In Marina Meila and Tom Heskes (Eds.), Proceedings of the 31st Conference on
Uncertainty in Artificial Intelligence, Corvallis, OR: AUAI Press, 802--811, 2015.

TECHNICAL REPORT
R-454
July 2015

Missing Data as a Causal and Probabilistic Problem

Ilya Shpitser
Mathematical Sciences
University of Southampton
Southampton, UK SO14 6WD
i.shpitser@soton.ac.uk

Karthika Mohan
Dept. of Computer Science
Univ. of California, Los Angeles
Los Angeles, CA 90095
karthika@cs.ucla.edu

Judea Pearl
Dept. of Computer Science
Univ. of California, Los Angeles
Los Angeles, CA 90095
judea@cs.ucla.edu

Abstract

Causal inference is often phrased as a missing data problem – for every unit, only the response to observed treatment assignment is known, the response to other treatment assignments is not. In this paper, we extend the converse approach of [7] of representing missing data problems to causal models where only interventions on missingness indicators are allowed. We further use this representation to leverage techniques developed for the problem of identification of causal effects to give a general criterion for cases where a joint distribution containing missing variables can be recovered from data actually observed, given assumptions on missingness mechanisms. This criterion is significantly more general than the commonly used "missing at random" (MAR) criterion, and generalizes past work which also exploits a graphical representation of missingness. In fact, the relationship of our criterion to MAR is not unlike the relationship between the **ID** algorithm for identification of causal effects [22, 18], and conditional ignorability [13].

1 INTRODUCTION

Missing data is a ubiquitous problem in data analysis, and can arise due to imperfect data collection, or various types of censoring, for instance via loss to followup, or death. In addition, causal inference can be viewed as a missing data problem, since the fundamental problem of causal inference [4] is that for every unit only the response to observed treatment assignment is known, the responses to other, hypothetical treatment assignments are not known.

Handling missing data entails either dealing with a latent variable model or finding plausible assumptions under which *recoverability*, that is unbiased inferences about *all* cases from the *observed* cases, is possible. Well-known approaches of the former type include fitting a latent variable model via gradient descent [17], the EM algorithm [1], or performing Monte Carlo averaging via multiple imputation [16]. Well-known approaches of the latter type include the Kaplan-Meier estimator in survival analysis [5], and adjustments based on Missing Completely At Random (MCAR), and Missing At Random (MAR) assumptions [15].

While methods based on inference in a latent variable model are more generally applicable, they are also methodologically and computationally challenging. At the same time, recoverability methods based on MCAR and MAR rely on strong assumptions on how missingness comes about. When neither MCAR nor MAR holds, data is said to be Missing Not At Random (MNAR), and in this case a characterization of recoverability is an open problem, although many sufficient conditions for recoverability are known [7, 6].

In this paper, we take the *converse* view to "causality as missing data," and view missing data as a particular type of partly causal, and partly probabilistic inference problem [2, 7]. We then represent this problem using partly causal, and partly probabilistic graphical models, and exploit techniques developed for similar models in the context of identification of causal effects to develop a general algorithm for recoverability under MNAR. In fact, the relationship between our algorithm and MAR is not unlike the relationship between the **ID** algorithm for identification of causal effects [22, 18, 19], and the conditional ignorability assumption in causal inference [13].

The paper is organized as follows. We introduce the notation and concepts we will need in section 2. In section 3, we use missingness graphs and missingness models to formally define missing data as a type of causal inference problem where only interventions on certain variables are allowed. We introduce recoverability and give examples of where recoverability is possible in MNAR settings in section 4. We introduce a general algorithm for recoverability we call **MID** in section 5, and show it is sound. Section 6 illustrates a complex case where the entire recursive structure of **MID** is necessary. Section 7 discusses nonrecoverability, and section 8 contains our conclusions.

2 PRELIMINARIES

Variables are capital letters, values are small letters. Variable sets are bold capital letters, value sets are bold small letters. A state space for a variable A is $\mathfrak{X}_A$. A state space for a set of variables $\mathbf{A}$ is the Cartesian product of the individual state spaces: $\mathfrak{X}_{\mathbf{A}} \equiv \times_{A \in \mathbf{A}} \mathfrak{X}_A$. For a set of values $\mathbf{a}$, and $\mathbf{B} \subseteq \mathbf{A}$, denote by $\mathbf{a}_{\mathbf{B}}$ a projection of $\mathbf{a}$ to $\mathbf{B}$. Denote $\mathbf{a}_B$ as a shorthand for $\mathbf{a}_{\{B\}}$. We will denote a vector of 0s as $\mathbf{0}$. $\mathbf{0}_{\mathbf{B}}$ means "a set of 0 values to $\mathbf{B}$."

2.1 GRAPH THEORY AND NOTATION

A directed graph consists of a set of nodes and directed arrows ($\rightarrow$) connecting pairs of nodes. A mixed graph consists of a set of nodes and directed and/or bidirected arrows ($\leftrightarrow$) connecting pairs of nodes. A path is a sequence of distinct edges where any edge in a sequence that ends in a node A implies the subsequent edge must start with A, and each such node A may only occur at most once in this way in the sequence. A directed path from a node X to a node Y is a path consisting of directed edges where all edges on the path point away from X and towards Y.

If the edge $X \rightarrow Y$ exists in a graph $\mathcal{G}$, we say X is a parent of Y and Y is a child of X. If a directed path from X to Y exists in $\mathcal{G}$, we say X is an ancestor of Y, and Y is a descendant of X. We denote by $\text{pa}_{\mathcal{G}}(A), \text{ch}_{\mathcal{G}}(A), \text{de}_{\mathcal{G}}(A), \text{an}_{\mathcal{G}}(A), \text{nd}_{\mathcal{G}}(A)$ the sets of parents, children, descendants, ancestors, and non-descendants of A in $\mathcal{G}$, respectively. These are defined disjunctively for sets, e.g. $\text{pa}_{\mathcal{G}}(\mathbf{A}) = \bigcup_{A \in \mathbf{A}} \text{pa}_{\mathcal{G}}(A)$. Let $\text{fa}_{\mathcal{G}}(A) = \text{pa}_{\mathcal{G}}(A) \cup \{A\}$, $\text{pa}^s_{\mathcal{G}}(\mathbf{A}) = \text{pa}_{\mathcal{G}}(\mathbf{A}) \setminus \mathbf{A}$, $\text{ndp}_{\mathcal{G}}(A) = \text{nd}_{\mathcal{G}}(A) \setminus \text{pa}_{\mathcal{G}}(A)$. Given a graph $\mathcal{G}$, we say a vertex set $\mathbf{A}$ is *ancestral* if $\text{an}_{\mathcal{G}}(\mathbf{A}) = \mathbf{A}$. By convention, in any directed graph, $A \in \text{an}_{\mathcal{G}}(A) \cap \text{de}_{\mathcal{G}}(A)$. A directed graph is said to have a directed cycle if there is X, Y such that $X \in \text{an}_{\mathcal{G}}(Y) \cap \text{ch}_{\mathcal{G}}(Y)$. A directed graph without such cycles is called a directed acyclic graph (DAG).

A *conditional DAG* (CDAG) $\mathcal{G}(\mathbf{V} \mid \mathbf{W})$ is a DAG with vertices $\mathbf{V} \cup \mathbf{W}$ with the property that $\text{pa}_{\mathcal{G}}(\mathbf{W}) = \emptyset$. We will denote vertices in $\mathbf{V}$ as circles, and vertices in $\mathbf{W}$ as squares. Note that we do not require that all $V \in \mathbf{V}$ must have parents. We simply distinguish certain parentless nodes in $\mathcal{G}$ as $\mathbf{W}$. We will interpret vertices in $\mathbf{V}$ as associated with *random variables* and vertices in $\mathbf{W}$ as associated with variables that have been "set to a constant" in some way. One example of a CDAG is a *mutilated graph* that arises in the analysis of interventional distributions. When considering d-separation on vertices in $\mathbf{V}$ in a CDAG [9], we will treat it as ordinary d-separation in a DAG, except all nodes in $\mathbf{W}$ are implicitly conditioned on.

If vertices not in $\mathbf{W}$ in a CDAG correspond to a variable partition into observed and missing variables, we will explicitly denote the set of vertices corresponding to missing variables as $\mathbf{M}$, and the other vertices as $\mathbf{O}$, like so: $\mathcal{G}(\mathbf{O}, \mathbf{M} \mid \mathbf{W})$. A CDAG where $\mathbf{W}$ is empty is written as $\mathcal{G}(\mathbf{V})$ or $\mathcal{G}(\mathbf{O}, \mathbf{M})$ as a shorthand.

A conditional acyclic directed mixed graph (CADMG) $\mathcal{G}(\mathbf{V} \mid \mathbf{W})$ is a mixed graph with two types of edges $\rightarrow$ and $\leftrightarrow$ with no directed cycles, where no arrowhead may point to an element of $\mathbf{W}$. We will sometimes omit variables from CDAGs and CADMGs if they are obvious to avoid notation clutter, e.g. we will write $\mathcal{G}(\mathbf{V} \mid \mathbf{W})$ simply as $\mathcal{G}$. Given a CDAG $\mathcal{G}(\mathbf{O}, \mathbf{M} \mid \mathbf{W})$, define $\mathcal{G}_{\underline{\mathbf{B}}}(\mathcal{G})$ to be an edge subgraph obtained from $\mathcal{G}$ by removing all arrows pointing away from $\mathbf{B}$.

