Causal Effect Identifiability under Partial-Observability

Sanghack Lee [1] **Elias Bareinboim** [1]

Abstract

Causal effect identifiability is concerned with establishing the effect of intervening on a set of variables on another set of variables from observational or interventional distributions under causal assumptions that are usually encoded in the form of a causal graph. Most of the results of this literature implicitly assume that every variable modeled in the graph is measured in the available distributions. In practice, however, the data collections of the different studies considered do not measure the same variables, consistently. In this paper, we study the causal effect identifiability problem when the available distributions encompass different sets of variables, which we refer to as identification under partial-observability. We study a number of properties of the factors that comprise a causal effect under various levels of abstraction, and then characterize the relationship between them with respect to their status relative to the identification of a targeted intervention. We establish a sufficient graphical criterion for determining whether the effects are identifiable from partially-observed distributions. Finally, building on these graphical properties, we develop an algorithm that returns a formula for a causal effect in terms of the available distributions.

1. Introduction

One of the central goals in data sciences (the health and the social sciences), artificial intelligence, and machine learning is to discover cause and effect relationships. If the scientific study is performed appropriately, and causal relations are eventually discovered, the corresponding effects are more likely to hold under a broader set of conditions. Causal relations are usually more stable and generalizable across disparate conditions (Pearl, 2000; Pearl & Mackenzie, 2018).

[1]Department of Computer Science, Columbia University, New York, NY 10027, USA. Correspondence to: Sanghack Lee <sanghacklee@cs.columbia.edu>.

Proceedings of the 37^{th} *International Conference on Machine Learning*, Vienna, Austria, PMLR 119, 2020.

Causal inference provides a collection of principles and tools to help understand the conditions under which these extrapolations can take place (Pearl, 2000; Spirtes et al., 2001; Bareinboim & Pearl, 2016). For instance, a scientist may be able to use an observational study to infer the effect of a new intervention by leveraging knowledge encoded in its causal model (Pearl, 1995; Tian & Pearl, 2002; Tian, 2002; Shpitser & Pearl, 2006a; Huang & Valtorta, 2006b). Causal effects can also be inferred across a broad range of conditions, including across disparate populations (Bareinboim & Pearl, 2014; Lee et al., 2020), under selection bias (Bareinboim & Pearl, 2012b; Correa & Bareinboim, 2017), missing data (Mohan & Pearl, 2014), and in the absence of the causal graph (Jaber et al., 2019), to cite a few. Further, recent advances in causal inference lead to algorithmic solutions to combine data collected under multiple, disparate regimes (observational and interventional) to identify a causal effect (Lee et al., 2019). Despite all the generality and power provided by these results, by and large, they implicitly assume that every modeled variable is consistently available across the different data collections. Since each study is usually designed to fulfill its own objectives, datasets across studies tend to measure different sets of variables (i.e., the datasets have different columns). As a consequence, relevant data available to answer a question about a specific effect may be not usable in another, possibly very related study.

For concreteness, consider the setting where a researcher aims to understand the effect of physical exercise (X) on stroke (Y), written as $P_x(y)$, and is then analyzing two related datasets. The first is based on an experimental study estimating the effect of physical exercise itself (X) on blood pressure (C), collected from different age groups (A); commonly written as $P_X(A, C)$. The second dataset is based on an observational study about the association among body mass index (BMI, or B), blood pressure (C), and stroke (Y), i.e., $P(B, C, Y)$. The causal graphs representing these two studies are shown in Fig. 1a and 1b, respectively. After conducting standard causal analysis, the researcher realizes that $P_x(y)$, the effect of interest, cannot be inferred from the two datasets, separately. She then creates a common representation of their union, which is summarized in the causal graph shown in Fig. 1c. Unfortunately, the identification algorithms available today assume *full-observability*, which means that graphs and datasets

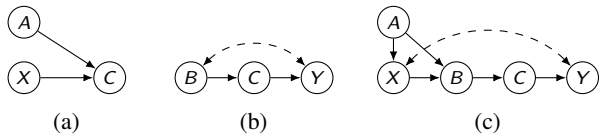


Figure 1. (a, b) Causal graphs representing experimental ($P_X(A, C)$) and observational studies ($P(B, C, Y)$), respectively. (c) The causal graph representing the union of both studies.

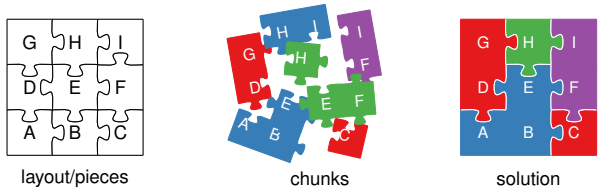


Figure 2. A high-level abstraction of our problem as a jigsaw puzzle. A causal query $P_{\mathbf{x}}(\mathbf{y})$ is a puzzle with how its pieces (factors) should be laid out. Each of the available distributions (represented as different colors) provides *chunks* of information, where a chunk is either a piece or the combination of multiple pieces. The solution for the puzzle is then putting the subset of chunks together.

defined over different sets of variables cannot be taken as input. On the other hand, the effect $P_x(Y)$ is inferable by the careful combination of these studies through the expression $\sum_{a,c} P_x(a, c) \sum_b P(Y|b, c)P(b)$. The first factor can be computed from the experimental study, while the other two can be obtained from the observational study.

Our goal in this paper is to understand under what conditions inferences such as this one are allowed from first principles. More broadly, and motivated by the lack of a systematic treatment to combining partially-observed datasets, we formally introduce and study the problem of *causal effect identifiability under partial-observability*. More specifically, the main contributions of this work are as follows: (i) We develop novel machinery to account for partial-observability constraints, including constructs for co-identification, embedding of critical identification factors, and formal understanding of minimum viable embeddings. Putting these results together, we derive a novel graphical condition for identification under partial-observability; (ii) We then develop the first general algorithm that avoids redundant computations and runs in polynomial time for the known subclasses of identifiability problems under full-observability. Furthermore, we provide a detailed discussion on the necessity of our algorithm and the NP-completeness status of this particular identifiability problem. Finally, we discuss the extension of this work to the transportability setting (Bareinboim & Pearl, 2014) in which the corresponding datasets may come from multiple, heterogeneous domains.

One metaphor that will be informative and facilitate the understanding of this paper is to compare the task of identification under partial-observability to a jigsaw puzzle (Fig. 2). The targeted causal query and underlying causal graph define the puzzle and its layout (Sec. 4.1), while each of the available observational and experimental distributions provides pieces and chunks (pieces tied together) for the puzzle (Sec. 4.1.1 and 4.1.2). Then, the given pieces and chunks are combined, without overlapping each other, to complete the puzzle, eventually forming the targeted effect (Sec. 4.2). For simplicity, we start in Sec. 3 the discussion with a relatively simple class of puzzles, and then build over it, refining its understanding (just after the preliminaries).

2. Preliminaries

Following conventions in the field, a variable is denoted by an uppercase letter, e.g., Z, and its value is denoted by the corresponding lowercase letter, $z \in \mathfrak{X}_Z$, where $\mathfrak{X}_Z$ is the state space of Z. Bold letters are for a set of variables or values, e.g., $\mathbf{z} \in \mathfrak{X}_{\mathbf{Z}} = \times_{Z \in \mathbf{Z}} \mathfrak{X}_Z$. For simplicity, we may omit curly braces, e.g., $f(\{x\})$ versus $f(x)$, for a singleton set when it is used as an argument.

This paper builds on the language of Structural Causal Models (SCM) (Pearl, 2000). Each SCM $\mathcal{M}$ is a quadruple, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{F}$, and $P(\mathbf{U})$. The set of unobserved variables $\mathbf{U}$ follows the joint probability distribution $P(\mathbf{U})$. The set of observed variables $\mathbf{V}$ are specified through the set of structural functions $\mathbf{F} = \{f_V\}_{V \in \mathbf{V}}$, where each function is of the form $f_V(\mathbf{pa}_V, \mathbf{u}_V)$ such that $\mathbf{pa}_V$ and $\mathbf{u}_V$ are the values for $\mathbf{PA}_V \subseteq \mathbf{V} \setminus \{V\}$ and $\mathbf{U}_V \subseteq \mathbf{U}$, respectively. The SCM $\mathcal{M}$ induces a causal graph $\mathcal{G}$ over $\mathbf{V}$ where there are directed edges $W \to V$ if $W \in \mathbf{PA}_V$ and bidirected ones $W \leftrightarrow V$ if there exists an unobserved confounder (UC, for short) $U \in \mathbf{U}_W \cap \mathbf{U}_V$. Further, $\mathcal{M}$ induces a set of observational and interventional distributions. One can intervene on $\mathbf{X}$, setting them to $\mathbf{x}$, which yields a submodel $\mathcal{M}_{\mathbf{x}}$, where the function for $X \in \mathbf{X}$ in $\mathcal{M}$ is replaced by constants $x \in \mathbf{x}$. The distribution generated by $\mathcal{M}_{\mathbf{x}}$ is denoted by $P_{\mathbf{x}}(\mathbf{V})$ (or $P(\mathbf{V} \mid do(\mathbf{x}))$). For simplicity, $P_{\mathbf{Z}}(\mathbf{W})$ denotes a collection of probabilities $\{P_{\mathbf{z}}(\mathbf{w})\}_{\mathbf{z} \in \mathfrak{X}_{\mathbf{Z}}, \mathbf{w} \in \mathfrak{X}_{\mathbf{W}}}$, which we may call $P_{\mathbf{Z}}(\mathbf{W})$ just a distribution. For a more detailed discussion on SCMs, please refer to (Pearl, 2000, Ch. 7).

