Journal of Machine Learning Research 27 (2026) 1-56 Submitted 5/25; Revised 6/26; Published 8/26

Clustering and Pruning in Causal Data Fusion

Otto Tabell OTTO.H.E.TABELL@JYU.FI
Santtu Tikka SANTTU.TIKKA@JYU.FI
Juha Karvanen JUHA.T.KARVANEN@JYU.FI
Department of Mathematics and Statistics
FI-40014 University of Jyvaskyla, Finland

Editor: Jin Tian

Abstract

Data fusion—the process of combining observational and experimental data—can enable the identification of causal effects that would otherwise remain non-identifiable. Although identification algorithms have been developed for specific scenarios, do-calculus remains the only general-purpose tool for causal data fusion, particularly when variables are present in some data sources but not others. However, approaches based on do-calculus may encounter computational challenges as the number of variables increases and the causal graph grows in complexity. Consequently, there exists a need to reduce the size of such models while preserving the essential features. For this purpose, we propose pruning (removing unnecessary variables) and clustering (combining variables) as preprocessing operations for causal data fusion. We generalize earlier results on a single data source and derive conditions for applying pruning and clustering in the case of multiple data sources. We give sufficient conditions for inferring the identifiability or non-identifiability of a causal effect in a larger graph based on a smaller graph and show how to obtain the corresponding identifying functional for identifiable causal effects. Examples from epidemiology and social science demonstrate the use of the results.

Keywords: causal inference, graph theory, identifiability, identification invariance, multiple data sources

1. Introduction

The modern data landscape integrates data from multiple sources such as experiments, surveys, registries, and operational systems. In causal data fusion, the goal is to combine these data sources to identify a causal effect that cannot be identified from a single input distribution (Bareinboim and Pearl, 2016; Shi et al., 2023; Colnet et al., 2024). We call this the general causal effect identification problem. The potential application areas include, for instance, medicine (Dahabreh et al., 2023; Li et al., 2020), ecology (Spake et al., 2022) and econometrics (Hünermund and Bareinboim, 2023).

The theoretical understanding of the general causal effect identification problem is still partial. While do-calculus (Pearl, 1995) is complete for causal effect identification with a single observational data source (Shpitser and Pearl, 2006; Huang and Valtorta, 2006), a similar result has not been derived in the context of multiple data sources of unrestricted form.

Algorithms for special cases of causal data fusion have been reviewed by Tikka et al. (2021, Table 1). The more recent developments include the corrected g-identifiability algorithm with a positivity assumption (Kivva et al., 2022) and an algorithm for data fusion under systematic selection (Lee et al., 2024). All of these algorithms have important restrictions regarding their input distributions. For instance, g-identifiability (Lee et al., 2020b; Kivva et al., 2022) and g-transportability (Lee et al., 2020a) are constrained by the assumption that entire distributions are available, meaning that the union of observed and intervened variables must always be the set of all endogenous variables in the model. As an example, this means that besides direct application of do-calculus, there does not currently exist an identification algorithm that could determine the identifiability of $p(y \mid \mathrm{do}(x))$ in the graph $X \to Z \to Y$ from the input distributions $p(x, z)$ and $p(z, y)$.

As a real-data example of these kinds of partially overlapping input distributions, Karvanen et al. (2020) studied a causal data fusion problem where one data source contains information on salt-adding behavior and salt intake, while another data source contains information on salt intake and blood pressure. The goal was to estimate the causal effect of salt-adding behavior on blood pressure. Similar examples on partially overlapping data sources can be found in the Maelstrom Catalogue (Maelstrom Research, 2025; Bergeron et al., 2018) that provides metadata on epidemiological studies. For instance, the catalogue lists 25 studies where some information on chronic obstructive pulmonary disease (COPD) has been measured. Out of these studies, 20 have data on physical measurements, 22 have data from biosamples, 11 have data on cognitive measures, and 10 have data from administrative databases; only two studies have data from all four categories.

For more general cases with no constraints on the input distributions, Do-search (Tikka et al., 2021) has been suggested. Do-search is an algorithm that applies do-calculus and standard probability manipulations to determine the identifiability of a causal effect with given input distributions without constraints. However, as a consequence of the search-based nature of the algorithm, it becomes computationally expensive as the number of variables increases. The poor scalability of search-based identification algorithms also motivates the preprocessing efforts to reduce the size of the causal graph and thus decrease the number of variables.

Importantly, if a causal graph is modified in any way, there is no guarantee that inferences made in the modified graph are transferable back to the original graph. Ideally, a method to reduce the size of causal graphs should not affect the identifiability of the causal effect of interest but such methods are scarce. We focus on two approaches for graph-size reduction: clustering and pruning.

In clustering, the aim is to represent a larger set of, typically related, vertices in a graph as a smaller set of vertices (Schaeffer, 2007; Malliaros and Vazirgiannis, 2013). There are several examples of finding clusters and implementing clustering in causal graphs (Anand et al., 2023; Zeng et al., 2025; Niu et al., 2022). However, due to the coarse graphical representations of clusters, these approaches may lead to a loss of identifiability, i.e., a causal effect may be identifiable in the clustered graph but not in the original graph. To remedy this issue, Tikka et al. (2023) introduced transit clusters which were proven to preserve the identifiability under certain conditions.

In addition, Tikka and Karvanen (2018) introduced pruning, which refers to the process of completely removing vertices that are irrelevant for identification. Such vertices can, for

instance, include variables that are descendants of the response variable or vertices that are connected to the causal graph through only one edge. Illustrations of pruning and clustering are presented in Figure 1. The existing identifiability results regarding pruning and clustering have been presented in the context of a single observational data source, and are thus not directly applicable to causal data fusion.

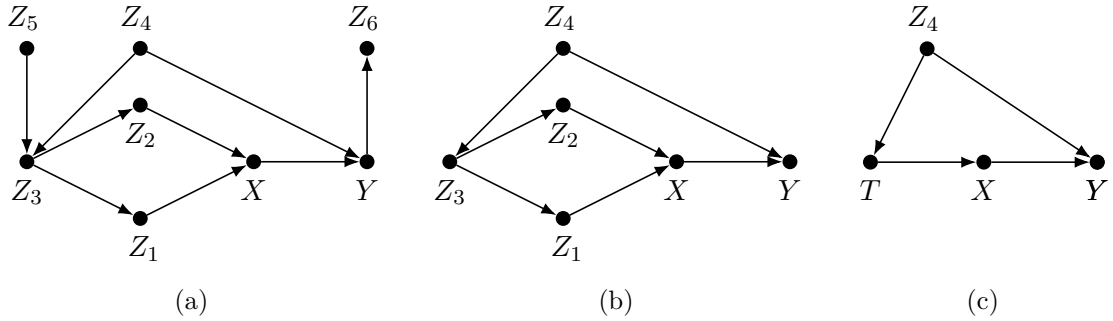

Figure 1: Illustrations of size-reduction techniques for causal graphs: (a) the original graph, (b) a pruned graph where Z_5 and Z_6 have been deemed irrelevant for the identification of $p(y \mid \mathrm{do}(x))$ and thus removed, (c) a clustered graph obtained from the pruned graph where T represents the set $\{Z_1, Z_2, Z_3\}$.

In this paper, we derive sufficient conditions under which pruning and clustering can be applied within the general causal effect identification framework. We also show how identifying functionals obtained from a pruned or clustered graph can be used to derive identifying functionals for the original graph. Our results generalize the results by Tikka and Karvanen (2018) and Tikka et al. (2023) to the case of multiple data sources. These generalizations are not straightforward because previous results on identification of causal effects regarding pruning and clustering rely on concepts that necessitate a clear distinction between observed and unobserved variables, such as c-components (districts), hedges, and the latent projection. These concepts are not well-defined when entire distributions are not available.

By pruning and clustering, the computational load of algorithms used in the causal effect identification can be reduced, and the identifying functionals for the causal effects can be presented more concisely. Another perspective to clustering is top-down causal modeling where causal relations are first modeled between clusters of variables. This kind of approach is intuitive and commonly applied in practice (Tennant et al., 2021, and references therein). Top-down causal modeling has also been applied with multiple data sources (Valkonen et al., 2024; Youngmann et al., 2023; Karvanen et al., 2020). The results derived in this paper can be used to determine when non-identifiability in a clustered graph implies non-identifiability in a graph without clusters, and when more detailed modeling of the internal structure of the clusters could be beneficial for the identification.

The paper is organized as follows. In Section 2, we introduce the notations and core definitions such as the structural causal model and the identifiability of causal effects. In Sections 3 and 4, we present our main results on causal effect identifiability with multiple data sources under pruning and clustering, respectively. In Section 5, we show that identifying

functionals obtained from pruned or clustered causal graphs are applicable in the original causal graph. In Section 6, we perform a simulation study to assess the impact of pruning and clustering on running time performance of Do-search. In Section 7, we illustrate our results via examples that originate from epidemiology and social science. Finally, Section 8 offers some concluding remarks. All identifying functionals of the paper were obtained using R Statistical Software version 4.5.1 (R Core Team, 2025) and the `dosearch` package (Tikka et al., 2021, 2024). The R codes for the simulation study and the examples of the paper are available at `https://github.com/ottotabell/clustering-pruning`.

2. Notations and Definitions

We use upper case letters to denote random variables and vertices, and lower case letters to denote values. The shortcut notation $p(x)$ is used for the probability $P(X = x)$. We use bold letters to denote sets. We assume the reader is familiar with basic concepts of graph theory such as directed acyclic graphs (DAGs). We make no distinction between random variables and vertices and use the same symbols to refer to both. A DAG over a set of vertices $\mathbf{V}$ is denoted by $\mathcal{G}(\mathbf{V})$. We omit braces for sets consisting of only one element and write e.g., $\mathbf{V} \setminus T$ as a shortcut notation for $\mathbf{V} \setminus \{T\}$.

Let $\mathcal{G}(\mathbf{V})$ be a DAG and let $\mathbf{W} \subset \mathbf{V}$. The induced subgraph $\mathcal{G}[\mathbf{W}]$ is a DAG obtained from $\mathcal{G}$ by keeping all vertices in $\mathbf{W}$ and all edges between members of $\mathbf{W}$ in $\mathcal{G}$. We use the notation $\mathcal{G}[\overline{\mathbf{W}}, \underline{\mathbf{Z}}]$ to denote the DAG obtained from $\mathcal{G}$ by removing all incoming edges of $\mathbf{W}$ and outgoing edges of $\mathbf{Z}$.

Following (Pearl, 2009), we define structural causal models as follows:

Definition 1 (Structural causal model)**.** *A* structural causal model *(SCM) $\mathcal{M}$ is a tuple $(\mathbf{V}, \mathbf{U}, \mathbf{F}, p)$, where*

- $\mathbf{V}$ *and* $\mathbf{U}$ *are disjoint sets of random variables.*
- $\mathbf{F}$ *is a set of functions such that each $f_i \in \mathbf{F}$ is a mapping from $\mathbf{U}[V_i] \cup \mathrm{pa}(V_i)$ to V_i, where $\mathbf{U}[V_i] \subset \mathbf{U}$ and $\mathrm{pa}(V_i) \subset \mathbf{V} \setminus V_i$, such that $\mathbf{F}$ forms a mapping from $\mathbf{U}$ to $\mathbf{V}$. In other words, each V_i is defined by the structural equation $V_i = f_i(\mathrm{pa}(V_i), \mathbf{U}[V_i])$.*
- *p is the joint probability distribution of $\mathbf{U}$.*

We only consider recursive SCMs, i.e., models where the unique solution of $\mathbf{F}$ in terms of $\mathbf{U}$ is obtained by recursively substituting each $V_j \in \mathrm{pa}(V_i)$ in each function f_i with their corresponding function f_j. It is common to characterize members of $\mathbf{V}$ as "observed" or "endogenous", but because we consider scenarios with multiple data sources where the set of observations is not necessarily the same between the inputs, we will not use these terms. Similarly, members of $\mathbf{U}$ are commonly referred to as "unobserved", "latent" or "exogenous", which we will generally avoid. In other words, the only distinction between $\mathbf{V}$ and $\mathbf{U}$ is that the former is defined by the structural equations of $\mathbf{F}$ and the latter by the joint distribution $p(\mathbf{u})$. Finally, we extend the notation $\mathbf{U}[V_i]$ to subsets $\mathbf{W} \subset \mathbf{V}$ disjunctively as $\mathbf{U}[\mathbf{W}] = \bigcup_{W_i \in \mathbf{W}} \mathbf{U}[W_i]$.

The DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{U})$ induced by an SCM $\mathcal{M}$ is called a *causal graph* and it has a vertex for each member of $\mathbf{V} \cup \mathbf{U}$, and there is a directed edge from each member of $\mathrm{pa}(V_i) \cup \mathbf{U}[V_i]$

to V_j for each V_j. A member of $\mathbf{U}$ is called a *dedicated error term* if it is present in exactly one structural equation, i.e., it has exactly one child in the corresponding causal graph (Karvanen et al., 2024). The set of dedicated error terms is denoted as $\overline{\mathbf{U}}$, and complementarily $\widetilde{\mathbf{U}} = \mathbf{U} \setminus \overline{\mathbf{U}}$ denotes members of $\mathbf{U}$ with more than one child. The dedicated error term for a variable $V_i \in \mathbf{V}$ is denoted by $\overline{U}[V_i]$. In all figures, dedicated error terms are omitted for clarity. Moreover, members of $\widetilde{\mathbf{U}}$ that have exactly two children are represented by dashed bi-directed edges in the figures. Starting from Section 4 where we introduce clustering, we will also consider causal graphs where the dedicated error terms are explicitly omitted, i.e., DAGs of the form $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$, because it is natural to assume that the dedicated error terms are always clustered together with their children.

The sets of parents and ancestors of a vertex $V_i \in \mathbf{V} \cup \mathbf{U}$ in a causal graph $\mathcal{G}(\mathbf{V} \cup \mathbf{U})$ are denoted by $\mathrm{pa}^+_{\mathcal{G}}(V_i)$ and $\mathrm{an}^+_{\mathcal{G}}(V_i)$, respectively. When we wish to include only those members of the aforementioned sets that are members of $\mathbf{V}$, we use notations $\mathrm{pa}_{\mathcal{G}}(V_i)$ and $\mathrm{an}_{\mathcal{G}}(V_i)$, respectively. The sets of children and descendants are denoted by $\mathrm{ch}_{\mathcal{G}}(V_i)$ and $\mathrm{de}_{\mathcal{G}}(V_i)$, respectively. All of these notations are also defined for sets of vertices by taking the union over the set elements, e.g., $\mathrm{an}_{\mathcal{G}}(\mathbf{W}) = \bigcup_{W_i \in \mathbf{W}} \mathrm{an}_{\mathcal{G}}(W_i)$.

The d-separation criterion can be used to infer conditional independence constraints of a joint distribution that is Markov-relative to a DAG $\mathcal{G}$ (Pearl, 1995, 2009), such as the joint distribution induced by a recursive SCM.

Definition 2 (d-separation)**.** *Let $\mathcal{G}(\mathbf{V})$ be a DAG and let $X, Y \in \mathbf{V}$. A path $\mathcal{P}$ between X and Y in $\mathcal{G}$ is* d-separated *by $\mathbf{Z} \subset (\mathbf{V} \setminus \{X, Y\})$ in $\mathcal{G}$ if*

- *$\mathcal{P}$ contains a chain $A \to M \to B$ or a fork $A \leftarrow M \to B$ such that $M \in \mathbf{Z}$, or*
- *$\mathcal{P}$ contains a collider $A \to M \leftarrow B$ such that $(M \cup \mathrm{de}_{\mathcal{G}}(M)) \cap \mathbf{Z} = \emptyset$.*

In addition, X and Y are d-separated by $\mathbf{Z}$ if $\mathbf{Z}$ d-separates every path between X and Y. If X and Y are not d-separated by $\mathbf{Z}$, they are d-connected*. For sets, $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$ if $\mathbf{Z}$ d-separates every path between the members of $\mathbf{X}$ and $\mathbf{Y}$.*

We denote the d-separation of $\mathbf{X}$ and $\mathbf{Y}$ given $\mathbf{Z}$ in a DAG $\mathcal{G}$ as $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_{\mathcal{G}}$ and the implied conditional independence as $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}$.

Interventions by the do-operator (Pearl, 2009) induce new causal models, called submodels.

Definition 3 (Submodel)**.** *A* submodel *$\mathcal{M}_{\mathbf{x}}$ of an SCM $\mathcal{M}$ obtained from an intervention* $\mathrm{do}(\mathbf{x})$ *is an SCM $(\mathbf{V}, \mathbf{U}, \mathbf{F}', p)$ where $\mathbf{F}'$ is obtained from $\mathbf{F}$ by replacing $f_X \in \mathbf{F}$ for each $X \in \mathbf{X}$ with a constant function that outputs the corresponding value in $\mathbf{x}$.*

An *interventional distribution* $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is the marginal distribution of $\mathbf{Y}$ in $\mathcal{M}_{\mathbf{x}}$, which we also refer to as the *causal effect* of $\mathbf{X}$ on $\mathbf{Y}$.

In the general causal effect identification problem, the available information can be represented symbolically as a set of *input distributions* $\mathcal{I} = \{p(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)\}_{i=1}^n$ where the sets $\mathbf{A}_i, \mathbf{B}_i$, and $\mathbf{C}_i$ are disjoint and $\mathbf{A}_i \cup \mathbf{B}_i \cup \mathbf{C}_i \subset \mathbf{V}$ for each $i = 1, \ldots, n$, where n is the number of input distributions. In this notation, $\mathbf{A}_i$ represents the measured variables, $\mathbf{B}_i$ represents the intervened variables, and $\mathbf{C}_i$ represents the conditioning variables. For input

distributions, the notation $p(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$ implies that the probabilities are known for all possible values of $\mathbf{A}_i$, $\mathbf{B}_i$ and $\mathbf{C}_i$.

For settings that involve multiple data sources, we define the identifiability of a causal effect as follows:

Definition 4 (Identifiability)**.** *Let $\mathcal{G}(\mathbf{V} \cup \mathbf{U})$ be a causal graph and let $\mathbf{Y}, \mathbf{X} \subset \mathbf{V}$ be disjoint. The causal effect $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is* identifiable *from $(\mathcal{G}, \mathcal{I})$ if it is uniquely computable from $\mathcal{I}$ in every SCM that induces $\mathcal{G}$ and fulfills the positivity assumption $p(\mathbf{v}) > 0$.*

The positivity assumption $p(\mathbf{v}) > 0$ ensures that all input distributions and causal effects are always well-defined. The typical way to apply Definition 4 is to demonstrate non-identifiability by constructing two models that agree on $\mathcal{I}$ but disagree on $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$. In many proofs of this paper, we construct two models, $\mathcal{M}_1$ and $\mathcal{M}_2$ for a graph $\mathcal{G}$ based on models $\mathcal{M}'_1$ and $\mathcal{M}'_2$ that are known to exist in an altered graph $\mathcal{G}'$ due to the non-identifiability of the causal effect in this altered graph or vice versa.

Manual identification of causal effects relies on *do-calculus* (Pearl, 1995), which consists of three rules for manipulating interventional distributions:

1. Insertion and deletion of observations:

$$p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}) = p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}), \mathbf{w}), \text{ if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \,|\, \mathbf{X}, \mathbf{W})_{\mathcal{G}[\overline{\mathbf{X}}]}.$$

2. Exchanging actions and observations:

$$p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}, \mathbf{z}), \mathbf{w}) = p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}), \text{ if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \,|\, \mathbf{X}, \mathbf{W})_{\mathcal{G}[\overline{\mathbf{X}}, \underline{\mathbf{Z}}]}.$$

3. Insertion and deletion of actions:

$$p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}, \mathbf{z}), \mathbf{w}) = p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}), \mathbf{w}), \text{ if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \,|\, \mathbf{X}, \mathbf{W})_{\mathcal{G}[\overline{\mathbf{X}}, \overline{Z(\mathbf{W})}]},$$

   where $Z(\mathbf{W}) = \mathbf{Z} \setminus \mathrm{an}(\mathbf{W})_{\mathcal{G}[\overline{\mathbf{X}}]}$.

When we consider the identifiability of causal effects outside the restricted settings of known identifiability algorithms, do-calculus remains the only tool available for demonstrating that a causal effect is identifiable. The positivity assumption of Definition 4 also ensures the validity of the rules of do-calculus.

Clustering and pruning are operations that transform the causal graph and the input distribution in a specific way. We are interested in finding conditions under which these operations preserve identifiability and non-identifiability. We define this property formally as follows:

Definition 5 (Identification invariance)**.** *Let $\mathcal{G}$ be a causal graph and let $\mathcal{I}$ be a set of input distributions. An operation Ω is* identification invariant *for a causal effect $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ if $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is identifiable from $(\mathcal{G}, \mathcal{I})$ if and only if $p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is identifiable from $(\mathcal{G}', \mathcal{I}')$ where $(\mathcal{G}', \mathcal{I}') = \Omega(\mathcal{G}, \mathcal{I})$.*

In the definition, p refers to any distribution that is Markov-relative to $\mathcal{G}$, and p' refers to any distribution that is Markov-relative to $\mathcal{G}'$. We note that identification invariance establishes an equivalence class of causal graphs and input distributions in the sense that if we are able to determine whether a causal effect is identifiable from $(\mathcal{G}', \mathcal{I}')$, then we will also know whether it is identifiable from any $(\mathcal{G}, \mathcal{I})$ such that $(\mathcal{G}', \mathcal{I}') = \Omega(\mathcal{G}, \mathcal{I})$. In this paper, we focus on pruning and clustering which are examples of operations that produce coarser graphical representations of the system under study. We note that there are also operations that produce more refined graphical representations, the peripheral extension (Tikka et al., 2023) being an example.

3. Preprocessing by Pruning

It is sometimes possible to simplify the causal model by completely removing variables that are unnecessary for the identification of a causal effect of interest. This operation is called *pruning* by Tikka and Karvanen (2018) who studied pruning in the case of a single observational data source. Ideally, the pruning operation should be identification invariant.

We present three results regarding pruning in the case of multiple data sources. The first theorem states that non-ancestors of $\mathbf{Y}$ can be removed without affecting the identifiability of $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$, as long as none of the input distributions are conditioned or intervened by non-ancestors of $\mathbf{Y}$. The second theorem states conditions for removing variables that are d-separated from $\mathbf{Y}$ after an intervention on $\mathbf{X}$. The third theorem gives similar conditions for removing variables that are connected to the rest of the graph only through a single variable. The two latter theorems assume that non-ancestors of $\mathbf{Y}$ have already been pruned. Our results generalize those by Tikka and Karvanen (2018) to multiple data sources. The proofs of this section are presented in Appendix A.

First we define how pruning changes the input distributions. If the pruning reduces the set of variables from $\mathbf{V}$ to $\mathbf{W} \subset \mathbf{V}$, the input distributions are restricted in the following way:

Definition 6 (Restricted input distributions)**.** *Let* $\mathcal{I} = \{p(\mathbf{a}_i \mid \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)\}_{i=1}^n$ *be a set of input distributions and let* $\mathbf{W} \subset \mathbf{V}$*. The set of* input distributions restricted to $\mathbf{W}$ *is*

$$\mathcal{I}[\mathbf{W}] = \{p(\mathbf{a}_i \cap \mathbf{w} \mid \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) : p(\mathbf{a}_i \mid \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) \in \mathcal{I},\ \mathbf{A}_i \cap \mathbf{W} \neq \emptyset,\ \mathbf{B}_i \cup \mathbf{C}_i \subseteq \mathbf{W}\}.$$

Definition 6 ensures that the restricted input distributions are well-defined and excludes inputs that are not obtainable from the original inputs $\mathcal{I}$ directly through marginalization.

Definition 6 may appear too limited at first glance because $\mathcal{I}$ does not necessarily uniquely represent the available information. For example, if $\mathcal{I}_1 = \{p(x), p(y \mid x)\}$, $\mathcal{I}_2 = \{p(x \mid y), p(y)\}$, and $\mathcal{I}_3 = \{p(x, y)\}$, then each of these three sets of input distributions represents the same information (the joint distribution $p(x, y)$) but according to Definition 6 we have $\mathcal{I}_1[\{Y\}] = \emptyset$ while $\mathcal{I}_2[\{Y\}] = \mathcal{I}_3[\{Y\}] = \{p(y)\}$. However, there are good reasons why Definition 6 is given as it is. Assume that we would consider all distributions that can be derived from the original input distributions via standard probability calculus. The number of these distributions increases quickly when the number of inputs and the number of variables increase. Even more distributions can be derived if we consider conditional independence relations implied by the graph and apply the rules of do-calculus. Another argument for the current definition

is that we wish pruning to be a genuine preprocessing step meaning that there is no reason to use the variables that are removed in pruning. For example, if we have datasets 1 and 2 that provide information on $p(x)$ and $p(y \mid x)$, respectively, we do not have direct access to data from $p(y)$ but we should, for instance, resample the rows of dataset 2 using the probabilities $\hat{p}(x)$ estimated from dataset 1 as weights. To summarize, an analyst wishing to extend the possibilities for pruning has to explicitly derive new input distributions before the pruning is done. The formulation of Definition 6 is a design choice that deliberately favors simplicity. Alternative views are also possible, such as considering the input distributions from a purely symbolic perspective and allowing pruning whenever a joint distribution is derivable from the inputs. Such an approach would extend the applicability of pruning in theory while ignoring aspects related to practical data analysis.

Theorem 7 (Pruning non-ancestors)**.** *Let $\mathcal{G}(\mathbf{V} \cup \mathbf{U})$ be a causal graph. If $\mathbf{B}_i \cup \mathbf{C}_i \subset \mathrm{an}_{\mathcal{G}}(\mathbf{Y})$ for all $i = 1, \ldots, n$, then pruning $(\mathcal{G}, \mathcal{I}) \to (\mathcal{G}', \mathcal{I}')$, where $\mathcal{I}' = \mathcal{I}[\mathrm{an}_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y}]$, $\mathcal{G}' = \mathcal{G}[\mathrm{an}^{+}_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y}]$ and $\mathbf{X}' = \mathbf{X} \cap \mathrm{an}_{\mathcal{G}}(\mathbf{Y})$, is identification invariant for $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$, and $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) = p'(\mathbf{y} \mid \mathrm{do}(\mathbf{x}'))$.*

Theorem 7 allows us to only consider graphs $\mathcal{G} = \mathcal{G}[\mathrm{an}^{+}_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y}]$ where the non-ancestors of the response variables $\mathbf{Y}$ have been removed. In the remainder of the paper, we assume that the graphs have already been pruned this way.

We highlight the nuances of the assumption $\mathbf{B}_i \cup \mathbf{C}_i \subset \mathrm{an}_{\mathcal{G}}(\mathbf{Y})$ via two examples. First, consider the graph of Figure 2a where Z is not an ancestor of Y. The variable Z is necessary for the identification of $p(y \mid \mathrm{do}(x))$ if the input distributions are $p(z)$ and $p(x, y \mid z)$. On the other hand, one might be tempted to represent the input distributions simply as $p(x, y, z)$, which would satisfy the assumption and thus seems to permit the pruning of Z. This is, however, a misguided conclusion because forming the input distribution $p(x, y)$ requires the use of Z. In the graph of Figure 2b, $p(y \mid \mathrm{do}(x))$ can be identified from the input distribution $p(x, y \mid z)$ directly as $p(y \mid \mathrm{do}(x)) = p(y \mid x, z)$ because Z and Y are independent when X is intervened. Here, the input distribution $p(x, y \mid z)$ is necessary for the identification but the actual value of Z in the identifying functional $p(y \mid x, z)$ is irrelevant.