Define a *latent projection of* $\mathcal{G}(\mathbf{O}, \mathbf{M} \mid \mathbf{W})$ *onto* $\mathbf{O} \cup \mathbf{W}$ [23] to be a CADMG $\mathcal{G}_{(\mathbf{O})}(\mathbf{O}, \mathbf{M} \mid \mathbf{W}) \equiv \mathcal{G}^{\dagger}(\mathbf{O} \mid \mathbf{W})$ such that for any $V_1, V_2 \in \mathbf{O} \cup \mathbf{W}$:

- There is an edge $V_1 \rightarrow V_2$ if and only if there is a directed path $V_1 \rightarrow \ldots \rightarrow V_2$ in $\mathcal{G}(\mathbf{O}, \mathbf{M} \mid \mathbf{W})$ with all intermediate nodes in $\mathbf{M}$.
- There is an edge $V_1 \leftrightarrow V_2$ if and only if there is a marginally d-connected path $V_1 \leftarrow \ldots \rightarrow V_2$ in $\mathcal{G}(\mathbf{O}, \mathbf{M} \mid \mathbf{W})$ with all intermediate nodes in $\mathbf{M}$.

Latent projections are a simplified representation of an infinitely large class of hidden variable CDAGs with structural features in common. In this paper, we use them only to simplify the statements and proofs of our results. The results themselves will always be about models represented by DAGs (and CDAGs).

Given a CDAG $\mathcal{G}(\mathbf{V} \mid \mathbf{W})$, and $\mathbf{A} \subseteq \mathbf{V} \cup \mathbf{W}$, define $\mathcal{G}_{\mathbf{A}}(\mathbf{V} \mid \mathbf{W}) \equiv \mathcal{G}(\mathbf{V} \cap \mathbf{A} \mid \mathbf{W} \cap \mathbf{A})$ be a subgraph of $\mathcal{G}$ containing the vertex set $\mathbf{A}$ and any edge in $\mathcal{G}$ between elements in $\mathbf{A}$.

Given a CADMG $\mathcal{G}(\mathbf{V} \mid \mathbf{W})$, and $V \in \mathbf{V}$, define *the district (or c-component [22, 18]) of* V *in* $\mathcal{G}(\mathbf{V} \mid \mathbf{W})$ to be $\text{dis}_{\mathcal{G}}(V) = \{A \in \mathbf{V} \mid V \leftrightarrow \ldots \leftrightarrow A\}$. The set of districts of $\mathcal{G}(\mathbf{V} \mid \mathbf{W})$ is denoted by $\mathcal{D}(\mathcal{G}(\mathbf{V} \mid \mathbf{W}))$, and it partitions $\mathbf{V}$.

For any $V \in \mathbf{O}$ in a CDAG $\mathcal{G}(\mathbf{O}, \mathbf{M} \mid \mathbf{W})$ where for every $M \in \mathbf{M}$, $\text{de}_{\mathcal{G}}(M) \cap \mathbf{O} \neq \emptyset$, define the *clan of* V as $\text{cla}_{\mathcal{G}}(V) \equiv \text{an}_{\mathcal{G}_{\mathbf{D}_V \cup \mathbf{M}}}(\mathbf{D}_V)$, where $\mathbf{D}_V = \text{dis}_{\mathcal{G}_{(\mathbf{O})}}(V)$. For example, in $\mathcal{G}$ shown in Fig. 1 (c), where $\{X, W\}$ are missing, $\text{cla}_{\mathcal{G}}(R_X) = \text{cla}_{\mathcal{G}}(S_W) = \{W, R_X, S_W\}$, and $\text{cla}_{\mathcal{G}}(R_W) = \text{cla}_{\mathcal{G}}(S_X) = \{X, R_W, S_X\}$.

For any $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{(\mathbf{O})}(\mathbf{O}, \mathbf{M} \mid \mathbf{W}))$, and $D_1, D_2 \in \mathbf{D}$, $\text{cla}_{\mathcal{G}}(D_1) = \text{cla}_{\mathcal{G}}(D_2)$. Thus we will write $\text{cla}_{\mathcal{G}}(\mathbf{D}) \equiv \text{cla}_{\mathcal{G}}(D)$, for any $D \in \mathbf{D}$. In fact, the set of clans partitions $\mathbf{O} \cup \mathbf{M}$ in $\mathcal{G}$ with the property above.

Given a CDAG $\mathcal{G}$, a total ordering $\prec$ on vertices in $\mathcal{G}$ is *topological given* $\mathcal{G}$ if $A \prec B$ implies $A \notin \text{de}_{\mathcal{G}}(B)$. Given an ordering $\prec$ topological given $\mathcal{G}$, define for any vertex V

in $\mathcal{G}$, $\mathrm{pre}_{\mathcal{G},\prec}(V) = \{W \neq V \mid W \prec V\}$. Given $\prec$ topological for $\mathcal{G}$ with a vertex set $\mathbf{V}$, if there is a subgraph $\mathcal{G}'$ of $\mathcal{G}$ with a vertex set $\mathbf{V}' \subset \mathbf{V}$, we will view $\prec$ with respect to $\mathcal{G}'$ as the natural subordering restricted to $\mathbf{V}'$. Note that this subordering will also be topological with respect to $\mathcal{G}'$.

A counterfactual (potential outcome) $Y(\mathbf{a})$ [8, 14] is a response Y to a hypothetical assignment of a set of treatments $\mathbf{A}$ to values $\mathbf{a}$. Given a set of potential outcomes $Y_1(\mathbf{a}), \ldots Y_k(\mathbf{a})$, where $\mathbf{Y} = \{Y_1, \ldots Y_k\}$, we may consider a joint distribution

$$p(\{Y_1, \ldots Y_k\}(\mathbf{a})) \equiv p(\mathbf{Y}(\mathbf{a})) \equiv p(\mathbf{Y} \mid \mathrm{do}(\mathbf{a})).$$

The do(.) notation is discussed extensively in [10].

3 MISSING GRAPHS AND MISSINGNESS MODELS

Given a CDAG $\mathcal{G}(\mathbf{V} \mid \mathbf{W})$, we say $p_{\mathbf{W}}(\mathbf{V})$ (a mapping from $\mathfrak{X}_{\mathbf{W}}$ to $p(\mathbf{V})$) is Markov relative to $\mathcal{G}$ if

$$p_{\mathbf{W}}(\mathbf{V}) = \prod_{V \in \mathbf{V}} p_{\mathbf{W}}(V \mid \mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{W}), \qquad (1)$$

and each term $p_{\mathbf{W}}(V \mid \mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{W})$ only depends on $\mathbf{W} \cap \mathrm{pa}_{\mathcal{G}}(V)$.

Definition 1 (missingness graph) *Given a DAG $\mathcal{G}(\mathbf{O}, \mathbf{M})$, a DAG $\mathcal{G}^m$ is called a* missingness graph *for $\mathcal{G}$ if $\mathcal{G}^m$ has the vertex set $\mathbf{O} \cup \mathbf{M} \cup \mathbf{R}_{\mathbf{M}} \cup \mathbf{S}_{\mathbf{M}}$, where $\mathbf{R}_{\mathbf{M}} = \{R_M \mid M \in \mathbf{M}\}$, $\mathbf{S}_{\mathbf{M}} = \{S_M \mid M \in \mathbf{M}\}$, $\mathcal{G} = \mathcal{G}^m_{\mathbf{O}\cup\mathbf{M}}$, and for all M in $\mathbf{M}$, $\mathrm{pa}_{\mathcal{G}^m}(S_M) = \{M, R_M\}$, $\mathrm{ch}_{\mathcal{G}^m}(S_M) = \emptyset$, and $\mathrm{ch}_{\mathcal{G}^m}(R_M) \cap (\mathbf{O} \cup \mathbf{M}) = \emptyset$.*

By convention, if $\mathbf{M} = \emptyset$, then $\mathbf{S}_{\emptyset} = \mathbf{R}_{\emptyset} = \emptyset$. We will refer to $\mathbf{O} \cup \mathbf{R}_{\mathbf{M}} \cup \mathbf{S}_{\mathbf{M}}$ as $\mathbf{V}$, and to $\mathbf{V} \cup \mathbf{M}$ as $\mathbf{A}$. We call elements of $\mathbf{R}_{\mathbf{M}}$ indicators, and elements of $\mathbf{S}_{\mathbf{M}}$ proxies.

