension II, Hypertensive Crisis, i.e., $v(\text{Blood Pressure}) = 5$, and Bad Cholesterol: Optimal, Above Optimal, Borderline High, High, Very High, i.e., $v(\text{Bad Cholesterol}) = 5$. Thus, we need to form $2 \times 4 \times 5 \times 5 \times 5 = 1000$ blocks. Performing a randomized experiment with a large number of blocks can be very costly. Sometimes, it may not be feasible as the number of blocks can be larger than the number of subjects. This would cause some of the blocks to be empty. For instance, there may not be any male subjects under the age of 25 who are Obese II with a Hypertensive Crisis and Optimal Cholesterol level.

Other than the economic aspects and some blocks being empty, there are reasons why some variables should never be used for forming blocks. Refer to Section 3.4 for details.

2.3. Problem Statement

So far, we observed that blocking improves the precision of causal effect estimation. However, in situations when the number of blocks to be formed is very large, blocking becomes costly and/or infeasible. One possible way to reduce the number of blocks is to form blocks using some but not all covariates. But the question is which covariates should be preferred over others while forming blocks? One possible way is to select the (smallest) set of covariates that leads to maximum precision

in causal effect estimation. In the context of estimating the effect of treatment in the population, β, (discussed in Section 2.1), it means finding the smallest set, $\mathcal{Z}$, that minimizes Var $(\hat{\beta}_{\mathcal{Z}})$.

We next formalize the problem of interest in this paper as follows. We are given a directed acyclic causal graph, $\mathcal{G}$, that depicts the causal relationships among some observed variables, $\mathcal{V}$, and unobserved variables, $\mathcal{U}$. We are interested in studying the causal effect of $X \in \mathcal{V}$ (treatment) on $Y \in \mathcal{V}$ (response) by performing a randomized block experiment. The set of observed covariates is $\mathcal{C} := \mathcal{V} \backslash \{X, Y\}$.

Studying the causal effect of X on Y means learning the (interventional) probability distribution, $\mathbb{P}(Y|do(X))$. We can write $\mathbb{P}(Y = y|do(X = x))$ as

$$\mathbb{P}(Y = y|do(X = x)) = \frac{\mathbb{P}\left(Y = y \cap do(X = x)\right)}{\mathbb{P}\left(do(X = x)\right)} \tag{13}$$

For estimating the above probability, we perform an RBD by forming blocks using covariates in $\mathcal{Z} \subseteq \mathcal{V} \backslash \{X, Y\}$, and similar to the estimation of $\mathbb{E}\left(Y|do(X)\right)$, we define an estimate of $\mathbb{P}(Y = y|do(X = x))$ as

$$\hat{\mathbb{P}}_{\mathcal{Z}}(Y = y|do(X = x)) := \sum_{\mathcal{Z}=z} \frac{\hat{\mathbb{P}}\left(Y = y \cap do(X = x) \cap \mathcal{Z} = z\right)}{\hat{\mathbb{P}}\left(do(X = x) \cap \mathcal{Z} = z\right)} \hat{\mathbb{P}}\left(\mathcal{Z} = z\right) \tag{14}$$

where $\hat{\mathbb{P}}\left(\cdot\right)$ are the sample relative frequencies.

It is desirable to select the $\mathcal{Z}$ such that Var $(\hat{\mathbb{P}}_{\mathcal{Z}}(Y = y|do(X = x)))$ is minimized, i.e., maximum gain in precision. For ease of notation, we define Var $(\hat{\mathbb{P}}_{\mathcal{Z}}(Y|do(X))) :=$ Var $(\hat{\mathbb{P}}_{\mathcal{Z}}(Y = y|do(X = x)))$.

Problem 1 *Given a directed acyclic graph, $\mathcal{G}$, that depicts the causal relationships among some observed variables, $\mathcal{V}$, and unobserved variables, $\mathcal{U}$; treatment, $X \in \mathcal{V}$, and response, $Y \in \mathcal{V}$, obtain a smallest subset, $\mathcal{Z}^*$, of the set of observed covariates, $\mathcal{C} := \mathcal{V} \backslash \{X, Y\}$, such that*

$$\mathcal{Z}^* \in \arg\min_{\mathcal{Z}} \textit{Var}\left(\hat{\mathbb{P}}_{\mathcal{Z}}(Y|do(X))\right). \tag{15}$$

Note 2 It is reasonable to assume that using some expert knowledge, we have access to the underlying causal graph $\mathcal{G}$ and do not know the strengths/weights of the causal connections depicted in this graph (Kyriacou et al., 2022). For instance, experts may tell us that alcohol consumption causes high blood pressure but may not have the knowledge of the magnitude of this causation. We can also construct the underlying causal graph using some observational data if it is available (Nordon et al., 2019).

Note 3 While performing blocking for the minimization of the variance of the estimate of the causal effect, it is important to ensure that we are still able to accurately estimate the causal effect of interest. As introduced in Section 1.1, we want the chosen set of covariates to be *stable*, i.e. we exclude the covariates that one should never be used to create blocks. We formally motivate and define the notion of stability of the solution to Problem 1 in Section 3.4. We also provide a method of achieving stability, i.e. finding and excluding the covariates that lead to inaccurate causal effect estimation.

Note 4 In experimental design, the term 'block' refers to a subset of the experimental units that are homogenous with respect to some covariates while in causal graphical models, it refers to an obstruction in some causal path. For instance, for an experiment, a group of smokers can be a block while in causal graphical models, a common cause (confounder) blocks/obstructs a causal path.

3. Methodology

In this section, we develop a methodology to find a solution to Problem 1. We first examine Var $(\hat{\mathbb{P}}_{\mathcal{Z}}(Y|do(X)))$ as a function of $\mathcal{Z}$ to obviate some edges and nodes from the causal graph, $\mathcal{G}$. We next discuss some motivating examples and make key observations that lead to several lemmas for our main result. In the end, we provide our main result and develop an efficient algorithm for solving Problem 1.

3.1. Examining Var $(\hat{\mathbb{P}}_{\mathcal{Z}}(Y|do(X)))$ as a function of $\mathcal{Z}$

In Problem 1, our interest is to minimize Var $(\hat{\mathbb{P}}_{\mathcal{Z}}(Y|do(X)))$ as a function of $\mathcal{Z} \subseteq \mathcal{C}$. $\mathcal{Z}$ can affect Var $(\hat{\mathbb{P}}_{\mathcal{Z}}(Y|do(X)))$ through the causal relationships from $\mathcal{Z}$ to Y, and from $\mathcal{Z}$ to X. Note that, X is the treatment on which we are performing the intervention, i.e., we set the levels/values of X. This is equivalent to saying that X is no longer affected by the rest of the variables. Therefore, $\mathcal{Z}$ also cannot affect X when an intervention is performed on X. Therefore, $\mathcal{Z}$ can affect Var $(\hat{\mathbb{P}}_{\mathcal{Z}}(Y|do(X)))$ only through the causal relationships from $\mathcal{Z}$ to Y. Therefore, a smallest subset of $\mathcal{C}$ that blocks all causal paths from $\mathcal{C}$ to Y gives a solution to 1.