We denote by $\mathbf{V}_{\mathcal{H}}$ the vertices of a graph $\mathcal{H}$. We often use $\mathbf{V}$ as the vertices of $\mathcal{G}$ where no ambiguity arises. We denote by $pa(\mathbf{W})_{\mathcal{G}}$ the union of parents of $W \in \mathbf{W}$ in $\mathcal{G}$. Similarly, ch, an, and de are children, ancestors, and descendants. Additionally, Ch, An, De include their arguments as well, e.g., $Ch(\mathbf{W})_{\mathcal{G}} = ch(\mathbf{W})_{\mathcal{G}} \cup \mathbf{W}$. A subgraph of $\mathcal{G}$ over $\mathbf{V}' \subseteq \mathbf{V}$ is denoted by $\mathcal{G}[\mathbf{V}']$. Further, $\mathcal{G}_{\overline{\mathbf{X}}}$ and $\mathcal{G}_{\underline{\mathbf{Z}}}$ denote $\mathcal{G}$ with edges incoming to $\mathbf{X}$ and going out from $\mathbf{Z}$ removed, respectively.

The *latent projection* (or projection, for short) of a causal

graph is defined in a way to retain the causal relationships among a subset of variables. This concept is crucial in understanding partial-observability. We define projection as below, adopted from (Tian & Pearl, 2003).

Definition 1 (Projection)**.** The projection of a causal graph $\mathcal{G}$ over $\mathbf{V}$ on $\mathbf{V}' \subseteq \mathbf{V}$, denoted by $\mathcal{G}\langle\mathbf{V}'\rangle$, is a causal graph over $\mathbf{V}'$ such that for every pair of vertices X and Y:

1. There exists a directed edge $X \to Y$ in $\mathcal{G}\langle\mathbf{V}'\rangle$ if there exists a directed path from X to Y in $\mathcal{G}$ such that every vertex other than X and Y on the path is not in $\mathbf{V}'$.
2. There exists a bidirected edge $X \leftrightarrow Y$ to $\mathcal{G}\langle\mathbf{V}'\rangle$ if there exists a divergent path[1] between X and Y in $\mathcal{G}$ such that every vertex other than X and Y on the path is not in $\mathbf{V}'$.

We denote by $\mathcal{G}\langle\text{-}\mathbf{W}\rangle$ a causal graph with $\mathbf{W}$ *projected out*, i.e., $\mathcal{G}\langle\text{-}\mathbf{W}\rangle = \mathcal{G}\langle\mathbf{V} \setminus \mathbf{W}\rangle$. Conditional independence statements and the rules of do-calculus (Pearl, 1995) on a projection are valid in $\mathcal{G}$, vice versa. We often simplify notation involving projections by letting $\mathcal{G}' = \mathcal{G}\langle\mathbf{V}'\rangle$. Similarly, $\mathcal{G}''$ and $\mathcal{G}_j$ is defined for $\mathbf{V}''$ and $\mathbf{V}_j$, respectively. An illustration for how projection induces new edges and some remarks are provided in Appendix A.

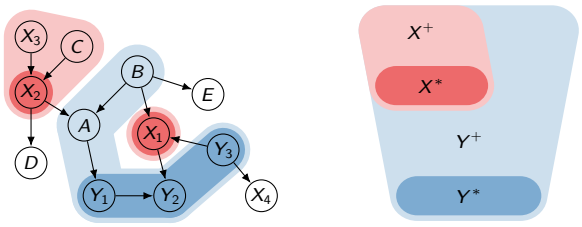
Graphical Constructs for Identifiability Throughout out the paper, we fix the use of a few symbols. A query $P_{\mathbf{x}}(\mathbf{y})$ is defined on a causal graph $\mathcal{G}$ of an unknown SCM $\mathcal{M}$ where its vertices (or variables) are $\mathbf{V}$ where $\mathbf{X}$ and $\mathbf{Y}$ are possibly empty disjoint subsets of $\mathbf{V}$. Further, we define a few new symbols that would greatly help understanding the anatomy of a given graph with respect to the identification of a query, namely:

$$\mathbf{X}^* = \mathbf{X} \cap An(\mathbf{Y})_{\mathcal{G}_{\overline{\mathbf{X}}}} \qquad \mathbf{X}^+ = An(\mathbf{X}^*)_{\mathcal{G}} \setminus \mathbf{Y}^+$$
$$\mathbf{Y}^* = \mathbf{Y} \qquad \mathbf{Y}^+ = An(\mathbf{Y})_{\mathcal{G}_{\underline{\mathbf{X}^*}}}$$
$$\mathbf{V}^* = \mathbf{X}^* \cup \mathbf{Y}^* \qquad \mathbf{V}^+ = \mathbf{X}^+ \cup \mathbf{Y}^+.$$

The left column presents essential parts of the query and the right column describes relevant variables in identifying the query; Fig. 3 provides an illustration of these notions. Their meanings and relationships will become more clear when they appear in lemmas and theorems, especially in Sec. 4. We will use the same color scheme to visualize graphs.

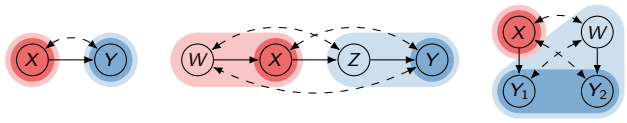
Concepts for Non-identifiability One of crucial building blocks for characterizing our problem of interest is a graphical structure articulating the non-identifiability of a causal effect. A graph $\mathcal{H}$ is said to be a *c-component* if a subset of its bidirected arcs forms a spanning tree over all vertices in $\mathcal{H}$ (Tian & Pearl, 2002; Tian, 2002). The definition can be understood as a set of vertices that are connected via bidirected edges. We utilize the notion of the decomposition of a set of vertices in a graph with respect to its maximal c-components, which we call c-component decomposition. In a special type of c-component, a graph $\mathcal{H}$ with root-set (sink nodes) $\mathbf{R}$ is said to be an $\mathbf{R}$-rooted *c-forest* if $\mathcal{H}$ is a c-component with a minimal number of edges.

[1] A divergent path between X and Y exists if two directed paths $(X, \dots, W_1)$ and $(W_2, \dots, Y)$ in $\mathcal{G}$ towards X and Y, respectively, such that $W_1 = W_2$ or $W_1 \leftrightarrow W_2$.



(a) a causal graph (part) (b) abstract illustration

Figure 3. Illustrative example for $\mathbf{X}^+$ (light red), $\mathbf{Y}^+$ (light blue), $\mathbf{X}^*$ (dark red), and $\mathbf{Y}^*$ (dark blue). Bidirected edges are not relevant in defining those symbols and omitted on purpose.



(a) (b) (c)

Figure 4. (a, b) Hedges for $P_x(y)$ where, for both cases, $\{Y\}$ is the sole root set and the only element of $\mathcal{F}'$ with $\mathcal{F} = \mathcal{G}$. (c) is not a hedge for $P_x(\mathbf{y})$ but $\{Y_1, W\}$-rooted c-forests $\mathcal{F} = \mathcal{G}[\{Y_1, W, X\}]$ and $\mathcal{F}' = \mathcal{G}[\{Y_1, W\}]$ form a hedge for $P_x(\mathbf{y})$.