Define $\mathcal{M}(\mathcal{G}^m(\mathbf{A}))$ to be the *missingness model* for a missingness graph $\mathcal{G}^m(\mathbf{A})$ as a set of distributions $\{p(\mathbb{A})\}$ over the following set of counterfactuals $\mathbb{A} \equiv \{\mathbf{A}(\mathbf{r}) | \mathbf{R} \subseteq \mathbf{R}_{\mathbf{M}}, \mathbf{r} \in \mathfrak{X}_{\mathbf{R}}\}$, such that $(\forall M \in \mathbf{M})\ \mathfrak{X}_{R_M} = \{0, 1\}$, $\mathfrak{X}_{S_M} = \mathfrak{X}_M \cup \{\textbf{missing}\}$, and the missingness mechanism that determines the value of S_M is as follows: $S_M(0_{R_M}) = M$ and $S_M(1_{R_M}) = \textbf{missing}$. In addition: $(\forall \mathbf{R} \subseteq \mathbf{R}_{\mathbf{M}}, \mathbf{r} \in \mathfrak{X}_{\mathbf{R}}, V \in \mathbf{A})$,

$$V(\mathbf{r}) \perp\!\!\!\perp \{\mathrm{ndp}_{\mathcal{G}^m_{\underline{\mathbf{R}}}}(V)\}(\mathbf{r}) \mid \{\mathrm{pa}_{\mathcal{G}^m_{\underline{\mathbf{R}}}}(V)\}(\mathbf{r}). \qquad (2)$$

To obtain the set $\mathbb{A}$, we first define

$$\{\mathbf{A}(\mathbf{r}) | \mathbf{r} \in \mathfrak{X}_{\mathbf{R}_{\mathbf{M}}}\} \equiv \{\mathbf{S}_{\mathbf{M}}(\mathbf{r}), \mathbf{O}, \mathbf{M} | \mathbf{r} \in \mathfrak{X}_{\mathbf{R}_{\mathbf{M}}}\},$$

and obtain the others via modified recursive substitution as in definition 43 in [11], pp. 100-101.

A missingness model is thus really a particular type of a graphical causal model where we only define interventions on a subset of variables [11]. In particular, we allow $\mathcal{G}(\mathbf{O}, \mathbf{M})$ to represent an ordinary hidden variable statistical model. (2) is just the DAG local Markov property linking $p(\mathbf{A}(\mathbf{r}))$ and $\mathcal{G}^m_{\underline{\mathbf{R}}}$, for every $\mathbf{r}$. If we had chosen to split variables in $\mathbf{R}$ into random and intervened versions, and display both explicitly in the graph rather than only displaying the random version of variables, and keeping intervened versions implicit, as we do in $\mathcal{G}^m_{\underline{\mathbf{R}}}$, we would end up with Single World Intervention Graphs (SWIGs), and the appropriate local Markov property for those graphs, as discussed in [11].

Standard results on DAG models imply (2) is equivalent to (1) for $p(\mathbf{A}(\mathbf{r}))$ and $\mathcal{G}^m_{\underline{\mathbf{R}}}$ (if we let $\mathbf{W} = \emptyset$, and keep fixed versions of $\mathbf{R}$ implicit in the graph). We may also let $\mathbf{W} = \mathbf{R}$, and treat $\mathbf{R}$ as a split node as in a SWIG.

4 RECOVERABILITY

We call $p(\mathbf{V})$ the *manifest distribution*. A functional of $p(\mathbb{A})$, $f(p(\mathbb{A}))$ is said to be *recoverable* given $p(\mathbf{V})$ in $\mathcal{G}^m$ if there is a functional g of $p(\mathbf{V})$, such that $f(p(\mathbb{A})) = g(p(\mathbf{V}))$ for every element of $\mathcal{M}(\mathcal{G}^m)$. In this paper, we will concentrate on recoverability of $p(\mathbf{O} \cup \mathbf{M})$, although many other kinds of recoverability problems are also interesting, for instance recovering the causal effect in a causal model with missingness.

We explicitly represent missingness as a causal inference problem because this allows us to rephrase recoverability as identifiability of causal effects. If we were allowed to assign $\mathbf{R}_{\mathbf{M}}$ without affecting other variables, we could use proxies $\mathbf{S}_{\mathbf{M}}$ to recover the behavior of the underlying missing variables $\mathbf{M}$, due to the following result.

Lemma 1 *In a DAG $\mathcal{G}$ where $\mathbf{M} \neq \emptyset$, for any $p(\mathbb{A}) \in \mathcal{M}(\mathcal{G}^m(\mathbf{V}, \mathbf{M}))$, and $R_M \in \mathbf{R}_{\mathbf{M}}$, $p(\mathbb{Y}) \in \mathcal{M}(\mathcal{G}^m(\mathbf{V} \cup \{M\}, \mathbf{M} \setminus \{M\})_{\mathbf{V}\cup\mathbf{M}\setminus\{S_M, R_M\}})$, where $\mathbb{Y}$ is*

$$\left\{\{\mathbf{V} \cup \mathbf{M} \setminus \{R_M\}\}(\mathbf{r}, 0_{R_M}) \middle| \mathbf{R} \subseteq \mathbf{R}_{\mathbf{M}\setminus\{M\}}, \mathbf{r} \in \mathfrak{X}_{\mathbf{R}}\right\}.$$

Proof: $\{\mathbf{V} \cup \mathbf{M}\}(\mathbf{r}, 0_{R_M})$ obeys (2) for $\mathcal{G}^m_{\underline{\mathbf{R}\cup\{R_M\}}}$. Since $\mathbf{A} \setminus \{R_M\}$ is ancestral in $\mathcal{G}^m_{\underline{\mathbf{R}\cup\{R_M\}}}$, $\{\mathbf{V} \cup \mathbf{M} \setminus \{R_M\}\}(\mathbf{r}, 0_{R_M})$ obeys (2) for $(\mathcal{G}^m_{\underline{\mathbf{R}\cup\{R_M\}}})_{\mathbf{A}\setminus\{R_M\}}$. Our conclusion follows since $M = S_M(0_{R_M})$. □

In other words, fixing R_M to 0 gives a new model where M is effectively observed since $M = S_M(0_{R_M})$. This implies that if we were able to fix all of $\mathbf{R}_{\mathbf{M}}$, we could recover $p(\mathbf{O} \cup \mathbf{M})$.

Corollary 1 $p(\{\mathbf{O}, \mathbf{S}_{\mathbf{M}}\}(\mathbf{0}_{\mathbf{R}_{\mathbf{M}}})) = p(\mathbf{O} \cup \mathbf{M})$ *for any $\mathcal{G}^m$, and any $p(\mathbb{A}) \in \mathcal{M}(\mathcal{G}^m)$.*

This corollary implies that our recoverability problem is solved by expressing a particular interventional distribution as a function of the manifest in a restricted causal model. We will attack this problem via two standard results for causal models that hold in restricted causal models as well, as shown in [11], propositions 45 and 46.

Theorem 1 *For any* $p(\mathbb{A}) \in \mathcal{M}(\mathcal{G}^m(\mathbf{V}, \mathbf{M}))$*, and* $(\forall \mathbf{R} \subseteq \mathbf{R_M}, \mathbf{r} \in \mathfrak{X}_\mathbf{R})$*,*

$$p(\mathbf{A}(\mathbf{r})) = \prod_{V \in \mathbf{A}} p(V \mid \text{pa}_{\mathcal{G}^m}(V) \setminus \mathbf{R}, \mathbf{r}_{\text{pa}_{\mathcal{G}^m}(V) \cap \mathbf{R}}). \quad (3)$$

Theorem 2 *For any* $p(\mathbb{A}) \in \mathcal{M}(\mathcal{G}^m(\mathbf{V}, \mathbf{M}))$*, and* $(\forall \mathbf{R} \subseteq \mathbf{R_M}, \mathbf{r} \in \mathfrak{X}_\mathbf{R})$*,*

$$p(\{(\mathbf{V} \cup \mathbf{M}) \setminus \mathbf{R}\}(\mathbf{r}) \mid \mathbf{r}) = p((\mathbf{V} \cup \mathbf{M}) \setminus \mathbf{R} \mid \mathbf{r}). \quad (4)$$

(3) is known as the truncated factorization [10], manipulated distribution [21], or the g-formula [12]. (4) is known as the *consistency* property.

We now illustrate how constraints of the missingness model encoded by $\mathcal{G}^m$, as well as (3) and (4) lead to recoverability.

4.1 EXAMPLES OF RECOVERABILITY

Consider Fig. 1, where X, C, W may possibly be high-dimensional. In Fig. 1 (a), X is missing according to a mechanism governed by an independent proxy R_X, so

$$p(X) = p(S_X(0_{R_X})) = p(S_X \mid R_X = 0).$$

The assumption present in this model which allows us to recover the underlying missing variable, namely $(S_X(0_{R_X}) \perp\!\!\!\perp R_X)$ is known as *missing completely at random* (MCAR) assumption.[1] This assumption is the missingness analogue of *ignorability* (lack of confounding between the missingness indicator R_X and the proxy $S_X(r)$ under assignment r to R_X).

In Fig. 1 (b), X is missing according to a mechanism governed by a proxy R_X which has a (statistical) dependence on X through C, which is a fully observed variable. In this case,

$$p(X, C) = p(S_X(0_{R_X}) \mid C)p(C) = p(S_X \mid R_X = 0, C)p(C).$$

The assumption present in this model which allows us to recover the underlying missing variable, namely $(S_X(0_{R_X}) \perp\!\!\!\perp R_X \mid C)$ is known as the *missing at random* (MAR) assumption. This assumption is the missingness analogue of *conditional ignorability* (lack of confounding between the indicator R_X and the proxy $S_X(r)$ under assignment r to R_X given that we conditioned on a set of variables C).