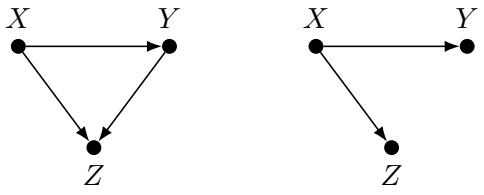

Figure 2: Causal graphs demonstrating the necessity of the assumption $\mathbf{B}_i \cup \mathbf{C}_i \subset \mathrm{an}_{\mathcal{G}}(\mathbf{Y})$ for pruning non-ancestors.

The second pruning theorem considers the removal of variables that are independent of the response after an intervention.

Theorem 8 (Pruning non-ancestors after intervention)**.** *Let $\mathcal{G}(\mathbf{V} \cup \mathbf{U})$ be a causal graph such that $\mathcal{G} = \mathcal{G}[\mathrm{an}^+_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y}]$ and let $\mathbf{Z} = \{Z \in \mathbf{V} \setminus \mathbf{X} : Z \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{X}$ in $\mathcal{G}[\overline{\mathbf{X}}]\}$, $\mathbf{V}' = \mathbf{V} \setminus \mathbf{Z}$ and $\mathbf{U}' = \mathbf{U}[\mathbf{V}']$. If all of the following conditions hold*

(a) $\mathbf{Z} \cap \mathrm{de}_{\mathcal{G}}(\mathbf{X}) = \emptyset$,

(b) $\mathbf{B}_i \cup \mathbf{C}_i \subset \mathbf{V}'$ *for all* $i = 1, \ldots, n$, *and*

(c) for all members of $\mathbf{U}[\mathbf{Z}]$ *it holds that if* $U \in \mathbf{U}[\mathbf{Z}]$ *is an ancestor of all variables in* $\mathbf{X}_S \subseteq (\mathbf{X} \cap \mathrm{ch}_{\mathcal{G}}(\mathbf{Z} \cup \mathbf{U}[\mathbf{Z}]))$, *then there exists* $U_S \in \mathbf{U}$ *that is a parent of all members of* $\mathbf{X}_S$,

then pruning $(\mathcal{G}, \mathcal{I}) \to (\mathcal{G}', \mathcal{I}')$, where $\mathcal{I}' = \mathcal{I}[\mathbf{V}']$ and $\mathcal{G}' = \mathcal{G}[\mathbf{V}' \cup \mathbf{U}']$, is identification invariant for $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$, and $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) = p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$.

In Theorem 8, the set of pruned variables, $\mathbf{Z}$, is defined to be d-separated from the response $\mathbf{Y}$ when $\mathbf{X}$ is intervened. The conditions (a), (b), and (c) are needed to ensure that the essential properties of the graph do not change when $\mathbf{Z}$ is pruned. Condition (a) ensures that the members of $\mathbf{Z}$ cannot be on a directed path between members of $\mathbf{X}$ because then the removal of $\mathbf{Z}$ could change the relationships between members of $\mathbf{X}$. Condition (b) says that $\mathbf{Z}$ cannot contain variables that are a part of intervention or conditioning in some input distribution. Condition (c) ensures that the d-separation properties between the members of $\mathbf{X}$ do not change when $\mathbf{Z}$ and $\mathbf{U}[\mathbf{Z}]$ are removed.

Our final pruning result relates to a scenario where a set of vertices is connected to the rest of the causal graph via a single vertex.

Theorem 9 (Pruning isolated vertices)**.** *Let $\mathcal{G}(\mathbf{V} \cup \mathbf{U})$ be a causal graph such that $\mathcal{G} = \mathcal{G}[\mathrm{an}^+_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y}]$ and let $W \in \mathbf{V}$. If there exists a set $\mathbf{Z} \subset \mathbf{V}$ such that $\mathbf{Z} \cap (\mathbf{X} \cup \mathbf{Y}) = \emptyset$, $\mathbf{Z}$ is connected to $\mathbf{V} \setminus \mathbf{Z}$ only through W, and $\mathbf{B}_i \cup \mathbf{C}_i \subset \mathbf{V} \setminus \mathbf{Z}$ for all $i = 1, \ldots, n$, then pruning $(\mathcal{G}, \mathcal{I}) \to (\mathcal{G}', \mathcal{I}')$, where $\mathcal{I}' = \mathcal{I}[\mathbf{V}']$, $\mathcal{G}' = \mathcal{G}[\mathbf{V}' \cup \mathbf{U}']$ and $\mathbf{V}' = \mathbf{V} \setminus \mathbf{Z}$ and $\mathbf{U}' = \mathbf{U}[\mathbf{V}']$, is identification invariant for $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$, and $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) = p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$.*

In some instances, it is possible to identify a causal effect using only a subset of the input distributions. In such cases, the conditions of Theorems 7–9 need to hold only for the members of this subset, thus expanding the applicability of these theorems. We will explain this feature in greater detail in Section 5.

To illustrate the results of this section, we consider the graph of Figure 3a with input distributions of $p(y, z_3, z_4, z_5 \,|\, \mathrm{do}(w_1), w_2)$, $p(w_1, z_1, z_2, z_5, z_7 \,|\, \mathrm{do}(x_1, x_2), w_2)$, and $p(w_2, z_6, z_7)$, and aim to identify $p(y \,|\, \mathrm{do}(x_1, x_2))$. First, Theorem 7 is applied to remove Z_4 and Z_5 because they are not ancestors of Y. We can see that Z_1, Z_2, and Z_3 are d-separated from Y when the incoming edges to X_1 and X_2 are removed. As $\mathbf{Z} = \{Z_1, Z_2, Z_3\}$ fulfills the conditions of Theorem 8, we can prune this set next. Note that the unobserved confounder between X_1 and X_2 is necessary for the application of Theorem 8 because X_1 and X_2 have common ancestors in $\mathbf{U}[\mathbf{Z}]$. Finally, we apply Theorem 9 to remove Z_6 and Z_7. Figure 3b presents the pruned version of the graph.

Applying do-calculus to the pruned graph and the corresponding input distributions $p(y \,|\, \mathrm{do}(w_1), w_2)$, $p(w_1 \,|\, \mathrm{do}(x_1, x_2), w_2)$ and $p(w_2)$, we obtain the identifying functional

$$p(y \,|\, \mathrm{do}(x_1, x_2)) = \sum_{w_1, w_2} p(y \,|\, \mathrm{do}(w_1), w_2) p(w_1 \,|\, \mathrm{do}(x_1, x_2), w_2) p(w_2).$$

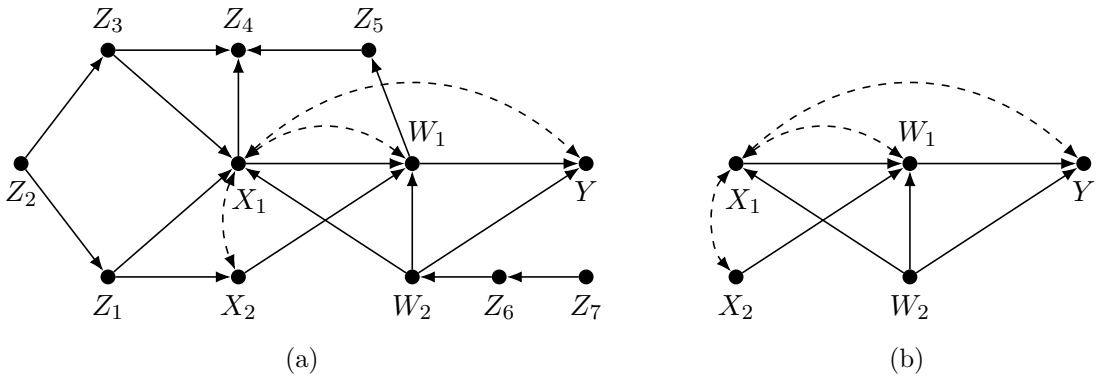

Figure 3: Graphs related to the example on pruning with multiple data sources: (a) the original graph, (b) the pruned graph obtained by applying Theorems 7, 8, and 9.

In this example, all three data sources are necessary, as the causal effect is not identifiable if any of the input distributions is unavailable. It is easy to see that $p(y \,|\, \mathrm{do}(x_1, x_2))$ cannot be identified if Y, X_1, or X_2 is not present in any input distribution, and the necessity of $p(w_2)$ follows from the fact that the other two input distributions are conditioned on W_2 and cannot be made independent of W_2.

4. Clustering Variables in Causal Graphs

Clustering is a method of preprocessing, in which a group of vertices in a causal graph is instead represented by a smaller set of vertices. In particular, we are interested in clusters that retain identifiability and non-identifiability of causal effects between the original graph and the clustered graph. Clustering can be especially useful when the causal graph consists of multiple vertices with similar "roles" regarding the causal effect to be identified, such as variables related to the environment or demographics.

4.1 Basic Concepts of Clustering

We define clustering as an operation that takes a causal graph $\mathcal{G}$ and a set of vertices $\mathbf{T}$ as inputs and produces a new graph where $\mathbf{T}$ is instead represented by a single vertex T. In addition, as we are clustering DAGs induced by causal models, a variable being clustered must always be clustered together with its dedicated error term. For this purpose, we will explicitly consider graphs where the dedicated error terms have been omitted, i.e., causal graphs over $\mathbf{V} \cup \widetilde{\mathbf{U}}$.

Definition 10 (Clustering)**.** *Let $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ be a causal graph and let $\mathbf{T} \subset \mathbf{V} \cup \widetilde{\mathbf{U}}$.* Clustering *of $\mathbf{T}$ in $\mathcal{G}$ induces a graph $\mathcal{G}'(\mathbf{V}' \cup \widetilde{\mathbf{U}}')$ where $\mathbf{V}' = (\mathbf{V} \setminus \mathbf{T}) \cup T$ and $\widetilde{\mathbf{U}}' = (\widetilde{\mathbf{U}} \setminus \mathbf{T})$, that is obtained from $\mathcal{G}$ by replacing $\mathbf{T}$ with a new vertex T whose parents and children in $\mathcal{G}'$ are $\mathrm{pa}^{+}_{\mathcal{G}}(\mathbf{T}) \setminus \mathbf{T}$ and $\mathrm{ch}_{\mathcal{G}}(\mathbf{T}) \setminus \mathbf{T}$, respectively.*

We denote the graph $\mathcal{G}'$ induced by clustering $\mathbf{T}$ into T in $\mathcal{G}$ by $\mathcal{G}[\mathbf{T} \to T]$. Importantly, we always consider T as a member of $\mathbf{V}'$ in the causal model whose induced causal graph is

$\mathcal{G}[\mathbf{T} \to T]$ regardless of whether $\mathbf{T}$ contains members of $\widetilde{\mathbf{U}}$. We also note that the above definition is very general, and the resulting graph may not be a DAG. For these reasons, it is important to focus on a restricted class of clusters.

We concentrate on transit clusters (Tikka et al., 2023) which can be intuitively explained as structures where information is transmitted from the receivers of the cluster to the emitters of the cluster. We define receivers and emitters as follows.

Definition 11 (Receiver)**.** *For a set of vertices $\mathbf{T} \subset \mathbf{V} \cup \widetilde{\mathbf{U}}$ in a causal graph $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$, the set of* receivers *is the set*

$$\mathrm{re}_{\mathcal{G}}(\mathbf{T}) = \{R \in \mathbf{T} \mid \mathrm{pa}^{+}_{\mathcal{G}}(R) \cap ((\mathbf{V} \cup \widetilde{\mathbf{U}}) \setminus \mathbf{T}) \neq \emptyset\}.$$

Definition 12 (Emitter)**.** *For a set of vertices $\mathbf{T} \subset \mathbf{V} \cup \widetilde{\mathbf{U}}$ in a causal graph $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$, the set of* emitters *is the set*

$$\mathrm{em}_{\mathcal{G}}(\mathbf{T}) = \{E \in \mathbf{T} \mid \mathrm{ch}_{\mathcal{G}}(E) \cap ((\mathbf{V} \cup \widetilde{\mathbf{U}}) \setminus \mathbf{T}) \neq \emptyset\}.$$

In other words, a receiver is a vertex in $\mathbf{T}$ that has parents outside of $\mathbf{T}$, while an emitter is a vertex in $\mathbf{T}$ that has children outside of $\mathbf{T}$. Transit clusters can now be defined as follows.

Definition 13 (Transit cluster)**.** *A non-empty set $\mathbf{T} \subset \mathbf{V} \cup \widetilde{\mathbf{U}}$ is a* transit cluster *in a connected causal graph $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ if the following conditions hold*

(a) $\mathrm{pa}^{+}_{\mathcal{G}}(R_i) \setminus \mathbf{T} = \mathrm{pa}^{+}_{\mathcal{G}}(R_j) \setminus \mathbf{T}$ *for all pairs* $R_i, R_j \in \mathrm{re}_{\mathcal{G}}(\mathbf{T})$,

(b) $\mathrm{ch}_{\mathcal{G}}(E_i) \setminus \mathbf{T} = \mathrm{ch}_{\mathcal{G}}(E_j) \setminus \mathbf{T}$ *for all pairs* $E_i, E_j \in \mathrm{em}_{\mathcal{G}}(\mathbf{T})$,

(c) For all vertices $T_i \in \mathbf{T}$, there exists a receiver R or an emitter E such that T_i and R or T_i and E are connected via an undirected path in a subgraph of $\mathcal{G}$ where all incoming edges to receivers of $\mathbf{T}$ and all outgoing edges from emitters of $\mathbf{T}$ are removed.

(d) If $\mathrm{em}_{\mathcal{G}}(\mathbf{T}) \neq \emptyset$ *then for all* $R \in \mathrm{re}_{\mathcal{G}}(\mathbf{T})$ *there exists* $E \in \mathrm{em}_{\mathcal{G}}(\mathbf{T})$ *such that* $E \in \mathrm{de}_{\mathcal{G}}(R)$,

(e) If $\mathrm{re}_{\mathcal{G}}(\mathbf{T}) \neq \emptyset$ *then for all* $E \in \mathrm{em}_{\mathcal{G}}(\mathbf{T})$ *there exists* $R \in \mathrm{re}_{\mathcal{G}}(\mathbf{T})$ *such that* $R \in \mathrm{an}_{\mathcal{G}}(E)$.

Conditions (a) and (b) determine that all receivers in $\mathbf{T}$ must have the same parents outside of $\mathbf{T}$, and all emitters must have the same children outside of $\mathbf{T}$. Condition (c) ensures that within $\mathbf{T}$, each vertex is connected to a receiver or an emitter. Conditions (d) and (e) state that if the receivers and emitters are non-empty sets, then each receiver is connected to an emitter via a directed path, and each emitter is connected to a receiver via a directed path. Importantly, $\mathcal{G}[\mathbf{T} \to T]$ is a DAG when $\mathcal{G}$ is a DAG and $\mathbf{T}$ is a transit cluster.

As an illustration of transit clusters, consider the graph in Figure 4a. The set $\mathbf{S} = \{R, S_1, S_2, E_1, E_2\}$ fulfills the conditions of Definition 13, and is thus a transit cluster. The set $\mathbf{T} = \{T_1, T_2\}$ also fulfills these conditions. The graph obtained by clustering both $\mathbf{S}$ and $\mathbf{T}$ is presented in Figure 4b.

Clustering a causal graph not only has an impact on the graph itself but also on the available input distributions. We define clustered input distributions for identifiability considerations as follows.

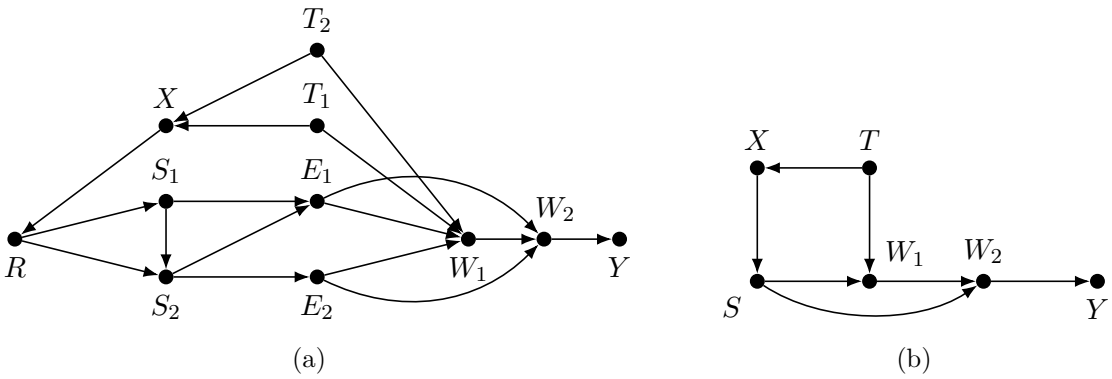

Figure 4: Graphs for the example on clustering with multiple data sources: (a) the original graph, (b) the clustered graph where the sets $\{R, S_1, S_2, E_1, E_2\}$ and $\{T_1, T_2\}$ have been clustered as S and T, respectively.

Definition 14 (Clustered input distributions)**.** *Let* $\mathcal{I} = \{p(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)\}_{i=1}^n$ *be a set of input distributions and let* $\mathbf{T} \subset \mathbf{V} \cup \widetilde{\mathbf{U}}$ *be a transit cluster such that* $\mathrm{em}_{\mathcal{G}}(\mathbf{T}) \cap \widetilde{\mathbf{U}} = \emptyset$ *in a causal graph* $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$*. If for each* $i = 1, \ldots, n$*, exactly one of the following conditions holds*

(a) $\mathrm{em}_{\mathcal{G}}(\mathbf{T}) \subseteq \mathbf{A}_i$ *and* $\mathbf{T} \cap (\mathbf{B}_i \cup \mathbf{C}_i) = \emptyset$*,*

(b) $\mathrm{em}_{\mathcal{G}}(\mathbf{T}) \subseteq \mathbf{B}_i$ *and* $\mathbf{T} \cap (\mathbf{A}_i \cup \mathbf{C}_i) = \emptyset$*,*

(c) $\mathrm{em}_{\mathcal{G}}(\mathbf{T}) \subseteq \mathbf{C}_i$ *and* $\mathbf{T} \cap (\mathbf{A}_i \cup \mathbf{B}_i) = \emptyset$*,*

(d) $\mathbf{T} \cap (\mathbf{A}_i \cup \mathbf{B}_i \cup \mathbf{C}_i) = \emptyset$*,*

and at least one input distribution satisfies one of the conditions (a)–(c) then $\mathcal{I}$ *is* compatible *with* $\mathbf{T}$ *and the set of* clustered input distributions *with respect to* $\mathbf{T}$ *is* $\mathcal{I}[\mathbf{T} \to T] = \{p(\mathbf{a}'_i \,|\, \mathrm{do}(\mathbf{b}'_i), \mathbf{c}'_i)\}_{i=1}^n$ *where*

$$\mathbf{W}'_i = \begin{cases} \mathbf{W}_i & \text{if } \mathbf{W}_i \cap \mathbf{T} = \emptyset, \\ (\mathbf{W}_i \cup T) \setminus \mathbf{T} & \text{otherwise,} \end{cases}, \text{ for } i = 1 \ldots, n \text{ and } \mathbf{W}_i = \mathbf{A}_i, \mathbf{B}_i, \mathbf{C}_i,$$

and T *is a new variable that represents* $\mathbf{T}$ *after clustering.*

Conditions (a)–(c) can be summarized as a requirement that the cluster is not split among the sets $\mathbf{A}_i$, $\mathbf{B}_i$, and $\mathbf{C}_i$, and that at least the emitters of the cluster are present in one of these sets. We exclude the case where $\mathbf{T}$ is not present in any input distribution by requiring that one of the conditions (a)–(c) holds for at least one input distribution.

If the emitters are not entirely present in $\mathbf{A}_i$, $\mathbf{B}_i$, or $\mathbf{C}_i$, condition (d) states that no other member of $\mathbf{T}$ can belong to $\mathbf{A}_i$, $\mathbf{B}_i$, or $\mathbf{C}_i$. If exactly one of the conditions holds for each input distribution, we can replace the members of $\mathbf{T}$ in each input distribution by T. We highlight that considering only the emitters of $\mathbf{T}$, instead of $\mathbf{T} \cap \mathbf{V}$, is sufficient to represent a set of input distributions as a set of clustered input distributions. Intuitively, this can be understood by noting that information flows out from a transit cluster only

through the emitters. This sufficiency is further demonstrated by the results of Section 4.2. We note that there exists a one-to-one mapping between the clustered input distributions and their original counterparts. This property is important for using identifying functionals obtained from the clustered causal graph in the original graph, as we discuss in Section 5.

Importantly, Definition 14 rules out scenarios where only some, but not all emitters are observed in a data source. For example, consider the causal graph of Figure 5a where $U \in \widetilde{\mathbf{U}}$. If we ignore compatibility and consider $\{Z, U\}$ as a transit cluster, we can obtain the graph of Figure 5b, where by Definition 10, $T \in \mathbf{V}'$. It is clear that $p(y \,|\, \mathrm{do}(x))$ is not identifiable from $p(x, y, z)$ in the original graph but is identifiable from $p(x, y, t)$ in the clustered graph. It is crucial that scenarios like this are ruled out.

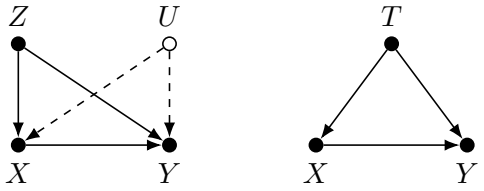

Figure 5: Causal graphs demonstrating the importance of the compatibility of the input distributions for the identification invariance of clustering.

We present the following two lemmas that characterize transit clusters and are needed in later results that concern identification invariance. The first lemma states that a transit cluster remains a transit cluster after the removal of incoming or outgoing edges of a set of vertices.

Lemma 15. *A transit cluster* $\mathbf{T}$ *in* $\mathcal{G}$ *is also a transit cluster in* $\mathcal{G}[\overline{\mathbf{Z}}, \underline{\mathbf{W}}]$ *where* $\mathbf{Z}$ *and* $\mathbf{W}$ *are disjoint, possibly empty, subsets of* $\mathbf{V}$ *and* $\mathbf{T}$ *is either a subset of* $\mathbf{Z}$*, a subset of* $\mathbf{W}$*, or disjoint from both* $\mathbf{Z}$ *and* $\mathbf{W}$*.*

The second lemma concerns d-separation in clustered and unclustered graphs.

Lemma 16. *Let* $\mathbf{T}$ *be a transit cluster in* $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ *and let* $\mathcal{G}'(\mathbf{V}' \cup \widetilde{\mathbf{U}}') = \mathcal{G}[\mathbf{T} \to T]$*. Let* $\mathbf{X}'$*,* $\mathbf{Y}'$ *and* $\mathbf{W}'$ *be disjoint subsets of* $\mathbf{V}'$ *and define* $\mathbf{X}$*,* $\mathbf{Y}$ *and* $\mathbf{W}$ *to be the corresponding subsets of* $\mathbf{V}$ *when* T *is replaced by* $\mathbf{T}$*. Then* $\mathbf{X}'$ *and* $\mathbf{Y}'$ *are d-separated by* $\mathbf{W}'$ *in* $\mathcal{G}'$ *if and only if* $\mathbf{X}$ *and* $\mathbf{Y}$ *are d-separated by* $\mathbf{W}$ *in* $\mathcal{G}$*.*

In Lemma 16, the sets $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{W}$ are obtained from disjoint sets $\mathbf{X}'$, $\mathbf{Y}'$ and $\mathbf{W}'$, respectively. This rules out cases where the transit cluster is divided between the sets $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{W}$ because T can be a member of only one of the sets $\mathbf{X}'$, $\mathbf{Y}'$ and $\mathbf{W}'$ in $\mathcal{G}'$. Together Lemmas 15 and 16 imply that the rules of do-calculus are applicable in a clustered graph $\mathcal{G}'$ if and only if they are applicable in the unclustered graph $\mathcal{G}$. More precisely, whenever a transit cluster $\mathbf{T}$ is not split among the sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and $\mathbf{W}$ that appear in the rules, and whenever $\mathbf{T}$ is present for a rule that applies in $\mathcal{G}$, then the rule applies with $\mathbf{T}$ replaced by T in $\mathcal{G}'$ and vice versa. If $\mathbf{T}$ is not present, then the applicability of the rule is always preserved between the graphs.

4.2 Identification Based on Clustered Causal Graphs

We present conditions for determining the identification invariance of clustering with multiple data sources. We consider the identification invariance separately based on the identifiability of the causal effect in the clustered graph. Theorems 17 and 18 consider identifiable and non-identifiable scenarios, respectively. In cases where the causal effect is not identifiable in the clustered graph, the input distributions must satisfy additional conditions. These conditions are verified by Algorithm 1. The proofs of this section are presented in Appendix B.