Denote the subgraph of $\mathcal{G}$ where all edges coming into X were deleted as $\mathcal{G}_{\widetilde{X}}$.

Lemma 1 *For finding a solution to Problem 1, it is sufficient to work with the subgraph $\mathcal{G}_{\widetilde{X}}$.*

Proof The proof follows from the fact that when an intervention is performed on X then it is no longer affected by the rest of the variables. In terms of the causal graph, $\mathcal{G}$, it is the same as deleting all edges coming into X. ■

Definition 2 *$\mathcal{S}, \mathcal{T} \subseteq \mathcal{C}$ are called equivalent sets (denoted as $\mathcal{S} \equiv \mathcal{T}$), if Var $(\hat{\mathbb{P}}_{\mathcal{S}}(Y|do(X))) =$ Var $(\hat{\mathbb{P}}_{\mathcal{T}}(Y|do(X)))$.*

Lemma 3 *If $\mathcal{F}$ is the set of all covariates that do not have causal paths to Y, then $\mathcal{C}$ and $\mathcal{C} \backslash \mathcal{F}$ are equivalent sets.*

Proof The proof follows from the fact that a set of covariates can affect Var $(\hat{\mathbb{P}}_{\mathcal{Z}}(Y|do(X)))$ only through the causal relationships from itself to Y. ■

Following Theorem 1, we restrict ourselves to the subgraph $\mathcal{G}_{\widetilde{X}}$. Following Theorem 3, we further delete the covariates that do not have any causal paths to Y. Denote this new subgraph as $\mathcal{G}'_{\widetilde{X}}$ and $\mathcal{C}' = \mathcal{C} \backslash \mathcal{F}$. Therefore, a smallest subset of $\mathcal{C}$ that blocks all causal paths from $\mathcal{C}'$ to Y in $\mathcal{G}'_{\widetilde{X}}$ gives a solution to 1.

3.2. Motivating Examples and Key Observations

All causal graphs in Figures 1 to 5 represent $\mathcal{G}'_{\widetilde{X}}$. Denote the set of parents of W in $\mathcal{G}'_{\widetilde{X}}$ as $\mathcal{P}(W)$.

Note 5 In all the graphs provided in this paper, a bi-directed edge between any two observed variables is the same as those two variables having one or more unobserved ancestors. For example $V_3 \dashleftarrow\dashrightarrow V_4$ is the same as $V_3 \dashleftarrow U \dashrightarrow V_4$ for some unobserved variable, U.

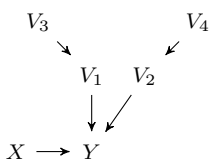

Figure 1: Example Markovian model (no latent structures).

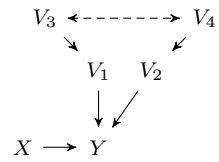

Figure 2: Example semi-Markovian model with latent variables involving only $\mathcal{C}'$.

Example 1 For Figures 1 and 2, $\mathcal{P}(Y)\backslash\{X\} = \{V_1, V_2\}$ is sufficient to block all causal paths from $\mathcal{C}' = \{V_1, V_2, V_3, V_4\}$ to Y. Therefore, $\mathcal{Z}^* = \mathcal{P}(Y)\backslash\{X\}$.

Example 2 For Figures 3a and 3b, $\mathcal{P}(Y)\backslash\{X\} = \{V_1, V_2\}$ is not sufficient to block all causal paths from $\mathcal{C}' = \{V_1, V_2, V_3, V_4\}$ to Y. In Figure 3a, the reason is that blocking using V_2 opens the path $V_4 \rightarrow V_2 \leftarrow\!\!\dashrightarrow Y$ for V_4 to cause variability in Y. In Figure 3b, the reason is that the latent structure $Y \leftarrow\!\!\dashrightarrow V_4$ enables V_4 to cause Y even when the path $V_4 \rightarrow V_2 \rightarrow Y$ is blocked. For both Figures 3a and 3b, $\mathcal{Z}^* = (\mathcal{P}(Y)\backslash\{X\}) \cup \{V_4\}$.

(a) (b)

Figure 3: Example semi-Markovian model with latent variables involving $\mathcal{C}'$ and Y.

Definition 4 *(Pearl, 2009; Rohrer, 2018) A variable is a collider or an inverted fork when it is causally influenced by two or more variables.*

Note 6 If we condition/block using a collider (and/or it's descendants) then the common causes of the collider become statistically dependent (Berkson, 1946; Cole et al., 2010). For example, if alcohol and smoking are common causes of high blood pressure then it is not necessary that these common causes are correlated in the population at large. However, if we restrict our attention to (condition on) those with high blood pressure then we introduce a non-causal dependence between alcohol and smoking. [1]

Example 3 For Figure 4, despite Y having descendants, $(\mathcal{P}(Y)\backslash\{X\}) \cup \{V_4\}$ blocks all causal paths from $\mathcal{C}'$ to Y. This is because descendants of Y can never be a cause of variability in Y. In Figures 4a and 4b, V_5 (a descendant of Y) cannot cause Y through the edge $Y \rightarrow V_5$. Moreover, V_4 (an ancestor of Y) cannot cause Y through V_5 because V_5 is a collider on the path $V_4 \rightarrow V_5 \leftarrow Y$.

(a) (b)

Figure 4: Example Semi-Markovian Models with latent variables involving $\mathcal{C}'$ and Y, and Y having descendants.

Denote the set of ancestors (including itself) of W in $\mathcal{G}'_{\widetilde{X}}$ as $\mathcal{A}(W)$. Denote the subgraph of $\mathcal{G}'_{\widetilde{X}}$ restricted to $\mathcal{A}(Y)$ as $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$. Based on the insights from Examples 1 to 3, we write Theorem 5.

Lemma 5 *For finding a solution to Problem 1, it is sufficient to work with the subgraph $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$.*

Proof The proof follows from the fact the graph under consideration is acyclic. Hence, the descendants of a node can never be its causes even in the presence of latent structures. ■

3.3. Solution to Problem 1

We now combine the insights from Theorems 1, 3 and 5 to provide a solution to Problem 1.

Theorem 6 *The smallest $\mathcal{Z}$ such that $Y \perp (\mathcal{A}(Y)\backslash\{X,Y\})\backslash\mathcal{Z}|\mathcal{Z}$ in $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$ is a solution to Problem 1.*

Proof The proof follows from the fact that $\mathcal{Z}$ such that $Y \perp (\mathcal{A}(Y)\backslash\{X,Y\})\backslash\mathcal{Z}|\mathcal{Z}$ in $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$ blocks all causal paths from $\mathcal{C}'$ to Y. ■

Theorem 6 provides a sufficient condition for a set $\mathcal{Z}$ to be a solution to Problem 1. We next provide a method to construct a set $\mathcal{Z}$ that satisfies this sufficient condition.

Let a path composed entirely of bi-directed edges be called a *bi-directed path*.