In Fig. 1 (c), it is not the case that

$$\{S_W(0_{R_W}), S_X(0_{R_X})\} \perp\!\!\!\perp \{R_X, R_W\}.$$

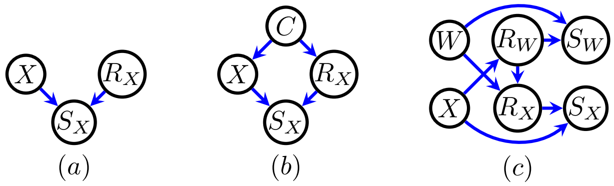

Figure 1: (a) A missingness model satisfying the *missing completely at random* (MCAR) assumption. (b) A missingness model satisfying the *missing at random* (MAR) assumption. (c) A missingness model where missingness is *not at random* (MNAR), but where recoverability is nevertheless possible.

That is, data on X, W is not missing completely at random (nor at random, since there is no fully observed variable to screen off the dependence of proxies under indicator assignment from indicators.) Nevertheless, despite the fact that data on $p(X, W)$ is missing not at random (MNAR), we now show that $p(X, W)$ is recoverable. We will exploit the fact that the missingness model implies

$$\{S_W(0_{R_W}), R_X(0_{R_W})\} \perp\!\!\!\perp \{S_X(0_{R_X}), R_W\}. \quad (5)$$

It is not difficult to show that $p(R_W, R_X, S_W, S_X)$ is equal to

$$p(\{S_W, R_X\}(R_W)) \cdot p(S_X(R_X), R_W) =$$
$$(p(S_W \mid R_X, R_W)p(R_X \mid R_W)) \cdot (p(S_X \mid R_X, R_W)p(R_W))$$

This implies $p(X, W) = p(X)p(W)$ is equal to

$$\left(\sum_{R_X} p(\{S_W, R_X\}(0_{R_W}))\right) \cdot \left(\sum_{R_W} p(S_X(0_{R_X}), R_W)\right) =$$
$$p(S_W \mid 0_{R_W}) \cdot \left(\sum_{R_W} p(S_X \mid 0_{R_X}, R_W)p(R_W)\right)$$

The key to this example is the joint independence (5); independences of this type arise in hidden variable DAG models. We give an example later where recoverability is based not on an ordinary independence, but on a generalized independence, or Verma constraint [23, 20]. In the following sections, we give a general recursive scheme for solving recoverability problems under MNAR using these types of constraints.

4.2 KNOWN RESULTS FOR MISSINGNESS GRAPHS

Recently [7] and [6] have used missingness graphs to derive conditions for recoverability when data is MNAR. In particular, the following characterization appears in [7] (as theorem 2).

[1] $\perp\!\!\!\perp$ is the independence symbol.

Theorem 3 *For any $p(\mathbb{A}) \in \mathcal{M}(\mathcal{G}^m(\mathbf{V},\mathbf{M}))$, if no elements of $\mathbf{R_M}$ are adjacent in $\mathcal{G}^m(\mathbf{V},\mathbf{M})$, then $p(\mathbf{O} \cup \mathbf{M})$ is recoverable from $p(\mathbf{O},\mathbf{S_M},\mathbf{0_{R_M}})$ if and only if $M \notin \text{pa}_{\mathcal{G}}(R_M)$ for any $M \in \mathbf{M}$. Moreover, $p(\mathbf{O} \cup \mathbf{M})$ is equal to*

$$\frac{p(\mathbf{O},\mathbf{S_M},\mathbf{0_{R_M}})}{\prod\limits_{R_M \in \mathbf{R_M}} p\left(0_{R_M} \middle| \text{pa}_{\mathcal{G}}(R_M) \setminus \mathbf{M}, \mathbf{S}_{\text{pa}_{\mathcal{G}}(R_M) \cap \mathbf{M}}, \mathbf{0}_{\mathbf{R}_{\text{pa}_{\mathcal{G}}(R_M)\cap\mathbf{M}}}\right)}.$$

This result can be generalized in three directions. We may consider cases where variables are unobserved and no missingness mechanism exists. We may consider recoverability of other queries than $p(\mathbf{O} \cup \mathbf{M})$, for instance causal effects or marginal distributions. Finally, we may consider cases where elements of $\mathbf{R_M}$ are adjacent. This case is important because it represents important classes of missingness such as monotonic missingness due to loss to followup. A unit that drops out of a longitudinal study at time t often remains dropped out at times $t+1, \ldots$. In our framework, we would code this by requiring that for all $t' > t$, $R_{M_{t'}} = 1$ if $R_{M_{t'-1}} = 1$, where M_t is unit's status at time t. But this coding is only possible if indicators are allowed to be adjacent in the graph. In addition, allowing indicators to be adjacent allows us to model *non-monotone missing data*, where a unit may be missing at a particular time t, but then becomes observed at a later time $t+k$.

In this paper, we consider the problem of recovering $p(\mathbf{O} \cup \mathbf{M})$ given that every missing variable has an indicator and a proxy (e.g. no completely hidden variables), and that indicators $\mathbf{R_M}$ are allowed to be adjacent. We give a recoverability algorithm that generalizes earlier work in this setting.

5 A GENERAL RECOVERABILITY ALGORITHM

The algorithm, which we call **MID**, work as follows. It tries, for every $R_M \in \mathbf{R_M}$, to recover

$$p(0_{R_M} \mid \text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{R_M}, \mathbf{0}_{\text{pa}_{\mathcal{G}^m}(R_M)\cap\mathbf{R_M}})$$

via a subroutine **DIR**. If every such conditional distribution is recovered, **MID** recovers $p(\mathbf{O} \cup \mathbf{M})$ via (3), otherwise **MID** fails.

The subroutine **DIR** (so named for its resemblance to the way the **ID** algorithm operates when identifying controlled direct effects) has three cases. The first case, which is sufficient for obtaining the soundness part of Theorem 3, attempts to check if indicators for missing parents of R_M are non-parental non-descendants of R_M, in which case recoverability of the conditional distribution for R_M is immediate.

Otherwise, **DIR** uses the other two cases to isolate R_M and its parents into smaller subproblems based on a particular type of ancestral set $\mathbf{A}^\dagger$, or the clan $\mathbf{D}^\dagger$ of R_M. **DIR** is recursive, which means the input must also keep track of a set $\mathbf{W}$ representing variables the clan subproblem ends up depending on.

The situation is somewhat analogous to the way in which the **ID** algorithm attempts to identify controlled direct effects $p(Y \mid \text{do}(\mathbf{v}_{\text{pa}_{\mathcal{G}}(Y)})) = p(Y(\mathbf{v}_{\text{pa}_{\mathcal{G}}(Y)}))$, with three major differences. First, we are attempting identification in a setting where some variables start off being treated as hidden, but in the course of the recursion of **DIR** become observed due to fixing indicators to 0. In **ID** variables are always either hidden or observed and do not change status. Second, since we are only allowed to intervene on indicators, we are attempting to identify

$$p(R_M(\mathbf{0}_{\mathbf{R_M}\cap\text{pa}_{\mathcal{G}}(R_M)}) \mid \{\text{pa}_{\mathcal{G}}(R_M) \setminus \mathbf{R_M}\}(\mathbf{0}_{\mathbf{R_M}\cap\text{pa}_{\mathcal{G}}(R_M)})).$$

Finally, there is not necessarily a fully interventional interpretation for the intermediate objects $p_{\mathbf{W}}(.)$ that arise during the execution of **DIR**, since $\mathbf{W}$ may contain elements outside $\mathbf{R_M}$. This is a necessary consequence of our insistence on not imposing a causal model on $p(\mathbf{M} \cup \mathbf{O})$. Intermediate objects that arise during the execution of **ID** can always be interpreted as interventional distributions.

5.1 SOUNDNESS

MID and its subroutine **DIR** appear below as algorithm 1. In this section, we prove that **MID** is sound.

Corollary 1 implies that if were able to express $p(\{\mathbf{O},\mathbf{S_M}\}(\mathbf{0_{R_M}}))$ as a function of the manifest distribution, we would solve the recoverability problem for $p(\mathbf{O} \cup \mathbf{M})$. If we happen to know

$$p(0_{R_M} \mid \text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{R_M}, \mathbf{0}_{\mathbf{R_M}\cap\text{pa}_{\mathcal{G}^m}(R_M)})$$

for every $R_M \in \mathbf{R_M}$ as a function of the manifest, this would suffice due to the following result.

Lemma 2 *Under $\mathcal{M}(\mathcal{G}^m)$, if for every $R_M \in \mathbf{R_M}$,*

$$p(0_{R_M} \mid \text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{R_M}, \mathbf{0}_{\text{pa}_{\mathcal{G}^m}(R_M)\cap\mathbf{R_M}})$$

is a functional $f_{R_M}(.)$ of $p(\mathbf{O},\mathbf{S_M},\mathbf{0_{R_M}})$, then

$$p(\{\mathbf{O},\mathbf{S_M}\}(\mathbf{0_{R_M}})) = \frac{p(\mathbf{O},\mathbf{S_M},\mathbf{0_{R_M}})}{\prod_{R_M \in \mathbf{R_M}} f_{R_M}(p(\mathbf{O},\mathbf{S_M},\mathbf{0_{R_M}}))}.$$

Proof: $p(\{\mathbf{O},\mathbf{S_M}\}(\mathbf{0_{R_M}})) = \sum_{\mathbf{M}} p(\{\mathbf{A} \setminus \mathbf{R_M}\}(\mathbf{0_{R_M}}))$.

$$p(\{\mathbf{A} \setminus \mathbf{R_M}\}(\mathbf{0_{R_M}})) = \frac{p(\mathbf{A} \setminus \mathbf{R_M}, \mathbf{0_{R_M}})}{\prod_{R_M \in \mathbf{R_M}} f_{R_M}(p(\mathbf{O},\mathbf{S_M},\mathbf{0_{R_M}}))}$$

is implied by (3). But no denominator is a function of $\mathbf{M}$, so we can apply the sum to the numerator first. □ Finding functionals $f_{R_M}(.)$ for every R_M in order to apply Lemma 2 is the job of the subroutine **DIR**.