Algorithm 1 Verify the clustered input distributions to assess identification invariance of clustering $\mathbf{T}$ for a non-identifiable causal effect $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$, where $\mathbf{X} \cap \mathbf{T} = \emptyset$, $\mathbf{Y} \cap \mathbf{T} = \emptyset$. The inputs are a clustered causal graph $\mathcal{G}' = \mathcal{G}[\mathbf{T} \to T]$, a set of clustered input distributions $\mathcal{I}' = \mathcal{I}[\mathbf{T} \to T]$, and the cluster vertex T that represents the cluster $\mathbf{T}$.

function VerifyInputs($\mathcal{G}', \mathcal{I}', T$)
  if $\mathrm{pa}^{+}_{\mathcal{G}'}(T) = \emptyset$ **then return TRUE**
  for all $\mathcal{I}'_i \in \mathcal{I}'$ **do**
    if $T \in \mathbf{A}'_i \cup \mathbf{B}'_i \cup \mathbf{C}'_i$ **then**
      if $\mathbf{C}'_i \cap \mathrm{de}_{\mathcal{G}'}(T) \neq \emptyset$ **then return FALSE**
      $\mathbf{D} \leftarrow \mathrm{de}_{\mathcal{G}'}(T) \cap \mathbf{A}'_i$
      if $T \in \mathbf{A}'_i$ **and** $(\mathbf{D} \perp\!\!\!\perp T \mid \mathbf{B}'_i, \mathbf{C}'_i, \mathbf{A}'_i \setminus (\mathbf{D} \cup T))_{\mathcal{G}'[\overline{\mathbf{B}'_i}, \underline{T}]}$ **then continue**
      if $T \in \mathbf{B}'_i$ **then continue**
      if $T \in \mathbf{C}'_i$ **and** $(\mathbf{A}'_i \perp\!\!\!\perp T \mid \mathbf{B}'_i, \mathbf{C}'_i \setminus T)_{\mathcal{G}'[\overline{\mathbf{B}'_i}, \underline{T}]}$ **then continue**
    else
      if $\exists \mathcal{I}'_j \in \mathcal{I}'$ s.t. $T \in \mathbf{A}'_j$ **and** $\mathbf{B}'_j = \emptyset$ **and** $\mathbf{C}'_j = \emptyset$ **then**
        if $(\mathbf{A}'_i \perp\!\!\!\perp T \mid \mathbf{B}'_i, \mathbf{C}'_i)_{\mathcal{G}'[\overline{\mathbf{B}'_i}, \underline{T}]}$ **and** $(T \perp\!\!\!\perp \mathbf{C}'_i \mid \mathbf{B}'_i)_{\mathcal{G}'[\overline{\mathbf{B}'_i}]}$ **and** $(T \perp\!\!\!\perp \mathbf{B}'_i)_{\mathcal{G}'[\overline{\mathbf{B}'_i}]}$ **then continue**
      if $(\mathbf{A}'_i \perp\!\!\!\perp T \mid \mathbf{B}_i, \mathbf{C}_i)_{\mathcal{G}'[\overline{\mathbf{B}'_i}, \overline{T(\mathbf{C}'_i)}]}$ **then continue**
      if $\exists \mathcal{I}'_k \in \mathcal{I}'$ s.t. $\mathbf{A}'_i \subseteq \mathbf{A}'_k$ **and** $\mathbf{B}'_k = \mathbf{B}'_i \cup T$ **and** $\mathbf{C}'_k = \mathbf{C}'_i$ **then continue**
      return FALSE
  return TRUE

Theorem 17. *Let $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ be a causal graph where $\mathbf{T} \subset \mathbf{V} \cup \widetilde{\mathbf{U}}$ is a transit cluster and let $\mathcal{I}$ be a set of input distributions compatible with $\mathbf{T}$. Let $\mathcal{I}' = \mathcal{I}[\mathbf{T} \to T]$ in $\mathcal{G}' = \mathcal{G}[\mathbf{T} \to T]$. If $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ is identifiable from $\mathcal{I}'$ in $\mathcal{G}'$, then it is identifiable from $\mathcal{I}$ in $\mathcal{G}$, where $\mathbf{X} \cup \mathbf{Y} \subset \mathbf{V}$, $\mathbf{Y} \cap \mathbf{X} = \emptyset$, $\mathbf{X} \cap \mathbf{T} = \emptyset$, $\mathbf{Y} \cap \mathbf{T} = \emptyset$.*

Theorem 18. *Let $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ be a causal graph where $\mathbf{T} \subset \mathbf{V} \cup \widetilde{\mathbf{U}}$ is a transit cluster and let $\mathcal{I}$ be a set of input distributions compatible with $\mathbf{T}$. Let $\mathcal{I}' = \mathcal{I}[\mathbf{T} \to T]$ in $\mathcal{G}' = \mathcal{G}[\mathbf{T} \to T]$. If $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ is not identifiable from $\mathcal{I}'$ in $\mathcal{G}'$ and if VerifyInputs($\mathcal{G}', \mathcal{I}', T$) returns* **TRUE**, *then $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ is not identifiable from $\mathcal{I}$ in $\mathcal{G}$, where $\mathbf{X} \cup \mathbf{Y} \subset \mathbf{V}$, $\mathbf{Y} \cap \mathbf{X} = \emptyset$, $\mathbf{X} \cap \mathbf{T} = \emptyset$, $\mathbf{Y} \cap \mathbf{T} = \emptyset$.*

The leading idea of VerifyInputs is to validate properties related to the clustered causal graph and the clustered input distributions that ensure identification invariance with respect to any unclustered graph. Some of these properties are conditional independence

constraints derived via d-separation, while others require the existence of another clustered input distribution with specific features. The condition on line 2 considers a special case where the cluster has no external parents in the causal graph, from which we can directly conclude identification invariance without any other requirements.

Violations of the conditions checked by VerifyInputs may lead to clustering not being identification invariant for the causal effect of interest, as we demonstrate via examples in Appendix D. Theorem 18 implies that VerifyInputs is sound, that is, if the algorithm returns **TRUE**, then the clustering operation is identification invariant. The completeness of the algorithm remains unknown. Furthermore, even if the algorithm returns **FALSE**, verification may still succeed for a subset of the original inputs. We will describe this point in greater detail in Section 5. The simulation study in Section 6 provides information on how often VerifyInputs returns **TRUE** for random input distributions.

We highlight that Theorems 17 and 18 encompass as a special case the identification invariance result for a single data source by Tikka et al. (2023), who showed that if the original input distribution contains the set of emitters and the parents of the transit cluster, then the transit cluster can be replaced by a single vertex without affecting the identifiability of the causal effect. VerifyInputs also determines such scenarios as identification invariant, because when $\mathrm{em}(\mathbf{T}) \cup (\mathrm{pa}(\mathbf{T}) \setminus \mathbf{T}) \subseteq \mathbf{A}_i$, all backdoor paths from any vertex of $\mathbf{T}$ are blocked by the parents of the cluster. By Lemmas 15 and 16, the same d-separation also holds for T in the clustered input distributions, and thus the condition of line 7 is satisfied. In addition, because $\mathbf{C}'_i = \emptyset$ for a single observation data source, the condition of line 5 is satisfied, guaranteeing identification invariance. We note that our results are more general even in the case of a single observational data source than the previous result, as the parents of the cluster need not all be observed.

Theorems 17 and 18 and VerifyInputs make no assumptions about the internal structure of the transit cluster $\mathbf{T}$. However, there are instances where additional information about the transit cluster guarantees identification invariance while only requiring the compatibility of the input distributions. One such scenario is characterized by the following theorem.

Theorem 19. *Let $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ be a causal graph where $\mathbf{T} \subset \mathbf{V} \cup \widetilde{\mathbf{U}}$ is a transit cluster and let $\mathcal{I}$ be a set of input distributions compatible with $\mathbf{T}$. If $\mathrm{re}_{\mathcal{G}}(\mathbf{T}) \cap \mathrm{em}_{\mathcal{G}}(\mathbf{T}) \neq \emptyset$, then clustering $(\mathcal{G}, \mathcal{I}) \to (\mathcal{G}', \mathcal{I}')$, where $\mathcal{I}' = \mathcal{I}[\mathbf{T} \to T]$ in $\mathcal{G}' = \mathcal{G}[\mathbf{T} \to T]$, is identification invariant for any causal effect $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$, where $\mathbf{X} \cup \mathbf{Y} \subset \mathbf{V}$, $\mathbf{Y} \cap \mathbf{X} = \emptyset$, $\mathbf{Y} \cap \mathbf{T} = \emptyset$, $\mathbf{X} \cap \mathbf{T} = \emptyset$.*

As an example, the transit cluster $\mathbf{T} = \{T_1, T_2, T_3\}$ of Figure 6a satisfies the conditions of Theorem 19, because the vertex T_1 serves both as a receiver and as an emitter. In this case, clustering $\mathbf{T}$ is identification invariant for the causal effect $p(y \,|\, \mathrm{do}(x))$, and its identifiability can be assessed based on the clustered graph of Figure 6b for all combinations of possible input distributions that are compatible with $\mathbf{T}$.

Next, we present an example on the use of Theorems 17 and 18. Consider again the graph of Figure 4a, for which we detected the transit cluster $\mathbf{S} = \{R, S_1, S_2, E_1, E_2\}$ and the transit cluster $\mathbf{T} = \{T_1, T_2\}$. We consider four different combinations of input distributions to identify the causal effect $p(y \,|\, \mathrm{do}(x))$:

(i) $p(x, e_1, e_2, s_1, r)$ and $p(y, e_1, e_2, t_1, t_2)$,

(ii) $p(x, e_1, e_2, r, w_1)$ and $p(y, w_2 \,|\, \mathrm{do}(w_1))$,

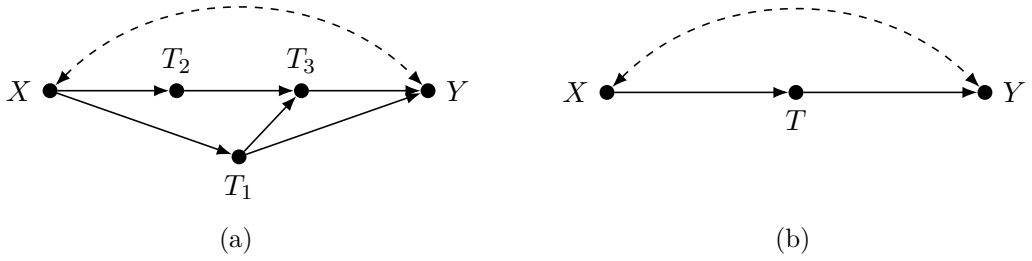

Figure 6: An example where the transit cluster $\{T_1, T_2, T_3\}$ in panel (a) can be clustered to a single vertex T as depicted in panel (b) without any additional information on the input distributions when considering the identification invariance of clustering regarding the causal effect $p(y \mid \mathrm{do}(x))$.

(iii) $p(y \mid \mathrm{do}(t_1, t_2), e_1, e_2, s_2)$ and $p(x, t_1, t_2)$,

(iv) $p(y, t_1, t_2, r, e_1, e_2)$, $p(x, w_1 \mid \mathrm{do}(e_1, e_2))$ and $p(x, w_1)$.

First, we can notice that all sets of input distributions are compatible with respect to $\mathbf{S}$ and $\mathbf{T}$. This applies because $\mathrm{em}_{\mathcal{G}}(\mathbf{S}) = \{E_1, E_2\}$ and $\mathrm{em}_{\mathcal{G}}(\mathbf{T}) = \{T_1, T_2\}$ are always entirely present in an input distribution if any vertex within the cluster is present in the input distribution. The corresponding clustered input distributions for each case are:

(i) $p(x, s)$ and $p(y, s, t)$,

(ii) $p(x, s, w_1)$ and $p(y, w_2 \mid \mathrm{do}(w_1))$,

(iii) $p(y \mid \mathrm{do}(t), s)$ and $p(x, t)$,

(iv) $p(y, t, s)$, $p(x, w_1 \mid \mathrm{do}(s))$ and $p(x, w_1)$.

In case (i) the causal effect $p(y \mid \mathrm{do}(x))$ is identifiable in the clustered graph. This means that according to Theorem 17, the causal effect $p(y \mid \mathrm{do}(x))$ is also identifiable in the original graph. We can therefore use the clustered graph and input distributions to obtain the identifying functional

$$p(y \mid \mathrm{do}(x)) = \sum_{s,t} p(s|x)p(t)p(y|s,t),$$

which also provides us with an identifying functional for the causal effect in $\mathcal{G}$ as we will show in Section 5.

In cases (ii)–(iv), the causal effect $p(y \mid \mathrm{do}(x))$ is not identifiable in the clustered graph. To assess whether the identification invariance holds in these cases, by Theorem 18, the conditions regarding the input distributions must be checked by VerifyInputs.

We can first see that T does not have any incoming edges in the clustered graph, which means that VerifyInputs returns TRUE in all cases per line 2. Clustering $\mathbf{T}$ is therefore identification invariant for the causal effect $p(y \mid \mathrm{do}(x))$. Next, we will show that the clustering is also identification invariant with respect to the transit cluster $\mathbf{S}$ by going through the remaining conditions of VerifyInputs for the remaining clustered input distributions.

In case (ii), for the input distribution $p(x, s, w_1)$, vertex X blocks all backdoor paths between S and its descendants, which satisfies the condition of line 7. In addition, $\mathbf{C}'_1 = \emptyset$ and therefore the condition for FALSE on line 5 is not triggered. In the first input distribution $S \in \mathbf{A}'_1$, and $\mathbf{B}'_1$ and $\mathbf{C}'_1$ are empty, allowing us to proceed to the line 12. Next, the input distribution $p(y, w_2 \,|\, \mathrm{do}(w_1))$ does not contain any $S \in \mathbf{S}$. We can see that $(S \perp\!\!\!\perp Y, W_2 \,|\, W_1)_{\mathcal{G}'[\overline{W_1}, \underline{S}]}$. In addition $(S \perp\!\!\!\perp W_1)_{\mathcal{G}'[\overline{W_1}]}$, and thus all conditions of line 12 are satisfied. We can conclude that in case (ii) clustering $\mathbf{S}$ is identification invariant for $p(y \,|\, \mathrm{do}(x))$.

In case (iii), for the input distribution $p(y \,|\, \mathrm{do}(t), s)$, it applies that $S \in \mathbf{C}'_2$. Now $T \in \mathbf{B}'_2$ blocks all backdoor paths between S and $\mathbf{A}'_2 = \{Y\}$, satisfying the condition of line 9. In addition, $\mathbf{C}'_2$ does not contain any descendants of S, and therefore all conditions for this input distribution are satisfied. Finally, we can see that no member of $\mathbf{S}$ is present in the clustered input distribution $p(x, t)$. The condition on line 12 is not satisfied but as $\mathbf{A}'_3$ contains only non-descendants of S the condition $(\mathbf{A}'_3 \perp\!\!\!\perp S)_{\mathcal{G}'[\overline{S}]}$ on line 13 is fulfilled. All input distributions satisfy the conditions of VerifyInputs, and therefore clustering $\mathbf{S}$ is identification invariant for the causal effect $p(y \,|\, \mathrm{do}(x))$.

In case (iv), T blocks the backdoor paths between the cluster vertices and Y, and $\mathbf{C}'_1 = \emptyset$, meaning that lines 5 and 7 are satisfied for the first clustered input distribution $p(y, t, s)$. For the second input distribution $p(x, w_1 \,|\, \mathrm{do}(s))$, it applies that $T \in \mathbf{B}'_2$. In this case the condition of line 8 applies directly. Finally, we can see that no cluster vertex is present in $p(x, w_1)$. Now, we see that the second distribution, $p(x, w_1 \,|\, \mathrm{do}(t))$, has the same variables, but with T inside the do-operator. This satisfies the condition of line 14 for the third input distribution. Therefore, also in case (iv), clustering is identification invariant with respect to $\mathbf{S}$.

In all cases (ii)–(iv) we were able to establish the identification invariance. Using Theorem 18, we can conclude that the causal effect $p(y \,|\, \mathrm{do}(x))$ is not identifiable in the graph where $\mathbf{S}$ is not clustered either.

5. Obtaining Identifying Functionals

Theorems 7, 8, 9, 17, 18 and 19 show that identifiability is neither lost nor gained by pruning or clustering under specific assumptions. Except for Theorem 7, these results do not provide a way to obtain an identifying functional for the causal effect in the original graph. Intuitively, pruning should not alter the identifying functional, i.e., an identifying functional obtained in the pruned graph should be applicable in the original. Similarly for clustering, we should simply be able to replace the clustered input distributions by their original counterparts in the identifying functional. Fortunately, both of these intuitive notions apply to a large class of identifying functionals, as we will demonstrate with the following two theorems for pruning and clustering, respectively. The proofs of this section are presented in Appendix C.

Theorem 20 (Identifying functional after pruning)**.** *Let $\mathcal{G}(\mathbf{V} \cup \mathbf{U})$ be a causal graph and let $\mathcal{I}$ be a set of input distributions. Let $\mathcal{G}'$ and $\mathcal{I}'$ be the causal graph and the set of input distributions obtained respectively by applying Theorem 7, 8, or 9 to $\mathcal{G}$ and $\mathcal{I}$. If $p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is identifiable from $\mathcal{I}'$ in $\mathcal{G}'$ by do-calculus with the identifying functional $g(\mathcal{I}')$,*

then $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ *is identifiable from* $\mathcal{I}$ *in* $\mathcal{G}$ *by do-calculus with the identifying functional* $f(\mathcal{I})$ *where* $f(\mathcal{I}) = g(\mathcal{I}')$.

Theorem 21 (Identifying functional after clustering)**.** *Let* $\mathbf{T}$ *be a transit cluster in a causal graph* $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ *and let* $\mathcal{I}$ *be a set of input distributions compatible with* $\mathbf{T}$*. If* $p'(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ *is identifiable from* $\mathcal{I}' = \mathcal{I}[\mathbf{T} \to T]$ *in* $\mathcal{G}' = \mathcal{G}[\mathbf{T} \to T]$ *by do-calculus with the identifying functional* $g(\mathcal{I}')$*, then* $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ *is identifiable from* $\mathcal{I}$ *in* $\mathcal{G}$ *by do-calculus with the identifying functional* $f(\mathcal{I})$ *obtained from* $g(\mathcal{I}')$ *by replacing every instance of* T *by the subset of* $\mathbf{T}$ *that is present in the corresponding input distribution in* $\mathcal{I}$*.*

In the case of a single observational data source, Theorems 20 and 21 apply to all identifying functionals due to the completeness of do-calculus (Shpitser and Pearl, 2006; Huang and Valtorta, 2006). Do-calculus has also been shown to be complete for specific constrained instances of the general causal effect identifiability problem, such as g-identifiability, where the set of input distributions $\mathcal{I}$ is of the form $\{p(\mathbf{v} \mid \mathrm{do}(\mathbf{z}_i))\}_{i=1}^n$ for a collection of subsets $\{\mathbf{Z}_i\}_{i=1}^n$ of $\mathbf{V}$ (Kivva et al., 2022). Note that in the g-identifiability problem, the set of inputs can also include $p(\mathbf{v})$ if $\mathbf{Z}_i = \emptyset$ for some i, but excludes several other types of input distributions such as conditional distributions, conditional interventional distributions, and joint distributions of subsets of variables. It is not known whether do-calculus is complete for the unconstrained case of the general causal effect identifiability problem.

We note that it is sometimes beneficial to exclude specific inputs from consideration if only a subset of inputs is necessary for identification. Suppose that we have a set of input distributions $\mathcal{I}$ and a subset $\mathcal{J} \subset \mathcal{I}$ such that Theorem 20 or Theorem 21 applies to $\mathcal{J}$ but not $\mathcal{I}$. If a causal effect of interest is identifiable in the pruned or clustered graph from $\mathcal{J}'$, where $\mathcal{J}'$ is the corresponding set of restricted input distributions or clustered input distributions, we can readily apply the relevant theorem to $\mathcal{J}$ instead to obtain the identifying functional. For instance, consider a simple graph $X \to Y \to Z$, with input distributions $\{p(x, y), p(x \mid z)\}$. As the causal effect $p(y \mid \mathrm{do}(x))$ is identifiable from $p(x, y)$ alone in this graph, we can remove Z by Theorem 7, even though the input $p(x \mid z)$ does not satisfy the conditions by applying the theorem with $p(x, y)$ as the sole input distribution. We note that the number of possible subsets $\mathcal{J}$ is $2^{|\mathcal{I}|}$, meaning that finding a suitable subset may be computationally challenging when the number of input distributions is large. However, it is often not necessary to check all subsets for the following reasons. In the case of pruning, the conditions in Theorems 7–9 can be checked independently for each input distribution. Similarly in the case of clustering, the validity of a specific clustered input distribution $p(\mathbf{a}'_i \mid \mathrm{do}(\mathbf{b}'_i), \mathbf{c}'_i)$ is independent of other clustered inputs whenever $T \in \mathbf{A}'_i \cup \mathbf{B}'_i \cup \mathbf{C}'_i$ as can be observed from the conditions checked by Algorithm 1 in the if-block starting on line 4.

6. Simulations

We performed a simulation study to assess the practical computational impact of the presented graph reduction operations for causal effect identification. We studied the impact of pruning and clustering both together and separately. In this section, our main focus is on the simulation performed for pruning and clustering together; if we discuss the simulations for pruning and clustering separately, we mention it explicitly. The details and the results regarding the simulations for pruning and clustering separately are presented in Appendix E.

We started by generating random causal graphs and random sets of input distributions where each input distribution is of the form $p(\mathbf{a}_i \mid \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$ with sets $\mathbf{B}_i$ and $\mathbf{C}_i$ potentially being empty. In addition, the graphs were generated such that each graph contained at least one transit cluster and one non-ancestor of the response. To our knowledge, the only general-purpose algorithm able to handle such inputs in causal effect identification is Do-search, implemented in the R package `dosearch` (Tikka et al., 2021, 2024). We compare two strategies

- **Direct strategy:** Apply Do-search directly in the original graph and record the time elapsed.
- **Reduction strategy:**
  1. Prune the original graph according to the Theorems 7–9 if the respective conditions of each theorem are satisfied. For the sake of the comparison, pruning is omitted if the set to be pruned intersects with the transit cluster.
  2. Find the transit clusters of the (possibly) pruned graph and determine the identifiability of the causal effect in the pruned and clustered graph by applying Do-search.
  3. If the causal effect is identifiable in the reduced graph, it is also identifiable in the original graph by Theorem 17. Record the time elapsed so far and stop (Setting A).
  4. Otherwise, use VerifyInputs (Algorithm 1) to determine the identification invariance using the pruned and clustered graph, and pruned and clustered input distributions.
  5. If the identification invariance holds, the causal effect is not identifiable in the original graph by Theorem 18. Record the time elapsed so far and stop (Setting B).
  6. If the graph size was reduced by pruning in step 1, apply Do-search in the pruned graph, record the time elapsed so far and stop (Setting D; see Appendix E for the results on this setting).
  7. Otherwise, apply Do-search in the original graph, record the time elapsed so far and stop (Setting C).

Ideally, we expect benefits from the reduction strategy in the cases of settings A (the causal effect is identifiable in the clustered graph) and B (the causal effect is not identifiable in the clustered graph and clustering is identification invariant), compared to applying Do-search directly in the original graph. Additionally, in the worst case of setting C, where the graph reduction does not help in identification, we expect that assessing the graphical conditions to verify the identification invariance does not induce a major overhead. To avoid extensive running times, we imposed a time limit of 15 minutes for Do-search. The generation of one simulation instance is defined in Algorithm 2 in Appendix E.

In total, we obtained 108 933 instances. 6.6% of the instances exceeded the time limit of 15 minutes when using Do-search in the original graph. Table 1 reports the differences and ratios of the running times between the direct strategy and the reduction strategy for settings A–C. For running times that exceeded 15 minutes, we applied a conservative

approach and fixed these times to 15 minutes. Additionally, we visualized the results for settings A–B with scatter plots in Figure 7, and for setting C with a box plot in Figure 8.

When identification with clustering was applicable (Settings A–B), the benefits of the graph reduction increased considerably as the original graph sizes increased. For instance, when the original graph size was 12, the median difference was over 12 minutes to the benefit of the reduction strategy in the case of setting B. In contrast, for small graphs with seven vertices, the reduction strategy performed worse for setting A. However, in this case the median time lost was less than a tenth of a second. In the case of setting C, where the original graph must be used for causal effect identification after checking if reduction is applicable, it is naturally faster to use the original graph in identification directly. Reassuringly, the median time lost for checking whether graph reduction is applicable was less than five seconds for all graph sizes, with the ratio approaching one as the graph size increases.

VerifyInputs returned **TRUE** for 52.0% of the instances that reached step 4 of the reduction strategy, which indicates that the conditions of the algorithm are not too restrictive for practical use. Additionally, it appears that the benefits of the clustering operations were higher for setting B than for setting A. This is likely explained by the search-based nature of Do-search. When the causal effect is identifiable (as is the case for setting A), Do-search does not have to explore its entire search space, unlike in setting B. Therefore, reduction of the search space by clustering benefits setting B more than setting A.

Overall, the results show that when graph reduction is applicable, it can offer significant performance benefits in general causal effect identification. Additionally, in the cases where the graph reduction does not offer improved running time performance, the disadvantage of using the operations is minimal in practice. Across all instances, the median time spent finding transit clusters was 0.13 seconds (IQR 0.10 to 0.17), median time for checking the identification invariance was 0.004 seconds (IQR 0.001 to 0.010) and the median time for performing the pruning operations was 0.024 seconds (IQR 0.017 to 0.038). We also highlight that graph reduction does not have to be automatically applied if the researcher views it as redundant—as can be the case with small graphs or small cluster sizes—for example.

7. Demonstrations

To show how the results of this paper can be used in practice, we consider two example cases that are selected from the review by Tennant et al. (2021). For the first example, we use the causal graph of the study on infant mortality by Filippidis et al. (2017). As the second example, we consider the ELSA-Brasil study (Camelo et al., 2015) on the effect of life-course socioeconomic position on a marker of atherosclerosis. The focus in the first example is on pruning while the second example focuses on clustering.

7.1 Causal Model from an Infant Mortality Study

Filippidis et al. (2017) studied the associations and potential causal pathways of cigarette prices and infant mortality in 23 countries of the European Union. To select an appropriate set of covariates for the analyses, Filippidis et al. (2017) used the causal graph depicted in Figure 9a. To demonstrate the uses of pruning in the case of multiple data sources, we focus on various potential causal effects that can be assessed based on Figure 9a. The variables

Graph size	Setting	Median difference (IQR)		Median ratio (IQR)		N
7	A	−0.03	(−0.07, 0.03)	0.66	(0.26, 1.28)	918
7	B	0.00	(−0.04, 0.06)	1.00	(0.56, 1.62)	875
7	C	−0.12	(−0.14, −0.11)	0.40	(0.30, 0.54)	122
8	A	0.20	(−0.02, 0.60)	2.57	(0.85, 5.63)	3 751
8	B	0.50	(0.19, 1.00)	4.73	(2.50, 8.80)	3 830
8	C	−0.17	(−0.22, −0.14)	0.72	(0.61, 0.81)	674
9	A	1.52	(0.34, 3.78)	9.72	(3.15, 23.46)	6 875
9	B	3.54	(1.57, 7.00)	20.98	(9.45, 41.81)	8 059
9	C	−0.26	(−0.42, −0.19)	0.89	(0.82, 0.95)	1 895
10	A	9.54	(2.68, 24.11)	36.60	(11.94, 99.00)	7 782
10	B	21.65	(9.89, 45.57)	75.46	(32.12, 184.04)	10 950
10	C	−0.51	(−1.21, −0.28)	0.96	(0.91, 0.98)	3 262
11	A	49.07	(14.72, 134.37)	109.78	(30.70, 338.38)	5 912
11	B	122.88	(54.19, 268.86)	241.27	(83.99, 603.91)	10 592
11	C	−1.57	(−4.54, −0.61)	0.98	(0.95, 0.99)	3 852
12	A	278.47	(72.58, 736.50)	272.05	(69.12, 888.89)	3 453
12	B	725.57	(302.57, 898.71)	584.86	(150.68, 1517.28)	7 418
12	C	−4.28	(−11.42, −1.64)	0.99	(0.97, 0.99)	3 505

Table 1: Comparison of the direct and the reduction strategies. Median differences (in seconds) of running times between the strategies, and median ratios (the running time for the direct strategy divided by the running time of the reduction strategy) are reported. The interquartile ranges (IQR) are included in parentheses. The column N tells the number of instances for each graph size and setting (A: identifiable in the clustered graph, B: non-identifiable in the clustered graph but identification invariant, C: no graph reduction operation was identification invariant, original graph had to be used).

represented by the causal graph are presented in Table 2. For more detailed definitions of these variables, see (Filippidis et al., 2017).

Assume we have two sources of data available. The first one provides the observational distribution $p(c, o, r, e, m)$ of cigarette consumption, tobacco control policies, tobacco price, education level, and maternal age. The second one provides the experimental distribution $p(c, s, g, d, b \,|\, \mathrm{do}(o))$ on the effect of tobacco control policies on cigarette consumption, second-hand smoking in pregnancy, smoking during pregnancy, pre-term birth, and small-for-gestational-age infants.