Definition 7 *(Tian and Pearl, 2002). For a graph $\mathcal{G}$, the set of observed variables $\mathcal{V}$, can be partitioned into disjoint groups by assigning two variables to the same group if and only if they are connected by a bi-directed path. Assume that V is therefore partitioned into k groups $\mathcal{S}_1, \ldots, \mathcal{S}_k$. We will call each $\mathcal{S}_j; j = 1, \ldots, k$, a c-component of V in $\mathcal{G}$ or a c-component (abbreviating confounded component) of $\mathcal{G}$. The $\mathcal{S}_j$ such that $W \in S_j$ is called the c-component of W in $\mathcal{G}$ and is denoted as $C_{W,\mathcal{G}}$. As $\{\mathcal{S}_1, \ldots, \mathcal{S}_k\}$ is a partition of $\mathcal{V}$, c-component of a variable always exists and is unique.*

In Example 2, we observed that due to the presence of latent structures involving Y, $\mathcal{P}(Y)$ was not sufficient to block all causal paths from $\mathcal{C}'$ to Y. However the parents of c-component of Y in $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$ would be sufficient for that purpose.

1. The reader may refer to the tutorials Collectively Exhaustive and Causal Graphs by David Childers for understanding the basic concepts related to causal graphs.

Theorem 8 *The smallest $\mathcal{Z}$ such that $Y \perp (\mathcal{A}(Y)\backslash\{X,Y\})\backslash\mathcal{Z}|\mathcal{Z}$ is the set of parents of c-component of Y in $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$ excluding X denoted as $\mathcal{P}(C_{Y,\mathcal{G}_{\widetilde{X},\mathcal{A}(Y)}})\backslash\{X\}$.*

Proof The proof follows from Corollary 1 (Tian, 2002), which states that a node is independent of its ancestors excluding the parents of its c-component given the parents of its c-component. ■

Note 7 In a Markovian model there are no bi-directed paths, Therefore, the elementary partition of $\mathcal{A}(Y)$ is the c-component of $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$. Therefore, $C_Y = \{Y\}$, and $\mathcal{P}(C_{Y,\mathcal{G}_{\widetilde{X},\mathcal{A}(Y)}})\backslash\{X\} = \mathcal{P}(Y)\backslash\{X\}$ in $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$. This matches with our insight in Example 1, where we observed that $Y \perp (\mathcal{A}(Y)\backslash\{X,Y\})\backslash(\mathcal{P}(Y)\backslash\{X\})|(\mathcal{P}(Y)\backslash\{X\})$.

3.4. Stability of the Solution to Problem 1

We next discuss some concerns with using $\mathcal{Z}^*$ (a solution to Problem 1) as a blocking set. We provide a method to address those concerns.

Definition 9 *Post-treatment ancestors of (excluding X) of the response are those ancestors of (excluding X) of the response who are also the descendants of the treatment.*

Denote the post-treatment ancestors of the response while studying the causal effect $Y|do(X)$ as $\mathcal{M}(Y|do(X))$, and the set of descendants (including itself) of W in $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$ as $\mathcal{D}(W)$.

Lemma 10 $\mathcal{M}(Y|do(X)) = (\mathcal{D}(X) \cap \mathcal{A}(Y))\backslash\{X,Y\}$.

Proof The proof follows directly from Theorem 9. ■

Example 4 In Figures 5a and 5b, there exist post-treatment ancestors (excluding itself) of the response. Blocks created using these covariates are not well-defined. For instance, we are interested in studying the causal effect of a drug on blood pressure. Suppose anxiety level mediates the effect of drug on blood pressure, i.e., drug $\rightarrow$ anxiety $\rightarrow$ blood pressure. If we create blocks using anxiety then the blocks will change (i.e., become unstable) during the experiment because change in the drug level will cause the anxiety level to change. Therefore, it is reasonable not to create blocks using these covariates. With this insight, for Figure 5a, $(\mathcal{P}(Y)\backslash\{X\}) \cup \{V_4\}$ blocks all (pre-treatment) causal paths from $\mathcal{C}'$ to Y. However, we cannot simply delete the post-treatment ancestors (excluding itself) of the response from the subgraph $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$ of interest. This is because the presence of latent structures between pre- and post-treatment ancestors of the response affects blocking set of interest. For instance, in Figure 5b, V_1 is a collider on the path $V_3 \rightarrow V_1 \leftarrow V_6$. Hence, keeping V_1 in the blocking set will open this path for V_3 to cause variability in Y. Therefore, for Figure 5b, we need $(\mathcal{P}(Y)\backslash\{X\}) \cup \{V_3, V_4\}$ to block all (pre-treatment) causal paths from $\mathcal{C}'$ to Y.

Definition 11 *For studying the causal effect of X on Y, a blocking set $\mathcal{Z}$, is said to be stable if $\mathcal{Z} \cap \mathcal{M}(Y|do(X)) = \phi$.*

Theorem 12 *A stable solution to Problem 1 is $\mathcal{Z}^* = (\mathcal{P}(C_{Y,\mathcal{G}_{\widetilde{X},\mathcal{A}(Y)}}\backslash\{X\})\backslash\mathcal{M}(Y|do(X))$.*

Proof The proof follows from Theorem 10. ■

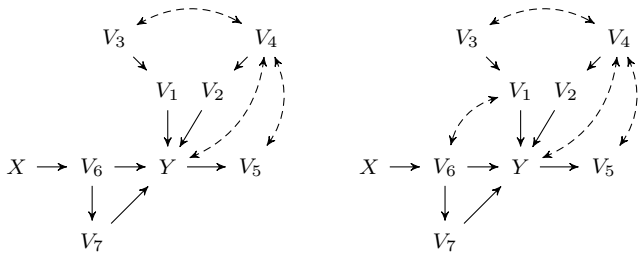

Figure 5: Example semi-Markovian models with post-treatment ancestors of Y.

3.5. Final Algorithm

We next provide steps to obtain a stable solution to Problem 1 based on Theorems 6, 8 and 12. First, we reduce $\mathcal{G}$ to $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$. Next, we obtain the set $\mathcal{Z} = \mathcal{P}(C_{Y,\mathcal{G}_{\widetilde{X},\mathcal{A}(Y)}})\backslash\{X\}$ which is the smallest set such that $Y \perp (\mathcal{A}(Y)\backslash\{X,Y\})\backslash\mathcal{Z}|\mathcal{Z}$ in $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$. Finally, we drop $\mathcal{M}(Y|do(X))$ from $\mathcal{Z}$ to get a stable solution to Problem 1.

Denote the edge $W_1 \rightarrow W_2$ as *directed-edge*(W_1, W_2), a path consisting of directed edges from W_1 to W_2 as *directed-path*(W_1, W_2), and a path consisting of all bi-directed edges from W_1 to W_2 as *bi-directed-path*(W_1, W_2). Algorithm 1 outlines the psuedocode of the proposed algorithm to obtain a stable solution to Problem 1.