A hedge is said to be formed for $P_{\mathbf{x}}(\mathbf{y})$ if $\mathbf{R} \subseteq An(\mathbf{Y})_{\mathcal{G}_{\underline{\mathbf{X}}}}$, $\mathcal{F} \cap \mathbf{X} \neq \emptyset$, and $\mathcal{F}' \cap \mathbf{X} = \emptyset$, which implies the non-identifiability of $P_{\mathbf{x}}(\mathbf{y})$ from $P(\mathbf{V})$ in $\mathcal{G}$ (Shpitser & Pearl, 2006a). We prefer to separate the graphical definition of hedge from the specific syntactic goal/task, following discussion in (Lee et al., 2019). Such hedge satisfies that $\mathbf{V}_{\mathcal{F}\setminus\mathcal{F}'} \subseteq \mathbf{V}^+$ intersects with $\mathbf{X}^*$. Further, $\mathbf{V}_{\mathcal{F}'} \subseteq \mathbf{Y}^+$ since $\mathbf{X}^+$ being the part of $\mathcal{F}'$ implies that $X \in \mathbf{X}^*$ is in $\mathcal{F}'$, which violates the definition. Examples are illustrated in Fig. 4 where the first two causal graphs are hedges for $P_x(y)$ but the last one isn't. Omitted proofs and other supporting materials are provided in Appendix.

3. Identifiability with a Single Partially-Observed Distribution

We begin with a simpler, yet crucial identifiability problem that concerns with identifying a causal effect given a single

partially-observed distribution. Using the jigsaw metaphor discussed earlier, this task can be seen as a puzzle where all the available chunks are of the same color, as defined next.

Definition 3 (Causal Effect Identifiability under a Partially-Observed Distribution)**.** Given a causal graph $\mathcal{G}$, let $\mathbf{Z}'$, $\mathbf{V}', \mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ where $\mathbf{Z}' \cap \mathbf{V}' = \emptyset$ and $\mathbf{X} \cap \mathbf{Y} = \emptyset$. The causal effect $P_{\mathbf{x}}(\mathbf{y})$ is said to be identifiable from $P_{\mathbf{Z}'}(\mathbf{V}')$ in $\mathcal{G}$ if $P_{\mathbf{x}}(\mathbf{y})$ is (uniquely) computable from $P_{\mathbf{Z}'}(\mathbf{V}')$ in any causal model which induces $\mathcal{G}$ and $P_{\mathbf{Z}'}(\mathbf{V} \setminus \mathbf{Z}') > 0$.

The definition is similar to a number of identifiability problems with full-observability except for the given data being partially-observed. The positivity assumption is imposed on $P_{\mathbf{Z}'}(\mathbf{V} \setminus \mathbf{Z}')$ (i.e., the full joint before the projection) instead of the given distribution $P_{\mathbf{Z}'}(\mathbf{V}')$. This is to ensure that the partial-observability is correctly responsible for the non-identifiability of a causal effect especially when the effect is identifiable with $P_{\mathbf{Z}'}(\mathbf{V} \setminus \mathbf{Z}')$.

We show multiple necessary criteria as follows.

Proposition 1. *A causal effect $P_{\mathbf{x}}(\mathbf{y})$ is identifiable from $P_{\mathbf{Z}'}(\mathbf{V}')$ in $\mathcal{G}$ only if*

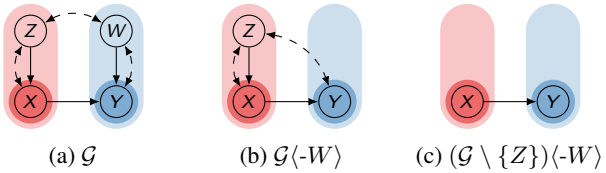
- $\mathbf{Y} \subseteq \mathbf{V}'$ *(inclusion of outcomes),*
- $\mathbf{X}^* \subseteq \mathbf{V}' \cup \mathbf{Z}'$ *(inclusion of minimal treatments), and*
- $\mathbf{Z}' \cap \mathbf{Y}^+ = \emptyset$ *(undisturbed mechanisms).*

Under full-observability, the first condition is implied by the third condition, and the second condition holds trivially since $\mathbf{Z}' \cup \mathbf{V}' = \mathbf{V}$ and $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$. The third condition highlights that an intervention prohibits the understanding of the underlying natural mechanisms relevant to the distribution over $\mathbf{Y}$ (Lee et al., 2019). When these criteria hold, the problem is reducible to the classic identifiability.

Lemma 1. *Given $P_{\mathbf{x}}(\mathbf{y})$ and $P_{\mathbf{Z}'}(\mathbf{V}')$ in $\mathcal{G}$ satisfying the three criteria in Prop. 1, $P_{\mathbf{x}}(\mathbf{y})$ is identifiable from $P_{\mathbf{Z}'}(\mathbf{V}')$ in $\mathcal{G}$ if and only if $Q_{\mathbf{x}^* \setminus \mathbf{Z}'}(\mathbf{y})$ is identifiable in $(\mathcal{G} \setminus \mathbf{Z}')\langle \mathbf{V}' \cap \mathbf{V}^+ \rangle$ where $Q = P_{\mathbf{z}'}$ with $\mathbf{z}'$ consistent with $\mathbf{x}^+ \cap \mathbf{Z}'$.*

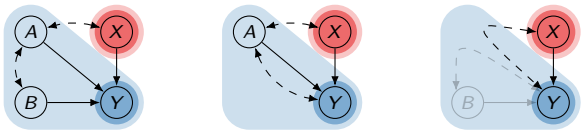
Roughly speaking, Lemma 1 implies that the effect is identifiable when there is no structure that embeds a hedge under the projection onto the partially-observed variables. A simple example is provided (Fig. 5) showing the identifiability of $P_x(y)$ given $P_Z(X, Y)$. The effect is not identifiable with $P(\mathbf{V})$ (the existence of a hedge with $\mathcal{F} = \mathcal{G}$, $\mathcal{F}' = \mathcal{G}[\{W, Y\}]$), $P_W(X, Y, Z)$ (a disturbed mechanism), nor $P_Y(W, X, Z)$ (exclusion of outcomes).

Corollary 1 (Soundness and Completeness)**.** *$P_{\mathbf{x}}(\mathbf{y})$ is identifiable from $P_{\mathbf{Z}'}(\mathbf{V}')$ in $\mathcal{G}$ if and only if $\mathbf{Y} \subseteq \mathbf{V}'$, $\mathbf{X}^* \subseteq \mathbf{V}' \cup \mathbf{Z}'$, $\mathbf{Z}' \cap \mathbf{Y}^+ = \emptyset$, and $Q_{\mathbf{x}^* \setminus \mathbf{Z}'}(\mathbf{y})$ is identifiable in $(\mathcal{G} \setminus \mathbf{Z}')\langle \mathbf{V}' \cap \mathbf{V}^+ \rangle$ where $Q = P_{\mathbf{z}'}$ with $\mathbf{z}'$ consistent with $\mathbf{x}^+ \cap \mathbf{Z}'$.*



(a) $\mathcal{G}$ (b) $\mathcal{G}\langle\text{-}W\rangle$ (c) $(\mathcal{G} \setminus \{Z\})\langle\text{-}W\rangle$

Figure 5. Causal diagrams representing the identifiability of $P_x(y)$ given $P_Z(X, Y)$ as $P_x(y) = P_z(y|x)$ for any z.



(a) $\mathcal{G}$ (b) $\mathcal{G}\langle\text{-}B\rangle$ (c) $\mathcal{G}\langle\text{-}A\rangle$

Figure 6. A causal graph $\mathcal{G}$ (a) which embeds hedges under different projections onto (b) $\{A, X, Y\}$ and (c) $\{B, X, Y\}$.

4. Identifiability with Multiple Partially-Observed Distributions

We now investigate the main task of this paper, i.e., how to systematically use multiple distributions with different levels of observability, as defined next.

Definition 4 (Causal Effect Identifiability under Partially-Observed Distributions (GID-PO))**.** Let $\mathcal{G}$ be a causal graph, and $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ be sets corresponding to the treatment and outcomes variables, respectively. Further, let $\mathbb{P} = \{P_{\mathbf{Z}_i}(\mathbf{V}_i)\}_{i=1}^m$ be a collection of partially observable distributions such that $\mathbf{Z}_i$ and $\mathbf{V}_i$ are disjoint subsets of $\mathbf{V}$ for $1 \leq i \leq m$. The causal effect $P_{\mathbf{x}}(\mathbf{y})$ is identifiable from the graph $\mathcal{G}$ and $\mathbb{P}$ if it is uniquely computable from $\mathbb{P}$ in any model that induces $\mathcal{G}$ where $\{P_{\mathbf{Z}_i}(\mathbf{V} \setminus \mathbf{Z}_i)\}_{i=1}^m$ are positive distributions.