We consider three different causal effects: tobacco price on second-hand smoking in pregnancy, tobacco price on pre-term birth, and cigarette consumption on gestational age, denoted as $p(s \,|\, \mathrm{do}(r))$, $p(b \,|\, \mathrm{do}(r))$ and $p(g \,|\, \mathrm{do}(c))$, respectively. In each setting, we can prune the graph differently.

The pruned graph for the causal effect $p(s \,|\, \mathrm{do}(r))$ is presented in Figure 9b. Using Theorem 7, we pruned J, N, B, I, D, G, H and A which are non-ancestors of S. Variable W was pruned according to Theorem 8, as it is connected to S only through R after pruning the non-ancestors. In addition, having pruned A, G, J and N, the vertices E and M have

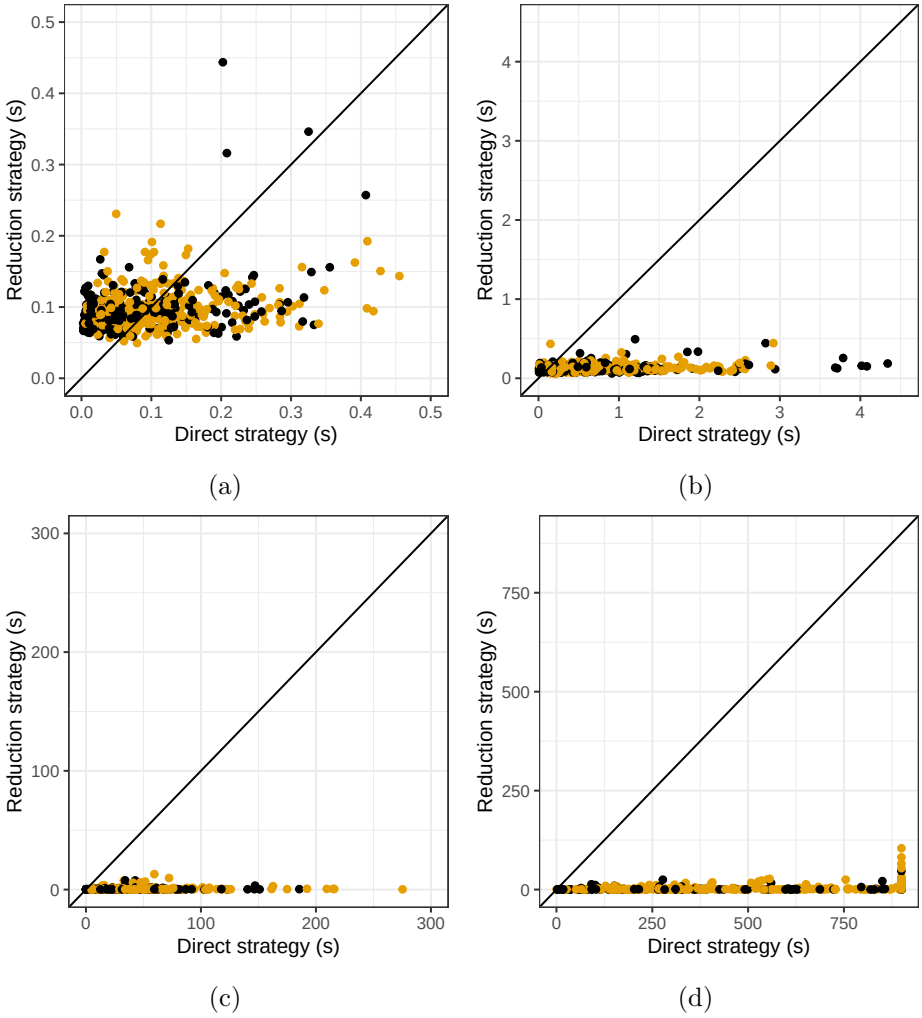

Figure 7: Comparison of the direct and the reduction strategies when reductions are possible. The scatter plots show the running times for both strategies in settings A (black) and B (orange). The panels depict different original graph sizes: 7 (a), 8 (b), 10 (c) and 12 (d). A random sample of 500 instances is used for each panel. In Panel (d), the heap of points at 900 seconds is due to the time limit of 15 minutes.

the same parents and children, meaning that $\{E, M\}$ is a transit cluster that satisfies the conditions of Theorem 19. This cluster is depicted as vertex T in Figure 9b. Now, using the pruned and clustered graph and the input distributions $p(c, o, r, t)$ and $p(c, s \,|\, \mathrm{do}(o))$, we can obtain an identifying functional

$$p(s \,|\, \mathrm{do}(r)) = \sum_{c,t,o} p(s \,|\, \mathrm{do}(o), c) p(c \,|\, r, t, o) p(t, o),$$

where the first term is derived from $p(c, s \,|\, \mathrm{do}(o))$ and the other terms from $p(c, o, r, t)$.

Regarding the causal effect $p(b \,|\, \mathrm{do}(r))$, we can perform the same pruning operations as in the case of the causal effect $p(s \,|\, \mathrm{do}(r))$, with the exception that D and B cannot be

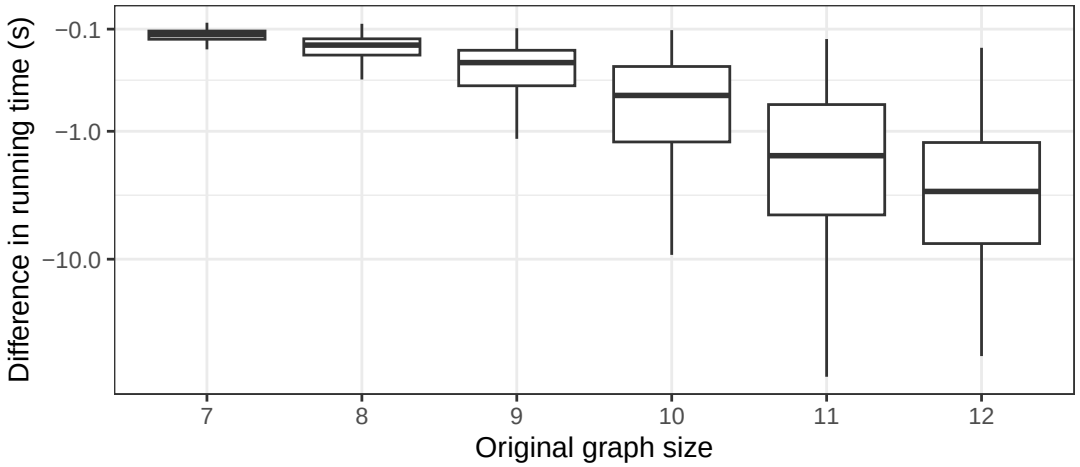

Figure 8: Overhead of the reduction strategy when no reductions are possible. The box plots show the differences in running times between the direct strategy and the reduction strategy for setting C by the size of the original graph. Negative differences indicate that more time was spent with the reduction. The vertical axis uses logarithmic scaling.

Variable	Symbol
GDP unemployment rate	W
Health expenditure	H
Access to healthcare	A
Tobacco control policies	O
Smoking during pregnancy	D
Congenital anomalies	N
Tobacco price	R
Cigarette consumption	C
Second-hand smoking in pregnancy	S
Pre-term birth	B
Education	E
Maternal age	M
Small-for-gestational-age infant	G
Social and cultural factors	F
Ethnicity	Q
Second-hand smoking in infancy	J
Infant mortality	I

Table 2: Variables in the DAG of (Filippidis et al., 2017) and their corresponding symbols in Figure 9a.

pruned this time. This also means that we are not able to cluster E and M as they no longer share the same children. The pruned graph is presented in Figure 9c. However, with the given data sources, the causal effect $p(b \,|\, \mathrm{do}(r))$ is not identifiable. Relying on Theorems 7

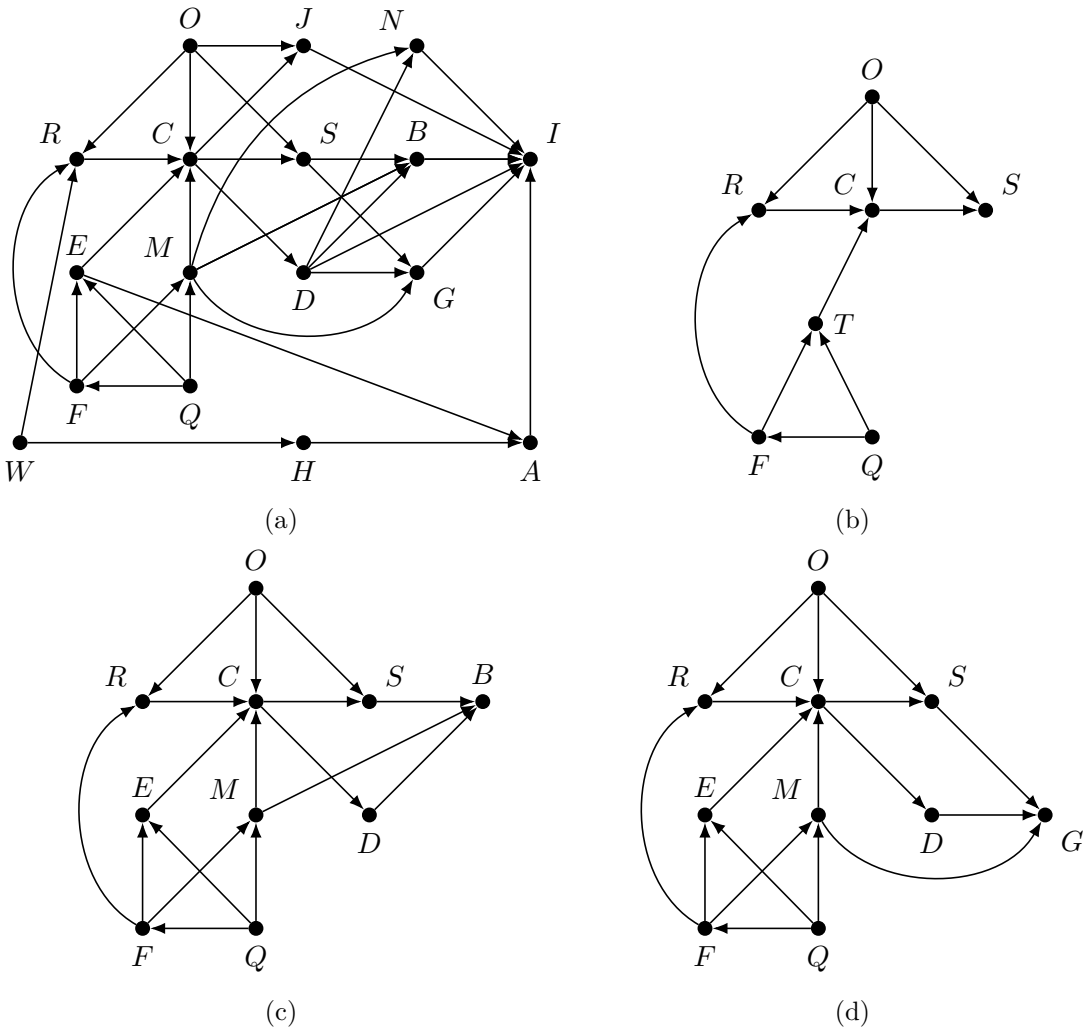

Figure 9: Causal graphs related to (Filippidis et al., 2017): (a) the original graph, (b) the pruned graph for the causal effect $p(s \,|\, \mathrm{do}(r))$, (c) the pruned graph for the causal effect $p(b \,|\, \mathrm{do}(r))$, and (d) the pruned graph for the causal effect $p(g \,|\, \mathrm{do}(c))$.

and 8, we conclude that the causal effect is also not identifiable using a graph where none of the variables are removed.

Lastly, the pruned graph for the causal effect $p(g \,|\, \mathrm{do}(c))$ is presented in Figure 9d. Again, variables J, N, B, I, A and H can be pruned by Theorem 7. Now, W is still connected to G even after an intervention to C. However, after the first pruning, W is connected to G only through R, and can therefore be pruned using Theorem 9. Using the pruned graph and the restricted input distributions $p(c, o, r, e, m)$ and $p(c, s, g, d \,|\, \mathrm{do}(o))$, we can identify the causal effect as

$$p(g \,|\, \mathrm{do}(c)) = \sum_{s,d,o} p(o)p(s, d \,|\, \mathrm{do}(o), c) \sum_{c'} p(c' \,|\, \mathrm{do}(o))p(g \,|\, \mathrm{do}(o), c', s, d),$$

where the first term is obtained from $p(c, o, r, e, m)$ and the other terms are derived from $p(c, s, g, d \mid \text{do}(o))$. All three settings are summarized in Table 3.

Causal effect	Fig.	Pruned vertices	Input distributions	ID
$p(s \mid \text{do}(r))$	9b	$J, N, B, I, D, G, H, A, W$	$p(c, o, r, t), p(c, s \mid \text{do}(o))$	Yes
$p(b \mid \text{do}(r))$	9c	J, N, I, G, H, A, W	$p(c, o, r, e, m), p(c, s, d, b \mid \text{do}(o))$	No
$p(g \mid \text{do}(c))$	9d	J, N, B, I, H, A, W	$p(c, o, r, e, m), p(c, s, d, g \mid \text{do}(o))$	Yes

Table 3: Summary of the results of Section 7.1. The table presents the causal effects, the graphs obtained by pruning, vertices that were pruned, the pruned input distributions, and whether the causal effect is identifiable.

7.2 Causal Model from an Atherosclerosis Study

We consider the ELSA-Brasil study (Camelo et al., 2015) which investigated the effect of life-course socioeconomic position on carotid intima-media thickness, a marker of atherosclerosis (Björkegren and Lusis, 2022). The variables of interest are presented in Table 4.

Variable	Symbol
Baseline variables	$\mathbf{B}$
Life-course socioeconomic position	L
Job stress	S
Health-related behavior	$\mathbf{H}$
Metabolic, endocrine, and immune dysregulation	$\mathbf{M}$
Intima-media thickness	Y

Table 4: Variables in the causal graph of (Camelo et al., 2015) and their corresponding symbols in Figure 10a.

Camelo et al. (2015) suggested the causal graph of Figure 10a for the study setting. For this demonstration, we assume that the causal graph has been constructed in accordance with top-down causal modeling and consider variables $\mathbf{B}$, $\mathbf{H}$, and $\mathbf{M}$ as transit clusters, with otherwise unspecified internal structures. Hence, we show the identification invariance for all relevant clusters in the examples.

Camelo et al. (2015) assumed that the data have been observed conditionally on the baseline variables, i.e., we have the input distribution of $p(y, l, \mathbf{h}, s, \mathbf{m} \mid \mathbf{b})$ available. With this data alone, the causal effect $p(y \mid \text{do}(l))$ is not identifiable. Given the input distribution, identification invariance holds when $\mathbf{M}$, $\mathbf{H}$ and $\mathbf{B}$ are clustered in any order. This follows because in the input distribution L blocks all backdoor paths between $\mathbf{M}$ and its children, while L and S block all backdoor paths between $\mathbf{H}$ and its children. In addition, the data source is not conditioned by a descendant of any cluster. Finally, transit cluster $\mathbf{B}$ does not have any paths entering the cluster, satisfying the condition of line 2. Therefore, the necessary conditions of VerifyInputs apply to each cluster. Using Theorem 18, we can

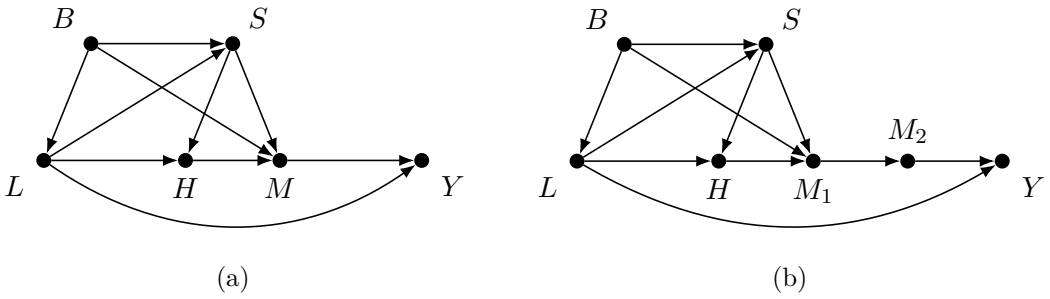

Figure 10: Causal graphs related to (Camelo et al., 2015): (a) the original graph and (b) a graph where we assume that the cluster $\mathbf{M}$ consists of two vertices M_1 and M_2.

conclude that the causal effect is non-identifiable regardless of the internal structures of the clusters $\mathbf{B}$, $\mathbf{H}$ and $\mathbf{M}$.

Now, suppose we have additional data on the baseline variables in the Brazilian population. We can denote the combination of this information as a single input distribution $p(y, l, \mathbf{h}, s, \mathbf{m}, \mathbf{b})$. By Theorem 17, we obtain the identification formula $p(y \,|\, \mathrm{do}(l)) = \sum_b p(b)p(y \,|\, l, b)$ and by Theorem 21 we may write

$$p(y \,|\, \mathrm{do}(l)) = \sum_{\mathbf{b}} p(\mathbf{b})p(y \,|\, l, \mathbf{b}).$$

Finally, we consider the effect of all health-related variables on intima-media thickness, i.e., $p(y \,|\, \mathrm{do}(\mathbf{h}))$, with input distributions of $p(\mathbf{b}, \mathbf{m}, s, y)$ and $p(\mathbf{b}, \mathbf{h}, \mathbf{m}, s)$. Now, we cannot identify the causal effect in the clustered graph. In addition, Theorem 18 does not apply to this setting because for the first data source there exists an open backdoor path between $\mathbf{M}$ and Y that is $Y \leftarrow L \rightarrow \mathbf{H} \rightarrow \mathbf{M}$. Therefore, it is possible that $p(y \,|\, \mathrm{do}(\mathbf{h}))$ could be identifiable in a graph where the variables in $\mathbf{M}$ are not clustered. To show that this indeed is the case, we consider the graph of Figure 10b where the transit cluster $\mathbf{M}$ consists of two vertices M_1 and M_2. Now, using the same input distributions $p(\mathbf{b}, m_1, m_2, s, y)$ and $p(\mathbf{b}, \mathbf{h}, m_1, m_2, s)$, where $\mathbf{M}$ is not clustered, the causal effect can be identified as

$$p(y \,|\, \mathrm{do}(\mathbf{h})) = \sum_{s, m_2, \mathbf{b}} \left(\sum_{m_1} p(s, \mathbf{b}) p(m_1, m_2 \,|\, \mathbf{h}, s, \mathbf{b}) \right) \left(\sum_{m_1} p(m_1 \,|\, s, \mathbf{b}) p(y \,|\, s, m_1, m_2, \mathbf{b}) \right).$$

This also demonstrates the necessity of line 7 in VerifyInputs. All four settings are summarized in Table 5.

8. Conclusion

We have proposed pruning and clustering as preprocessing operations for causal data fusion and derived sufficient conditions for identification invariance under these operations. Preprocessing enables the presentation of causal graphs in a more concise manner and leads to more efficient application of identification methods and simplified identifying functionals. In addition, pruning and clustering can serve as valuable tools in the planning of data

Causal effect	Fig.	Input distributions	Clusters	Invariance	ID
$p(y \mid \mathrm{do}(l))$	10a	$p(y, l, \mathbf{h}, s, \mathbf{m} \mid \mathbf{b})$	**B**, **H**, **M**	**B**, **H**, **M**	No
$p(y \mid \mathrm{do}(l))$	10a	$p(y, l, \mathbf{h}, s, \mathbf{m}, \mathbf{b})$	**B**, **H**, **M**	**B**, **H**, **M**	Yes
$p(y \mid \mathrm{do}(\mathbf{h}))$	10a	$p(\mathbf{b}, \mathbf{m}, s, y), p(\mathbf{b}, \mathbf{h}, \mathbf{m}, s)$	**B**, **M**	**B**	No
$p(y \mid \mathrm{do}(\mathbf{h}))$	10b	$p(\mathbf{b}, m_1, m_2, s, y), p(\mathbf{b}, \mathbf{h}, m_1, m_2, s)$	**B**	**B**	Yes

Table 5: Summary of the results of Section 7.2. The table presents the causal effects, the graphs used in the identification, the available input distributions, the clusters considered and for which clusters the identification invariance holds according to Theorems 17 and 18, and whether the causal effect is identifiable in the clustered graph.

collection as it is not necessary to measure variables that can be pruned, and for transit clusters it suffices to measure only the emitters of the cluster.

The practical benefits of pruning and clustering were demonstrated in Section 7 with graphs that have been previously used in applied works. The benefits in identification are the most obvious when the variables in different data sources are partially overlapping because the current identification algorithms with polynomial complexity are not applicable in such settings. The simulation study presented in Section 6 indicates that in moderately sized graphs, causal effect identification with clustering and pruning can be substantially faster, and even a small reduction in the graph size can decrease the running time considerably. The study also showed that when the graph could not be reduced, i.e., identification invariance could not be established for any graph reduction operation, the overhead for assessing the graphical conditions for identification invariance is insignificant compared to running the Do-search algorithm for the full graph.

We focused solely on marginal causal effects, but similar pruning and clustering results could likely be derived for conditional causal effects as well. Given a conditional causal effect of interest $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}), \mathbf{z})$, we can still apply the presented results to $p(\mathbf{y}, \mathbf{z} \mid \mathrm{do}(\mathbf{x}))$ as sufficient criteria. In addition to transit clusters, there may also exist other types of clusters for which identification invariance can be established, but we emphasize that transit clusters can be found efficiently (Tikka et al., 2023) and they have an intuitive interpretation.

Apart from Theorem 7 being the first pruning operation, we do not impose any particular order for performing the rest of the pruning operations or clustering. However, it is intuitive to think that the vertices that are completely irrelevant regarding the causal effect identification are removed first, after which we aim to find the transit clusters. Performing pruning first also removes the possibility that we end up pruning the clusters, which would render the clustering redundant.

We emphasize that applying pruning and clustering is a decision made by the researcher and there might be reasons to avoid these operations even when they are technically possible. We also acknowledge that there may be multiple ways to cluster a graph. In such situations, one approach could be to start with the largest cluster. Should the identification invariance be established, this approach would offer the largest benefits with respect to identification. Another approach could be to incorporate domain knowledge into clustering so that the cluster has a straightforward interpretation.

The theoretical results were derived without concepts such as c-components, hedges, or the latent projection which are essential when operating with a single observational data source but are not well-defined for the general causal effect identification problem where the variables may be observed in one data source but unobserved in another. Instead, to prove identification invariance between two causal graphs, we relied directly on the definition of identifiability and constructed the models that demonstrate non-identifiability in one graph starting from models with the same property that exist for the other graph by assumption. Kivva et al. (2022) employed a similar proof strategy for g-identifiability by establishing a connection between the original graph and a modified graph.

Identifying functionals obtainable via do-calculus in the pruned or clustered causal graphs are directly applicable to the original graphs as demonstrated by Theorem 20 and Theorem 21. We note that the limitation to functionals obtainable specifically via do-calculus is merely technical because in practice no other methods to identify causal effects are known for the general non-parametric setting. Nonetheless, it is of theoretical interest whether do-calculus is complete for the general causal effect identifiability problem.

Acknowledgments

The authors wish to thank Jouni Helske and the anonymous referees for their comments that helped to improve the paper. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources. This work was supported by the Finnish Doctoral Program Network in Artificial Intelligence (AI-DOC), Decision VN/3137/2024-OKM-6 and by the Research Council of Finland under grant number 368935.

A. Proofs for Section 3 (Preprocessing by Pruning)

Theorem 7 (Pruning non-ancestors). *Let $\mathcal{G}(\mathbf{V} \cup \mathbf{U})$ be a causal graph. If $\mathbf{B}_i \cup \mathbf{C}_i \subset \mathrm{an}_{\mathcal{G}}(\mathbf{Y})$ for all $i = 1, \ldots, n$, then pruning $(\mathcal{G}, \mathcal{I}) \to (\mathcal{G}', \mathcal{I}')$, where $\mathcal{I}' = \mathcal{I}[\mathrm{an}_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y}]$, $\mathcal{G}' = \mathcal{G}[\mathrm{an}^{+}_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y}]$ and $\mathbf{X}' = \mathbf{X} \cap \mathrm{an}_{\mathcal{G}}(\mathbf{Y})$, is identification invariant for $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$, and $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) = p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}'))$.*

Proof Let $\mathbf{W} = \mathbf{V} \setminus (\mathrm{an}_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y})$. By applying rule 3 of do-calculus twice and then noting that a causal model that induces $\mathcal{G}'$ has the same joint distribution over $\mathbf{W}$ as the submodel obtained from the intervention $\mathrm{do}(\mathbf{W} = \mathbf{w})$, we have that

$$p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) = p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}')) = p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}', \mathbf{w})) = p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}')).$$

This shows that the causal effects are the same in $\mathcal{G}$ and $\mathcal{G}'$. It remains to show that the identifiability of the causal effect is not affected by the removal of the non-ancestors.

Assume first that $p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is identifiable from $\mathcal{I}'$ in $\mathcal{G}'$. Then, members of $\mathcal{I}'$ are directly identifiable from $\mathcal{I}$ through marginalization, because of the assumption $\mathbf{B}_i \cup \mathbf{C}_i \subset \mathrm{an}_{\mathcal{G}}(\mathbf{Y})$ and because no member of $\mathbf{V} \setminus (\mathrm{an}_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y})$ appears in the structural equations of members of $\mathrm{an}_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y}$. Because of this, and because the causal effects are the same, $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is identifiable from $\mathcal{I}$ in $\mathcal{G}$ with the same identifying functional as in $\mathcal{G}'$.