3.6. Time Complexity of Algorithm 1

We next show that Algorithm 1 is an efficient polynomial-time algorithm. Finding if there exists a path between any two nodes in a directed graph can be done using breath-first-search which has the worst-case time complexity of $O(|\mathcal{V}|^2)$. Algorithm 1 first reduces $\mathcal{G}$ to $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$ which involves finding variables with edges going into X taking $O(|\mathcal{V}|)$ steps, finding variables with no path into Y taking $O(|\mathcal{V}|^3)$ steps, and finding all ancestors of Y taking $O(|\mathcal{V}|^3)$ steps. Algorithm 1 next finds the c-component of Y involves traversing all pairs of variables: $O(|\mathcal{V}|^2)$, and finding if there exists a bi-directed path between any pair of variables: $O(|\mathcal{V}|^2)$. Therefore, the total time at this step is $O(|\mathcal{V}|^4)$. Finally Algorithm 1 finds $\mathcal{M}(Y|do(X))$ which involves finding descendants of X taking $O(|\mathcal{V}|^3)$ steps, and finding ancestors of Y taking $O(|\mathcal{V}|^3)$ steps. Therefore, the time complexity of Algorithm 1 is $O(|\mathcal{V}|^4)$.

3.7. Implementing Algorithm 1

We demonstrate the implementation of Algorithm 1 to obtain a stable solution to Problem 1 for a general semi-Markovian causal graph, $\mathcal{G}$ given in Figure 6.

We first obtain the graph $G_{An_{\widetilde{X}}(Y)}$. We next obtain the c-components (following Theorem 7) of $\mathcal{G}_{\widetilde{X},\mathcal{A}(Y)}$ are $\{X\}$, $\{\text{Alcohol}, \text{Anxiety}, Y, \text{Cholesterol}\}$, $\{\text{Food-Habits}\}$, $\{\text{Age}\}$, and the c-component of Y is $\{\text{Alcohol}, \text{Anxiety}, Y, \text{Cholesterol}\}$. We next obtain the set of parents of the c-component of Y as $\{X, \text{Alcohol}, \text{Cholesterol}, \text{Food-Habits}, \text{Age}, \text{Anxiety}\}$. The set $\mathcal{D}(X) \cap \mathcal{A}(Y)$ is $\{X, Y, \text{Anxiety}, \text{Sleep-Quality}\}$.

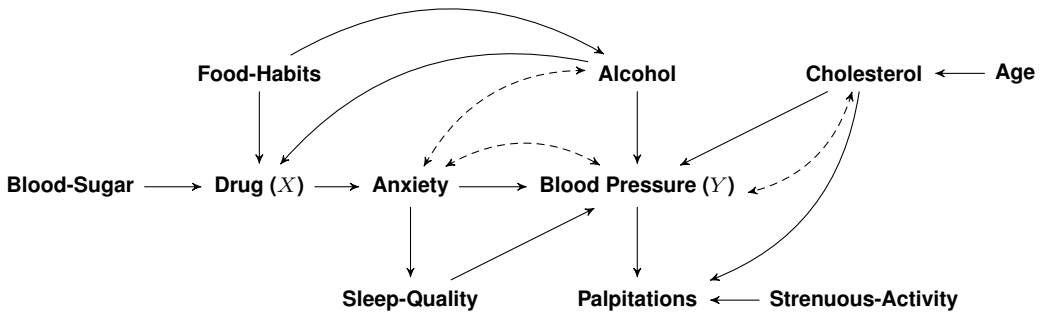

Figure 6: The Original Graph $\mathcal{G}$.

Table 1: Causal effect estimate and variability in the response.

Blocking set	Causal effect	Variance
ϕ, i.e. no blocking	0.6910	0.2126
{Food Habits}	0.6950	0.2098
{Food Habits, Alcohol}	0.6960	0.2028
{Food Habits, Alcohol, Cholesterol}	0.6760	0.2001
{Food Habits, Alcohol, Cholesterol, Age}	0.7150	0.1754

Finally, we obtain the final set as:

$$(\{X, \text{Alcohol}, \text{Cholesterol}, \text{Food-Habits}, \text{Age}, \text{Anxiety}\}\backslash\{X\})\backslash(\{X, Y, \text{Anxiety}, \text{Sleep-Quality}\}\backslash\{X, Y\})$$

$$= \{\text{Alcohol}, \text{Cholesterol}, \text{Food-Habits}, \text{Age}\}.$$

4. Experimental Evaluations

In this section, we provide experimental evaluations to compare our method with other choices of sets of covariates to create blocks. For the causal graph, $\mathcal{G}$ provided in Figure 6, we are interested in studying the causal effect of Drug (X) on Blood-Pressure (Y).

We considered a simple setting where all variables under consideration are binary variables taking values in {0,1}. We used a simple joint probability distribution that factorizes according to the given causal graph. We then performed simulated randomized experiments for different choices of sets of covariates to create blocks. The estimated causal effect and the variability in the response for different choices of blocking sets are provided in Table 1. We observe that the variability in the response is the smallest for the choice of covariates obtained using the algorithm proposed in this paper. For details of the experimental evaluations, refer to Appendix B.

5. Conclusion and Future Work

We investigated the problem of finding a stable set of covariates to be used for forming blocks, that minimizes the variance of the causal effect estimates. By leveraging the underlying causal graph, we provided an efficient algorithm to obtain such a set for a general semi-Markovian causal model.

In the future, we are interested in extending our work towards further pruning the stable causal blocking set obtained in this paper for the reasons of economic cost. If the experimenter has a constraint on the cardinality (size) of the blocking set due to budget limitations, i.e. $|\mathcal{Z}| \leq k_1$ where k_1 is specified by the experimenter. Furthermore, if blocking variables have different economic costs associated with them, for example, getting data on cholesterol level may be costlier (or difficult) as compared to the age of the patient, then we can introduce a knapsack budget constraint. Let $e(Z)$ be the cost associated with $Z \in \mathcal{Z}$ then the knapsack constraint would be $\sum_{Z \in \mathcal{Z}} e(Z) \leq k_2$ where k_2 is specified by the experimenter. We are interested in solving the problem of finding stable solutions to Problem 1 with additional cardinality or knapsack constraint on the feasible solutions. Furthermore, we are interested in extending our work for the cases when the treatment and response variables are sets of variables. For the future direction to use the proposed algorithm for different real-world applications, refer to Appendix C.