This definition generalizes g-identifiability (GID, Lee et al., 2019) with partial-observability. An example of the problem is shown in Fig. 7a, where $P(B, C, Y)$ and $P_X(A, C)$ are given to identify $P_x(y)$ (the same as Fig. 1c). Its formula can be derived as (see Appendix D for the detailed derivation):

$$P_x(y) = \sum_{a,c} P_{x'}(a) P_x(c|a) \sum_b P(y|b, c) P(b),$$

where x' can be any value in $\mathfrak{X}_X$. Naively incorporating Cor. 1 into GID, which looks for colored pieces but not chunks, fails to identify the query since one of the pieces is not identifiable with any of the given datasets (Fig. 7c).

Lemma 2. *Let $\mathcal{G}$ be a causal graph and $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$. A query $P_{\mathbf{x}}(\mathbf{y})$ is not identifiable if there exist two causal models $\mathcal{M}^1$ and $\mathcal{M}^2$ compatible with $\mathcal{G}$ such that $P^1_{\mathbf{Z}_i}(\mathbf{V}_i) = P^2_{\mathbf{Z}_i}(\mathbf{V}_i)$, for every $P_i \in \mathbb{P}$, but $P^1_{\mathbf{x}}(\mathbf{y}) \neq P^2_{\mathbf{x}}(\mathbf{y})$ and $P^1_{\mathbf{Z}_i}(\mathbf{V} \setminus \mathbf{Z}_i) = P^2_{\mathbf{Z}_i}(\mathbf{V} \setminus \mathbf{Z}_i) > 0$.*

Consider a causal graph (Fig. 6a) where two observational distributions $P(X, Y, A)$ and $P(B, X, Y)$ are available. A

causal query $P_x(y)$ is identifiable given $P(X, Y, A, B)$ but not from *each* of $P(X, Y, A)$ and $P(B, X, Y)$ due to the existence of a hedge (Cor. 1). Further, one can show that the query is not identifiable taking *both* into account: Let $f_X = U_1$, $f_A = U_1 \oplus U_2$, $f_B = U_2$ be the common functions between $\mathcal{M}^1$ and $\mathcal{M}^2$, and $f^1_Y = X \oplus A \oplus B \oplus U_Y$ and $f^2_Y = U_Y$. Further, let U_1 and U_2 be two fair coins, and $P(U_Y = 1) = 0.1$ for both models. Then, $\mathcal{M}^1$ and $\mathcal{M}^2$ will agree on both $P(A, X, Y)$ and $P(B, X, Y)$, while $P^1_{X=0}(Y = 0) = 0.5$ and $P^2_{X=0}(Y = 0) = 0.9$.

4.1. Characterization of Factors under Projection

To develop an algorithm capable of combining the different parts of the available distributions, we briefly review approaches to the related problem of decomposing distributions currently known in the literature. A joint probability distribution $P(\mathbf{V})$ in $\mathcal{G}$ can be seen as the product of causal effects in $\mathcal{G}$. Tian & Pearl (2002) introduced the *Q-decomposition* that expresses $P(\mathbf{v})$ using *c-factors* as follows: $P(\mathbf{v}) = \prod_i Q_i = \prod_i P_{\mathbf{v}\setminus\mathbf{s}_i}(\mathbf{s}_i)$, where $\mathbf{S}_i$ is a c-component of $\mathbf{V}$.[2] We will characterize here such factorization under projections. As a first step, we define the key notion of generalized c-factors:

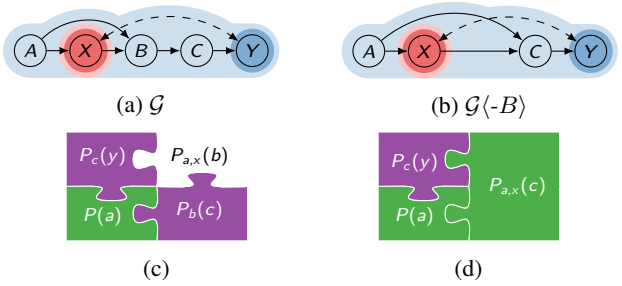
Definition 5 (gc-factors)**.** Let $\mathbf{V}'$ be a subset of variables such that $\mathbf{V}^* \subseteq \mathbf{V}' \subseteq \mathbf{V}^+$ and let $\mathcal{G}' = \mathcal{G}\langle\mathbf{V}'\rangle$. Let $\mathbb{S} = \{\mathbf{S}_i\}_{i=1}^k$ be the c-components of $\mathcal{G}'[\mathbf{Y}^{+\prime}]$ where $\mathbf{Y}^{+\prime} = An(\mathbf{Y})_{\mathcal{G}'_{\underline{\mathbf{X}^*}}}$. Then, the *gc-factors* of $P_{\mathbf{x}}(\mathbf{y})$ in $\mathcal{G}$ with respect to $\mathbf{V}'$ is defined as

$$\mathbb{F}^{\mathbf{X},\mathbf{Y}}_{\mathcal{G},\mathbf{V}'} = \{\langle pa(\mathbf{S}_i)_{\mathcal{G}'} \setminus \mathbf{S}_i, \mathbf{S}_i\rangle\}_{i=1}^k.$$

For example, consider a causal graph $\mathcal{G}$ in Fig. 6a and a query $P_x(y)$, where $\mathbf{V}^* = \{X, Y\}$ and $\mathbf{V}^+ = \mathbf{V}$. With $\mathbf{V}' = \mathbf{V}$, $\mathbf{Y}^{+\prime} = \{A, B, Y\}$ (i.e., variables in blue), and $\mathbb{S} = \{\{A, B\}, \{Y\}\}$ (i.e., the two c-components in the blue area). Hence, the gc-factors $\mathbb{F}^{X,Y}_{\mathcal{G},\mathbf{V}} = \{\langle\emptyset, \{A, B\}\rangle, \langle\{A, B, X\}, \{Y\}\rangle\}$. With $\mathbf{V}' = \{A, X, Y\}$, $\mathcal{G}'$ is shown in Fig. 6b. In this case, $\mathbf{Y}^{+\prime} = \{A, Y\}$, $\mathbb{S} = \{\{A, Y\}\}$, and $\mathbb{F}^{X,Y}_{\mathcal{G},\{A,X,Y\}} = \{\langle\{X\}, \{A, Y\}\rangle\}$.

A gc-factor (or, simply, a factor), which is represented as a pair of sets of variables, can be used to present a set of distributions, e.g., $\langle\mathbf{X}_i, \mathbf{Y}_i\rangle \in \mathbb{F}^{\mathbf{X},\mathbf{Y}}_{\mathcal{G},\mathbf{V}}$ and $P_{\mathbf{X}_i}(\mathbf{Y}_i)$. We may call the first and second elements as $\mathbf{X}$- and $\mathbf{Y}$-side of the gc-factor, respectively. Graphically, this decomposes Y'^+ in $\mathcal{G}'$ with respect to c-components so that the $\mathbf{Y}$-side of every gc-factor is maximally confounded with respect to the underlying projection. The $\mathbf{X}$-side resides in $\mathbf{Y}'^+ \cup \mathbf{X}^*$.

[2] Such a factorization leads to a natural divide-and-conquer approach to the problem of causal identification, which underpins, sometimes more or less explicitly, most of the results in this literature (e.g., (Huang & Valtorta, 2006a; Shpitser & Pearl, 2006a; Bareinboim & Pearl, 2012a; 2014; Lee et al., 2019), to cite a few). Depending on the identification task, one can prove that the Q-decomposition leads to a complete characterization.



(a) $\mathcal{G}$ (b) $\mathcal{G}\langle\text{-}B\rangle$ (c) (d)

Figure 7. Causal graphs (a) $\mathcal{G}$ and (b) $\mathcal{G}\langle\text{-}B\rangle$. Corresponding gc-factors (c, d) represented as puzzle where identified factors are colored with green, $P_X(A, C)$, and purple, $P(B, C, Y)$.[4]

For brevity, we simplify the notation by $\mathbb{F} = \mathbb{F}^{\mathbf{X},\mathbf{Y}}_{\mathcal{G},\mathbf{V}}$ and a superscript is delegated to $\mathbf{V}$ so that $\mathbb{F}' = \mathbb{F}^{\mathbf{X},\mathbf{Y}}_{\mathcal{G},\mathbf{V}'}$. Also we interchangeably use a factor $\langle\mathbf{Z}, \mathbf{W}\rangle$ with $P_{\mathbf{Z}}(\mathbf{W})$ or with $P_{\mathbf{z}}(\mathbf{w})$ for an arbitrary assignment. Probabilities associated with the gc-factors of $P_{\mathbf{x}}(\mathbf{y})$ in $\mathcal{G}$ with respect to $\mathbf{V}'$ can form an expression for $P_{\mathbf{x}}(\mathbf{y})$.