Assume then that $p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is not identifiable. It follows that there exist two models $\mathcal{M}'_1$ and $\mathcal{M}'_2$ for which the input distributions $\mathcal{I}'_1$ and $\mathcal{I}'_2$ agree but $p'_1(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) \neq p'_2(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ for some value assignment $\mathbf{x}$. The goal is to construct two models $\mathcal{M}_1$ and $\mathcal{M}_2$ for $\mathcal{G}$ for which the input distributions $\mathcal{I}$ agree but $p_1(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) \neq p_2(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ for some value assignment $\mathbf{x}$. We construct these models as follows. The functions for the members of $\mathrm{an}_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y}$ are copied from $\mathcal{M}'_1$ to $\mathcal{M}_1$ and analogously from $\mathcal{M}'_2$ into $\mathcal{M}_2$. This guarantees that the causal effects differ between $\mathcal{M}_1$ and $\mathcal{M}_2$. The functions for the members of $\mathbf{V} \setminus (\mathrm{an}_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y})$ are defined the same way for both $\mathcal{M}_1$ and $\mathcal{M}_2$ so that variables $\mathbf{V} \setminus (\mathrm{an}_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y})$ are Bernoulli distributed random variables independent of all other variables in $\mathbf{V} \cup \mathbf{U}$ but their dedicated error terms. This guarantees that $\mathcal{I}_1 = \mathcal{I}_2$, and together with the difference in the causal effects, it applies that $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is not identifiable. ■

Theorem 8 (Pruning non-ancestors after intervention). *Let $\mathcal{G}(\mathbf{V} \cup \mathbf{U})$ be a causal graph such that $\mathcal{G} = \mathcal{G}[\mathrm{an}^{+}_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y}]$ and let $\mathbf{Z} = \{Z \in \mathbf{V} \setminus \mathbf{X} : Z \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{X}$ in $\mathcal{G}[\overline{\mathbf{X}}]\}$, $\mathbf{V}' = \mathbf{V} \setminus \mathbf{Z}$ and $\mathbf{U}' = \mathbf{U}[\mathbf{V}']$. If all of the following conditions hold*

(a) $\mathbf{Z} \cap \mathrm{de}_{\mathcal{G}}(\mathbf{X}) = \emptyset$,

(b) $\mathbf{B}_i \cup \mathbf{C}_i \subset \mathbf{V}'$ *for all* $i = 1, \ldots, n$, *and*

(c) for all members of $\mathbf{U}[\mathbf{Z}]$ it holds that if $U \in \mathbf{U}[\mathbf{Z}]$ is an ancestor of all variables in $\mathbf{X}_S \subseteq (\mathbf{X} \cap \mathrm{ch}_{\mathcal{G}}(\mathbf{Z} \cup \mathbf{U}[\mathbf{Z}]))$, then there exists $U_S \in \mathbf{U}$ that is a parent of all members of $\mathbf{X}_S$,

then pruning $(\mathcal{G}, \mathcal{I}) \to (\mathcal{G}', \mathcal{I}')$, where $\mathcal{I}' = \mathcal{I}[\mathbf{V}']$ and $\mathcal{G}' = \mathcal{G}[\mathbf{V}' \cup \mathbf{U}']$, is identification invariant for $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$, and $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) = p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$.

Proof By assumption, $\mathcal{G}$ only contains $\mathbf{Y}$ and the ancestors of $\mathbf{Y}$. The definition of $\mathbf{Z}$ allows us to use rule 3 of do-calculus, and thus we have $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) = p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}, \mathbf{z}))$. Applying the truncated factorization formula (Pearl, 2009) we can write

$$\begin{aligned} p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) &= p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}, \mathbf{z})) \\ &= \sum_{\mathbf{u}} \sum_{\mathbf{v} \setminus (\mathbf{y} \cup \mathbf{x} \cup \mathbf{z})} \prod_{\mathbf{v} \setminus (\mathbf{x} \cup \mathbf{z})} p(v_i \,|\, \mathrm{pa}(v_i), \mathbf{u}[v_i]) \prod_{\mathbf{u}} p(u_i) \\ &= \sum_{\mathbf{u} \setminus \mathbf{u}'} \sum_{\mathbf{u}'} \sum_{\mathbf{v} \setminus (\mathbf{y} \cup \mathbf{x} \cup \mathbf{z})} \prod_{\mathbf{v} \setminus (\mathbf{x} \cup \mathbf{z})} p(v_i \,|\, \mathrm{pa}(v_i), \mathbf{u}[v_i]) \prod_{\mathbf{u}'} p(u_i) \prod_{\mathbf{u} \setminus \mathbf{u}'} p(u_i). \end{aligned}$$

As no children of $\mathbf{U} \setminus \mathbf{U}'$ are present in the expression, we can sum $\mathbf{U} \setminus \mathbf{U}'$ out and obtain

$$\begin{aligned} p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) &= \sum_{\mathbf{u}'} \sum_{\mathbf{v} \setminus (\mathbf{y} \cup \mathbf{x} \cup \mathbf{z})} \prod_{\mathbf{v} \setminus (\mathbf{x} \cup \mathbf{z})} p(v_i \,|\, \mathrm{pa}(v_i), \mathbf{u}[v_i]) \prod_{\mathbf{u}'} p(u_i) \\ &= \sum_{\mathbf{u}'} \sum_{\mathbf{v}' \setminus (\mathbf{y} \cup \mathbf{x})} \prod_{\mathbf{v}' \setminus \mathbf{x}} p'(v_i \,|\, \mathrm{pa}(v_i), \mathbf{u}[v_i]) \prod_{\mathbf{u}'} p'(u_i) \\ &= p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})). \end{aligned}$$

In other words, pruning does not change the causal effect $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$. It remains to show that the identifiability of the causal effect is preserved.

Assume first that causal effect $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is not identifiable from $\mathcal{I}$ in $\mathcal{G}$. It follows that there exist two models $\mathcal{M}_1$ and $\mathcal{M}_2$ for which $p_1(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) \neq p_2(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ but the input distributions $\mathcal{I}_1$ and $\mathcal{I}_2$ are the same. We may divide $\mathbf{U}[\mathbf{Z}]$ into partition $\mathcal{U}$ in such a way that all members of $\mathbf{U}[\mathbf{Z}]$ that have the same set of descendants in $\mathbf{X} \cap \mathrm{ch}_{\mathcal{G}}(\mathbf{Z} \cup \mathbf{U}[\mathbf{Z}])$ belong to the same class, i.e. U_i and U_j belong to the same class if $U_i, U_j \in \mathbf{U}[\mathbf{Z}]$ and $\mathrm{de}_{\mathcal{G}}(U_i) \cap (\mathbf{X} \cap \mathrm{ch}_{\mathcal{G}}(\mathbf{Z} \cup \mathbf{U}[\mathbf{Z}])) = \mathrm{de}_{\mathcal{G}}(U_j) \cap (\mathbf{X} \cap \mathrm{ch}_{\mathcal{G}}(\mathbf{Z} \cup \mathbf{U}[\mathbf{Z}]))$. The class whose children in the set $\mathbf{X} \cap \mathrm{ch}_{\mathcal{G}}(\mathbf{Z} \cup \mathbf{U}[\mathbf{Z}])$ are $\mathbf{X}_S$ is denoted by $\mathcal{U}[\mathbf{X}_S]$. It follows from condition (c) that there exists a surjective mapping g from $\mathcal{U}$ to $\mathbf{U}$ so that $U_S = g(\mathcal{U}[\mathbf{X}_S])$ is a parent of all members of $\mathbf{X}_S$. This is fulfilled, for instance, if there exists $U \in \mathbf{U}$ that is a parent of all members of $\mathbf{X} \cap \mathrm{ch}_{\mathcal{G}}(\mathbf{Z} \cup \mathbf{U}[\mathbf{Z}])$. Models $\mathcal{M}_1'$ and $\mathcal{M}_2'$ are obtained from $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively, by the following process which is presented only for $\mathcal{M}_1'$:

1. The structural equations for variables $\mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Z})$ are copied from $\mathcal{M}_1$ to $\mathcal{M}_1'$. The assumption $\mathcal{G} = \mathcal{G}[\mathrm{an}^+_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y}]$, the definition of $\mathbf{Z}$, and condition (a) guarantee that these structural equations do not contain members of $\mathbf{Z}$.

2. The variable U_S is redefined to be a vector-valued variable that contains all members of $\mathbf{U}[\mathbf{Z}]$ that belong to any class $\mathcal{U}[\mathbf{X}_S]$ for which $U_S = g(\mathcal{U}[\mathbf{X}_S])$. The distribution of U_S in $\mathcal{M}_1'$ is defined to be equal to the joint distribution of the members of all classes $\mathcal{U}[\mathbf{X}_S]$ in $\mathcal{M}_1$ for which $U_S = g(\mathcal{U}[\mathbf{X}_S])$.

3. The structural equations for variables in $\mathbf{X}$ are copied from $\mathcal{M}_1$ into $\mathcal{M}_1'$ and then redefined by the following procedure:

   3.1. If $X \in \mathbf{X}$ is a function of a variable $Z \in \mathbf{Z}$, variable Z is replaced by its function $f_Z(\mathrm{pa}(Z), \mathbf{U}[Z])$. From condition (a) it follows that $\mathrm{pa}(Z)$ cannot contain members of $\mathbf{X}$.

3.2. Step 3.1 is repeated until the structural equation does not contain any members of $\mathbf{Z}$.

3.3. If X is a function of a variable U_j and U_j belongs to class $\mathcal{U}[\mathbf{X}_S]$, then U_j is replaced by $U_S = g(\mathcal{U}[\mathbf{X}_S])$ and X becomes a function of the component of U_S that corresponds to U_j.

As a result of these steps, the distribution of $\mathbf{X}$ is equal between $\mathcal{M}'_1$ and $\mathcal{M}_1$, and between $\mathcal{M}'_2$ and $\mathcal{M}_2$. It follows from this and the first step of the procedure that the input distributions are the same between $\mathcal{M}'_1$ and $\mathcal{M}'_2$. Condition (b) guarantees that the input distributions $\mathcal{I}'$ are well-defined. Since it was shown above that $p_1(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) = p'_1(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ and $p_2(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) = p'_2(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$, the causal effects also differ between $\mathcal{M}'_1$ and $\mathcal{M}'_2$.

Assume then that causal effect $p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is not identifiable from $\mathcal{I}'$ in $\mathcal{G}'$. It follows that there exist models $\mathcal{M}'_1$ and $\mathcal{M}'_2$ for which $p'_1(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) \neq p'_2(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ but the input distributions $\mathcal{I}'_1$ and $\mathcal{I}'_2$ are the same. We construct models $\mathcal{M}_1$ and $\mathcal{M}_2$ from $\mathcal{M}'_1$ and $\mathcal{M}'_2$, respectively, by defining variables $\mathbf{Z}$ as Bernoulli variables independent of other variables except for their dedicated error terms. Now $p_1(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) \neq p_2(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ but the input distributions $\mathcal{I}_1$ and $\mathcal{I}_2$ are the same. ■

Theorem 9 (Pruning isolated vertices)**.** *Let $\mathcal{G}(\mathbf{V} \cup \mathbf{U})$ be a causal graph such that $\mathcal{G} = \mathcal{G}[\mathrm{an}^+_{\mathcal{G}}(\mathbf{Y}) \cup \mathbf{Y}]$ and let $W \in \mathbf{V}$. If there exists a set $\mathbf{Z} \subset \mathbf{V}$ such that $\mathbf{Z} \cap (\mathbf{X} \cup \mathbf{Y}) = \emptyset$, $\mathbf{Z}$ is connected to $\mathbf{V} \setminus \mathbf{Z}$ only through W, and $\mathbf{B}_i \cup \mathbf{C}_i \subset \mathbf{V} \setminus \mathbf{Z}$ for all $i = 1, \ldots, n$, then pruning $(\mathcal{G}, \mathcal{I}) \to (\mathcal{G}', \mathcal{I}')$, where $\mathcal{I}' = \mathcal{I}[\mathbf{V}']$, $\mathcal{G}' = \mathcal{G}[\mathbf{V}' \cup \mathbf{U}']$ and $\mathbf{V}' = \mathbf{V} \setminus \mathbf{Z}$ and $\mathbf{U}' = \mathbf{U}[\mathbf{V}']$, is identification invariant for $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$, and $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) = p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$.*

Proof Applying the truncated factorization formula, we obtain

$$p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) = \sum_{\mathbf{u}} \sum_{\mathbf{v} \setminus (\mathbf{y} \cup \mathbf{x})} p(w \,|\, \mathrm{pa}(w), \mathbf{u}[w]) \prod_{\mathbf{v} \setminus (\mathbf{x} \cup w)} p(v_i \,|\, \mathrm{pa}(v_i), \mathbf{u}[v_i]) \prod_{\mathbf{u}} p(u_i).$$

By definition, members of $\mathbf{Z}$ and $\mathbf{U}[\mathbf{Z}]$ are connected to $\mathbf{V} \setminus (\mathbf{Z} \cup W)$ only through W. By assumption, $\mathcal{G}$ does not contain any non-ancestors of $\mathbf{Y}$, which means that all members of $\mathbf{Z}$ must be ancestors of W. Thus, we can complete the marginalization over $\mathbf{Z}$ and $\mathbf{U}[\mathbf{Z}] \setminus \mathbf{U}[W]$ to obtain

$$\begin{aligned}
&p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) \\
&= \sum_{(\mathbf{u} \setminus \mathbf{u}[\mathbf{z}]) \cup \mathbf{u}[w]} \sum_{\mathbf{v} \setminus (\mathbf{y} \cup \mathbf{x} \cup \mathbf{z})} p(w | \mathrm{pa}(w) \setminus \mathbf{z}, \mathbf{u}[w]) \prod_{\mathbf{v} \setminus (\mathbf{x} \cup \mathbf{z} \cup w)} p(v_i | \mathrm{pa}(v_i), \mathbf{u}[v_i]) \prod_{(\mathbf{u} \setminus \mathbf{u}[\mathbf{z}]) \cup \mathbf{u}[w]} p(u_i) \\
&= \sum_{\mathbf{u}'} \sum_{\mathbf{v}' \setminus (\mathbf{y} \cup \mathbf{x})} p'(w \,|\, \mathrm{pa}'(w), \mathbf{u}'[w]) \prod_{\mathbf{v}' \setminus (\mathbf{x} \cup w)} p'(v_i \,|\, \mathrm{pa}'(v_i), \mathbf{u}'[v_i]) \prod_{\mathbf{u}'} p'(u_i) \\
&= \sum_{\mathbf{u}'} \sum_{\mathbf{v}' \setminus (\mathbf{y} \cup \mathbf{x})} \prod_{\mathbf{v}' \setminus \mathbf{x}} p'(v_i \,|\, \mathrm{pa}'(v_i), \mathbf{u}'[v_i]) \prod_{\mathbf{u}'} p'(u_i) \\
&= p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))
\end{aligned}$$

Thus, the causal effect is unaltered by the pruning. It remains to show that the identifiability of the causal effect is preserved.

Assume first that $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ is not identifiable from $\mathcal{I}$ in $\mathcal{G}$. It follows that there exist models $\mathcal{M}_1$ and $\mathcal{M}_2$ for which $p_1(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) \neq p_2(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ but the input distributions $\mathcal{I}_1$ and $\mathcal{I}_2$ are the same. We construct $\mathcal{M}_1'$ and $\mathcal{M}_2'$ from $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. First, the structural equations for variables in $\mathbf{V} \setminus (\mathbf{Z} \cup W)$ are copied from $\mathcal{M}_1$ into $\mathcal{M}_1'$ and from $\mathcal{M}_2$ into $\mathcal{M}_2'$. Next, the structural equation for W in $\mathcal{M}_1'$ is defined by first copying the function of W in $\mathcal{M}_1$ and then replacing every instance of members of $\mathrm{pa}_{\mathcal{G}}(W) \cap \mathbf{Z}$ by its corresponding function recursively until the function of W is expressed in terms of $\mathbf{U}$ only. The same process is carried out for $\mathcal{M}_2'$. This ensures that $\mathcal{I}_1' = \mathcal{I}_2'$, but $p_1'(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) \neq p_2'(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$.

Assume then that $p'(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ is not identifiable from $\mathcal{I}'$ in $\mathcal{G}'$. It follows that there exist models $\mathcal{M}_1'$ and $\mathcal{M}_2'$ for which $p_1'(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) \neq p_2'(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ but the input distributions $\mathcal{I}_1'$ and $\mathcal{I}_2'$ are the same. We construct models $\mathcal{M}_1$ and $\mathcal{M}_2$ from $\mathcal{M}_1'$ and $\mathcal{M}_2'$, respectively, by defining variables $\mathbf{Z}$ as Bernoulli variables independent of other variables except for their dedicated error terms in both models. The functions for W are constructed so that the value of W does not depend on the parents of W in $\mathbf{Z}$. Now $p_1(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) \neq p_2(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ but the input distributions $\mathcal{I}_1$ and $\mathcal{I}_2$ are the same. ■

B. Proofs for Section 4 (Clustering Variables in Causal Graphs)

Lemma 15. *A transit cluster* $\mathbf{T}$ *in* $\mathcal{G}$ *is also a transit cluster in* $\mathcal{G}[\overline{\mathbf{Z}}, \underline{\mathbf{W}}]$ *where* $\mathbf{Z}$ *and* $\mathbf{W}$ *are disjoint, possibly empty, subsets of* $\mathbf{V}$ *and* $\mathbf{T}$ *is either a subset of* $\mathbf{Z}$*, a subset of* $\mathbf{W}$*, or disjoint from both* $\mathbf{Z}$ *and* $\mathbf{W}$*.*

Proof By definition, $\mathbf{T}$ is either a subset of $\mathbf{Z}$, a subset of $\mathbf{W}$, or disjoint from both $\mathbf{Z}$ and $\mathbf{W}$. If $\mathbf{T} \subseteq \mathbf{Z}$ then $\mathbf{T}$ does not have parents in $\mathcal{G}[\overline{\mathbf{Z}}, \underline{\mathbf{W}}]$ and the set of receivers of $\mathbf{T}$ is empty. If $\mathbf{T} \subseteq \mathbf{W}$ then $\mathbf{T}$ does not have children in $\mathcal{G}[\overline{\mathbf{Z}}, \underline{\mathbf{W}}]$ and the set of emitters of $\mathbf{T}$ is empty. In both cases, $\mathbf{T}$ is a transit cluster in $\mathcal{G}[\overline{\mathbf{Z}}, \underline{\mathbf{W}}]$ after these operations. Finally, we consider the case where $\mathbf{T}$ is disjoint from both $\mathbf{Z}$ and $\mathbf{W}$. The transit cluster remains unchanged if $\mathbf{Z}$ does not contain children of $\mathbf{T}$ and $\mathbf{W}$ does not contain parents of $\mathbf{T}$. If $\mathbf{Z}$ contains a child of $\mathbf{T}$, all edges between $\mathbf{T}$ and this child are removed. If $\mathbf{W}$ contains a parent of $\mathbf{T}$, all edges between this parent and $\mathbf{T}$ are removed. In these cases, there are no other changes and $\mathbf{T}$ remains as a transit cluster after these operations. We conclude that the parents and the children of $\mathbf{T}$ may change when edges are removed but in all cases $\mathbf{T}$ is a transit cluster in $\mathcal{G}[\overline{\mathbf{Z}}, \underline{\mathbf{W}}]$. ■

Lemma 16. *Let* $\mathbf{T}$ *be a transit cluster in* $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ *and let* $\mathcal{G}'(\mathbf{V}' \cup \widetilde{\mathbf{U}}') = \mathcal{G}[\mathbf{T} \to T]$*. Let* $\mathbf{X}'$*,* $\mathbf{Y}'$ *and* $\mathbf{W}'$ *be disjoint subsets of* $\mathbf{V}'$ *and define* $\mathbf{X}$*,* $\mathbf{Y}$ *and* $\mathbf{W}$ *to be the corresponding subsets of* $\mathbf{V}$ *when* T *is replaced by* $\mathbf{T}$*. Then* $\mathbf{X}'$ *and* $\mathbf{Y}'$ *are d-separated by* $\mathbf{W}'$ *in* $\mathcal{G}'$ *if and only if* $\mathbf{X}$ *and* $\mathbf{Y}$ *are d-separated by* $\mathbf{W}$ *in* $\mathcal{G}$*.*

Proof We choose arbitrary $V_1 \in \mathbf{X}'$ and arbitrary $V_2 \in \mathbf{Y}'$ and consider a path s between V_1 and V_2 in $\mathcal{G}$. First, we study the case where s does not contain members of $\mathbf{T}$. This means that the vertices of the path remain unchanged after clustering $\mathbf{T}$ and we may denote $s' = s$ where s' is the corresponding path in $\mathcal{G}_*'$. If $T \notin \mathbf{W}'$, clustering $\mathbf{T}$ as T does not affect

whether s is open or not. If $T \in \mathbf{W}'$, the properties of the transit cluster ensure that s is blocked by $\text{em}_{\mathcal{G}}(\mathbf{T})$ in $\mathcal{G}_*$ if and only if it is blocked by T in $\mathcal{G}'_*$.

Next, we study the more complicated case where s contains members of $\mathbf{T}$. First, we assume that $V_1 \neq T$, and $V_2 \neq T$. In this case, there may not be a corresponding path in $\mathcal{G}'_*$ because path s can go through $\mathbf{T}$ multiple times in $\mathcal{G}_*$ which is not possible in $\mathcal{G}'_*$. To address this challenge, we show that there exists in $\mathcal{G}_*$ a path s^* from V_1 to V_2 that goes through $\mathbf{T}$ only once. To construct the path s^*, we start by writing the path s as follows

$$s = V_1, s_1, T_1, s_0, T_2, s_2, V_2,$$

where s_0, s_1, and s_2 are paths, T_1 is the first member of $\mathbf{T}$ on path s when starting from V_1, and T_2 is the first member of $\mathbf{T}$ on path s when starting from V_2. In other words, s_1 and s_2 do not contain members of $\mathbf{T}$ but s_0 may contain members and non-members of $\mathbf{T}$. We define path s^* as follows

$$s^* = \begin{cases} V_1, s_1, T_1, s_0^*, T_2, s_2, V_2 & \text{if } (T_1 \in \text{re}_{\mathcal{G}_*}(\mathbf{T}) \text{ and } T_2 \in \text{em}_{\mathcal{G}_*}(\mathbf{T})) \text{ or} \\ & \quad (T_1 \in \text{em}_{\mathcal{G}_*}(\mathbf{T}) \text{ and } T_2 \in \text{re}_{\mathcal{G}_*}(\mathbf{T})), \\ V_1, s_1, T_1, s_2, V_2 & \text{if } T_1, T_2 \in \text{re}_{\mathcal{G}_*}(\mathbf{T}) \text{ or } T_1, T_2 \in \text{em}_{\mathcal{G}_*}(\mathbf{T}), \end{cases}$$

where the path s_0^* contains only members of $\mathbf{T}$. The existence of s_0^* follows from conditions (d) and (e) of Definition 13. In the case $T_1, T_2 \in \text{re}_{\mathcal{G}_*}(\mathbf{T})$ or $T_1, T_2 \in \text{em}_{\mathcal{G}_*}(\mathbf{T})$, the existence of edge from T_1 to path s_2 follows from conditions (a) and (b) of Definition 13. If path s^* is blocked, then path s is blocked as well because the vertices of s^* that are outside $\mathbf{T}$ are also in s. This means we may consider only paths like s^* that go through $\mathbf{T}$ only once.

For each path s^* in $\mathcal{G}_*$, there exists the corresponding path s' in $\mathcal{G}'_*$ defined as follows

$$s' = V_1, s_1, T, s_2, V_2.$$

The vertices other than T and members of $\mathbf{T}$ are the same on the paths s^* and s'. If s' is blocked by a vertex that is not T, s^* is blocked by the same vertex. If s' is blocked by T and T is a collider then s^* is blocked by a receiver that is a collider in $\mathcal{G}_*$. If s' is blocked by T when T is not a collider then s^* is blocked by $\text{em}_{\mathcal{G}}(\mathbf{T})$ and by Definition 14, no emitter can be a member of $\widetilde{\mathbf{U}}$. We conclude that if s' is blocked in $\mathcal{G}'_*$ then s^* is blocked in $\mathcal{G}$. If all paths between V_1 and V_2 given $\mathbf{W}'$ are blocked in $\mathcal{G}'_*$ then all paths between V_1 and V_2 given $\mathbf{W}$ are blocked in $\mathcal{G}_*$. By the properties of a transit cluster, all paths between V_1 and V_2 are blocked also if conditioning set $\mathbf{W}$ contains other members of $\mathbf{T}$ in addition to $\text{em}_{\mathcal{G}}(\mathbf{T})$.

Finally, we consider the case $V_1 = T$. Let $T_1 \in \mathbf{T}$ be the first vertex on path s that can be now written as

$$s = T_1, s_0, T_2, s_2, V_2,$$

where s_0 and s_2 are paths, and T_2 is the first member of $\mathbf{T}$ on path s when starting from V_2. The conditions (d) and (e) of Definition 13 allow us to define path s^* as

$$s^* = T_1, s_0^*, T_2, s_2, V_2,$$

where the path s_0^* contains only members of $\mathbf{T}$. Noting that $\mathbf{T} \cap \mathbf{W} = \emptyset$, we conclude s^* is blocked by $\mathbf{W}$ in $\mathcal{G}$ if and only if $s' = T, s_2, V_2$ is blocked by $\mathbf{W}' = \mathbf{W}$ in $\mathcal{G}'$. The case $V_2 = T$ is analogous.

In all cases, V_1 and V_2 are d-separated given $\mathbf{W}'$ in $\mathcal{G}'$ if and only if V_1 and V_2 are d-separated given $\mathbf{W}$ in $\mathcal{G}$. As V_1, V_2 and $\mathbf{W}'$ were chosen arbitrarily, the same conclusion applies to all choices. The claim follows because sets $\mathbf{X}$ and $\mathbf{Y}$ are by definition d-separated if and only if all paths between them are d-separated. ■

Theorem 17. *Let $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ be a causal graph where $\mathbf{T} \subset \mathbf{V} \cup \widetilde{\mathbf{U}}$ is a transit cluster and let $\mathcal{I}$ be a set of input distributions compatible with $\mathbf{T}$. Let $\mathcal{I}' = \mathcal{I}[\mathbf{T} \to T]$ in $\mathcal{G}' = \mathcal{G}[\mathbf{T} \to T]$. If $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is identifiable from $\mathcal{I}'$ in $\mathcal{G}'$, then it is identifiable from $\mathcal{I}$ in $\mathcal{G}$, where $\mathbf{X} \cup \mathbf{Y} \subset \mathbf{V}$, $\mathbf{Y} \cap \mathbf{X} = \emptyset$, $\mathbf{X} \cap \mathbf{T} = \emptyset$, $\mathbf{Y} \cap \mathbf{T} = \emptyset$.*

Proof Assume that $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is not identifiable in $\mathcal{G}$ from $\mathcal{I}$. By Definition 4, non-identifiability implies that there exist two models $\mathcal{M}_1$ and $\mathcal{M}_2$ for $\mathcal{G}$ for which the causal effects differ, but the input distributions agree. To show the non-identifiability in $\mathcal{G}'$, we need to find two models $\mathcal{M}_1'$ and $\mathcal{M}_2'$ with the same properties. We aim to construct these models $\mathcal{M}_1'$ and $\mathcal{M}_2'$ based on $\mathcal{M}_1$ and $\mathcal{M}_2$. We show only how $\mathcal{M}_1'$ is constructed from $\mathcal{M}_1$ because the construction is done similarly for $\mathcal{M}_1'$ and $\mathcal{M}_2'$. According to Definition 10, we have $\mathbf{V}' = (\mathbf{V} \setminus \mathbf{T}) \cup T$ and $\mathbf{U}' = (\mathbf{U} \setminus \mathbf{T}) \cup U[T]$. We define $U[T]$ as a vector-valued random variable whose components are $\mathbf{U}[\mathbf{T}]$ and set $p'(u[T])$ in $\mathcal{M}_1'$ to be equal to $p(\mathbf{t} \cap \mathbf{u})$ in $\mathcal{M}_1$. Other members of $\mathbf{U}'$ are distributed as they are distributed in $\mathcal{M}_1$. Since $\mathbf{T}$ is a transit cluster, we may factorize the distribution $p'(\mathbf{u}')$ as $p'(\mathbf{u}') = p(\mathbf{u} \setminus \mathbf{t})p'(u[T])$.