Algorithm 1 Stable-Causal-Blocking $(\mathcal{G}, X, Y)$

for i in $1, \ldots, |\mathcal{V}|$ **do** ▷ Reducing $\mathcal{G}$ to $\mathcal{G}_{\widetilde{X}}$
  if $\exists$ *directed-edge*(V_i, X) **then**
    delete *directed-edge*(V_i, X) from $\mathcal{G}$
  end if
end for
for i in $1, \ldots, |\mathcal{V}|$ **do** ▷ Reducing $\mathcal{G}_{\widetilde{X}}$ to $\mathcal{G}'_{\widetilde{X}}$
  if $\nexists$ *directed-path*(V_i, Y) **then**
    delete V_i from $\mathcal{G}$
  end if
end for
$\mathcal{A}(Y) \leftarrow \{Y\}$
for i in $1, \ldots, |\mathcal{V}|$ **do** ▷ Reducing $\mathcal{G}'_{\widetilde{X}}$ to $\mathcal{G}'_{\widetilde{X},\mathcal{A}(Y)}$
  if $\exists$ path(V_i, Y) **then**
    $\mathcal{A}(Y) \leftarrow \mathcal{A}(Y) \cup \{V_i\}$
  end if
end for
$\mathcal{S}_i = \{V_i\}, i = 1, \ldots, |\mathcal{V}|$ ▷ Finding $C_{Y,\mathcal{G}_{\widetilde{X},\mathcal{A}(Y)}}$
for i in $1, \ldots, |\mathcal{V}|$ **do**
  for j in $1, \ldots, i$ **do**
    if $\exists$ *bi-directed-path*(V_i,V_j) **then**
      $\mathcal{S}_i \leftarrow \mathcal{S}_i \cup \mathcal{S}_j; \mathcal{S}_j \leftarrow \phi$
    end if
  end for
end for
$i \leftarrow 1; \mathcal{S} \leftarrow \mathcal{S}_1$
while $Y \notin \mathcal{S}$ **do**
  $i \leftarrow i + 1; \mathcal{S} \leftarrow \mathcal{S}_i$
end while
$\mathcal{P}(\mathcal{S}) = \phi$ ▷ Finding $\mathcal{P}(C_{Y,\mathcal{G}_{\widetilde{X},\mathcal{A}(Y)}})$
for i in $1, \ldots, |\mathcal{V}|$ **do**
  for W in $\mathcal{S}$ **do**
    if $\exists$ *directed-edge*(V_i, W) **then**
      $\mathcal{P}(\mathcal{S}) \leftarrow \mathcal{P}(\mathcal{S}) \cup \{V_i\}$
    end if
  end for
end for
$\mathcal{D}(X) \leftarrow \{X\}$ ▷ Finding $\mathcal{D}(X)$
for i in $1, \ldots, |\mathcal{V}|$ **do**
  if $\exists$ *directed-path*(X, V_i) **then**
    $\mathcal{D}(X) \leftarrow \mathcal{D}(X) \cup \{V_i\}$
  end if
end for
$\mathcal{Z} \leftarrow (\mathcal{P}(\mathcal{S}) \backslash \{X\}) \backslash (\mathcal{D}(X) \cap \mathcal{A}(Y)) \backslash \{X, Y\}$.
return $\mathcal{Z}$

References

Mridul Agarwal, Vaneet Aggarwal, Abhishek K Umrawal, and Christopher J Quinn. Stochastic top k-subset bandits with linear space and non-linear feedback with applications to social influence maximization. *ACM/IMS Transactions on Data Science (TDS)*, 2(4):1–39, 2022.

Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

Nancy Clements, James J Heckman, and Jeffrey A Smith. Making the most out of social experiments: Reducing the intrinsic uncertainty in evidence from randomized trials with an application to the jtpa exp, 1994.

William Gemmell Cochran and Gertrude M Cox. Experimental designs. Technical report, North Carolina State University. Dept. of Statistics, 1948.

Stephen R Cole, Robert W Platt, Enrique F Schisterman, Haitao Chu, Daniel Westreich, David Richardson, and Charles Poole. Illustrating bias due to conditioning on a collider. *International journal of epidemiology*, 39(2):417–420, 2010.

RA Fisher. The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33:503–515, 1926.

Dan Geiger and Judea Pearl. On the logic of causal models. In *Machine Intelligence and Pattern Recognition*, volume 9, pages 3–14. Elsevier, 1990.

Christopher Harshaw, Fredrik Sävje, Daniel Spielman, and Peng Zhang. Balancing covariates in randomized experiments with the gram–schmidt walk design. *arXiv preprint arXiv:1911.03071*, 2019.

James J Heckman. Randomization and social policy evaluation. *Evaluating welfare and training programs*, 1:201–30, 1992.

James J Heckman and Jeffrey A Smith. Assessing the case for social experiments. *Journal of economic perspectives*, 9(2):85–110, 1995.

Leonard Henckel, Emilija Perković, and Marloes H Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:1907.02435*, 2019.

Michael J Higgins, Fredrik Sävje, and Jasjeet S Sekhon. Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences*, 113(27):7369–7376, 2016.

Nathan Kallus. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):85–112, 2018.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146. ACM, 2003.

Oscar Kempthorne. The design and analysis of experiments. 1952.

Oscar Kempthorne. The randomization theory of experimental inference. *Journal of the American Statistical Association*, 50(271):946–967, 1955.

Demetrios N Kyriacou, Philip Greenland, and Mohammad A Mansournia. Using causal diagrams for biomedical research. *Annals of Emergency Medicine*, 2022.

Xinran Li and Peng Ding. Rerandomization and regression adjustment. *arXiv preprint arXiv:1906.11291*, 2019.

Chris Meek. Strong-completeness and faithfulness in belief networks. 1995.

Ryan T Moore. Multivariate continuous blocking to improve political science experiments. *Political Analysis*, 20(4):460–479, 2012.

Guanyu Nie, Mridul Agarwal, Abhishek Kumar Umrawal, Vaneet Aggarwal, and Christopher John Quinn. An explore-then-commit algorithm for submodular maximization under full-bandit feedback. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Galia Nordon, Gideon Koren, Varda Shalev, Benny Kimelfeld, Uri Shalit, and Kira Radinsky. Building causal graphs from medical literature and electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1102–1109, 2019.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Maya L Petersen, Sandra E Sinisi, and Mark J van der Laan. Estimation of direct causal effects. *Epidemiology*, pages 276–284, 2006.

Julia M Rohrer. Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in methods and practices in psychological science*, 1(1):27–42, 2018.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *J. Mach. Learn. Res.*, 21(188):1–86, 2020.

E Smucler, F Sapienza, and A Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *Biometrika*, 109(1):49–65, 03 2021. ISSN 1464-3510. doi: 10.1093/biomet/asab018. URL https://doi.org/10.1093/biomet/asab018.

Jin Tian. *Studies in causal reasoning and learning*. University of California, Los Angeles, 2002.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, 2002.

Abhishek K. Umrawal and Vaneet Aggarwal. A community-aware framework for social influence maximization. *arXiv preprint*, 2022. arXiv link.

Frank Yates. Complex experiments. *Supplement to the Journal of the Royal Statistical Society*, 2(2): 181–247, 1935.

Appendix A. Expectation and Variance of Causal Effect without and with Blocking

In this section, we calculate the expectation and variance of the causal effect without blocking (defined in (2)) and with blocking (defined in (8)).

A.1. Expectation and Variance of $\hat{\beta}$

We calculate the expectation of $\hat{\beta}$ as

$$\mathbb{E}\left(\hat{\beta}\right) = \mathbb{E}\left(\bar{Y}(1)\right) - \mathbb{E}\left(\bar{Y}(0)\right). \tag{16}$$

In general, due to non-homogeneous experimental units, $\mathbb{E}\left(\bar{Y}(x)\right) \neq \mathbb{E}\left(Y(x)\right), x = 0, 1$. Let $\mathcal{Z}$ be the covariates causing non-homogeneity, rewrite (16) as

$$\mathbb{E}\left(\hat{\beta}\right) = \mathbb{E}_{\mathcal{Z}}\left(\mathbb{E}\left(\bar{Y}(1)|\mathcal{Z}\right)\right) - \mathbb{E}_{\mathcal{Z}}\left(\mathbb{E}\left(\bar{Y}(0)|\mathcal{Z}\right)\right). \tag{17}$$