Algorithm 1 $\mathcal{G}^m(\mathbf{V},\mathbf{M})$ a missingness graph, $p(\mathbf{V})$ a manifest distribution from $p(\mathbb{A}) \in \mathcal{M}(\mathcal{G}^m(\mathbf{V},\mathbf{M}))$, $p_{\mathbf{W}}(\mathbf{V})$ a family of manifest distributions from elements of $p(\mathbb{A}) \in \mathcal{M}(\mathcal{G}^m(\mathbf{V},\mathbf{M}))$, $\prec$ a topological order on $\mathcal{G}^m$.

```
procedure MID($\mathcal{G}^m(\mathbf{V},\mathbf{M}), p(\mathbf{V})$)
    for each $R_M \in \mathbf{R_M}$,
        $\tilde{p}(0_{R_M} \mid \text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{R_M}, \mathbf{0}_{\text{pa}_{\mathcal{G}^m}(R_M)\cap\mathbf{R_M}})$
          $\leftarrow$ DIR($\mathcal{G}^m, p, R_M$)
    if $(\exists R_M \in \mathbf{R_M})$, s.t. DIR($\mathcal{G}^m, p, R_M) = \emptyset$,
            return "cannot recover."
    else return
        $\frac{p(\mathbf{O},\mathbf{S_M},\mathbf{0_{R_M}})}{\prod_{R_M\in\mathbf{R_M}} \tilde{p}(0_{R_M} \mid \text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{R_M}, \mathbf{0}_{\text{pa}_{\mathcal{G}^m}(R_M)\cap\mathbf{R_M}})}$.
end procedure
procedure DIR($\mathcal{G}^m(\mathbf{V},\mathbf{M} \mid \mathbf{W}), p_{\mathbf{W}}(\mathbf{V}), R_M$)
    if $\mathbf{R}_{\mathbf{M}\cap(\text{pa}_{\mathcal{G}^m}(R_M)\setminus\mathbf{W})} \subseteq \text{ndp}_{\mathcal{G}^m}(R_M)$, return
        $p_{\mathbf{W}}\left(0_{R_M} \middle| \begin{array}{c} \text{pa}_{\mathcal{G}^m}(R_M) \setminus (\mathbf{M}\cup\mathbf{W}\cup\mathbf{R_M}) \\ \mathbf{0}_{\mathbf{R}_{\mathbf{M}\cap\text{pa}_{\mathcal{G}^m}(R_M)}}, \mathbf{S}_{\mathbf{M}\cap\text{pa}_{\mathcal{G}^m}(R_M)} \\ \mathbf{0}_{\mathbf{R_M}\cap(\text{pa}_{\mathcal{G}^m}(R_M)\setminus\mathbf{W})} \end{array}\right)$.
    else $\mathbf{A}^\dagger \leftarrow \{R_M\}$.
    while $\left(\text{an}_{\mathcal{G}^m}(\mathbf{A}^\dagger) \cup \mathbf{S}_{\text{an}_{\mathcal{G}^m}(\mathbf{A}^\dagger)\cap\mathbf{M}} \not\subseteq \mathbf{A}^\dagger\right)$ do
        $\mathbf{A}^\dagger \leftarrow \text{an}_{\mathcal{G}^m}(\mathbf{A}^\dagger) \cup \mathbf{S}_{\text{an}_{\mathcal{G}^m}(\mathbf{A}^\dagger)\cap\mathbf{M}}$.
    if $\mathbf{A}^\dagger \subset \mathbf{A}$,
        return DIR($\mathcal{G}^m_{\mathbf{A}^\dagger}, p_{\mathbf{W}\cap\mathbf{A}^\dagger}(\mathbf{V}\cap\mathbf{A}^\dagger), R_M$).
    $\mathbf{D} \leftarrow \text{dis}_{\mathcal{G}^m_{(\mathbf{V})}}(R_M)$, $\mathbf{D}^\dagger \leftarrow \text{cla}_{\mathcal{G}^m}(R_M)$.
    if $\mathbf{D} \subset \mathbf{V}$,
        $\mathbf{Z}^\dagger \leftarrow \text{pa}^s_{\mathcal{G}^m_{(\mathbf{V})}}(\mathbf{D}) \cap \mathbf{R_M}$
        $\mathbf{Y}^\dagger \leftarrow \text{pa}^s_{\mathcal{G}^m_{(\mathbf{V})}}(\mathbf{D}) \setminus \mathbf{R_M}$
        $\mathbf{M}^o_{\mathbf{D}^\dagger} \leftarrow \{M \in (\mathbf{M}\cap\mathbf{D}^\dagger) \mid R_M \in \mathbf{Z}^\dagger\}$
        $\mathbf{M}^h_{\mathbf{D}^\dagger} \leftarrow (\mathbf{M}\cap\mathbf{D}^\dagger) \setminus \mathbf{M}^o_{\mathbf{D}^\dagger}$
        $\mathbf{V}^\dagger \leftarrow \mathbf{D} \cup \mathbf{M}^o_{\mathbf{D}^\dagger}$
        $\tilde{\mathcal{G}}^m \leftarrow \mathcal{G}^m_{\mathbf{D}^\dagger}(\mathbf{V}^\dagger, \mathbf{M}^h_{\mathbf{D}^\dagger} | \mathbf{Y}^\dagger)$
        $p_{\mathbf{Y}^\dagger}(\mathbf{D}) \leftarrow \prod_{V\in\mathbf{D}} p_{\mathbf{W}}\left(V \middle| \begin{array}{c} \text{pre}_{\mathcal{G}^m_{(\mathbf{V})},\prec}(V) \setminus \mathbf{Z}^\dagger, \\ \mathbf{0}_{\text{pre}_{\mathcal{G}^m_{(\mathbf{V})},\prec}(V)\cap\mathbf{Z}^\dagger} \end{array}\right)$
        return DIR($\tilde{\mathcal{G}}^m, p_{\mathbf{Y}^\dagger}(\mathbf{D}), R_M$)
    end if
    return $\emptyset$.
end procedure
```

Soundness of DIR

The subroutine **DIR** invoked by **MID** aims to recover $f_{R_M}(p) = p(0_{R_M} \mid \text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{R_M}, \mathbf{0}_{\text{pa}_{\mathcal{G}^m}(R_M)\cap\mathbf{R_M}})$ by recursively attempting to restrict R_M and $\text{pa}_{\mathcal{G}^m}(R_M)$ to either an appropriate ancestral subset containing these vertices, or an appropriate clan of $\mathcal{G}^m$, and, in the base case, exploiting the independence structure, and properties of the subproblem that is left.

To prove the soundness of **DIR**, we must establish, by induction on algorithm structure, certain results about the subproblems it considers. We will represent subproblems as a pair consisting of a CDAG $\tilde{\mathcal{G}}^m$ that is a subgraph of the original graph $\mathcal{G}^m$, and a *conditional fragment* of the missingness model which can be viewed as a set of all interventional distributions relevant to the subproblem, which also possibly depend on variables $\mathbf{W}$ from larger subproblems.

Given an element $p(\mathbb{A})$ of $\mathcal{M}(\mathcal{G}^m(\mathbf{A}))$, $\mathbf{B} \subseteq \mathbf{A}$, and $\mathbf{W} \subseteq \mathbf{A} \setminus \mathbf{B}$, *a conditional fragment of* $p(\mathbb{A})$ *with respect to* $\mathbf{B}$ *and* $\mathbf{W}$, denoted by $\mathcal{F}_{\mathbf{W},\mathbf{B}}$, is a mapping from elements $\mathbf{w}$ in $\mathfrak{X}_{\mathbf{W}}$ to

$$\mathcal{F}_{\mathbf{w},\mathbf{B}} \equiv \{p_{\mathbf{w}}(\mathbf{B}(\mathbf{r})_{\mathbf{w}}) \mid \mathbf{R} \subseteq \mathbf{R_M} \cap \mathbf{B}, \mathbf{r} \in \mathfrak{X}_{\mathbf{R}}\}.$$

Note that we cannot view $p_{\mathbf{w}}(\mathbf{B}(\mathbf{r})_{\mathbf{w}})$ as a joint response of $\mathbf{B}$ to an intervention setting $\mathbf{R}\cup\mathbf{W}$ to $\mathbf{r}\cup\mathbf{w}$, because $\mathbf{W}$ may contain elements outside $\mathbf{R}$ that we are not allowed to intervene on.

For each call to **DIR**, we want to show that all interventional distributions in the input fragment are Markov with respect to the appropriately modified input graph, that we have enough information in the subproblem to possibly obtain $f_{R_M}(p)$, and that the manifest distribution of the fragment for the current (inner) call can be obtained from the manifest distribution of the fragment for the previous (outer) call.