We define every variable in $\mathbf{V}'$, apart from T and $\mathrm{ch}_{\mathcal{G}'}(T)$, using the same structural equations as in $\mathcal{M}_1$, i.e., for each $V_j \notin \mathrm{ch}_{\mathcal{G}'}(T) \cup T$, we set $f_j' = f_j$. The random vector $T = (T_1, \ldots, T_m)$ contains as many components as there are emitters in $\mathbf{T}$. The structural equations for the components of T are constructed by the following iterative procedure:

1. Copy the structural equation of emitter E_j as the structural equation for component T_j.

2. If T_j is a function of a variable $Z \in (\mathbf{T} \cap \mathbf{V})$, variable Z is replaced by the function $f_Z(\mathrm{pa}(Z), \mathbf{U}[Z])$.

3. Step 2 is repeated until the structural equation does not contain any members of $\mathbf{T} \cap \mathbf{V}$.

4. Each member of $\mathbf{U}[\mathbf{T}]$ in the structural equation is replaced by the corresponding component of $U[T]$.

As a result of the procedure, each component of T becomes a function of $U[T]$ and the parents of $\mathbf{T}$ that are not members of $\mathbf{T}$. The structural equations for $\mathrm{ch}_{\mathcal{G}'}(T)$ are otherwise the same as the structural equations for $\mathrm{ch}_{\mathcal{G}}(\mathbf{T})$ but each emitter E_j is replaced by the corresponding component T_j of T. By this construction, we have $p'(T) = p(\mathrm{em}_{\mathcal{G}}(\mathbf{T}))$ and $p'(\mathrm{ch}_{\mathcal{G}'}(T)) = p(\mathrm{ch}_{\mathcal{G}}(\mathbf{T}))$. As the structural equations were copied directly from $\mathcal{M}_1$ to $\mathcal{M}_1'$, the construction does not change the functional relationships. This implies that the causal effects differ between $\mathcal{M}_1'$ and $\mathcal{M}_2'$ while all input distributions are equal. Thus the claim follows by contraposition. ■

Theorem 18. *Let $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ be a causal graph where $\mathbf{T} \subset \mathbf{V} \cup \widetilde{\mathbf{U}}$ is a transit cluster and let $\mathcal{I}$ be a set of input distributions compatible with $\mathbf{T}$. Let $\mathcal{I}' = \mathcal{I}[\mathbf{T} \to T]$ in $\mathcal{G}' = \mathcal{G}[\mathbf{T} \to T]$. If $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is not identifiable from $\mathcal{I}'$ in $\mathcal{G}'$ and if* VerifyInputs$(\mathcal{G}', \mathcal{I}', T)$ *returns* **TRUE**, *then $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is not identifiable from $\mathcal{I}$ in $\mathcal{G}$, where $\mathbf{X} \cup \mathbf{Y} \subset \mathbf{V}$, $\mathbf{Y} \cap \mathbf{X} = \emptyset$, $\mathbf{X} \cap \mathbf{T} = \emptyset$, $\mathbf{Y} \cap \mathbf{T} = \emptyset$.*

Proof By Definition 4, non-identifiability implies that there exist two models $\mathcal{M}'_1$ and $\mathcal{M}'_2$ for $\mathcal{G}'$ for which the causal effects differ but the input distributions agree. Let $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ be an arbitrarily chosen causal graph such that $\mathbf{T}$ is a transit cluster in $\mathcal{G}$ and $\mathcal{G}' = \mathcal{G}[\mathbf{T} \to T]$. To ensure the positivity of the joint distribution, $\mathbf{U}[\mathbf{T}]$ must contain a dedicated error term for each member of $\mathbf{T} \cap \mathbf{V}$.

To show non-identifiability in $\mathcal{G}$, we need to find two models $\mathcal{M}_1$ and $\mathcal{M}_2$ for which the causal effects differ but the input distributions agree. We construct these models based on $\mathcal{M}'_1$ and $\mathcal{M}'_2$. As the proof for non-identifiability is long, it will be divided into several parts. First, we will show how models $\mathcal{M}_1$ and $\mathcal{M}_2$ are constructed. Next, we will show that the input distributions are the same for $\mathcal{M}_1$ and $\mathcal{M}_2$. Finally, we will show that the causal effects are different between $\mathcal{M}_1$ and $\mathcal{M}_2$.

Constructing the models: By Definition 4, non-identifiability implies that we have two SCMs, $\mathcal{M}'_1 = (\mathbf{V}', \mathbf{U}', \mathbf{F}'_1, p'_1)$ and $\mathcal{M}'_2 = (\mathbf{V}', \mathbf{U}', \mathbf{F}'_2, p'_2)$, for $\mathcal{G}'$ for which the causal effects differ but the input distributions are the same. We construct two models $\mathcal{M}_1$ and $\mathcal{M}_2$ for $\mathcal{G}$ so that for each input distribution $p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$ and $p_1(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x})) \neq p_2(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$. We show in detail how $\mathcal{M}_1 = (\mathbf{V}, \mathbf{U}, \mathbf{F}_1, p_1)$ is constructed based on $\mathcal{M}'_1$. Model $\mathcal{M}_2$ is defined analogously based on $\mathcal{M}'_2$ instead. Without loss of generality, we assume that in $\mathcal{M}'_1$ and $\mathcal{M}'_2$, all variables in $\mathbf{V}'$ are scalar-valued, including T. We let $\mathbf{V} = (\mathbf{V}' \setminus T) \cup \mathbf{T}$ and $\mathbf{U} = (\mathbf{U}' \setminus \overline{U}[T]) \cup \overline{\mathbf{U}}[\mathbf{T}] \cup \widetilde{\mathbf{U}}[\mathbf{T}]$. We let each variable in $\overline{\mathbf{U}}[\mathbf{T}] \cup \widetilde{\mathbf{U}}[\mathbf{T}]$ be independent of the other variables in $\mathbf{U}$ and uniformly distributed on the interval $(0, 1)$. We define $p_1(\mathbf{u}) = p'_1(\mathbf{u}')p_1(\overline{\mathbf{u}}[\mathbf{t}])p_1(\widetilde{\mathbf{u}}[\mathbf{t}])$. It remains to construct the functions $\mathbf{F}_1$ that define the structural equations for the variables in $\mathbf{V}$. From Definition 13 it follows $\mathrm{ch}_{\mathcal{G}'}(T) \setminus T = \mathrm{ch}_{\mathcal{G}}(\mathbf{T}) \setminus \mathbf{T} = \mathrm{ch}_{\mathcal{G}}(\mathrm{em}_{\mathcal{G}}(\mathbf{T})) \setminus \mathbf{T}$. We define every variable, apart from $\mathrm{ch}_{\mathcal{G}}(\mathbf{T}) \cup \mathbf{T}$, using the same structural equations as in $\mathcal{M}'_1$, i.e., for each $V_j \notin \mathrm{ch}_{\mathcal{G}}(\mathbf{T}) \cup \mathbf{T}$, we let $f_j = f'_j$. This is possible because these vertices have the same incoming edges before and after the clustering and therefore share the same set of parents in both cases.

Next, we define f_j for each $V_j \subset \mathbf{T} \cap \mathbf{V}$. If $\mathrm{re}_{\mathcal{G}}(\mathbf{T}) \neq \emptyset$, we select a receiver $R_1 \in \mathrm{re}_{\mathcal{G}}(\mathbf{T})$ that does not have other receivers of $\mathbf{T}$ as its parents. Definition 13 guarantees that if also $\mathrm{em}_{\mathcal{G}}(\mathbf{T}) \neq \emptyset$, there must be a directed path from R_1 to some emitter $E_1 \in \mathrm{em}_{\mathcal{G}}(\mathbf{T})$. If $\mathrm{re}_{\mathcal{G}}(\mathbf{T}) = \emptyset$, we choose any $T_j \in \mathrm{an}_{\mathcal{G}}(E_1) \cap \mathbf{T}$, for which $\mathrm{pa}_{\mathcal{G}}(T_j) = \emptyset$ as R_1, and if $\mathrm{em}_{\mathcal{G}}(\mathbf{T}) = \emptyset$ we choose any $T_j \in \mathrm{de}_{\mathcal{G}}(R_1) \cap \mathbf{T}$ as E_1. Without loss of generality, we may assume that the length of this directed path is greater than one, as the claim follows directly from Theorem 19 otherwise. If the cluster has no receivers, then we can follow similar reasoning as in the proof of Theorem 19, and simply select any emitter of $\mathbf{T}$ and copy the structural equation of T as the structural equation of this emitter, from which the equality of input distributions and the difference of causal effects follows directly.

The directed path to the emitter is written as $R_1 \to T_1 \to \ldots \to T_k \to E_1$. Now, let the dedicated error term for R_1 be $\overline{U}[T]$, i.e. $\overline{U}[R_1] = \overline{U}[T]$. Let R_1 have the same structural equation as T had in $\mathcal{M}'_1$, while other members of $\mathbf{U}$ that are parents of R_1 are set to not affect R_1, i.e. $R_1 = f_{R_1}(\mathrm{pa}(R_1), \mathbf{U}[R_1]) = f_{R_1}(\mathrm{pa}(T), \overline{U}[T]) = f_T(\mathrm{pa}(T), \overline{U}[T]) = T$. To

define the structural equations for the rest of the vertices in the directed path, we apply the following inheritance function

$$h(X,U) = \begin{cases} X & \text{if } U \leq \gamma, \\ g\left(\frac{U}{1-\gamma}\right) & \text{if } U > \gamma, \end{cases}$$

where $\gamma \in (0,1)$ is a fixed probability parameter and g is either the quantile function of the marginal distribution of T in $\mathcal{M}_1'$ if there exists $\mathcal{I}_i' \in \mathcal{I}'$, such that $T \in \mathbf{A}_i'$ and $\mathbf{B}_i' \cup \mathbf{C}_i' = \emptyset$. Otherwise, if such an input distribution does not exist, g is the quantile function of any arbitrary distribution over the domain of T in both models. If T was vector-valued, we would define g as a multivariate quantile function.

Let $\overline{U}[T_1], \ldots, \overline{U}[T_k]$ and $\overline{U}[E_1]$ be the dedicated error terms that are the parents of $T_1, \ldots, T_k$ and E_1, respectively. Now, we define $T_1 = h(R_1, \overline{U}[T_1])$, $T_\ell = h(T_{\ell-1}, \overline{U}[T_\ell])$, $2 \leq \ell \leq k$, and $E_1 = h(T_k, \overline{U}[E_1])$. In other words, there is a probability of γ that a vertex of the chain inherits the value from its parent vertex in the chain. If the value is not inherited, then the value is drawn from a distribution that has the same domain as T. This assures that every vertex in the chain has the same domain of values as T and any interventional distribution $p(\mathbf{y} \,|\, \mathrm{do}(T = t))$ in $\mathcal{G}'$ can have a well-defined counterpart in $\mathcal{G}$.

We define the members of $\mathbf{T}$ that are not on the path $R_1 \to T_1 \to \cdots \to T_k \to E_1$ to be Bernoulli-distributed random variables with a probability parameter θ that are independent of all other variables. The remaining undefined variables in $\mathcal{M}_1$ are the members of $\mathrm{ch}_{\mathcal{G}}(\mathbf{T}) \setminus \mathbf{T}$. In $\mathcal{M}_1'$, each $V_j \in \mathrm{ch}_{\mathcal{G}'}(T)$ is defined as a function of T and the other parents $\mathrm{pa}^+_{\mathcal{G}'}(V_j) \setminus T$. In $\mathcal{M}_1$, we let each V_j have the same function, but with T replaced by E_1, i.e., $f_{V_j'}(E_1 = t, \mathrm{pa}(V_j) \setminus E_1) = f_{V_j'}(T = t, \mathrm{pa}(V_j') \setminus T)$. By this maneuver, each $V_j \in \mathrm{ch}_{\mathcal{G}}(\mathbf{T}) \setminus \mathbf{T}$ now has the same set of possible values in $\mathcal{M}_1$ and $\mathcal{M}_1'$ and for all $V_j \in \mathrm{ch}_{\mathcal{G}}(\mathbf{T}) \setminus \mathbf{T}$ it holds

$$p_1(v_j \,|\, \mathrm{do}(E_1 = t, \mathbf{b}), \mathbf{c}) = p_1'(v_j \,|\, \mathrm{do}(T = t, \mathbf{b}), \mathbf{c}), \tag{1}$$

where $\mathbf{B}$ and $\mathbf{C}$ are disjoint subsets of $\mathbf{V}'$. As the structural equations are the same in $\mathcal{M}_1$ and $\mathcal{M}_1'$ for variables not in $\mathrm{ch}_{\mathcal{G}}(\mathbf{T}) \cup \mathbf{T}$, we can extend this conclusion to all $V_j \in \mathbf{V} \setminus \mathbf{T}$.

Equality of Input Distributions: In this part, we show that having constructed the models as was explained in the previous part, the input distributions are the same in $\mathcal{M}_1$ and $\mathcal{M}_2$. In other words, we must show that for each $i = 1, \ldots, n$ it holds that

$$p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i).$$

We divide the proof into four parts: (I) when $T \in \mathbf{A}_i'$, (II) when $T \in \mathbf{B}_i'$, (III) when $T \in \mathbf{C}_i'$, and (IV) when T is not a member of $\mathbf{A}_i'$, $\mathbf{B}_i'$ or $\mathbf{C}_i'$.

(I) Let $T \in \mathbf{A}_i'$. Because $T \in \mathbf{A}_i'$, by Definition 14 it must be the case that $\mathrm{em}_{\mathcal{G}}(\mathbf{T}) \subseteq \mathbf{A}_i$. Let us denote $\mathbf{D} = \mathrm{de}'_{\mathcal{G}}(T) \cap \mathbf{A}_i'$, $\mathbf{Q} = \mathbf{A}_i' \setminus (\mathbf{D} \cup T)$, and $\mathbf{T}^* = \mathbf{T} \cap \mathbf{A}_i$. Now, the clustered input distribution can be factorized in the following way:

$$\begin{aligned} p_1'(\mathbf{a}_i' \,|\, \mathrm{do}(\mathbf{b}_i'), \mathbf{c}_i') &= p_1'(t, \mathbf{q}, \mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i'), \mathbf{c}_i') \\ &= p_1'(\mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i'), \mathbf{c}_i', t, \mathbf{q}) p_1'(t \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}) p_1'(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i'), \mathbf{c}_i'). \end{aligned} \tag{2}$$

As $p'_1(\mathbf{a}'_i \,|\, \mathrm{do}(\mathbf{b}'_i), \mathbf{c}'_i) = p'_2(\mathbf{a}'_i \,|\, \mathrm{do}(\mathbf{b}'_i), \mathbf{c}'_i)$, it holds that

$$p'_1(t \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}) = p'_2(t \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}), \tag{3}$$
$$p'_1(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p'_2(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i), \text{ and} \tag{4}$$
$$p'_1(\mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, t, \mathbf{q}) = p'_2(\mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, t, \mathbf{q}). \tag{5}$$

When the condition $(\mathbf{D} \perp\!\!\!\perp T \,|\, \mathbf{B}'_i, \mathbf{C}'_i, \mathbf{Q})_{\mathcal{G}'[\overline{\mathbf{B}'_i}, \underline{T}]}$ on line 7 holds, we can use the second rule of do-calculus to obtain:

$$p'_1(\mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, t, \mathbf{q}) = p'_1(\mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i, t), \mathbf{c}_i, \mathbf{q}). \tag{6}$$

We can write similarly for M'_2. For $\mathcal{M}_1$, using the chain rule, we can write

$$\begin{aligned} p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) &= p_1(\mathbf{t}^*, \mathbf{q}, \mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) \\ &= p_1(\mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{t}^*, \mathbf{q}) \\ &\quad \times p_1(\mathbf{t}^* \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}) p_1(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i). \end{aligned}$$

By Lemmas 15 and 16, $(\mathbf{D} \perp\!\!\!\perp T \,|\, \mathbf{B}'_i, \mathbf{C}'_i, \mathbf{Q})_{\mathcal{G}'[\overline{\mathbf{B}'_i}, \underline{T}]}$ implies $(\mathbf{D} \perp\!\!\!\perp \mathbf{T}^* \,|\, \mathbf{B}_i, \mathbf{C}_i, \mathbf{Q})_{\mathcal{G}[\overline{\mathbf{B}_i}, \underline{\mathbf{T} \cap \mathbf{A}_i}]}$. Again, using the second rule of do-calculus we obtain

$$p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p_1(\mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i, \mathbf{t}^*), \mathbf{c}_i, \mathbf{q}) p_1(\mathbf{t}^* \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}) p_1(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i). \tag{7}$$

Now, we can derive $p_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$ similarly as $p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$ was derived, which yields similar formulas for the input distributions but with p_1 replaced by p_2. We will proceed by showing that each term in equation (7) is equal to the corresponding term in the factorization for $p_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$.

The condition of line 5 assures that $\mathbf{C}_i$ does not contain any descendants of the cluster. This means that we can use the rule 3 of do-calculus to conclude the following equalities

$$\begin{aligned} p'_1(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) &= p'_1(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i), \\ p'_2(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) &= p'_2(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i), \\ p_1(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) &= p_1(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i, E_1 = e_1), \mathbf{c}_i), \\ p_2(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) &= p_2(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i, E_1 = e_1), \mathbf{c}_i). \end{aligned}$$

Now, equations (1) and (4) assure that

$$\begin{aligned} p_1(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i, E_1 = e_1), \mathbf{c}_i) &= p'_1(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i, T = e_1), \mathbf{c}_i) \\ &= p'_2(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i, T = e_1), \mathbf{c}_i) \\ &= p_2(\mathbf{q} \,|\, \mathrm{do}(\mathbf{b}_i, E_1 = e_1), \mathbf{c}_i). \end{aligned}$$

Similarly, we can use equations (1) and (6) to have

$$\begin{aligned} p_1(\mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i, E_1 = e_1), \mathbf{c}_i, \mathbf{q}) &= p'_1(\mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i, T = e_1), \mathbf{c}_i, \mathbf{q}) \\ &= p'_2(\mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i, T = e_1), \mathbf{c}_i, \mathbf{q}) \\ &= p_2(\mathbf{d} \,|\, \mathrm{do}(\mathbf{b}_i, E_1 = e_1), \mathbf{c}_i, \mathbf{q}). \end{aligned} \tag{8}$$

For the vertices on the path $R_1 \to T_1 \to \cdots \to T_k \to E_1$, we can write

$$\begin{aligned} &p(r_1, t_1, \ldots, t_k, e_1 \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}) \\ &\quad = p_1(r_1 \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}) p(t_1 \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}, r_1) \cdots p(e_1 \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}, t_k). \end{aligned} \tag{9}$$

In the model construction, the structural equation of R_1 is the same as the structural equation of T. This, together with equation (3) assures that

$$p_1(r_1 \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}) = p_2(r_1 \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}). \tag{10}$$

Now, per the inheritance function h, T_1 obtains the value t by either inheriting from R_1 or by having it drawn from $p^g(t)$, meaning that

$$p_1(T_1 = t \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}, R_1 = r) = \begin{cases} \gamma + (1-\gamma) p_1^g(t) & \text{if } r = t, \\ p_1^g(t) & \text{otherwise.} \end{cases}$$

By equation (10), $p_1(R_1 = t \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}) = p_2(R_1 = t \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q})$. The equality $p_1^g(t) = p_2^g(t)$ is guaranteed by construction. Therefore

$$p_1(t_1 \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}, r_1) = p_2(t_1 \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}, r_1).$$

By induction, the equality holds for the remaining factors in equation (9) as well. Finally, the model construction assures that all vertices outside the path $R_1 \to T_1 \to \cdots \to T_k \to E_1$ are Bernoulli distributed with the same probability parameter between the two models. This, together with the equality of all the factors in equation (9) implies that

$$p_1(\mathbf{t}^* \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}) = p_2(\mathbf{t}^* \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, \mathbf{q}).$$

We have shown that all factors in the input distributions remain the same between $\mathcal{M}_1$ and $\mathcal{M}_2$, and therefore $p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$ in case (I).

(II) Let us assume that $T \in \mathbf{B}_i'$, and denote $\mathbf{T}^* = \mathbf{T} \cap \mathbf{B}_i$ and $\mathbf{W} = \mathbf{B}_i' \setminus T$. Now, for M_1' we can write

$$p_1'(\mathbf{a}_i' \,|\, \mathrm{do}(\mathbf{b}_i'), \mathbf{c}_i') = p_1'(\mathbf{a}_i' \,|\, \mathrm{do}(\mathbf{w}, T = t), \mathbf{c}_i'). \tag{11}$$

Similarly, for M_1, we have

$$p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{w}, E_1 = e_1), \mathbf{c}_i).$$

Using equations (1) and (11), we have

$$\begin{aligned} p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{w}_i, E_1 = e_1), \mathbf{c}_i) &= p_1'(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{w}, T = e_1), \mathbf{c}_i) \\ &= p_2'(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{w}, T = e_1), \mathbf{c}_i) \\ &= p_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{w}, E_1 = e_1), \mathbf{c}_i). \end{aligned}$$

(III) Let us assume that $T \in \mathbf{C}_i'$, and denote $\mathbf{T}^* = \mathbf{C}_i \cap \mathbf{T}$ and $\mathbf{Q} = \mathbf{C}_i' \setminus T$. Now, we can write:

$$p_1'(\mathbf{a}_i' \,|\, \mathrm{do}(\mathbf{b}_i'), \mathbf{c}_i') = p_1'(\mathbf{a}_i' \,|\, \mathrm{do}(\mathbf{b}_i'), \mathbf{q}, T = t).$$

When the condition $(\mathbf{A}'_i \perp\!\!\!\perp T \,|\, \mathbf{B}'_i, \mathbf{C}'_i \setminus T)_{\mathcal{G}'[\overline{\mathbf{B}'_i}, \underline{T}]}$ on line 9 holds, we can now use the second rule of do-calculus to obtain

$$p'_1(\mathbf{a}'_i \,|\, \mathrm{do}(\mathbf{b}'_i), \mathbf{q}, T = t) = p'_1(\mathbf{a}'_i \,|\, \mathrm{do}(\mathbf{b}'_i, T = t), \mathbf{q}).$$

From Lemma 15 and Lemma 16, it follows that $(\mathbf{A}_i \perp\!\!\!\perp (\mathbf{T} \cap \mathbf{C}_i) \,|\, \mathbf{B}_i, \mathbf{C}_i \setminus \mathbf{T})_{\mathcal{G}_{[\overline{\mathbf{B}_i}, \underline{\mathbf{T} \cap \mathbf{C}_i}]}}$. For $\mathcal{M}_1$, we obtain

$$p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{q}).$$

We now have the same factorizations for $p'_1(\mathbf{a}'_i \,|\, \mathrm{do}(\mathbf{b}'_i), \mathbf{c}'_i)$ and $p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$ as in part (II). We can therefore use the same steps to conclude

$$p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i).$$

(IV) Finally, we assume that $T \notin (\mathbf{A}'_i \cup \mathbf{B}'_i \cup \mathbf{C}'_i)$. For $\mathcal{M}'_1$ we can write

$$p'_1(\mathbf{a}'_i \,|\, \mathrm{do}(\mathbf{b}'_i), \mathbf{c}'_i) = \sum_t p'_1(\mathbf{a}'_i \,|\, \mathrm{do}(\mathbf{b}'_i), \mathbf{c}'_i, t) p'_1(t \,|\, \mathrm{do}(\mathbf{b}'_i), \mathbf{c}'_i),$$

and correspondingly for $\mathcal{M}_1$

$$p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = \sum_{e_1} p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, e_1) p_1(e_1 \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i).$$

As emitter E_1 obtains the value t by either inheriting it through the path $R_1 \to T_1 \to \cdots \to T_k \to E_1$ or by having it drawn by g we can write

$$p_1(E_1 = t \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = \gamma^{k+1} p'_1(T = t \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) + (1 - \gamma^{k+1}) p_1^g(t).$$

It then follows

$$\begin{aligned} p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = \gamma^{k+1} &\sum_t p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, E_1 = t) p'_1(T = t \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) \\ &+ (1 - \gamma^{k+1}) \sum_t p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, E_1 = t) p_1^g(t), \end{aligned}$$

In the case of inheritance, which has probability γ^{k+1}, the value of E_1 in $\mathcal{M}_1$ is the same as the value of T in $\mathcal{M}'_1$ and $p_1(\mathbf{a}_1 \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, E_1 = t) = p'_1(\mathbf{a}_1 \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, T = t)$. In the case of non-inheritance, which has probability $1 - \gamma^{k+1}$, the value of E_1 is generated either from $p'_1(t)$ or from a distribution over the domain of T that is the same between the models, i.e., the value of E_1 is set by a stochastic intervention. This allows us to write

$$\begin{aligned} &p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) \\ &\quad = \gamma^{k+1} \sum_t p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, T = t) p'_1(T = t \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) \\ &\qquad + (1 - \gamma^{k+1}) \sum_t p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i) p_1^g(T = t) \\ &\quad = \gamma^{k+1} p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) + (1 - \gamma^{k+1}) \sum_t p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i) p_1^g(T = t). \end{aligned}$$

Now it holds $p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p'_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$ but additional conditions are needed to conclude that

$$\sum_t p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i) p^g_1(T = t) = \sum_t p'_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i) p^g_2(T = t).$$

The lines 11–14 of VerifyInputs go through three separate cases that allow us to conclude this equality.

First, assume that conditions on lines 11 and 12 hold, i.e., there exists $\mathcal{I}'_j \in \mathcal{I}'$ such that $T \subseteq \mathbf{A}'_j$, $\mathbf{B}'_j = \emptyset$, and $\mathbf{C}'_j = \emptyset$, and $(\mathbf{A}'_i \perp\!\!\!\perp T \,|\, \mathbf{B}'_i, \mathbf{C}'_i)_{\mathcal{G}[\overline{\mathbf{B}'_i}, \underline{T}]}$, $(T \perp\!\!\!\perp \mathbf{C}'_i \,|\, \mathbf{B}'_i)_{\mathcal{G}[\overline{\mathbf{B}'_i}]}$ and $(T \perp\!\!\!\perp \mathbf{B}'_i)_{\mathcal{G}[\overline{\mathbf{B}'_i}]}$. We can then write

$$\begin{aligned}\sum_t p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, t), \mathbf{c}_i) p^g_1(t) &= \sum_t p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, t) p'_1(t \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) \\ &= p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i).\end{aligned}$$

where the equality $p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i, T = t) = p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i)$ follows from the second rule of do-calculus, and the equality $p'_1(t \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p'_1(t \,|\, \mathrm{do}(\mathbf{b}_i)) = p'_1(t) = p^g_1(t)$ follows from the first and the third rule of do-calculus, and from line 11, which ensures that $\mathcal{I}'_j = p'(\mathbf{a}'_j)$ exists such that $T \in \mathbf{A}'_j$. As $p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p'_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$, we have

$$\sum_t p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i) p^g_1(T = t) = \sum_t p'_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i) p^g_2(T = t).$$

Then, assume that the condition of line 13 holds, i.e., $(\mathbf{A}'_i \perp\!\!\!\perp T \,|\, \mathbf{B}'_i, \mathbf{C}'_i)_{\mathcal{G}[\overline{\mathbf{B}'_i}, \overline{T(\mathbf{C}'_i)}]}$.