For fixed covariates the experimental units are identical. Therefore, $\mathbb{E}\left(\bar{Y}(x)|\mathcal{Z}\right) = \mathbb{E}\left(Y(1)|\mathcal{Z}\right), x = 0, 1$. Thus,

$$\begin{aligned}\mathbb{E}\left(\hat{\beta}\right) &= \mathbb{E}_{\mathcal{Z}}\left(\mathbb{E}\left(Y(x)|\mathcal{Z}\right)\right) - \mathbb{E}_{\mathcal{Z}}\left(\mathbb{E}\left(Y(0)|\mathcal{Z}\right)\right), &(18)\\ &= \mathbb{E}\left(Y(1)\right) - \mathbb{E}\left(Y(0)\right) = \beta. &(19)\end{aligned}$$

We next calculate the variance of $\hat{\beta}$ as

$$\text{Var}\left(\hat{\beta}\right) = \text{Var}\left(\bar{Y}(1) - \bar{Y}(0)\right). \tag{20}$$

Let $\mathcal{Z}$ be the covariates causing non-homogeneity, we rewrite (20) as

$$\begin{aligned}\text{Var}\left(\hat{\beta}\right) &= \text{Var}_{\mathcal{Z}}\left(\mathbb{E}\left((\bar{Y}(1) - \bar{Y}(0))|\mathcal{Z}\right)\right) + \mathbb{E}_{\mathcal{Z}}\left(\text{Var}\left((\bar{Y}(1) - \bar{Y}(0))|\mathcal{Z}\right)\right), &(21)\\ &= \text{Var}_{\mathcal{Z}}\left(\mathbb{E}\left(\bar{Y}(1)|\mathcal{Z}\right) - \mathbb{E}\left(\bar{Y}(0)|\mathcal{Z}\right)\right) + \mathbb{E}_{\mathcal{Z}}\left(\text{Var}\left(\bar{Y}(1)|\mathcal{Z}\right) + \text{Var}\left(\bar{Y}(0)|\mathcal{Z}\right)\right). &(22)\end{aligned}$$

(22) uses linearity of expectation and randomization.

For fixed covariates the experimental units are identical. Therefore, $\mathbb{E}\left(\bar{Y}(x)|\mathcal{Z}\right) = Y(x)|\mathcal{Z}, x = 0, 1$, and $\text{Var}\left(\bar{Y}(x)|\mathcal{Z}\right) = \frac{1}{n_{x,z}}\text{Var}\left(Y(x)|\mathcal{Z}\right), x = 0, 1$. Therefore,

$$\text{Var}\left(\hat{\beta}\right) = \text{Var}_{\mathcal{Z}}\left(\mathbb{E}\left(Y(1)|\mathcal{Z}\right) - \mathbb{E}\left(Y(0)|\mathcal{Z}\right)\right) + \mathbb{E}_{\mathcal{Z}}\left(\frac{\text{Var}\left(Y(1)|\mathcal{Z}\right)}{n_{1,z}} + \frac{\text{Var}\left(Y(0)|\mathcal{Z}\right)}{n_{0,z}}\right) \tag{23}$$

Define $\beta(\mathcal{Z}) := \mathbb{E}\left(Y(1)|\mathcal{Z}\right) - \mathbb{E}\left(Y(0)|\mathcal{Z}\right)$. Therefore,

$$\begin{aligned}\text{Var}\left(\hat{\beta}\right) &= \text{Var}_{\mathcal{Z}}\left(\beta(\mathcal{Z})\right) + \mathbb{E}_{\mathcal{Z}}\left(\frac{\text{Var}\left(Y(1)|\mathcal{Z}\right)}{n_{1,z}} + \frac{\text{Var}\left(Y(0)|\mathcal{Z}\right)}{n_{0,z}}\right) &(24)\\ &= \mathbb{E}_{\mathcal{Z}}\left(\beta(\mathcal{Z}) - \mathbb{E}_{\mathcal{Z}}(\beta(\mathcal{Z}))\right)^2 + \mathbb{E}_{\mathcal{Z}}\left(\frac{\text{Var}\left(Y(1)|\mathcal{Z}\right)}{n_{1,z}} + \frac{\text{Var}\left(Y(0)|\mathcal{Z}\right.}{n_{0,z}})\right) &(25)\\ &= \mathbb{E}_{\mathcal{Z}}\left(\beta(\mathcal{Z}) - \beta\right)^2 + \mathbb{E}_{\mathcal{Z}}\left(\frac{\text{Var}\left(Y(1)|\mathcal{Z}\right)}{n_{1,z}} + \frac{\text{Var}\left(Y(0)|\mathcal{Z}\right)}{n_{0,z}}\right) &(26)\end{aligned}$$

(26) uses $\mathbb{E}_{\mathcal{Z}}(\beta(\mathcal{Z})) = \mathbb{E}_{\mathcal{Z}}(\mathbb{E}\left(Y(1)|\mathcal{Z}\right)) - \mathbb{E}_{\mathcal{Z}}(\mathbb{E}\left(Y(0)|\mathcal{Z}\right)) = \mathbb{E}\left(Y(1)\right) - \mathbb{E}\left(Y(0)\right) = \beta$.

A.2. Expectation and Variance of $\hat{\beta}_{\mathcal{Z}}$

We calculate the expectation of $\hat{\beta}_{\mathcal{Z}}$ as

$$\mathbb{E}\left(\hat{\beta}_{\mathcal{Z}}\right) = \mathbb{E}\left(\sum_{\mathcal{Z}} \hat{\mathbb{P}}(\mathcal{Z})\left(\bar{Y}(1)|\mathcal{Z} - \bar{Y}(0)|\mathcal{Z}\right)\right), \tag{27}$$

$$= \sum_{\mathcal{Z}} \mathbb{E}\left(\hat{\mathbb{P}}(\mathcal{Z})(\bar{Y}(1)|\mathcal{Z} - \bar{Y}(0)|\mathcal{Z})\right), \tag{28}$$

$$= \sum_{\mathcal{Z}} \mathbb{E}\left(\hat{\mathbb{P}}(\mathcal{Z})\right)\left(\mathbb{E}\left(\bar{Y}(1)|\mathcal{Z}\right) - \mathbb{E}\left(\bar{Y}(0)|\mathcal{Z}\right)\right), \tag{29}$$

$$= \sum_{\mathcal{Z}} \mathbb{P}(\mathcal{Z})\left(\mathbb{E}\left(\bar{Y}(1)|\mathcal{Z}\right) - \mathbb{E}\left(\bar{Y}(0)|\mathcal{Z}\right)\right), \tag{30}$$

$$= \mathbb{E}_{\mathcal{Z}}\left(\mathbb{E}\left(\bar{Y}(1)|\mathcal{Z}\right) - \mathbb{E}\left(\bar{Y}(0)|\mathcal{Z}\right)\right) = \beta. \tag{31}$$