Definition 2 $\mathcal{F}_{\mathbf{W},\mathbf{B}}$ *is* causal Markov relative to *a CDAG* $\mathcal{G}^m(\mathbf{B} \mid \mathbf{W})$ *if* $(\forall \mathbf{w} \in \mathfrak{X}_{\mathbf{W}}, p_{\mathbf{w}}(\mathbf{B}(\mathbf{r})_{\mathbf{w}}) \in \mathcal{F}_{\mathbf{w},\mathbf{B}})$, $p_{\mathbf{w}}(\mathbf{B}(\mathbf{r})_{\mathbf{w}})$ *is Markov relative to* $\mathcal{G}^m(\mathbf{B} \mid \mathbf{W})_{\underline{\mathbf{R}}}$.

This definition is how we will relate fragments and corresponding subgraphs, and the following two results establish this relationship for the two recursive cases relevant for **DIR**.

Lemma 3 *For* $\mathcal{F}_{\mathbf{W},\mathbf{A}}$ *causal Markov relative to* $\mathcal{G}^m(\mathbf{A} \mid \mathbf{W})$, *let* $\mathbf{D} \in \mathcal{D}(\mathcal{G}^m_{(\mathbf{V})})$, $\mathbf{D}^\dagger \equiv \text{cla}_{\mathcal{G}^m}(\mathbf{D})$, $\mathbf{W}^\dagger \equiv \text{pa}^s_{\mathcal{G}^m_{(\mathbf{V})}}(\mathbf{D})$, $\mathbf{W}^* \equiv \mathbf{W}^\dagger \setminus \mathbf{W}$. *Then for any* $\mathbf{w}^\dagger \in \mathfrak{X}_{\mathbf{W}^\dagger}$, $\mathcal{F}_{\mathbf{w}^\dagger,\mathbf{D}^\dagger} \equiv \left\{\tilde{p}_{\mathbf{w}^\dagger}(\mathbf{D}^\dagger(\mathbf{r})_{\mathbf{w}^\dagger}) \middle| \mathbf{r} \in \mathfrak{X}_{\mathbf{R}}, \mathbf{R} \subseteq \mathbf{R_M} \cap \mathbf{D}\right\}$ *is causal Markov relative to* $\mathcal{G}^m_{\text{fa}_{\mathcal{G}^m}(\mathbf{D}^\dagger)}(\mathbf{D}^\dagger \mid \mathbf{W}^\dagger)$, *where for any* $\mathbf{w}$ *consistent with* $\mathbf{w}^\dagger$, $\tilde{p}_{\mathbf{w}^\dagger}(\mathbf{D}^\dagger(\mathbf{r})_{\mathbf{w}^\dagger})$ *is*

$$\prod_{V\in\mathbf{D}^\dagger} p_{\mathbf{w}}(V|(\mathbf{r}\cup\mathbf{w}^\dagger)_{\text{pa}_{\mathcal{G}^m}(V)\cap(\mathbf{R}\cup\mathbf{W}^*)}, \text{pa}_{\mathcal{G}^m}(V) \setminus (\mathbf{R}\cup\mathbf{W}^\dagger))$$

Proof: For any CDAG $\mathcal{G}(\mathbf{O}, \mathbf{M} \mid \mathbf{W})$, $\text{fa}_{\mathcal{G}}(\text{cla}_{\mathcal{G}}(\mathbf{D}))$ is equal to $\text{cla}_{\mathcal{G}}(\mathbf{D}) \cup \text{pa}^{s}_{\mathcal{G}_{(\mathbf{O})}}(\mathbf{D})$ for any $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{(\mathbf{O})})$. The proof is now immediate. Elements $\tilde{p}_{\mathbf{w}^\dagger}(\mathbf{D}^\dagger(\mathbf{r})_{\mathbf{w}^\dagger})$ of each $\mathcal{F}_{\mathbf{w}^\dagger,\mathbf{D}^\dagger}$ are Markov relative to $(\mathcal{G}^m_{\text{fa}_{\mathcal{G}^m}(\mathbf{D}^\dagger)})_{\underline{\mathbf{R}}}$ by construction. The definition of $\tilde{p}_{\mathbf{w}^\dagger}(\mathbf{D}^\dagger(\mathbf{r})_{\mathbf{w}^\dagger})$ implies it is the same object for any $\mathbf{w}$ consistent with $\mathbf{w}^\dagger$. □

Lemma 4 *For $\mathcal{F}_{\mathbf{W},\mathbf{A}}$ causal Markov relative to $\mathcal{G}^m(\mathbf{A} \mid \mathbf{W})$, let $\mathbf{V}^\dagger \subseteq \mathbf{A} \cup \mathbf{W}$ be ancestral, $\mathbf{W}^\dagger \equiv \mathbf{W} \cap \mathbf{V}^\dagger$, $\mathbf{A}^\dagger \equiv \mathbf{A} \cap \mathbf{V}^\dagger$. Then for any $\mathbf{w}^\dagger \in \mathfrak{X}_{\mathbf{W}^\dagger}$, $\mathcal{F}_{\mathbf{w}^\dagger,\mathbf{D}^\dagger} \equiv \left\{\tilde{p}_{\mathbf{w}^\dagger}(\mathbf{A}^\dagger(\mathbf{r})_{\mathbf{w}^\dagger})\middle|\mathbf{r} \in \mathfrak{X}_{\mathbf{R}}, \mathbf{R} \subseteq \mathbf{R}_{\mathbf{M}} \cap \mathbf{A}^\dagger\right\}$ is causal Markov relative to $\mathcal{G}^m_{\mathbf{A}^\dagger}(\mathbf{A}^\dagger \mid \mathbf{W}^\dagger)$, where for any $\mathbf{w}$ consistent with $\mathbf{w}^\dagger$, $\tilde{p}_{\mathbf{w}^\dagger}(\mathbf{A}^\dagger(\mathbf{r})_{\mathbf{w}^\dagger})$ is*

$$\prod_{V \in \mathbf{A}^\dagger} p_{\mathbf{w}}(V \mid \mathbf{r}_{\text{pa}_{\mathcal{G}^m}(V) \cap \mathbf{R}}, \text{pa}_{\mathcal{G}^m}(V) \setminus (\mathbf{R} \cup \mathbf{W}^\dagger))$$

Proof: Immediate. Elements $p_{\mathbf{w}^\dagger}(\mathbf{A}^\dagger(\mathbf{r})_{\mathbf{w}^\dagger})$ of each $\mathcal{F}_{\mathbf{w}^\dagger,\mathbf{A}^\dagger}$ are Markov relative to $(\mathcal{G}^m_{\mathbf{A}^\dagger})_{\underline{\mathbf{R}}}$ by construction. The definition of $\tilde{p}_{\mathbf{w}^\dagger}(\mathbf{D}^\dagger(\mathbf{r})_{\mathbf{w}^\dagger})$ implies it is the same object for any $\mathbf{w}$ consistent with $\mathbf{w}^\dagger$. □

The next two results re-express $p(R_M \mid \text{pa}_{\mathcal{G}^m}(R_M))$ from a function of the larger fragment of the outer recursive call to a function of the smaller fragment of the inner call.

Lemma 5 *Assume $\mathcal{F}_{\mathbf{W},\mathbf{A}}$ is causal Markov relative to $\mathcal{G}^m(\mathbf{A} \mid \mathbf{W})$, and $\mathcal{F}_{\mathbf{W}^\dagger,\mathbf{D}^\dagger}$ is defined as in Lemma 3. Then for any $R_M \in \mathbf{D}^\dagger$, $p_{\mathbf{W}}(R_M \mid \text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{W})$ is equal to $p_{\mathbf{W}^\dagger}(R_M \mid \text{pa}_{\mathcal{G}^m_{\text{fa}_{\mathcal{G}^m}(\mathbf{D}^\dagger)}}(R_M) \setminus \mathbf{W}^\dagger)$.*

Proof: Since $R_M \in \mathbf{D}^\dagger$, this follows by Lemma 3. That is, $p_{\mathbf{W}^\dagger}(R_M \mid \text{pa}_{\mathcal{G}^m_{\text{fa}_{\mathcal{G}^m}(\mathbf{D}^\dagger)}}(R_M) \setminus \mathbf{W}^\dagger)$ is equal to $p_{\mathbf{W}}(R_M \mid \mathbf{W}^\dagger_{\text{pa}_{\mathcal{G}^m}(R_M) \cap (\mathbf{W}^\dagger \setminus \mathbf{W})}, \text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{W}^\dagger)$, which is equal to $p_{\mathbf{W}}(R_M \mid \text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{W})$. □

Lemma 6 *Assume $\mathcal{F}_{\mathbf{W},\mathbf{A}}$ is causal Markov relative to $\mathcal{G}^m(\mathbf{A} \mid \mathbf{W})$, and $\mathcal{F}_{\mathbf{W}^\dagger,\mathbf{A}^\dagger}$ is defined as in Lemma 4. Then for any $R_M \in \mathbf{A}^\dagger$, $p_{\mathbf{W}}(R_M \mid \text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{W})$ is equal to $p_{\mathbf{W}^\dagger}(R_M \mid \text{pa}_{\mathcal{G}^m_{\mathbf{A}^\dagger}}(R_M) \setminus \mathbf{W}^\dagger)$.*

Proof: Since $R_M \in \mathbf{A}^\dagger$, this follows by Lemma 4. That is, $p_{\mathbf{W}^\dagger}(R_M \mid \text{pa}_{\mathcal{G}^m_{\mathbf{A}^\dagger}}(R_M) \setminus \mathbf{W}^\dagger)$ is equal to $p_{\mathbf{W}}(R_M \mid \text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{W})$. □

The next two results express the analogue of the manifest distribution of the smaller fragment as a function of the manifest distribution of the larger fragment. We assume $\mathbf{M}^o_{\mathbf{D}^\dagger}, \mathbf{M}^h_{\mathbf{D}^\dagger}, \mathbf{V}^\dagger, \mathbf{Z}^\dagger, \mathbf{Y}^\dagger$, and $\tilde{\mathcal{G}}^m$ are defined as in the district case of **DIR**. Let $\mathbf{W}^\dagger = \mathbf{Y}^\dagger \cup \mathbf{Z}^\dagger$, and $\mathbf{O}^\dagger = \mathbf{D} \cap \mathbf{O}$.