By the third rule of do-calculus, we can write

$$\begin{aligned}\sum_t p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i) p^g_1(T = t) &= \sum_t p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) p^g_1(T = t) \\ &= p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) \sum_t p^g_1(T = t) \\ &= p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i).\end{aligned}$$

As $p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p'_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$, we conclude

$$\sum_t p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i) p^g_1(T = t) = \sum_t p'_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i) p^g_2(T = t).$$

Finally, if the iteration continues, the condition of line 14 is checked. When this condition holds, there exists another input distribution $\mathcal{I}'_k = p'(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, t))$ which ensures that $p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, t), \mathbf{c}_i) = p'_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, t), \mathbf{c}_i)$. As $p^g_1(t) = p^g_2(t)$ by the construction, we then obtain directly

$$\sum_t p'_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i) p^g_1(T = t) = \sum_t p'_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i, T = t), \mathbf{c}_i) p^g_2(T = t).$$

Combining the results of parts (I), (II), (III), and (IV), we have shown that $p_1(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i) = p_2(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$.

Difference of Causal Effects: Finally, we show that if $p'_1(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) \neq p'_2(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ for $\mathcal{M}'_1$ and $\mathcal{M}'_2$ then the same also holds for $\mathcal{M}_1$ and $\mathcal{M}_2$. First, consider the case where members of $\mathbf{T} \cup \mathrm{ch}_{\mathcal{G}}(\mathbf{T})$ are not ancestors of $\mathbf{Y}$ in $\mathcal{G}[\overline{\mathbf{X}}]$. The model construction ensures that $f_i = f'_i$ and $p(\mathbf{U}[V_i]) = p'(\mathbf{U}[V'_i])$ for all variables $V_i \in \mathbf{V} \setminus (\mathbf{T} \cup \mathrm{ch}_{\mathcal{G}}(\mathbf{T}))$. Therefore, we have that $p_1(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) = p'_1(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ and $p_2(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) = p'_2(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ and thus $p_1(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) \neq p_2(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$. Next, suppose that members of $\mathbf{T} \cup \mathrm{ch}_{\mathcal{G}}(\mathbf{T})$ are ancestors of $\mathbf{Y}$ in $\mathcal{G}[\overline{\mathbf{X}}]$. Analogously to part (IV) of the proof, we can write

$$p_j(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) = \gamma^{k+1} p'_j(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) + (1 - \gamma^{k+1}) \sum_t p'_j(\mathbf{y} \mid \mathrm{do}(\mathbf{x}, t)) p^g_j(t),$$

for $j = 1, 2$. We can consider the difference

$$\begin{aligned}
&p_1(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) - p_2(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) \\
&= \gamma^{k+1} \left[p'_1(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) - p'_2(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))\right] \\
&\quad + (1 - \gamma^{k+1}) \left[\sum_t p'_1(\mathbf{y} \mid \mathrm{do}(\mathbf{x}, t)) p^g_1(t) - \sum_t p'_2(\mathbf{y} \mid \mathrm{do}(\mathbf{x}, t)) p^g_2(t), \right] \\
&= \gamma^{k+1} \alpha(\mathbf{x}, \mathbf{y}) + (1 - \gamma^{k+1}) \beta(\mathbf{x}, \mathbf{y})
\end{aligned}$$

where $\alpha(\mathbf{x}, \mathbf{y})$ and $\beta(\mathbf{x}, \mathbf{y})$ denote the differences under value assignments $\mathbf{x}$ to $\mathbf{X}$ and $\mathbf{y}$ to $\mathbf{Y}$, respectively. As the causal effects differ between $\mathcal{M}'_1$ and $\mathcal{M}'_2$, we know that $\alpha(\mathbf{x}, \mathbf{y}) \neq 0$ for some $\mathbf{x}$ and $\mathbf{y}$. Thus we can always choose the value of γ, regardless the value of $\beta(\mathbf{x}, \mathbf{y})$, so that $\gamma^{k+1} \alpha(\mathbf{x}, \mathbf{y}) + (1 - \gamma^{k+1}) \beta(\mathbf{x}, \mathbf{y}) \neq 0$ for at least one pair of value assignments $\mathbf{y}$ and $\mathbf{x}$ because the value of γ has no bearing on the other parts of the proof. Thus $p_1(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) \neq p_2(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$. ■

Theorem 19. *Let $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ be a causal graph where $\mathbf{T} \subset \mathbf{V} \cup \widetilde{\mathbf{U}}$ is a transit cluster and let $\mathcal{I}$ be a set of input distributions compatible with $\mathbf{T}$. If $\mathrm{re}_{\mathcal{G}}(\mathbf{T}) \cap \mathrm{em}_{\mathcal{G}}(\mathbf{T}) \neq \emptyset$, then clustering $(\mathcal{G}, \mathcal{I}) \to (\mathcal{G}', \mathcal{I}')$, where $\mathcal{I}' = \mathcal{I}[\mathbf{T} \to T]$ in $\mathcal{G}' = \mathcal{G}[\mathbf{T} \to T]$, is identification invariant for any causal effect $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$, where $\mathbf{X} \cup \mathbf{Y} \subset \mathbf{V}$, $\mathbf{Y} \cap \mathbf{X} = \emptyset$, $\mathbf{Y} \cap \mathbf{T} = \emptyset$, $\mathbf{X} \cap \mathbf{T} = \emptyset$.*

Proof Assume first that $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ is not identifiable from $\mathcal{I}$ in $\mathcal{G}$. Thus, there exist two models $\mathcal{M}_1$ and $\mathcal{M}_2$ for which the input distributions $\mathcal{I}_1$ and $\mathcal{I}_2$ agree but the causal effects differ. We construct $\mathcal{M}'_1$ and $\mathcal{M}'_2$ based on these two models. The structural equation for T is defined as $f_T = (f_{T_1}, \ldots, f_{T_m})$ in both models, where $\mathbf{T} \cap \mathbf{V} = \{T_1, \ldots, T_m\}$. In other words, T is defined as a random vector whose components are the members of $\mathbf{T}$. This is possible because $\mathcal{M}_1$ and $\mathcal{M}_2$ are recursive, and the construction is described in detail in the proof of Theorem 18. Likewise, for each variable in $\mathrm{ch}_{\mathcal{G}}(\mathbf{T}) \setminus \mathbf{T}$, we can simply copy the structural equations from $\mathcal{M}_1$ and $\mathcal{M}_2$ such that whenever any member of $\mathbf{T}$ appears in the structural equation, the corresponding functions in $\mathcal{M}_1$ and $\mathcal{M}_2$ simply select the corresponding component of T. The structural equations for the remaining variables in $\mathbf{V} \setminus (\mathbf{T} \cup \mathrm{ch}_{\mathcal{G}}(\mathbf{T}))$ are directly copied from $\mathcal{M}_1$ and $\mathcal{M}_2$ into $\mathcal{M}'_1$ and $\mathcal{M}'_2$, respectively. This ensures that $\mathcal{I}'_1 = \mathcal{I}'_2$ as $\mathcal{I}$ is assumed to be compatible with $\mathbf{T}$, but the causal effects differ between $\mathcal{M}'_1$ and $\mathcal{M}'_2$.

Assume then that $p'(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ is not identifiable from $\mathcal{I}'$ in $\mathcal{G}'$. Thus there exist two models $\mathcal{M}'_1$ and $\mathcal{M}'_2$ for which the input distributions $\mathcal{I}'_1$ and $\mathcal{I}'_2$ agree but the causal effects differ. We construct $\mathcal{M}_1$ and $\mathcal{M}_2$ by selecting one variable T_1 from the set $\mathrm{re}_{\mathcal{G}}(\mathbf{T}) \cap \mathrm{em}_{\mathcal{G}}(\mathbf{T})$ to be defined exactly as T is for $\mathcal{M}'_1$ and $\mathcal{M}'_2$. In other words $f_{T_1} = f_T$ and $\mathbf{U}'[T_1] = \mathbf{U}[T]$. This is possible, because T_1 is a receiver of $\mathbf{T}$. The other variables in $\mathbf{T} \setminus T_1$ are defined as Bernoulli distributed with the same probability parameter and as mutually independent. Finally, every other variable in $\mathcal{M}_1$ and $\mathcal{M}_2$ is defined as in $\mathcal{M}'_1$ and $\mathcal{M}'_2$, apart from $\mathrm{ch}_{\mathcal{G}}(\mathbf{T}) \setminus \mathbf{T}$, for which T is replaced by T_1 in their corresponding structural equations. This is possible, because T_1 is an emitter of $\mathbf{T}$. Therefore, $\mathbf{U} = \mathbf{U}' \cup \mathbf{U}[\mathbf{T} \setminus T_1]$, where $\mathbf{U}'$ denotes the error terms for the clustered model and $\mathbf{U}[\mathbf{T} \setminus T_1]$ are the Bernoulli-distributed error terms. Now $\mathcal{M}_1$ and $\mathcal{M}'_1$ are equivalent, as are $\mathcal{M}_2$ and $\mathcal{M}'_2$. As variables $\mathbf{T} \setminus T_1$ and $\mathbf{U}[\mathbf{T} \setminus T_1]$ are defined similarly between $\mathcal{M}_1$ and $\mathcal{M}_2$, and they have no effect on the remaining variables, it follows that $\mathcal{I}_1 = \mathcal{I}_2$ but $p_1(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) \neq p_2(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$. ■

C. Proofs for Section 5 (Obtaining Identifying Functionals)

Theorem 20 (Identifying functional after pruning)**.** *Let $\mathcal{G}(\mathbf{V} \cup \mathbf{U})$ be a causal graph and let $\mathcal{I}$ be a set of input distributions. Let $\mathcal{G}'$ and $\mathcal{I}'$ be the causal graph and the set of input distributions obtained respectively by applying Theorem 7, 8, or 9 to $\mathcal{G}$ and $\mathcal{I}$. If $p'(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ is identifiable from $\mathcal{I}'$ in $\mathcal{G}'$ by do-calculus with the identifying functional $g(\mathcal{I}')$, then $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ is identifiable from $\mathcal{I}$ in $\mathcal{G}$ by do-calculus with the identifying functional $f(\mathcal{I})$ where $f(\mathcal{I}) = g(\mathcal{I}')$.*

Proof For Theorem 7, the claim follows directly from its proof. For Theorems 8 and 9, we investigate all possible paths not only in $\mathcal{G}$ and $\mathcal{G}'$ but also in all possible edge subgraphs. This ensures that all d-separation conditions needed in any do-calculus derivation are valid. Let $\mathcal{G}_*$ be a graph obtained from $\mathcal{G}$ by removing edges $\mathbf{E}_*$ between vertices in $\mathbf{V}$ and let $\mathcal{G}'_*$ be a graph obtained from $\mathcal{G}'$ by removing all edges in $\mathbf{E}_*$ that are in $\mathcal{G}'$. The set $\mathbf{E}_*$ may be empty. Let $V_1, V_2 \in \mathbf{V}'$ be arbitrarily chosen variables, and let $\mathbf{W} \subset \mathbf{V}$ be an arbitrarily chosen set of variables for which $V_1, V_2 \notin \mathbf{W}$. The set $\mathbf{W}$ may be empty. We aim to show that V_1 and V_2 are d-separated in $\mathcal{G}_*$ by $\mathbf{W}$ if and only if they are d-separated in $\mathcal{G}'_*$ by $\mathbf{W} \cap \mathbf{V}'$. This result would allow us to conclude that operations licensed by the rules of do-calculus are equivalent for $p(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ and $p'(\mathbf{y} \mid \mathrm{do}(\mathbf{x}))$ in $\mathcal{G}$ and $\mathcal{G}'$, respectively.

It is assumed in Theorems 8 and 9 that the graph does not contain non-ancestors of $\mathbf{Y}$. Let $\mathbf{Z}$ be the pruned set of variables. We choose arbitrary $V_1, V_2 \in \mathbf{V}'$ and consider a path between V_1 and V_2 in $\mathcal{G}_*$. If the path does not contain members of $\mathbf{Z}$, the vertices of the path remain unchanged after pruning $\mathbf{Z}$. The definitions of $\mathbf{Z}$ in Theorems 8 and 9 guarantee that members of $\mathbf{Z}$ cannot be descendants of the vertices on the path, and thus conditioning on $\mathbf{Z}$ or its subset does not affect whether the path is open or not. In the case of Theorem 9, the path cannot contain a member of $\mathbf{Z}$ because the isolating vertex (denoted by W in Theorem 9) can be on the path only once, and V_1 and V_2 do not belong to $\mathbf{Z}$. We conclude that in the case of Theorem 9, the d-separation of V_1 and V_2 does not depend whether members of $\mathbf{Z}$ are conditioned on or not.

In the case of Theorem 8, the path may contain a member of $\mathbf{Z}$, which implies that the path must also contain two members of $\mathbf{X}$. This follows from the definition of $\mathbf{Z}$, condition (a) of Theorem 8, and the fact that V_1 and V_2 are ancestors of $\mathbf{Y}$. Let $X_1, X_2 \in \mathbf{X}$ be such vertices on the path such that there is an incoming edge from $\mathbf{Z}$ to X_1 and from $\mathbf{Z}$ to X_2. The definition of $\mathbf{Z}$ and the fact that $\mathbf{Z}$ is an ancestor of $\mathbf{Y}$ guarantees that such X_1 and X_2 exist. Further, for the same reason, there must be a member of $\mathbf{Z}$ that is an ancestor of X_1 and X_2. Now by the condition (c) of Theorem 8, there exists $U \in \mathbf{U}$ such that the structure $X_1 \leftarrow U \rightarrow X_2$ exists in the graph. This means that there is another path between V_1 and V_2 that does not contain members of $\mathbf{Z}$. Thus, in the case of Theorem 8, the d-separation of V_1 and V_2 does not depend on whether V_1 and V_2 are conditioned on members of $\mathbf{Z}$ or not.

As V_1, V_2, and $\mathbf{W}$ were chosen arbitrarily, the same conclusions hold for all choices. Further, the identifying functional $f(\mathcal{I})$ can be derived by the same sequence of do-calculus rules that was used to derive the identifying functional $g(\mathcal{I}')$. As the distributions of the variables that are present in the input distributions in $\mathcal{I}'$ remain unchanged after pruning, it holds that $f(\mathcal{I}) = g(\mathcal{I}')$. ■

Theorem 21 (Identifying functional after clustering)**.** *Let* $\mathbf{T}$ *be a transit cluster in a causal graph* $\mathcal{G}(\mathbf{V} \cup \widetilde{\mathbf{U}})$ *and let* $\mathcal{I}$ *be a set of input distributions compatible with* $\mathbf{T}$*. If* $p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ *is identifiable from* $\mathcal{I}' = \mathcal{I}[\mathbf{T} \rightarrow T]$ *in* $\mathcal{G}' = \mathcal{G}[\mathbf{T} \rightarrow T]$ *by do-calculus with the identifying functional* $g(\mathcal{I}')$*, then* $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ *is identifiable from* $\mathcal{I}$ *in* $\mathcal{G}$ *by do-calculus with the identifying functional* $f(\mathcal{I})$ *obtained from* $g(\mathcal{I}')$ *by replacing every instance of* T *by the subset of* $\mathbf{T}$ *that is present in the corresponding input distribution in* $\mathcal{I}$*.*

Proof When $p'(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is identifiable by do-calculus, there exists a sequence of operations licensed by the rules of do-calculus and probability calculus that derive the identifying functional $g(\mathcal{I}')$ from the input distributions $\mathcal{I}'$ in $\mathcal{G}'$. The main idea of the proof is to show that the same sequence of operations can be used to derive identifying functional $f(\mathcal{I})$ for $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ in $\mathcal{G}$. To do this, we have to show that the d-separation conditions required by the rules of do-calculus in $\mathcal{G}'$ are fulfilled also when the sequence of operations is applied to derive $f(\mathcal{I})$ in $\mathcal{G}$.

The rules of do-calculus operate with graphs where the incoming edges of a set of vertices and the outgoing edges of another set of vertices may have been removed. We investigate all possible paths not only in $\mathcal{G}$ and $\mathcal{G}'$ but also in all possible subgraphs that may occur in a do-calculus derivation. This ensures that all d-separation conditions needed in a do-calculus derivation are valid. Let $\mathbf{Z}'_I, \mathbf{Z}'_O \subset \mathbf{V}'$ be disjoint, possibly empty, sets of vertices in $\mathcal{G}'$ and let $\mathbf{Z}_I$ and $\mathbf{Z}_O$ be sets obtained from $\mathbf{Z}'_I$ and $\mathbf{Z}'_O$, respectively, by replacing T by $\mathbf{T}$. We define $\mathcal{G}'_* = \mathcal{G}'[\overline{\mathbf{Z}'_I}, \underline{\mathbf{Z}'_O}]$ and $\mathcal{G}_* = \mathcal{G}[\overline{\mathbf{Z}_I}, \underline{\mathbf{Z}_O}]$.

By Lemma 15, a transit cluster $\mathbf{T}$ in $\mathcal{G}$ is also a transit cluster in $\mathcal{G}_*$. Applying Lemma 16 we conclude that if two sets are d-separated in $\mathcal{G}'_*$, then they are d-separated also in $\mathcal{G}_*$. It follows that the sequence of do-calculus operations used to derive identifying functional $g(\mathcal{I}')$ is applicable in $\mathcal{G}$ when T is replaced by $\mathrm{em}_{\mathcal{G}}(\mathbf{T})$. The identifying functional $f(\mathcal{I})$ is derived from $g(\mathcal{I}')$ by replacing all input distributions in $\mathcal{I}'$ by the corresponding input distributions in $\mathcal{I}$ and replacing for all $i = 1, \ldots, n$ the instance of T in the input distribution i by the subset of $\mathbf{T}$ that is present in the input distribution i in $\mathcal{I}$. Due to compatibility of $\mathcal{I}$, all these subsets include $\mathrm{em}_{\mathcal{G}}(\mathbf{T})$. We conclude that $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{x}))$ is identifiable from $\mathcal{I}$ with

identifying functional $f(\mathcal{I})$. ■

D. Examples on Identification Non-invariance

We show that the d-separation conditions of lines 7 and 9 of VerifyInputs cannot be relaxed. If either condition is violated, then it is possible that a causal effect is identifiable in the original graph but non-identifiable in the clustered graph.

First, we consider the case where $T \in \mathbf{A}'_i$, but the condition of line 7 does not hold, i.e., $(\mathbf{D} \not\perp\!\!\!\perp T \,|\, \mathbf{B}_i, \mathbf{C}_i, \mathbf{A}_i \setminus (\mathbf{D} \cup T))_{\mathcal{G}'[\overline{\mathbf{B}_i}, \underline{T}]}$, where $\mathbf{D} = \mathrm{de}_{\mathcal{G}'}(T) \cap \mathbf{A}'_i$. Consider the graph of Figure 11a and suppose our input distributions are $p(x, z_1, z_2)$ and $p(y, z_1, z_2)$. The causal effect $p(y \,|\, \mathrm{do}(x))$ is identifiable and

$$p(y \,|\, \mathrm{do}(x)) = \sum_{z_2} p(z_2 \,|\, x) \sum_{z_1} p(z_1) p(y \,|\, z_1, z_2). \tag{12}$$

However, if we cluster $\{Z_1, Z_2\}$ as Z and consider the corresponding causal graph depicted in Figure 11b, the causal effect is no longer identifiable from the clustered input distributions $p(x, z)$ and $p(y, z)$. We can see that condition of line 7 is not satisfied for the second input distribution $p(y, z_1, z_2)$ because the input distribution does not contain a sufficient set to close the backdoor path $Y \leftarrow W \rightarrow X \rightarrow Z$. In contrast, if the second input contains X or W, such as $p(y, x, z_1, z_2)$ or $p(y, z_1, z_2 \,|\, w)$, then the aforementioned backdoor path can be closed, and the causal effect is identifiable both before and after clustering.

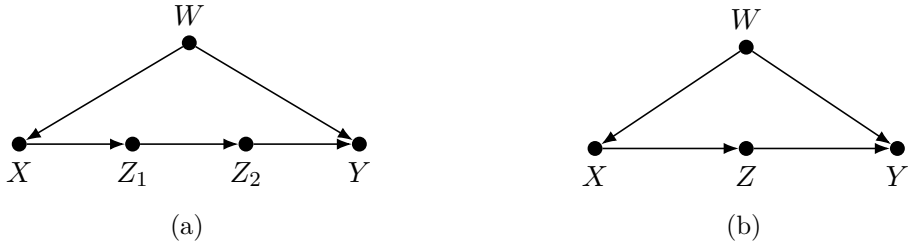

Figure 11: Causal graphs demonstrating the necessity of lines 7 and 9 in VerifyInputs.

Next, we consider the case where $T \in \mathbf{C}'_i$, but condition of line 9 is violated, i.e., $(\mathbf{A}'_i \not\perp\!\!\!\perp T \,|\, \mathbf{B}_i, \mathbf{C}_i \setminus T)_{\mathcal{G}'[\overline{\mathbf{B}'_i}, \underline{T}]}$. We continue with the graphs of Figure 11 and the transit cluster $\{Z_1, Z_2\}$. Suppose now that our input distributions are instead $p(x, z_1, z_2)$ and $p(y \,|\, z_1, z_2)$. The causal effect $p(y \,|\, \mathrm{do}(x))$ is again identifiable in the graph of Figure 11a with the same formula as in Equation 12. Again, the causal effect is no longer identifiable if we cluster $\{Z_1, Z_2\}$ as Z and consider the clustered input distributions $p(x, z)$ and $p(y \,|\, z)$. Similarly as before, line 9 is not satisfied for the second input distribution because the backdoor path between Z and Y is open. If the second input is replaced by $p(y \,|\, z_1, z_2, x)$, $p(y \,|\, \mathrm{do}(x), z_1, z_2)$, $p(y \,|\, z_1, z_2, w)$, or $p(y \,|\, \mathrm{do}(w), z_1, z_2)$, which would satisfy the condition, the causal effect is again identifiable in both graphs.

E. Further Details and Results on the Simulations

E.1 Generation of a Simulation Instance

In Algorithm 2, we present the process of generating one simulation instance when clustering and pruning are applied together. As a rough outline of the algorithm, we begin in lines 2–6 by creating the clustered graph $\mathcal{G}'$ which contains at least one non-ancestor of the response Y. We then select one vertex as the cluster vertex and construct the original graph $\mathcal{G}$ by replacing the cluster vertex in $\mathcal{G}'$ by the newly formed cluster set in lines 8–14. Lines 16–20 generate two or three input distributions, while ensuring that the inputs are compatible by Definition 14 and that no counterintuitive inputs are formed (such as inputs with no observed variables). Finally, lines 21–29 execute the direct and reduction strategies, as was explained in Section 6.

E.2 Benefits of Pruning When Clustering Is Not Helpful

In the simulation study of Section 6, we generated instances of four settings. However, in Table 1 and Figure 7 we concentrated on the most essential settings A, B and C. In setting D, the graph is pruned and clustered but it turns out that clustering is not helpful for identification. Nevertheless, the graph was reduced by pruning which decreases the running time of Do-search. The simulation results for setting D are summarized in Table 6.

Graph size	Setting	Median difference (IQR)		Median ratio (IQR)		N
7	D	−0.03	(−0.07, 0.02)	0.74	(0.40, 1.18)	397
8	D	0.30	(0.09, 0.69)	2.66	(1.57, 4.20)	1 692
9	D	2.47	(1.12, 4.88)	6.16	(4.07, 11.42)	4 045
10	D	15.34	(6.97, 31.79)	8.42	(5.93, 32.37)	6 277
11	D	85.11	(39.36, 179.14)	10.08	(6.71, 46.17)	7 219
12	D	442.94	(198.37, 752.05)	15.14	(5.51, 45.22)	5 578

Table 6: Benefits of pruning when clustering is not helpful (setting D). Median differences (in seconds) of running times between the direct and reduction strategies, and median ratios (the running time for the direct strategy divided by the running time of the reduction strategy) are reported. The interquartile ranges (IQR) are included in parentheses. The number of instances for each graph size is denoted by N.

We can note that, as was the case with settings A and B, setting D also provides clear benefits as the graph size increases. Compared to settings A and B, it appears that setting D was the least beneficial. However, this observation can be explained by the design of the study. In setting D, Do-search is applied twice: first for the clustered and pruned graph, and then for the pruned-only graph. In addition, pruning is likely to remove fewer vertices from the graph, due to limiting pruning to vertices that are neither treatment nor response variables, and that do not belong to the cluster. The benefits of pruning alone are illustrated more prominently in Table 7.

E.3 Simulations for Clustering and Pruning Separately

In addition to performing pruning and clustering together in the same instance, we also study their impacts separately. In such cases, some streamlining can be made to the procedures described earlier. In the pruning strategy, the following changes are made to the reduction strategy of Section 6:

- The steps 2–5 in the reduction strategy are dismissed, as they are related to clustering. This leaves us with two settings, E: apply pruning to the original graph and use Do-search in the pruned graph, and F: pruning operations are not identification invariant and Do-search is used for the original graph.
- On line 23, $\mathbf{T} = \emptyset$, i.e., we can also allow the pruning of the members in the generated cluster.
- Lines 24–27 in SimulateInstance are not executed, as they concern steps 2–5 of the reduction strategy. In this case $t_3 = t_4 = t_5 = 0$.

Correspondingly, the following changes are made in the clustering strategy:

- The steps 1 and 6 in the reduction strategy are dismissed, as they are related to pruning. In this case, we have three settings, G: the causal effect is identifiable in the clustered graph, H: the causal effect is non-identifiable, but the clustering operation is identification invariant and I: the clustering is not identification invariant, Do-search is used for the original graph.
- In SimulateInstance, line 6 that generates non-ancestors, and lines 22, 23 and 28 that assess the identification invariance of the pruned graph, are not executed. In this case $t_2 = t_6 = 0$.

The results regarding pruning only (N = 120 000) are summarized in Table 7 and Figures 12 and 13. The clustering-only results (N = 105 772) are summarized in Table 8 and Figures 14 and 15. In brief, the results are in line with the simulation conducted in Section 6, demonstrating the benefits of both pruning and clustering.

Algorithm 2 Generate a single simulation instance with a random unreduced and reduced causal graph and random input distributions. The function returns the running times for causal effect identification with direct and reduction strategies.