Proposition 2 (gc-decomposition)**.** *For* $\mathbf{V}^* \subseteq \mathbf{V}' \subseteq \mathbf{V}^+$,

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{y}^{+\prime}\setminus\mathbf{Y}} \prod_{\{i|\langle\mathbf{X}_i,\mathbf{Y}_i\rangle\in\mathbb{F}'\}} P_{\mathbf{x}_i}(\mathbf{y}_i) \qquad (1)$$

The algorithm for general-identifiability (Lee et al., 2019) makes the use of this decomposition based on $\mathbb{F}$ (i.e., restricted to $\mathbf{V}' = \mathbf{V}^+$) and identifies each factor using one of the available distributions. However, such strategy relying on the decomposition based on $\mathbb{F}$ is insufficient for handling partially-observed distributions. Recall Fig. 7a where $P_x(y)$ can be factorized as

$$P_x(y) = \sum_{a,b,c} P(a)P_{a,x}(b)P_b(c)P_c(y), \qquad (2)$$

following Prop. 2. Unfortunately, $P_{a,x}(b) = P(b|a, x)$ is not identifiable from each of available distributions since no distribution includes all A, B, X (Cor. 1). On the other hand, a gc-decomposition based on a projection onto $\mathbf{V}' = \{A, C, X, Y\}$ (Fig. 7b) yields

$$P_x(y) = \sum_{a,c} P(a)P_{a,x}(c)P_c(y), \qquad (3)$$

which allows each factor to be identified by at least one of the available distributions (i.e., $P(a) = P_{x'}(a)$, for any $x' \in \mathfrak{X}_X$, $P_{a,x}(c) = P_x(c|a)$, and $P_c(y) = \sum_b P(y|b, c)P(b)$). This provides a basis to the following sufficient condition.

Lemma 3 (Soundness)**.** *Let* $\mathcal{G}$, $P_{\mathbf{x}}(\mathbf{y})$, $\mathbb{P}$ *be the causal graph, the query, and the distributions forming a GID-PO instance. If there exists a subset of variables* $\mathbf{V}^* \subseteq \mathbf{V}' \subseteq \mathbf{V}^+$ *such that every term* $P_{\mathbf{x}_i}(\mathbf{y}_i)$ *in each gc-factor* $\langle\mathbf{X}_i, \mathbf{Y}_i\rangle$ *in* $\mathbb{F}'$ *is identifiable from* $P \in \mathbb{P}$*, then* $P_{\mathbf{x}}(\mathbf{y})$ *is identifiable from* $\mathbb{P}$ *in* $\mathcal{G}$.

Given Lemma 3, we are interested in finding $\mathbf{V}'$, a subset of $\mathbf{V}$, which would yield a gc-decomposition where each gc-factor is identifiable.[5] We note that a natural solution emerges since one could search over an exponential number of subsets of $\mathbf{V}^+$. While this would certainly lead to a sound procedure, it is clearly the case that this offers little to no insight into the problem. This motivates us to study the relationships among gc-factors at different levels of projections so as to develop an efficient solution while avoiding this naive, and clearly intractable solution.

4.1.1. Embedding Factor and Co-identification

In this section, we formally relate gc-factors before and after marginalizations through two new concepts called *embedding factor* and *co-identification*, which associate factors and their identification under various levels of projections so as for us to perceive the task of identification from a more comprehensive view.

Comparing the decompositions in Eqs. (2) and (3) based on $\mathcal{G}$ and $\mathcal{G}\langle\text{-}B\rangle$, respectively, we observe that two factors $P(a)$ and $P_c(y)$, which does not have B in it, are shared while other two factors $P_{a,x}(b)$ and $P_b(c)$ in Eq. (2) are replaced to $P_{a,x}(c)$ in Eq. (3) where we can elicit $\sum_b P_{a,x}(b)P_b(c) = P_{a,x}(c)$. We characterize such changes in gc-factors before and after a projection (equivalently, a marginalization).

Proposition 3 (Factors under Marginalization)**.** *Let $\mathbf{W} \subset \mathbf{V}'$ and $\mathbf{V}'' = \mathbf{V}' \setminus \mathbf{W}$ where $\mathbf{W} \cap \mathbf{V}^* = \emptyset$. Let $\mathcal{H}$ be an undirected graph where vertices are $\langle \mathbf{X}'_j, \mathbf{Y}'_j \rangle \in \mathbb{F}'$ and an edge exists if two factors satisfy their $\mathbf{Y}$-sides intersecting with $Ch(W)_{\mathcal{G}'}$ for some $W \in \mathbf{W}$. Then, vertices (i.e., gc-factors) in each connected component of $\mathcal{H}$ are merged to form $\langle \mathbf{X}''_k, \mathbf{Y}''_k \rangle \in \mathbb{F}''$ such that*

$$\mathbf{Y}''_k = \left(\cup_{j\in\mathbf{j}} \mathbf{Y}'_j\right) \setminus \mathbf{W}, \quad \mathbf{X}''_k = \left(\cup_{j\in\mathbf{j}} \mathbf{X}'_j\right) \setminus \mathbf{Y}''_k \setminus \mathbf{W},$$

where $\mathbf{j}$ is the set of indices of gc-factors in $\mathbb{F}'$ forming the connected component. Other gc-factors without $\mathbf{W}$ are remained intact, shared by both $\mathbb{F}'$ and $\mathbb{F}''$.

For instance, marginalizing B out in Eq. (2) will consolidate factors whose $\mathbf{Y}$-sides intersect with $Ch(B)_{\mathcal{G}} = \{B, C\}$, that is, $P_{a,x}(b)$ and $P_b(c)$. Further, $\sum_b P_{a,x}(b)P_b(c)$ will result in a new gc-factor that has its $\mathbf{Y}$-side, $(\{B\} \cup \{C\}) \setminus \{B\} = \{C\}$ and $\mathbf{X}$-side, $(\{A, X\} \cup \{B\}) \setminus \{C\} \setminus \{B\} = \{A, X\}$. We introduce an embedding relationship to describe the connection between the factors merged through a projection and a resulting factor.

Definition 6 (Embedding Factor)**.** A gc-factor $\langle \mathbf{X}', \mathbf{Y}' \rangle \in \mathbb{F}'$ is said to be *embedded in* a gc-factor $\langle \mathbf{X}'', \mathbf{Y}'' \rangle \in \mathbb{F}''$ for $\mathbf{V}^* \subseteq \mathbf{V}'' \subseteq \mathbf{V}'$ if $\mathbf{Y}''$ contains any variable in the $\mathbf{Y}$ parts of gc-factors in the connected component containing $\langle \mathbf{X}', \mathbf{Y}' \rangle$ in $\mathcal{H}$, which is constructed as follows. $\mathcal{H}$ is an undirected graph of gc-factors in $\mathbb{F}'$ where there exists an edge between two factors, say $\langle \mathbf{X}'_i, \mathbf{Y}'_i \rangle$, $\langle \mathbf{X}'_j, \mathbf{Y}'_j \rangle$, if both share $\mathbf{V}' \setminus \mathbf{V}''$, that is,

$$(\mathbf{X}'_i \cup \mathbf{Y}'_i) \cap (\mathbf{X}'_j \cup \mathbf{Y}'_j) \cap (\mathbf{V}' \setminus \mathbf{V}'') \neq \emptyset.$$

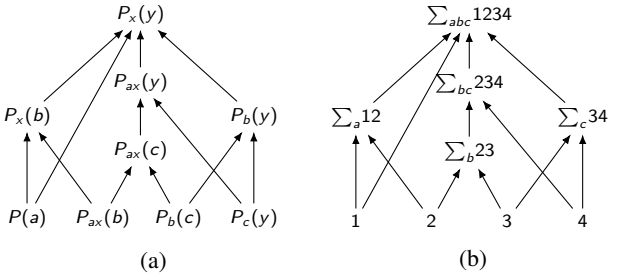


Figure 8. (a) embedding relationships among gc-factors where a directed edge $i \rightarrow j$ indicates that j embeds i. (b) the same embedding relationships with a sum-product notation.