We next calculate the variance of $\beta_{\mathcal{Z}}$ as

$$\text{Var}\left(\hat{\beta}_{\mathcal{Z}}\right) = \text{Var}_{\mathcal{Z}}\left(\mathbb{E}\left(\sum_{\mathcal{Z}}(\bar{Y(1)}|\mathcal{Z} - \bar{Y(0)}|\mathcal{Z})\hat{\mathbb{P}}\left(Z\right)\right)\right) + \mathbb{E}_{\mathcal{Z}}\left(\text{Var}\left(\sum_{\mathcal{Z}}(\bar{Y(1)}|\mathcal{Z} - \bar{Y(0)}|\mathcal{Z})\hat{\mathbb{P}}\left(Z\right)\right)\right), \tag{32}$$

$$= \text{Var}_{\mathcal{Z}}\left(\sum_{\mathcal{Z}}\mathbb{E}\left(\bar{Y(1)}|\mathcal{Z} - \bar{Y(0)}|\mathcal{Z}\right)\hat{\mathbb{P}}\left(Z\right)\right) + \mathbb{E}_{\mathcal{Z}}\left(\sum_{\mathcal{Z}}\text{Var}\left(\bar{Y(1)}|\mathcal{Z} - \bar{Y(0)}|\mathcal{Z}\right)\hat{\mathbb{P}}\left(Z\right)\right), \tag{33}$$

$$= \text{Var}_{\mathcal{Z}}\left(\sum_{\mathcal{Z}}\mathbb{E}\left(\bar{Y}(1)|\mathcal{Z}\right) - \mathbb{E}\left(\bar{Y}(0)|\mathcal{Z}\right)\hat{\mathbb{P}}\left(Z\right)\right) + \mathbb{E}_{\mathcal{Z}}\left(\sum_{\mathcal{Z}}\left(\text{Var}\left(\bar{Y}_1|\mathcal{Z}\right) + \text{Var}\left(\bar{Y}_0|\mathcal{Z}\right)\right)\hat{\mathbb{P}}\left(Z\right)\right), \tag{34}$$

(33) and (34) use linearity of expectation and randomization.

For fixed covariates the experimental units are identical. Therefore, $\mathbb{E}\left(\bar{Y}(x)|\mathcal{Z}\right) = \mathbb{E}\left(Y(x)|\mathcal{Z}\right), x = 0, 1$, and $\text{Var}\left(\bar{Y}(x)|\mathcal{Z}\right) = \frac{1}{n_{x,z}}\text{Var}\left(Y(x)|\mathcal{Z}\right), x = 0, 1$. Therefore,

$$\text{Var}\left(\hat{\beta}_{\mathcal{Z}}\right) = \text{Var}_{\mathcal{Z}}\left(\sum_{\mathcal{Z}}\beta(Z)\hat{\mathbb{P}}(\mathcal{Z})\right) + \mathbb{E}_{\mathcal{Z}}\left(\sum_{\mathcal{Z}}\left(\frac{\text{Var}\left(Y(1)|\mathcal{Z}\right)}{n_{1,z}} + \frac{\text{Var}\left(Y(0)|\mathcal{Z}\right)}{n_{0,z}}\right)\hat{\mathbb{P}}(\mathcal{Z})\right), \tag{35}$$

$$= \mathbb{E}_{\mathcal{Z}}\left(\sum_{\mathcal{Z}}\beta(Z)\hat{\mathbb{P}}(\mathcal{Z}) - \mathbb{E}_{\mathcal{Z}}\left(\sum_{\mathcal{Z}}\beta(Z)\hat{\mathbb{P}}(\mathcal{Z})\right)\right)^2 + \sum_{\mathcal{Z}}\frac{\text{Var}\left(Y(1)|\mathcal{Z}\right)}{n_{1,z}}\mathbb{E}_{\mathcal{Z}}(\hat{\mathbb{P}}(\mathcal{Z}))$$
$$+ \sum_{\mathcal{Z}}\frac{\text{Var}\left(Y(0)|\mathcal{Z}\right)}{n_{0,z}}\mathbb{E}_{\mathcal{Z}}(\hat{\mathbb{P}}(\mathcal{Z})), \tag{36}$$

$$= \mathbb{E}_{\mathcal{Z}}\left(\sum_{\mathcal{Z}} \beta(Z)\hat{\mathbb{P}}(\mathcal{Z}) - \beta\right)^2 + \sum_{\mathcal{Z}} \frac{\text{Var}\left(Y(1)|\mathcal{Z}\right)}{n_{1,z}}\mathbb{P}(Z) + \sum_{\mathcal{Z}} \frac{\text{Var}\left(Y(0)|\mathcal{Z}\right)}{n_{0,z}}\mathbb{P}(Z), \tag{37}$$

$$= \mathbb{E}_{\mathcal{Z}}\left(\sum_{\mathcal{Z}} \beta(Z)\hat{\mathbb{P}}(\mathcal{Z}) - \beta\right)^2 + \mathbb{E}_{\mathcal{Z}}\left(\frac{\text{Var}\left(Y(1)|\mathcal{Z}\right)}{n_{1,z}}\right) + \mathbb{E}_{\mathcal{Z}}\left(\frac{\text{Var}\left(Y(1)|\mathcal{Z}\right)}{n_{1,z}}\right), \tag{38}$$

$$= \mathbb{E}_{\mathcal{Z}}\left(\frac{\text{Var}\left(Y(1)|\mathcal{Z}\right)}{n_{1,z}} + \frac{\text{Var}\left(Y(0)|\mathcal{Z}\right)}{n_{0,z}}\right) + \mathbb{E}_{\mathcal{Z}}\left(\sum_{\mathcal{Z}} \beta(Z)\hat{\mathbb{P}}(\mathcal{Z}) - \beta\right)^2. \tag{39}$$

Appendix B. Details of the Experimental Evaluations

In this section, we provide details of the experimental evaluations.

B.1. Variable Definitions

For the semi-Markovian causal graph, $\mathcal{G}$ given in Figure 6, we define the treatment and response as follows.

$$\text{Drug} = \begin{cases} 1, & \text{if an individual receives the drug,} \\ 0, & \text{otherwise,} \end{cases}$$

$$\text{Blood-Pressure} = \begin{cases} 1, & \text{if an individual has high blood pressure,} \\ 0, & \text{otherwise,} \end{cases}$$

We define the observed covariates as follows.

$$\text{Food Habits} = \begin{cases} 1, & \text{if an individual's food habits are unhealthy,} \\ 0, & \text{otherwise,} \end{cases}$$

$$\text{Blood-Sugar} = \begin{cases} 1, & \text{if an individual has high blood sugar,} \\ 0, & \text{otherwise,} \end{cases}$$

$$\text{Age} = \begin{cases} 1, & \text{if an individual is above 40 years of age,} \\ 0, & \text{otherwise,} \end{cases}$$

$$\text{Alcohol} = \begin{cases} 1, & \text{if an individual consumes alcohol,} \\ 0, & \text{otherwise,} \end{cases}$$

$$\text{Cholesterol} = \begin{cases} 1, & \text{if an individual has high cholesterol,} \\ 0, & \text{otherwise,} \end{cases}$$

$$\text{Anxiety} = \begin{cases} 1, & \text{if an individual has anxiety,} \\ 0, & \text{otherwise,} \end{cases}$$

$$\text{Sleep Quality} = \begin{cases} 1, & \text{if an individual's sleep quality is bad,} \\ 0, & \text{otherwise,} \end{cases}$$

$$\text{Palpitations} = \begin{cases} 1, & \text{if an individual suffers from palpitations,} \\ 0, & \text{otherwise,} \end{cases}$$

$$\text{Strenuous Activity} = \begin{cases} 1, & \text{if an individual does strenuous activity,} \\ 0, & \text{otherwise.} \end{cases}$$

The graph $\mathcal{G}$ in Figure 6 has three unobserved variables. We label them as Anxiety $\dashleftarrow U_1 \dashrightarrow$ Alcohol, Anxiety $\dashleftarrow U_2 \dashrightarrow$ Blood-Pressure, and Blood-Pressure $\dashleftarrow U_3 \dashrightarrow$ Cholesterol. We consider each of U_1, U_2, and U_3 to be identically and independently distributed as Uniform$(-0.15, 0.15)$.