Lemma 7 *Assume $\mathcal{F}_{\mathbf{W},\mathbf{A}}$ is causal Markov relative to $\mathcal{G}^m(\mathbf{A} \mid \mathbf{W})$, and $\mathcal{F}_{\mathbf{W}^\dagger,\mathbf{D}^\dagger}$ is defined as in Lemma 3. Then the marginal $p_{\mathbf{Y}^\dagger,\mathbf{0}_{\mathbf{Z}^\dagger}}(\mathbf{O}^\dagger, \mathbf{M}^o_{\mathbf{D}^\dagger}, \mathbf{S}_{\mathbf{M}^h_{\mathbf{D}^\dagger}}, \mathbf{R}_{\mathbf{M}^h_{\mathbf{D}^\dagger}})$ of $p_{\mathbf{W}^\dagger}(\mathbf{D}^\dagger) \in \mathcal{F}_{\mathbf{W}^\dagger,\mathbf{D}^\dagger}$ is equal to $\prod_{V \in \mathbf{V}^\dagger} p_{\mathbf{W}}(V \mid \text{pre}_{\mathcal{G}^m_{(\mathbf{V})},\prec}(V) \setminus \mathbf{Z}^\dagger, \mathbf{0}_{\text{pre}_{\mathcal{G}^m_{(\mathbf{V})},\prec}(V) \cap \mathbf{Z}^\dagger})$.*

Proof: Fix $\mathbf{w}$ and $\mathbf{w}^\dagger$ consistent with $\mathbf{w}$, such that $\mathbf{w}^\dagger_{\mathbf{Z}^\dagger} = \mathbf{0}$. We get the following set of equalities, where the first is by assumption on missingness models, the second by (4), (3) and the definition of $\mathbf{M}^o_{\mathbf{D}^\dagger}$, the third by definition, the fourth by Lemma 3, and the last by standard results on district factorization of hidden variable DAG models found in [22]. If we range over all possible $\mathbf{w}^\dagger_{\mathbf{Y}^\dagger}$, the last expression reduces to $\prod_{V \in \mathbf{V}^\dagger} p_{\mathbf{W}}(V | \text{pre}_{\mathcal{G}^m_{(\mathbf{V})},\prec}(V) \setminus \mathbf{Z}^\dagger, \mathbf{0}_{\text{pre}_{\mathcal{G}^m_{(\mathbf{V})},\prec}(V) \cap \mathbf{Z}^\dagger})$.

$$\begin{aligned}
&p_{\mathbf{w}^\dagger}(\mathbf{O}^\dagger, \mathbf{M}^o_{\mathbf{D}^\dagger}, \mathbf{S}_{\mathbf{M}^h_{\mathbf{D}^\dagger}}, \mathbf{R}_{\mathbf{M}^h_{\mathbf{D}^\dagger}}) \\
&= p_{\mathbf{w}^\dagger}(\mathbf{O}^\dagger, \mathbf{S}_{\mathbf{M}^o_{\mathbf{D}^\dagger}}(\mathbf{R}_{\mathbf{M}^o_{\mathbf{D}^\dagger}} = \mathbf{0}), \mathbf{S}_{\mathbf{M}^h_{\mathbf{D}^\dagger}}, \mathbf{R}_{\mathbf{M}^h_{\mathbf{D}^\dagger}}) \\
&= p_{\mathbf{w}^\dagger}(\mathbf{O}^\dagger, \mathbf{S}_{\mathbf{M}^o_{\mathbf{D}^\dagger}}, \mathbf{S}_{\mathbf{M}^h_{\mathbf{D}^\dagger}}, \mathbf{R}_{\mathbf{M}^h_{\mathbf{D}^\dagger}}) \\
&= \sum_{\mathbf{M}^h_{\mathbf{D}^\dagger}} p_{\mathbf{w}^\dagger}(\mathbf{O}^\dagger, \mathbf{S}_{\mathbf{M}^o_{\mathbf{D}^\dagger}}, \mathbf{S}_{\mathbf{M}^h_{\mathbf{D}^\dagger}}, \mathbf{R}_{\mathbf{M}^h_{\mathbf{D}^\dagger}}, \mathbf{M}^h_{\mathbf{D}^\dagger}) \\
&= \sum_{\mathbf{M}^h_{\mathbf{D}^\dagger}} \prod_{V \in \mathbf{D}^\dagger} p_{\mathbf{w}}(V \mid \mathbf{w}^\dagger_{\text{pa}_{\mathcal{G}^m}(V) \cap (\mathbf{W}^\dagger \setminus \mathbf{W})}, \text{pa}_{\mathcal{G}^m}(V) \setminus \mathbf{W}^\dagger) \\
&= \prod_{V \in \mathbf{D}} p_{\mathbf{w}}(V | \text{pre}_{\mathcal{G}^m_{(\mathbf{V})},\prec}(V) \setminus \mathbf{W}^\dagger, \mathbf{w}^\dagger_{\text{pre}_{\mathcal{G}^m_{(\mathbf{V})},\prec}(V) \cap \mathbf{W}^\dagger})
\end{aligned}$$

□

Lemma 8 *Assume $\mathcal{F}_{\mathbf{W},\mathbf{A}}$ is causal Markov relative to $\mathcal{G}^m(\mathbf{A} \mid \mathbf{W})$, and $\mathcal{F}_{\mathbf{W}^\dagger,\mathbf{A}^\dagger}$ is defined as in Lemma 4. Then the element $p_{\mathbf{W}^\dagger}(\mathbf{V} \cap \mathbf{A}^\dagger)$ of $\mathcal{F}_{\mathbf{W}^\dagger,\mathbf{A}^\dagger}$ is equal to $\sum_{\mathbf{V} \setminus \mathbf{A}^\dagger} p_{\mathbf{W}}(\mathbf{V})$.*

Proof: $p_{\mathbf{W}^\dagger}(\mathbf{V} \cap \mathbf{A}^\dagger)$ is equal to $\sum_{\mathbf{A}^\dagger \setminus \mathbf{V}} p_{\mathbf{W}^\dagger}(\mathbf{A}^\dagger)$ (by definition), which is equal to $\sum_{\mathbf{A}^\dagger \setminus \mathbf{V}} \sum_{\mathbf{A} \setminus \mathbf{A}^\dagger} p_{\mathbf{W}}(\mathbf{A})$ by Lemma 4. But since both $\mathbf{V}, \mathbf{A}^\dagger$ are subsets of $\mathbf{A}$, this is just $\sum_{\mathbf{A} \setminus (\mathbf{V} \cap \mathbf{A}^\dagger)} p_{\mathbf{W}}(\mathbf{A})$, which is equal to $\sum_{\mathbf{V} \setminus \mathbf{A}^\dagger} \sum_{\mathbf{A} \setminus \mathbf{V}} p_{\mathbf{W}}(\mathbf{A}) = \sum_{\mathbf{V} \setminus \mathbf{A}^\dagger} p_{\mathbf{W}}(\mathbf{V})$. □

The following result establishes the validity of the base case of **DIR**, where $p_{\mathbf{W}}(R_M \mid \text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{W})$ is expressed in terms of the manifest distribution for the current fragment.