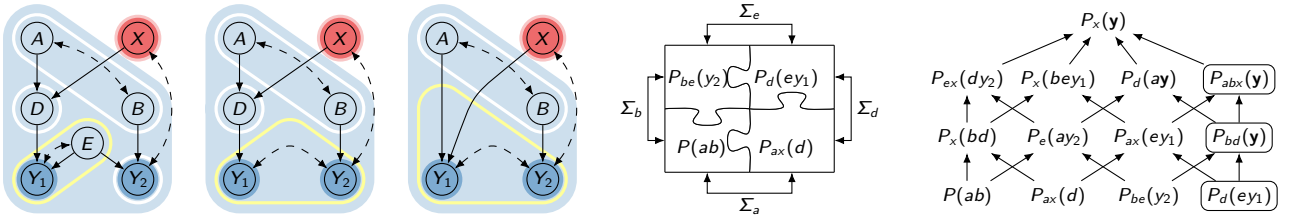
Note that a gc-factor is also a (non-proper) embedding factor of itself, and $P_{\mathbf{x}}(\mathbf{y})$, the query itself as $\langle \mathbf{X}, \mathbf{Y} \rangle$ is an embedding factor of every gc-factor. We illustrate embedding relationships among every gc-factor of $P_x(y)$ in $\mathcal{G}$ (Fig. 7a) in Fig. 8a. The four factors at the bottom correspond to $\mathbb{F}$ and the corresponding embedding relationships are represented as directed edges. For instance, $\langle \{A, X\}, \{C\} \rangle$ (i.e., $P_{a,x}(c)$) is an embedding factor of $\langle \{B\}, \{C\} \rangle$ and $\langle \{A, X\}, \{B\} \rangle$ (Fig. 7a). Quantitatively, embedding relationships can be understood as sum-product relationships (Fig. 7b), where projections are equivalent to marginalizations.

Although a gc-factor in $\mathbb{F}$ (a finest-grained factor) cannot be identified, it can be *identified as a group* under the summation, i.e., its embedding factor (a coarser-grained factor) is identified.

Definition 7 (Co-Identification)**.** If a factor Q is identifiable with a distribution, the factors in $\mathbb{F}$ embedded in Q are said to be *co-identified* with the distribution with respect to Q.

With the puzzle metaphor, co-identification of a factor refers to whether there will be a colored chunk which covers the piece. Following the strategy described in Lemma 3, $P_{\mathbf{x}}(\mathbf{y})$ will only be identified when every gc-factor in $\mathbb{F}$ is co-identified by one of the available distributions. However, there can be an exponential number of embedding factors for a gc-factor in $\mathbb{F}$. Hence, we investigate the relationship between (non-)identifiability of embedding factors.

[5] After submitting this work (in early February, 2020), Lee & Shpitser (2020) independently introduced the problem of 'mID', where 'm' stands for marginal distributions, which corresponds to the notion of partial-observability here (Def. 4). Even though there are subtle differences in terminology and notation, their main result in this context (Lemma 3) can be seen as our Lemma 3.



(a) $\mathcal{G}$ (b) $\mathcal{G}\langle\text{-}E\rangle$ (c) $\mathcal{G}\langle\text{-}\{E, D\}\rangle$ (d) $\mathbb{F}$ (e) Hasse diagram

Figure 9. (a, b, c) Illustrations of causal diagrams with c-components in $\mathbf{Y}'^{+}$ highlighted which correspond to $\mathbf{Y}$-side of gc-factors, (d) $\mathbb{F}$ as a puzzle and how pieces are merged by marginalization, (e) a Hasse diagram for the embedding relationships among factors (not exhaustive) where four vertices at the bottom are $\mathbb{F}$. The highlighted areas in (a, b, c) correspond to the right three vertices in (e).

4.1.2. Minimum Viable Embedding Factor

We introduce the crucial concept of *minimum viable embedding factor* (MVEF) that will help with the characterization of the co-identification of gc-factors in $\mathbb{F}$ with respect to a single available distribution. Its purpose is to find the finest-grained factors that a (partially-observed) distribution might be able to identify. In other words, as the default decomposition offers the factors ($\mathbb{F}$) of the right granularity for fully-observed distributions, we attempt to find factors of appropriate granularity with respect to a single partially-observed distribution.

Definition 8 (Minimum Viable Embedding Factor (MVEF))**.** An embedding factor $\langle \mathbf{X}^\dagger, \mathbf{Y}^\dagger \rangle \in \mathbb{F}^\dagger$ of a gc-factor $\langle \mathbf{X}_i, \mathbf{Y}_i \rangle \in \mathbb{F}$ is said to be a *minimum viable embedding factor* of $\langle \mathbf{X}_i, \mathbf{Y}_i \rangle$ with respect to $P_{\mathbf{Z}'}(\mathbf{V}')$ if the three criteria holds true: $\mathbf{Y}^\dagger \subseteq \mathbf{V}'$; $\mathbf{X}^\dagger \subseteq \mathbf{V}' \cup \mathbf{Z}'$; and $\mathbf{Z}' \cap An(\mathbf{Y}^\dagger)_{\mathcal{G}_{\underline{\mathbf{X}^\dagger}}} = \emptyset$, and $\mathbf{V} \setminus \mathbf{V}^\dagger$ is minimal. Further, $\mathbf{V} \setminus \mathbf{V}^\dagger$ is said to be a *MVEF-admissible set*.

Given one of available distributions $\mathbb{P}$, we want to check whether it co-identifies a gc-factor in $\mathbb{F}$. Among an exponential number of its embedding factors, we can choose an embedding factor, which satisfies the necessary conditions as specified in Prop. 1. A polynomial time algorithm for finding out an MVEF is depicted in Appendix D. The algorithm iteratively seeks variables to be projected out in order to satisfy the necessary conditions.

For example, take a look at a graph Fig. 9a where there are four factors in $\mathbb{F}$ (Fig. 9d). Consider co-identifying a factor $P_d(e, y_1)$ (with its $\mathbf{Y}$ side highlighted in Fig. 9a) with $P_B(A, X, \mathbf{Y}) \in \mathbb{P}$. Since D and E do not appear in $\mathbf{Z}' = \{B\}$ and $\mathbf{Z}' \cup \mathbf{V}' = \{B, A, X, \mathbf{Y}\}$, respectively (Prop. 1), we may project out both D and E in $\mathcal{G}$ at once, and examine the resulting embedding factor of $P_d(e, y_1)$ in $\mathcal{G}\langle\text{-}\{E, D\}\rangle$ (Fig. 9c), that is, $P_{a,b,x}(\mathbf{y})$. The puzzle diagram illustrates how projecting out D and E, or equivalently, $\sum_{d,e}$ yields a chunk with three pieces except $P(a, b)$. Further, a Hasse diagram (Fig. 9e) represents the transitive reduction of embedding relationships where one can examine that $P_{a,b,x}(\mathbf{y})$ embeds the three factors in $\mathbb{F}$ excluding $P(a, b)$ and other two intermediate factors. Another example is given in Appendix showing that obtaining an MVEF may take multiple steps.

An MVEF, if exists, is uniquely determined. Further, MVEFs obtained given a distribution for a subset of gc-factors are disjoint with respect to embedded gc-factors.

Proposition 4 (Uniqueness)**.** *If an MVEF exists for a gc-factor with respect to a distribution, then it is unique.*

Proposition 5 (Disjointness)**.** *Given a distribution, let Q and R be MVEFs of two gc-factors in $\mathbb{F}$, respectively. Then either $Q = R$ or there is no common gc-factor embedded in Q and R.*

MVEFs are only a subset of all factors at different levels. If they are identifiable with a distribution, then they correspond to same-colored chunks. Should we further examine the possibility of other chunks of the same color? If an MVEF is found and not identified, is it still possible for factors further embedding the MVEF to be identifiable? We show the non-identifiability of factors embedding a non-identified MVEF.

Lemma 4 (Hedge over Embedding Factors)**.** *Consider a gc-factor $\langle \mathbf{X}_i, \mathbf{Y}_i \rangle \in \mathbb{F}$ and a distribution $P_{\mathbf{Z}'}(\mathbf{V}')$. The gc-factor is not co-identifiable with $P_{\mathbf{Z}'}(\mathbf{V}')$ with respect to any of its embedding factors if its MVEF does not exist or there exists a hedge for $P_{\mathbf{x}^\dagger \setminus \mathbf{Z}'}(\mathbf{y}^\dagger)$ in $(\mathcal{G} \setminus \mathbf{Z}')\langle \mathbf{V}' \rangle$ given the MVEF $\langle \mathbf{X}^\dagger, \mathbf{Y}^\dagger \rangle \in \mathbb{F}^\dagger$.*

Hence, the failure to identify an MVEF informs us that none of its embedding factors needs to be examined for (non-)identifiability. Below is a complementing lemma that, with puzzle terms, any chunks of the same color is composed of MVEFs (i.e., the smallest chunks of the color).

Lemma 5 (Compositionality)**.** *Given a distribution $P_{\mathbf{Z}'}(\mathbf{V}')$, any identifiable embedding factor of a gc-factor in $\mathbb{F}$ can be represented as the summation over the product of a subset of identified MVEFs.*

For instance, given that $P(a)$ and $P_{a,x}(c)$ are identified with $P_X(A, C)$ (Fig. 7d), it is clear to see the identifiability

of $P_x(c)$ as $\sum_a P(a)P_{a,x}(c)$ and checking it is redundant. Thus, finding MVEFs and checking their identifiability are sufficient to comprehend how one distribution contributes to co-identification of a portion of factors.