B.2. Probability Distributions

For the observed variables, we consider the following probability distributions. For $x \in \{0, 1\}$,

$$\begin{aligned}
\mathbb{P}\left(\text{Food Habits} = x\right) &= 0.5, \\
\mathbb{P}\left(\text{Age} = x\right) &= 0.5, \\
\mathbb{P}\left(\text{Alcohol} = x\right) &= \text{Food-Habits}[(0.3 + U_1)(1 - x) + (0.7 - U_1)x] \\
&\quad + (1 - \text{Food-Habits})[(0.3 + U_1)(1 - x) + (0.7 - U_1)x], \\
\mathbb{P}\left(\text{Cholesterol} = x\right) &= \text{Age}[(0.3 + U_3)(1 - x) + (0.7 - U_3)x] \\
&\quad + (1 - \text{Age})((0.3 + U_3)(1 - x) + (0.7 - U_3)x), \\
\mathbb{P}\left(\text{Anxiety} = x\right) &= \text{Drug}[(0.3 + U_2)(1 - x) + (0.7 - U_2)x] \\
&\quad + (1 - \text{Drug})((0.3 + U_2)(1 - x) + (0.7 - U_2)x), \\
\mathbb{P}\left(\text{Sleep Quality} = x\right) &= \text{Anxiety}[(0.3 + U_2)(1 - x) + (0.7 - U_2)x] \\
&\quad + (1 - \text{Anxiety})((0.3 + U_2)(1 - x) + (0.7 - U_2)x), \\
\mathbb{P}\left(\text{Blood-Pressure} = x\right) &= W[(0.3 + U_1 - U_2)(1 - x) + (0.7 - U_1 + U_2)x] \\
&\quad + (1 - W)[(0.3 - U_1 + U_2)(1 - x) + (0.7 + U_1 - U_2)x],
\end{aligned}$$

where $W = \text{Alcohol} \times \text{Cholesterol} \times \text{Anxiety} \times \text{Sleep-Quality}$.

B.3. Baselines

For empirical comparisons, we constructed the following baselines by choosing some representative subsets of the observed covariates.

1. ϕ: No blocking, i.e., a completely randomized design. We use this as a baseline to compare if our blocking set performs at least better than no blocking.

2. {Food Habits}: Food Habits affect the response (Blood-Pressure) through a mediator, Alcohol. We use {Food Habits} as the blocking set to see how using only a non-parent ancestor of the response compares with our blocking set.

Table 2: p-values for the difference between causal effect estimates.

p-value	{Food Habits}	{Food Habits, Alcohol}	{Food Habits, Alcohol, Cholesterol}	{Food Habits, Alcohol, Cholesterol, Age}
ϕ	0.9511	0.9389	0.8196	0.7103
{Food Habits}		0.9877	0.7723	0.7565
{Food Habits, Alcohol}			0.7606	0.7682
{Food Habits, Alcohol, Cholesterol}				0.5490

3. {Food Habits, Alcohol}: As discussed above Food Habits affect the response through Alcohol. We use {Food Habits, Alcohol} as the blocking set to see how only this parent and non-parent ancestor pair compare with our blocking set.

4. {Food Habits, Alcohol, Cholesterol}: This subset includes the above parent and non-parent ancestor pair, {Food Habits, Alcohol} and one more parent of the response, Cholesterol while leaving out Cholesterol's parent, Age. We use {Food Habits, Alcohol, Cholesterol} as the blocking set to see how leaving out Age in the presence of a latent variable path compares with our blocking set.

B.4. Simulation and Estimation

We performed the experiment with no blocking by randomizing the treatment (Drug) 100 times. For performing the experiment with blocking, we first created blocks using the chosen set of covariates and then randomized the treatment within each block. Using the above probability distributions, we generated/simulated 100 samples of the response (Blood-Pressure). We calculated the causal effect of the drug on blood pressure as the average of the response. We repeated the above experiment 10 times. The estimated causal effect and the variability in the response (averaged across 10 runs) for different choices of blocking sets are provided in Table 1.

B.5. Testing for Statistical Significance

We tested for the statistical significance of all pairwise differences between the estimates of the causal effect using different blocking sets using a t-test. The p-values are provided in Table 2. From Table 2, we observe that the p-values for all pairwise differences in causal effect estimates are quite high. Hence, we conclude that these differences are not significant. This confirms the idea that blocking does not alter the causal effect estimates. We tested for the statistical significance of all pairwise ratios of the estimates of the variability in the response using different blocking sets using an F-test. The p-values are provided in Table 3. From the last column of Table 3, we observe that the p-values for comparing the variance using our blocking set with other blocking sets are fairly low (maximum being 0.2567). We can say that the differences in variance are significant at a 25.67% level of significance.

Table 3: p-values for the ratio of variance estimates.

p-value	{Food Habits}	{Food Habits, Alcohol}	{Food Habits, Alcohol, Cholesterol}	{Food Habits, Alcohol, Cholesterol, Age}
ϕ	0.4738	0.4074	0.3818	0.1701
{Food Habits}		0.4331	0.4071	0.1872
{Food Habits, Alcohol}			0.4735	0.2385
{Food Habits, Alcohol, Cholesterol}				0.2567

Appendix C. Future Direction to use the Proposed Algorithm for Different Real-World Applications

In this section, we discuss some future directions to use the proposed algorithm for different real-world applications. The proposed algorithm efficiently obtains a stable and variance-minimizing set of covariates to be used for forming blocks while performing random experiments. Random experiments are commonly used to study the causal effect. For instance in answering the following questions. Does a new seed variety help in improving crop yield? Does a new medicine help in alleviating muscle pain? Does offering discounts help in product adoption? Does recommending a set of products improve the user-satisfaction? From a reinforcement learning standpoint, an intervention like a new seed variety, a new medicine, a new promotional offer, a new product recommendation, etc. can be seen as an action, and the response like the crop yield, reduction in muscle pain, number of product adoptions, user-satisfaction, etc. can be regarded as the reward. A common question in reinforcement learning is to find a reward-maximizing action that is equivalent to finding good interventions in randomized experiments. Hence, the algorithm proposed in this paper can be used in the efficient and low-variance estimation of causal effects to identify good actions/interventions in different real-world applications. We are interested in the applications of the proposed algorithm to offline (Kempe et al., 2003; Umrawal and Aggarwal, 2022) and online (Agarwal et al., 2022; Nie et al., 2022) social influence maximization – the task of selecting a set of individuals on a social network to be given some discounts to cause a large cascade of further product adoptions.