1: **function** SimulateInstance
2: Let $\mathbf{V}' = \{X, Y, Z_1, \dots Z_n\}$, where $n \sim U(\{3, 4, 5, 6\})$.
3: Let π be a topological order for $\mathbf{V}'$ such that X is not first and Y is the last in π.
4: For all $V', W' \in \mathbf{V}'$ such that $V' \prec W'$ in π, generate a directed edge from V' to W' with probability 0.35 if $V' \prec X$ in π, otherwise with probability 0.5.
5: If $V' \in \mathbf{V}'$ is not an ancestor of Y, generate an edge to W' such that $V' \prec W'$ in π and $W' \in Y \cup \mathrm{an}(Y)$.
6: Let $k \sim U(\{1, 2\})$. If $k = 1$, let $\mathbf{V}' = \mathbf{V}' \cup \{W_1\}$, else $\mathbf{V}' = \mathbf{V}' \cup \{W_1, W_2\}$, and let there be directed edges from Y to W_1 and W_2.
7: Let $\mathcal{G}'[\mathbf{V}']$ be the clustered causal graph that is the result of the lines 2–6.
8: Select a random vertex $Z \in \{Z_1, Z_2, \dots Z_n\}$ as the cluster vertex.
9: Let $\mathbf{R} = \emptyset$. If $\mathrm{pa}(Z) \neq \emptyset$, let $\mathbf{R} = \{R_1\}$ or $\mathbf{R} = \{R_1, R_2\}$ with probabilities 0.5.
10: Let $\mathbf{E} = \emptyset$. If $\mathrm{ch}(Z) \neq \emptyset$, let $\mathbf{E} = \{E_1\}$ or $\mathbf{E} = \{E_1, E_2\}$ with probabilities 0.5.
11: Let $\mathbf{S} = \emptyset$ with probability 0.5, otherwise $\mathbf{S} = \{S\}$.
12: Let $\mathbf{T} = \mathbf{R} \cup \mathbf{S} \cup \mathbf{E}$. If $|\mathbf{T}| < 2$, go to line 8.
13: Let ϕ be a topological order of $\mathbf{T}$ such that $\mathbf{R} \prec \mathbf{S} \prec \mathbf{E}$.
14: For all $T_1, T_2 \in \mathbf{T}$ such that $T_1 \prec T_2$ in ϕ, generate a directed edge from T_1 to T_2 with probability 0.5. Repeat this step until all following conditions are met: 1) Each $R \in \mathbf{R}$ is an ancestor of some $E \in \mathbf{E}$, 2) Each $E \in \mathbf{E}$ is a descendant of some $R \in \mathbf{R}$, 3) If $\mathbf{S} \neq \emptyset$, then $\mathrm{pa}(S) \neq \emptyset$ and $\mathrm{ch}(S) \neq \emptyset$. After this step $\mathbf{T}$ is a transit cluster by Definition 13.
15: Let $\mathbf{V} = (\mathbf{V}' \cup \mathbf{T}) \setminus Z$. Expand $\mathcal{G}'[\mathbf{V}']$ to $\mathcal{G}[\mathbf{V}]$ where Z is replaced by $\mathbf{T}$.
16: Choose the number of input distributions in $\mathcal{I}'$ randomly as 2 or 3 with probabilities 0.5.
17: Each member $p(\mathbf{a}_i \,|\, \mathrm{do}(\mathbf{b}_i), \mathbf{c}_i)$ of $\mathcal{I}'$ is generated as follows: every member of $\mathbf{V}'$ is randomly assigned to the sets $\mathbf{A}_i$, $\mathbf{B}_i$, $\mathbf{C}_i$ and $\mathbf{D}_i$ with probabilities 0.35, 0.125, 0.125, and 0.40, respectively. Repeat this step until all following conditions are met: 1) $\mathbf{A}_i \neq \emptyset$ for all i, 2) if $Y \in \mathbf{A}_i$ then $X \notin \mathbf{B}_i$ for all i, 3) $Y \in \mathbf{A}_i$ for some i, 4) $X, Z \in \mathbf{A}_i \cup \mathbf{B}_i \cup \mathbf{C}_i$ for some i.
18: To generate original inputs, let $\mathcal{I} = \mathcal{I}'$. For all $\mathcal{I}_i \in \mathcal{I}$ such that $Z \in \mathbf{A}_i \cup \mathbf{B}_i \cup \mathbf{C}_i$, **do**
19: Let $\mathbf{Q} = \mathbf{E}$. For all $T \in \mathbf{T} \setminus \mathbf{E}$, append $\mathbf{Q}$ by T with probability 0.5.
20: If $\mathbf{Q} = \emptyset$, go to line 19. Otherwise, replace Z by $\mathbf{Q}$ in $\mathcal{I}_i$.
21: Use Do-search to determine if $p(y \,|\, \mathrm{do}(x))$ is identifiable from $\mathcal{I}$ in $\mathcal{G}$. Let t_1 be the running time for this operation. If more than 15 minutes have elapsed, terminate the operation.
22: Let $\mathcal{I}_P = I$ and $\mathcal{G}^P = \mathcal{G}$.
23: Check if Theorems 7–9 can be applied. Let $\mathbf{P}_7$, $\mathbf{P}_8$ and $\mathbf{P}_9$ be the corresponding pruned sets. If a pruned set intersects with $\mathbf{T}$, replace it with the empty set. Now it holds $(\mathbf{P}_7 \cup \mathbf{P}_8 \cup \mathbf{P}_9) \cap \mathbf{T} = \emptyset$. Let $\mathbf{P} = \mathbf{P}_7 \cup \mathbf{P}_8 \cup \mathbf{P}_9$ and let $\mathcal{I}^P = \mathcal{I}^P[\mathbf{V} \setminus \mathbf{P}]$, $\mathcal{G}^P = \mathcal{G}^P[\mathbf{V} \setminus \mathbf{P}]$ and $\mathcal{I}' = \mathcal{I}'[\mathbf{V}' \setminus \mathbf{P}]$, $\mathcal{G}' = \mathcal{G}'[\mathbf{V}' \setminus \mathbf{P}]$. Let t_2 be the running time for the pruning operation.
24: Use FindTrClust (Tikka et al., 2021), to find the transit clusters of $\mathcal{G}$. Let t_3 be the running time for this operation.
25: Use Do-search to determine if $p(y \,|\, \mathrm{do}(x))$ is identifiable from $\mathcal{I}'$ in $\mathcal{G}'$. Let t_4 be the running time for this operation.
26: If $p(y \,|\, \mathrm{do}(x))$ is identifiable from $\mathcal{I}'$ in $\mathcal{G}'$ **return** $(t_1, t_2 + t_3 + t_4)$.
27: Run VerifyInputs($\mathcal{G}'$, $\mathcal{I}'$, Z). Let t_5 be the running time for this operation. If VerifyInputs returns **TRUE**, **return** $(t_1, t_2 + t_3 + t_4 + t_5)$.
28: If $\mathcal{I}^P \neq \mathcal{I}$, use Do-search to determine if $p(y \,|\, \mathrm{do}(x))$ is identifiable from $\mathcal{I}^P$ in $\mathcal{G}^P$. Let t_6 be the running time for this operation and **return** $(t_1, t_2 + t_3 + t_4 + t_5 + t_6)$
29: Else **return** $(t_1, t_1 + t_2 + t_3 + t_4 + t_5 + t_6)$.

Graph size	Setting	Median difference (IQR)		Median ratio (IQR)		N
7	E	0.05	(0.01, 0.11)	2.68	(1.44, 4.55)	1 875
7	F	−0.02	(−0.03, −0.01)	0.82	(0.70, 0.89)	254
8	E	0.40	(0.15, 0.83)	7.45	(3.98, 18.14)	8 336
8	F	−0.02	(−0.03, −0.01)	0.96	(0.91, 0.98)	1 477
9	E	2.51	(1.03, 5.33)	18.65	(6.82, 58.07)	16 711
9	F	−0.02	(−0.03, −0.01)	0.99	(0.98, 1.00)	3 629
10	E	15.25	(6.19, 33.72)	32.29	(7.76, 116.89)	21 619
10	F	−0.02	(−0.04, −0.02)	1.00	(1.00, 1.00)	5 498
11	E	87.43	(35.36, 196.99)	34.72	(7.59, 176.70)	20 923
11	F	−0.02	(−0.04, −0.02)	1.00	(1.00, 1.00)	6 099
12	E	481.80	(194.75, 860.78)	39.77	(8.78, 265.42)	14 267
12	F	−0.03	(−0.05, −0.02)	1.00	(1.00, 1.00)	5 084

Table 7: Comparison of the direct strategy and the pruning strategy. The settings are E: graph was reduced by pruning, F: pruning was not possible and the original graph was used. Median differences (in seconds) of running times between the direct and reduction strategies, and median ratios (the running time for the direct strategy divided by the running time of the reduction strategy) are reported. The interquartile ranges (IQR) are included in parentheses. The column N tells the number of instances for each setting and original graph size.

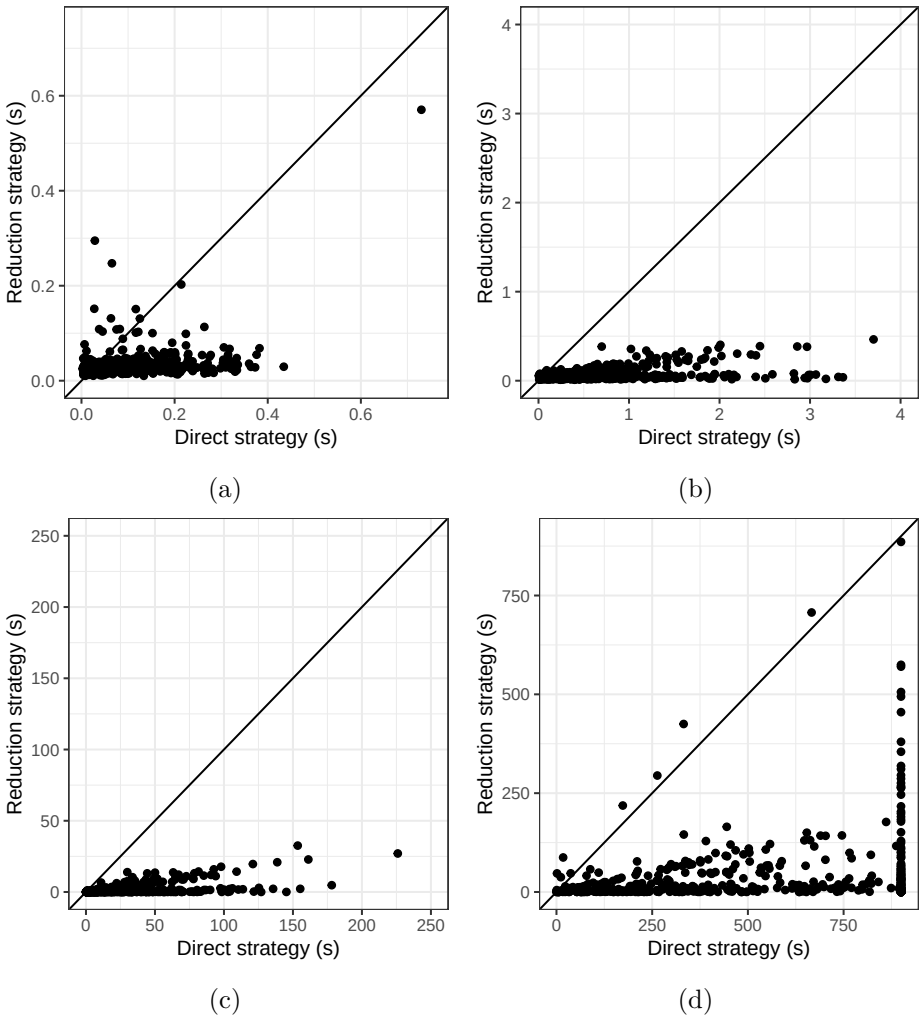

Figure 12: Comparison of the direct and the pruning strategies when pruning is possible. The scatter plots show the running times for the direct strategy and the pruning strategy for setting E. The panels depict different original graph sizes: 7 (a), 8 (b), 10 (c) and 12 (d). A random sample of 500 instances is used for each panel. In Panel (d), the heap of points at 900 seconds is due to the time limit of 15 minutes.

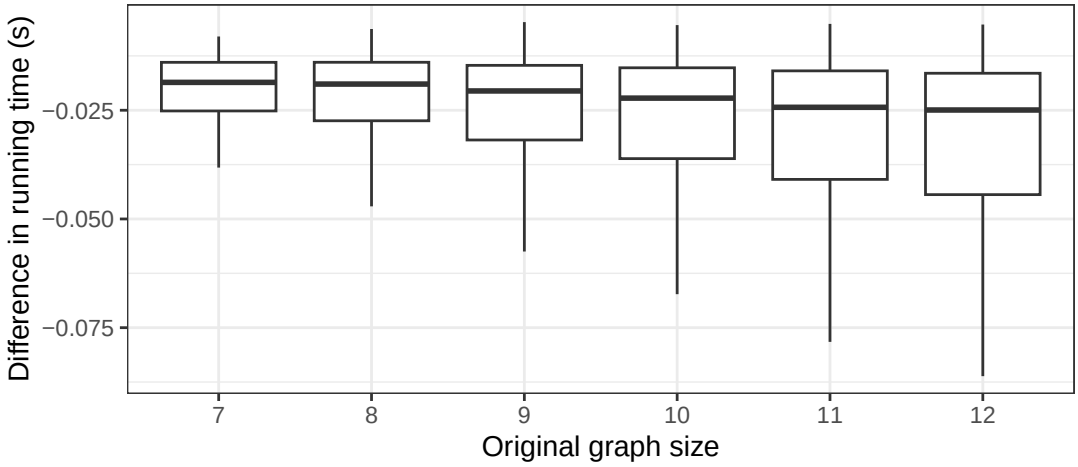

Figure 13: Overhead of the pruning strategy when pruning is not possible. The box plots show the differences in running times between the direct strategy and the pruning strategy for setting F by the size of the original graph. Negative differences indicate that more time was spent with the pruning strategy.

Graph size	Setting	Median difference (IQR)		Median ratio (IQR)		N
6	G	−0.06	(−0.07, −0.05)	0.17	(0.09, 0.30)	1 872
6	H	−0.05	(−0.07, −0.04)	0.23	(0.14, 0.37)	1 585
6	I	−0.09	(−0.11, −0.08)	0.13	(0.09, 0.20)	944
7	G	−0.04	(−0.08, 0.02)	0.57	(0.22, 1.17)	6 156
7	H	−0.01	(−0.05, 0.07)	0.90	(0.47, 1.66)	5 792
7	I	−0.13	(−0.15, −0.11)	0.37	(0.24, 0.52)	3 718
8	G	0.16	(−0.04, 0.57)	2.08	(0.70, 4.70)	8 726
8	H	0.45	(0.14, 1.03)	3.88	(1.95, 7.78)	9 954
8	I	−0.17	(−0.21, −0.14)	0.71	(0.56, 0.83)	7 635
9	G	1.48	(0.32, 3.85)	6.95	(2.62, 17.08)	7 691
9	H	3.36	(1.39, 7.27)	13.37	(6.08, 30.41)	11 191
9	I	−0.25	(−0.36, −0.20)	0.89	(0.81, 0.95)	10 735
10	G	9.23	(2.59, 23.30)	21.92	(9.22, 52.05)	5 059
10	H	21.55	(9.01, 45.88)	44.62	(21.05, 90.50)	9 877
10	I	−0.40	(−0.71, −0.28)	0.97	(0.94, 0.98)	10 917
11	G	49.07	(16.28, 129.81)	80.32	(33.69, 159.16)	2 347
11	H	122.58	(49.48, 289.03)	163.69	(88.25, 361.43)	5 312
11	I	−0.60	(−0.97, −0.39)	0.99	(0.99, 1.00)	6 661
12	G	243.13	(80.53, 590.55)	285.31	(129.83, 481.84)	549
12	H	681.84	(262.97, 898.71)	528.05	(351.77, 829.21)	1 429
12	I	−0.80	(−1.13, −0.55)	1.00	(1.00, 1.00)	1 848

Table 8: Comparison of the direct strategy and the clustering strategy. The settings are G: identifiable in the clustered graph, H: non-identifiable in the clustered graph but identification invariant, I: non-identifiable in the clustered graph and not identification invariant. Median differences (in seconds) of running times between the direct and reduction strategies, and median ratios (the running time for the direct strategy divided by the running time of the reduction strategy) are reported. The interquartile ranges (IQR) are included in parentheses. The column N tells the number of instances for each setting and original graph size.

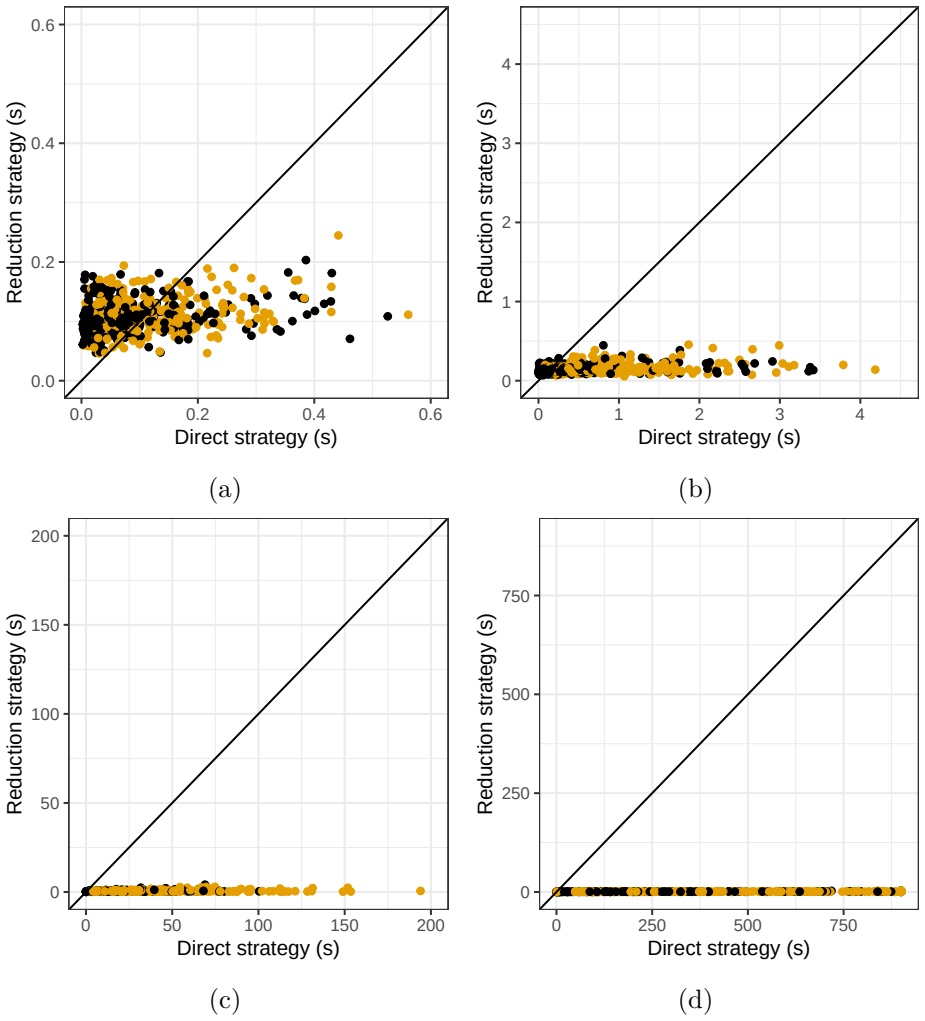

Figure 14: Comparison of the direct and the clustering strategies when clustering is helpful. The scatter plots show the running times for the direct strategy and the clustering strategy for settings G and H. The panels depict different original graph sizes: 7 (a), 8 (b), 10 (c) and 12 (d). A random sample of 500 instances is used for each panel. The points are colored by the setting (black for setting G, and orange for setting H).

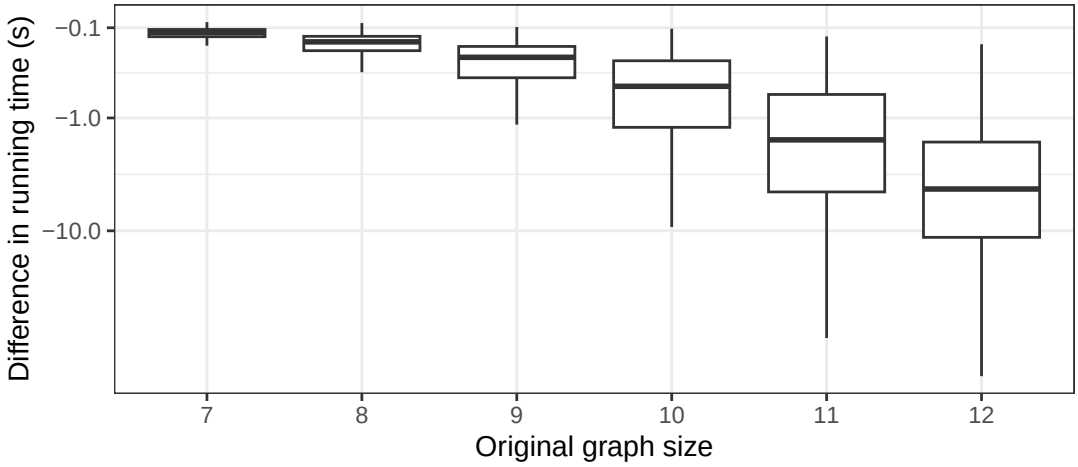

Figure 15: Overhead of the clustering strategy when clustering is not helpful for identification. The box plots show the differences in running times between the direct strategy and the clustering strategy for setting I. Negative differences indicate that more time was spent with the clustering strategy. The vertical axis uses logarithmic scaling.

References

T. V. Anand, A. H. Ribeiro, J. Tian, and E. Bareinboim. Causal effect identification in cluster DAGs. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 12172–12179, 2023.

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. doi: 10.1073/pnas.1510507113.

J. Bergeron, D. Doiron, Y. Marcon, V. Ferretti, and I. Fortier. Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit. *PLoS One*, 13(7): e0200926, 2018.

J. L. M. Björkegren and A. J. Lusis. Atherosclerosis: recent developments. *Cell*, 185(10): 1630–1645, 2022. doi: 10.1016/j.cell.2022.04.004.

L. V. Camelo, L. Giatti, D. Chor, R. H. Griep, I. M. Benseñor, I. S. Santos, I. Kawachi, and S. M. Barreto. Associations of life course socioeconomic position and job stress with carotid intima-media thickness. The Brazilian Longitudinal Study of Adult Health (ELSA-Brasil). *Social Science & Medicine*, 141:91–99, 2015. doi: 10.1016/j.socscimed.2015.07.032.

B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical Science*, 39(1):165–191, 2024.

I. J. Dahabreh, S. E. Robertson, L. C. Petito, M. A. Hernán, and J. A. Steingrimsson. Efficient and robust methods for causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population. *Biometrics*, 79(2):1057–1072, 2023.

F. T. Filippidis, A. A. Laverty, T. Hone, J. V. Been, and C. Millett. Association of cigarette price differentials with infant mortality in 23 European Union countries. *JAMA Pediatrics*, 171(11):1100–1106, 2017. doi: 10.1001/jamapediatrics.2017.2536.

Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 217–224. AUAI Press, 2006.

P. Hünermund and E. Bareinboim. Causal inference and data fusion in econometrics. *The Econometrics Journal*, 28:41–82, 2023. doi: 10.1093/ectj/utad008.

J. Karvanen, S. Tikka, and A. Hyttinen. Do-search: A tool for causal inference and study design with multiple data sources. *Epidemiology*, 32(1):111–119, 2020. doi: 10.1097/ede.0000000000001270.

J. Karvanen, S. Tikka, and M. Vihola. Simulating counterfactuals. *Journal of Artificial Intelligence Research*, 80:835–857, 2024. doi: 10.1613/jair.1.15579.

Y. Kivva, E. Mokhtarian, J. Etesami, and N. Kiyavash. Revisiting the general identifiability problem. In J. Cussens and K. Zhang, editors, *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 180 of *Proceedings of Machine Learning Research*, pages 1022–1030. PMLR, 2022. URL `https://proceedings.mlr.press/v180/kivva22a.html`.

J. J. R. Lee, A. Ghassami, and I. Shpitser. A general identification algorithm for data fusion problems under systematic selection. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 2188–2204, 2024.

S. Lee, J. Correa, and E. Bareinboim. Generalized transportability: Synthesis of experiments from heterogeneous domains. In *AAAI Conference on Artificial Intelligence*. AAAI Press, 2020a.

S. Lee, J. D. Correa, and E. Bareinboim. General identifiability with arbitrary surrogate experiments. In R. P. Adams and V. Gogate, editors, *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 115 of *Proceedings of Machine Learning Research*, pages 389–398. PMLR, 2020b.

H. Li, W. Miao, Z. Cai, X. Liu, T. Zhang, F. Xue, and Z. Geng. Causal data fusion methods using summary-level statistics for a continuous outcome. *Statistics in Medicine*, 39(8): 1054–1067, 2020.

Maelstrom Research. Maelstrom catalogue. `https://www.maelstrom-research.org/page/catalogue`, 2025. Accessed: 2025-12-17.

F. D. Malliaros and M. Vazirgiannis. Clustering and community detection in directed networks: A survey. *Physics Reports*, 533(4):95–142, 2013. doi: 10.1016/j.physrep.2013.08.002.

X. Niu, X. Li, and P. Li. Learning cluster causal diagrams: An information-theoretic approach. In *International Joint Conference on Artificial Intelligence (IJCAI-22)*, pages 4871–4877, 2022.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. doi: 10.1093/biomet/82.4.669.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025. URL `https://www.R-project.org/`.

S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007. doi: 10.1016/j.cosrev.2007.05.001.

X. Shi, Z. Pan, and W. Miao. Data integration in causal inference. *WIREs Computational Statistics*, 15(1):e1581, 2023. doi: 10.1002/wics.1581.

I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *AAAI Conference on Artificial Intelligence*, pages 1219–1226. AAAI Press, 2006.

R. Spake, R. E. O'dea, S. Nakagawa, C. P. Doncaster, M. Ryo, C. T. Callaghan, and J. M. Bullock. Improving quantitative synthesis to achieve generality in ecology. *Nature Ecology & Evolution*, 6(12):1818–1828, 2022.

P. W. Tennant, E. J. Murray, K. F. Arnold, L. Berrie, M. P. Fox, S. C. Gadd, W. J. Harrison, C. Keeble, L. R. Ranker, J. Textor, G. D. Tomova, M. S. Gilthorpe, and G. T. H. Ellison. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *International Journal of Epidemiology*, 50(2):620–632, 2021. doi: 10.1093/ije/dyaa213.

S. Tikka and J. Karvanen. Enhancing identification of causal effects by pruning. *Journal of Machine Learning Research*, 18(194):1–23, 2018.

S. Tikka, A. Hyttinen, and J. Karvanen. Causal effect identification from multiple incomplete data sources: A general search-based approach. *Journal of Statistical Software*, 99(5), 2021. doi: 10.18637/jss.v099.i05.

S. Tikka, J. Helske, and J. Karvanen. Clustering and structural robustness in causal diagrams. *Journal of Machine Learning Research*, 24(195):1–32, 2023.

S. Tikka, A. Hyttinen, and J. Karvanen. dosearch: Causal effect identification from multiple incomplete data sources, 2024. R package version 1.0.11.

L. Valkonen, S. Tikka, J. Helske, and J. Karvanen. Price optimization combining conjoint data and purchase history: A causal modeling approach. *Observational Studies*, 10(1): 37–53, 2024.

B. Youngmann, M. Cafarella, B. Salimi, and A. Zeng. Causal data integration. *Proceedings of the VLDB Endowment*, 16(10):2659–2665, 2023. doi: 10.14778/3603581.3603602.

A. Zeng, M. Cafarella, B. Kenig, M. Markakis, B. Youngmann, and B. Salimi. Causal DAG summarization. *Proceedings of the VLDB Endowment*, 18(6):1933–1947, 2025. doi: 10.14778/3725688.3725717.