4.2. Algorithm for GID-PO

Previous sections discussed what factors of different levels of projections will be available from each distribution. Hence, identified MVEFs from available distributions become chunks of different colors, which needs to be put together to complete the causal jigsaw puzzle $P_{\mathbf{x}}(\mathbf{y})$.

We devise a two-phase algorithm called GID-PO (Alg. 1), which first identifies MVEFs co-identifying gc-factors in $\mathbb{F}$ (colorful chunks), and then combine them to produce a formula for $P_{\mathbf{x}}(\mathbf{y})$. The implementation of the first phase is straightforward given the characterizations of MVEFs in the previous section. For every gc-factor in $\mathbb{F}$ and a distribution in $\mathbb{P}$ (Line 3), the existence of MVEF is first checked (Line 5). If it exists and identifiable, then record information including the MVEF and the gc-factors embedded in the MVEF (i.e., co-identified), the used distribution, and what variables are marginalized out (Lines 7–9).

With the colorful chunks of identified MVEFs, the second phase of the algorithm tries to complete the big picture. Based on Prop. 2, it attempts to select MVEFs whose co-identified gc-factors (represented as their indices in $\mathbb{F}$) do not overlap. This reduces to solving an *exact cover* (Line 13), an NP-complete problem (Karp, 1972): given a family $\mathbb{I}$ of subsets of a set $[n] = \{1, \dots, n\}$, whether there exists a subfamily $\mathbb{I}' \subseteq \mathbb{I}$ such that sets in $\mathbb{I}'$ are disjoint and $\cup\mathbb{I}' = \cup\mathbb{I} = [n]$. If a solution exists, then an expression for $P_{\mathbf{x}}(\mathbf{y})$ is written as the summation of product of sub-formulas corresponding to selected MVEFs where the variables to be marginalized follows Eq. (1) except that the part of variables are moved inside the sub-formulas (Line 16). Finally, we show next that this approach is indeed correct.

Theorem 1 (Soundness)**.** GID-PO *is sound.*

Remarkably, GID-PO is more efficient than a naive implementation of Lemma 3, which would require to run a traditional identifiability algorithm ($O(|\mathbf{V}|^4)$), for exponentially many gc-factors and for each dataset. On the other hand, GID-PO requires for the identifiability algorithm to run $|\mathbb{F}|$ for each distribution to collect the identified MVEFs. Then, an exact cover runs in exponential time in the number of uniquely identified MVEFs, which is upper bounded by $|\mathbb{F}|$. More specifically, GID-PO will run in $O(\ell mn^4 + 2^\ell)$ while a naive implementation for Lemma 3 runs in $O(mn^4 2^n)$, where $n = |V|$, $m = |\mathbb{P}|$, and $\ell = |\mathbb{F}|$. Further, if the problem instance is compatible with g-identifiability, where no partial-observability is involved, then the algorithm runs in a polynomial time in $|\mathbf{V}|$ since every identified MVEF will be a single-piece chunk and an exact cover algorithm will spot uncovered elements first.

Algorithm 1 GID-PO

1: **Input:** $\mathcal{G}$ a causal graph, $\mathbf{x}$ and $\mathbf{y}$ value assignments for a query, $\mathbb{P}$ a collection of available distributions
2: Prepare $\mathbb{J}$ an empty collection.
3: **for** $\langle P_{\mathbf{Z}'}(\mathbf{V}'), \langle \mathbf{X}_i, \mathbf{Y}_i \rangle \rangle \in \mathbb{P} \times \mathbb{F}$ **do**
4: **continue if** an MVEF that embeds $\langle \mathbf{X}_i, \mathbf{Y}_i \rangle$ is already found by the same data $P_{\mathbf{Z}'}(\mathbf{V}')$.
5: $\langle \mathbf{X}^\dagger, \mathbf{Y}^\dagger \rangle, \mathbf{W}^\dagger \leftarrow \text{MVEF}(\langle \mathbf{X}_i, \mathbf{Y}_i \rangle, P_{\mathbf{Z}'}(\mathbf{V}'))$.
6: **continue if** the MVEF is co-identified by some data.
7: **if** $\langle \mathbf{X}^\dagger, \mathbf{Y}^\dagger \rangle$ exists **and** identifiable with $P_{\mathbf{Z}'}(\mathbf{V}')$ **then**
8: $\mathbf{i} \leftarrow$ the indices of factors in $\mathbb{F}$ embedded in the MVEF.
9: Add $\langle \langle \mathbf{X}^\dagger, \mathbf{Y}^\dagger \rangle, \mathbf{W}^\dagger, \mathbf{i}, \langle \mathbf{Z}', \mathbf{V}' \rangle \rangle$ to $\mathbb{J}$.
10: **end if**
11: **end for**
12: Let $\mathbb{I}$ be the collection of indices in $\mathbb{J}$.
13: **if** $\mathbb{I}' \leftarrow$exact cover with $\{1, \dots, |\mathbb{F}|\}$ and $\mathbb{I}$ **then**
14: Let $\mathbb{J}' \subseteq \mathbb{J}$ be the subcollection matching the solution $\mathbb{I}'$.
15: Let $\mathbf{W}'$ be the union of all MVEF-admissible sets in $\mathbb{J}'$.
16: **return** $\sum_{\mathbf{y}^{+\prime} \setminus (\mathbf{Y} \cup \mathbf{W}')} \prod_{\mathbb{J}'} \text{GID}(P_{\mathbf{x}^\dagger}(\mathbf{y}^\dagger), \mathcal{G}\langle \mathbf{V}' \rangle, \{\mathbf{Z}'\})$.
17: **end if**
18: **return** NULL.

5. Discussions: Completeness, Complexity, Conditional Effects, and Transportability

We conjecture that Lemma 3 is also a necessary condition and, hence, our algorithm is sound and complete. We can show, under the completeness conjecture, that the decision version of our problem is NP-complete. Further, we discuss two possible extensions to this work.

On Completeness Our intuition for its completeness lies in the fact that each gc-factor in $\mathbb{F}$ provides non-decomposable information about the model. Formally speaking, a probability $P_{\mathbf{w}}(\mathbf{z}) \in \mathbb{F}$ is not identifiable from the combination of smaller pieces, i.e., $\mathbb{P} = \{P_{\mathbf{W}'}(\mathbf{Z}') \mid \mathbf{Z}' \subseteq \mathbf{Z}, \mathbf{W}' \subseteq \mathbf{W}\} \setminus \{P_{\mathbf{W}}(\mathbf{Z})\}$. Following Lemma 2, we can construct two models $\mathcal{M}^1$ and $\mathcal{M}^2$ where $\mathbf{W}$ and UCs are independent fair coins and each $Z \in \mathbf{Z}$ takes parents (including UCs) with exclusive-or for both models except the fact that one $Z' \in \mathbf{Z}$ flips its value for $\mathcal{M}^2$. Further, estimates for proper embedding factors and other gc-factors will not pinpoint a gc-factor of interest. For example, does knowing a chunk $\{f, g\}$ and piece $\{g\}$ allow us to infer what $\{f\}$ is? Consider an equation $h = \int_x f(x) \cdot g(x) \mathrm{d}x$. Generally, knowing h and g does not allow us to infer what f is. Having a number of such equations may help characterizing f but it won't identify f unless the domain of X is small and finite. For such reasons, we conjecture that $\mathbb{F}$ provides a fundamental, necessary, and sufficient building blocks for constructing $P_{\mathbf{x}}(\mathbf{y})$ given marginal interventional distributions. A rigorous proof for the completeness is still under investigation.

On NP-Completeness Assume that we have proved the completeness of the algorithm, which reflects Prop. 2. Then, we can show that the decision problem of the puzzle is indeed NP-complete. We present a polynomial reduction of an exact cover problem to a GID-PO problem instance.

Consider an arbitrary exact cover problem with a universe $[n]$ and a collection of its subsets.[6] We construct a causal graph $\mathcal{G}$ with $\mathbf{V} = \{V_{i,j}\}_{1 \le i < j \le n}$ and its direct edges are $V_{i,j} \to V_{j,i}$ for $i < j$. We add bidirected edges to connect $\mathbf{V}_{i,:} = \{V_{i,j}\}_{j \neq i}$ side by side. Examples of such causal graph with $n = 3$ and $n = 4$ are shown in Figs. 10a and 10b. We then fix a query as $P(\mathbf{V}')$ where $\mathbf{V}' = \{V_{j,i}\}_{j>i}$ are the sink nodes in $\mathcal{G}$. Each gc-factor $\langle pa(\mathbf{V}_{i,:}), \mathbf{V}_{i,:} \rangle \in \mathbb{F}$ represents element i in the universe $[n]$, e.g., $P(V_{1,2}, V_{1,3})$ for 1 and $P(V_{3,1}, V_{3,2} | do(V_{1,3}, V_{2,3}))$ for 3. Given a subset of integers $\mathbf{w} \subseteq [n]$, there exists a corresponding distribution $P_{\mathbf{X}^\dagger}(\mathbf{Y}^\dagger)$ in $\mathbb{P}$ such that $\langle \mathbf{X}^\dagger, \mathbf{Y}^\dagger \rangle \in \mathbb{F}^\dagger$ is an MVEF of any of gc-factor corresponding to $w \in \mathbf{w}$ where $\mathbf{V}^\dagger = \mathbf{V} \setminus \{V_{i,j}\}_{\{i \neq j | i,j \in \mathbf{w}\}}$. For instance, a subset of integers for an exact cover problem in Fig. 10a will be matched to the following distributions:

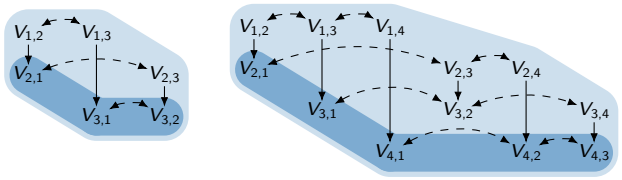
$$
\begin{aligned}
\{1\} &= P(V_{1,2}, V_{1,3}) \\
\{2\} &= P(V_{2,1}, V_{2,3} \mid do(V_{1,2})) \\
\{3\} &= P(V_{3,1}, V_{3,2} \mid do(V_{1,3}, V_{2,3})) \\
\{1,2\} &= P(V_{1,3}, V_{2,1}, V_{2,3}) \\
\{1,3\} &= P(V_{1,2}, V_{3,1}, V_{3,2} \mid do(V_{2,3})) \\
\{2,3\} &= P(V_{2,1}, V_{3,1}, V_{3,2} \mid do(V_{1,2}, V_{1,3})) \\
\{1,2,3\} &= P(V_{2,1}, V_{3,1}, V_{3,2})
\end{aligned}
$$

Hence, the construction can be done in a polynomial time of $n|\mathbb{I}|$, the size of the problem. Further, Alg. 1 yields an identification formula if and only if the reduced problem has an exact cover solution since each set as an available distribution can only identify itself but not others: Therefore, under the completeness of Lemma 3, we can show that the decision problem of general identifiability under partial-observability is NP-complete.

Generalization to Transportability and Conditional Distributions One of the possible extensions of this work is to adopting our result to *transportability* (Pearl & Bareinboim, 2011; Bareinboim & Pearl, 2014; Lee et al., 2020), which concerns about the identifiability when data comes from heterogeneous domains where mechanisms for some of the variables differ from a domain in which a causal effect is sought. Unless the differences between a source and a target (pieces' shapes are incompatible) are directly on the $\mathbf{Y}$-side of a factor, the same procedure can be applied.

Another generalization is identifying a *conditional* interventional distribution, e.g., $P_{\mathbf{x}}(\mathbf{y}|\mathbf{w})$ is delegated to identifying $P_{\mathbf{x}',\mathbf{w}'}(\mathbf{y}, \mathbf{w}'')$ (Tian, 2004; Shpitser & Pearl, 2006b; Lee et al., 2020). This change only requires an additional pre- and post-process of the result from GID-PO or its extension to transportability. We provide the generalization of our problem taking account both directions in Appendix F.

[6] An empty set can be simply ignored because it does not contribute to answering the exact cover problem.



(a) (b)

Figure 10. Causal diagrams for exact cover problems with universe (a) $\{1, 2, 3\}$ and (b) $\{1, 2, 3, 4\}$.

6. Conclusions

We introduced the general identifiability problem when the available distributions are only partially observable, which is named GID-PO. We investigated how a causal query can be factorized under different levels of projections, and then introduced new constructs called *embedding factors* and *co-identification*. These constructs make explicit the connection of the factors required to identify the targeted query and the available observed distributions, which allows a systematic view of the problem of identifiability under different granularities. We introduced a new graphical structure called *minimum viable embedding factor* (MVEF) and studied its properties, including its uniqueness, disjointedness, and compositionality. Putting these results together, we developed a new algorithm (GID-PO) that efficiently and systematically examines the identifiability of embedding factors and combines the identified MVEFs to compose the expression for a given query. Since each of the factors cannot be identified from smaller marginal interventional distributions, we conjecture that the procedure is also necessary. Assuming its completeness, we showed that the decision version of this new identifiability problem is NP-complete; yet, it does run in a polynomial time in the number of observed variables for the problems under full-observability (Lee et al., 2019). Noting that, in practice, available datasets are measured inconsistently with respect to the variables they cover – they usually have different columns – we hope the results in this paper can help data scientists tackle more challenging identification instances and determine causal effects in more intricate and realistic scenarios.

Acknowledgements

This research is supported in parts by grants from NSF (IIS-1704352 and IIS-1750807 (CAREER)).

References

Bareinboim, E. and Pearl, J. Causal inference by surrogate experiments: z-identifiability. In de Freitas, N. and Murphy, K. (eds.), *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 113–120, Corvallis, OR, 2012a. AUAI Press.

Bareinboim, E. and Pearl, J. Controlling selection bias in causal inference. In Girolami, M. and Lawrence, N. (eds.), *Proceedings of The Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, pp. 100–108. JMLR (22), 2012b.

Bareinboim, E. and Pearl, J. Transportability from multiple environments with limited experiments: Completeness results. In *Advances in Neural Information Processing Systems 27*, pp. 280–288. Curran Associates, Inc., 2014.

Bareinboim, E. and Pearl, J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113:7345–7352, 2016.

Correa, J. and Bareinboim, E. Causal effect identification by adjustment under confounding and selection biases. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 3740–3746. AAAI Press, 2017.

Huang, Y. and Valtorta, M. Identifiability in causal bayesian networks: A sound and complete algorithm. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*, pp. 1149–1156. AAAI Press, Menlo Park, CA, 2006a.

Huang, Y. and Valtorta, M. Pearl's calculus of intervention is complete. In Dechter, R. and Richardson, T. (eds.), *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, pp. 217–224. AUAI Press, Corvallis, OR, 2006b.

Jaber, A., Zhang, J., and Bareinboim, E. Causal identification under Markov equivalence: Completeness results. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2981–2989. PMLR, 2019.

Karp, R. M. Reducibility among combinatorial problems. In Miller R.E., Thatcher J.W., B. J. (ed.), *Complexity of Computer Computations*, The IBM Research Symposia Series, pp. 85–103. Plenum Press, New York, 1972.

Lee, J. J. R. and Shpitser, I. Identification methods with arbitrary interventional distributions as inputs, 2020. arXiv:2004.01157.

Lee, S., Correa, J. D., and Bareinboim, E. General identifiability with arbitrary surrogate experiments. In *Proceedings of the Thirty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence*, Corvallis, OR, 2019. AUAI Press.

Lee, S., Correa, J. D., and Bareinboim, E. Generalized transportability: Synthesis of experiments from heterogeneous domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI Press, 2020.

Mohan, K. and Pearl, J. On the testability of models with missing data. In *Proceedings of The Seventeenth International Conference on Artificial Intelligence and Statistics*. JMLR, 2014.

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

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

Pearl, J. and Bareinboim, E. Transportability of causal and statistical relations: A formal approach. Technical Report Technical Report r-372, Cognitive Systems Laboratory, Department of Computer Science, UCLA, 2011.

Pearl, J. and Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

Shpitser, I. and Pearl, J. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of The Twenty-First National Conference on Artificial Intelligence*, pp. 1219–1226. AAAI Press, 2006a.

Shpitser, I. and Pearl, J. Identification of conditional interventional distributions. In Dechter, R. and Richardson, T. (eds.), *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, pp. 437–444. AUAI Press, Corvallis, OR, 2006b.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2001.

Tian, J. *Studies in Causal Reasoning and Learning*. PhD thesis, Computer Science Department, University of California, Los Angeles, CA, November 2002.

Tian, J. Identifying conditional causal effects. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 561–568. AUAI Press, 2004.

Tian, J. and Pearl, J. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002)*, pp. 567–573, Menlo Park, CA, 2002. AAAI Press/The MIT Press.

Tian, J. and Pearl, J. On the identification of causal effects. Technical Report R-290-L, Department of Computer Science, University of California, Los Angeles, CA, 2003.