Lemma 9 *Assume $\mathcal{F}_{\mathbf{W},\mathbf{A}}$ is causal Markov relative to $\mathcal{G}^m(\mathbf{A} \mid \mathbf{W})$. Then if $\mathbf{R}_{\mathbf{M} \cap (\text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{W})} \subseteq \text{ndp}_{\mathcal{G}^m}(R_M)$, then*

$$p_{\mathbf{W}}(0_{R_M} \mid \text{pa}_{\mathcal{G}^m}(R_M) \setminus (\mathbf{W} \cup \mathbf{R}_{\mathbf{M}}), \mathbf{0}_{\mathbf{R}_{\mathbf{M}} \cap (\text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{W})})$$

is equal to $$p_{\mathbf{W}}\left(0_{R_M} \,\middle|\, \begin{array}{c} \text{pa}_{\mathcal{G}^m}(R_M) \setminus (\mathbf{M} \cup \mathbf{W} \cup \mathbf{R}_{\mathbf{M}}) \\ \mathbf{0}_{\mathbf{R}_{\mathbf{M} \cap \text{pa}_{\mathcal{G}^m}(R_M)}}, \mathbf{S}_{\mathbf{M} \cap \text{pa}_{\mathcal{G}^m}(R_M)} \\ \mathbf{0}_{\mathbf{R}_{\mathbf{M}} \cap (\text{pa}_{\mathcal{G}^m}(R_M) \setminus \mathbf{W})} \end{array}\right).$$

Proof: We get the following set of equalities, where the first follows by assumption, and the fact that $p_{\mathbf{W}}(\mathbf{A})$ is Markov relative to $\mathcal{G}^m$, the second is by the properties of the missingness model, and the third is by (4):

$$p_{\mathbf{W}}\left(0_{R_M}\middle|\begin{array}{c}\text{pa}_{\mathcal{G}^m}(R_M)\setminus(\mathbf{W}\cup\mathbf{R_M})\\ \mathbf{0}_{\mathbf{R_M}\cap(\text{pa}_{\mathcal{G}^m}(R_M)\setminus\mathbf{W})}\end{array}\right)=$$

$$p_{\mathbf{w}}\left(0_{R_M}\middle|\begin{array}{c}\text{pa}_{\mathcal{G}^m}(R_M)\setminus(\mathbf{M}\cup\mathbf{W}\cup\mathbf{R_M})\\ \text{pa}_{\mathcal{G}^m}(R_M)\cap\mathbf{M},\mathbf{0}_{\mathbf{R}_{\mathbf{M}\cap\text{pa}_{\mathcal{G}^m}(R_M)}}\\ \mathbf{0}_{\mathbf{R_M}\cap(\text{pa}_{\mathcal{G}^m}(R_M)\setminus\mathbf{W})}\end{array}\right)=$$

$$p_{\mathbf{W}}\left(0_{R_M}\middle|\begin{array}{c}\text{pa}_{\mathcal{G}^m}(R_M)\setminus(\mathbf{M}\cup\mathbf{W}\cup\mathbf{R_M}),\\ \mathbf{S}_{\mathbf{M}\cap\text{pa}_{\mathcal{G}^m}(R_M)}(\mathbf{0}_{\mathbf{R}_{\mathbf{M}\cap\text{pa}_{\mathcal{G}^m}(R_M)}})\\ \mathbf{0}_{\mathbf{R}_{\mathbf{M}\cap\text{pa}_{\mathcal{G}^m}(R_M)}},\mathbf{0}_{\mathbf{R_M}\cap(\text{pa}_{\mathcal{G}^m}(R_M)\setminus\mathbf{W})}\end{array}\right)=$$

$$p_{\mathbf{W}}\left(0_{R_M}\middle|\begin{array}{c}\text{pa}_{\mathcal{G}^m}(R_M)\setminus(\mathbf{M}\cup\mathbf{W}\cup\mathbf{R_M})\\ \mathbf{0}_{\mathbf{R}_{\mathbf{M}\cap\text{pa}_{\mathcal{G}^m}(R_M)}},\mathbf{S}_{\mathbf{M}\cap\text{pa}_{\mathcal{G}^m}(R_M)}\\ \mathbf{0}_{\mathbf{R_M}\cap(\text{pa}_{\mathcal{G}^m}(R_M)\setminus\mathbf{W})}\end{array}\right).$$

□

Before putting all these results together to show soundness of $\mathbf{DIR}$, we must prove one additional utility lemma that shows the set $\mathbf{A}^\dagger$ constructed by $\mathbf{DIR}$ is ancestral.

Define an automorphism from vertex sets in $\mathcal{G}^m$, $\rho_{M,\mathcal{G}^m}(\mathbf{B})$, as $\text{an}_{\mathcal{G}^m}(\{R_M\}\cup\mathbf{B})\cup\mathbf{S}_{\text{an}_{\mathcal{G}^m}(\{R_M\}\cup\mathbf{B})}$. Let $\mathbf{A}^\dagger$ be the fixed point of $\rho_{M,\mathcal{G}^m}$ with the starting input of the empty set.

Lemma 10 $\mathbf{A}^\dagger$ *is an ancestral set in* $\mathcal{G}^m$.

Proof: A simple proof by contradiction follows by definition of $\rho_{M,\mathcal{G}^m}$. □

We now show the main result of this paper.

Theorem 4 **MID** *is sound.*

Proof: Assuming $\mathbf{DIR}$ returns the answer for every $R_M\in\mathbf{R_M}$, Corollary 1, and Lemma 2 ensure that $\mathbf{MID}$ recovers $p(\mathbf{M}\cup\mathbf{O})$ from $p(\mathbf{V})$.

The soundness of $\mathbf{DIR}$ follows by induction on the recursive call structure. The inductive hypothesis is that the input conditional fragment $\mathcal{F}_{\widetilde{\mathbf{W}},\widetilde{\mathbf{A}}}$ is causal Markov relative to the appropriate graph derived from the input graph $\tilde{\mathcal{G}}^m$, that the input manifest $\tilde{p}_{\widetilde{\mathbf{W}}}(\mathbf{V})$ is the function of the original manifest $p(\mathbf{V})$, and that $\tilde{p}_{\widetilde{\mathbf{W}}}(R_M\mid\text{pa}_{\tilde{\mathcal{G}}^m}(R_M)\setminus\widetilde{\mathbf{W}})=p(R_M\mid\text{pa}_{\mathcal{G}^m}(R_M))$.

The base case trivially holds for the original inputs to $\mathbf{DIR}$. If the inductive hypothesis is true, and $\mathbf{DIR}$ returns after the first conditional, soundness follows by Lemma 9.

If $\mathbf{DIR}$ returns after the second conditional, then Lemma 10 ensures the constructed set $\mathbf{A}^\dagger$ is ancestral, and the induction for the following recursive call is maintained via Lemmas 4, 8 and 6.

If $\mathbf{DIR}$ returns after the third conditional, Lemma 3 ensures $\mathcal{F}_{\widetilde{\mathbf{w}},\mathbf{D}^\dagger}$ is causal Markov relative to $\mathcal{G}^m_{\text{fa}_{\mathcal{G}^m}(\mathbf{D}^\dagger)}$ for all values of $\widetilde{\mathbf{w}}$, including those that set $\mathbf{Z}^\dagger$ to $\mathbf{0}$. Lemma 5, and the inductive hypothesis ensures $p_{\mathbf{W}^\dagger}(R_M\mid\text{pa}_{\mathcal{G}^m_{\text{fa}_{\mathcal{G}^m}(\mathbf{D}^\dagger)}}(R_M)\setminus\mathbf{W}^\dagger)$ is equal to $p_{\mathbf{W}}(R_M\mid\text{pa}_{\mathcal{G}^m}(R_M)\setminus\mathbf{W})$. Finally, Lemma 7 ensures the manifest for the recursive call is a function of the input manifest. In fact, because we set $\mathbf{Z}^\dagger$ to $\mathbf{0}$, properties of missingness models ensure we can treat $\mathbf{M}^o_{\mathbf{D}^\dagger}$ as observed in subsequent recursive calls, which means we no longer need to consider $\mathbf{S}_{\mathbf{M}^o_{\mathbf{D}^\dagger}}$.

Since induction follows for all cases, so does our conclusion. □

6 A COMPLEX RECOVERABLE EXAMPLE

We now work through an example where all cases of **MID** and **DIR** are necessary. Consider the graph shown in Fig. 2 (a). Here C and D are shown in green to indicate that they are fully observed. This is a more complex version of the example in Fig. 1 (c). Unlike that case, here, there are no conditional independences that hold between proxies and indicators. However, if we were to divide by $p(D\mid C)$ and sum out C, in the resulting distribution $p_D(S_A,S_B,R_A,R_B,A,B)$, for any fixed value d of D, we would have

$$(\{S_A(0_{R_A}),R_B\}\perp\!\!\!\perp\{S_B(0_{R_B}),R_A(0_{R_B})\})_{p_d}$$

This is a type of Verma constraint [23] or generalized independence constraint [20].

Our goal is to recover $p(A,B,C,D)$ given the missingness model corresponding to this graph, and in particular the above constraint. We must recover $p(0_{R_B}\mid A,D)$ and $p(0_{R_A}\mid 0_{R_B},B,D)$ from $p(R_A,R_B,S_A,S_B,C,D)$. In either case, we note that D is not an element of $\text{cla}_{\mathcal{G}^m}(R_A)=\text{cla}_{\mathcal{G}^m}(R_B)$, which implies we can use the clan case of **DIR** and consider a subproblem shown in Fig. 2 (b), with the corresponding manifest $\tilde{p}_D(R_A,R_B,S_A,S_B,C)=p(S_A,S_B,R_A,R_B\mid D,C)p(C)$. In the new subproblem (for either R_A or R_B), C is not a part of the ancestral set $\mathbf{A}^\dagger$ constructed by **DIR** in the ancestral case, so we consider a new subproblem shown in Fig. 2 (c), with the corresponding manifest $\tilde{p}_D(R_A,R_B,S_A,S_B)=\sum_c\tilde{p}_D(S_A,S_B,R_A,R_B\mid D,c)\tilde{p}_D(c)$. This new subproblem now resembles the example in Fig. 1 (c), and is solved similarly. In particular, we recover $p(0_{R_B}\mid D,A)$ as

$$\frac{\tilde{p}_D(S_A\mid 0_{R_A},0_{R_B})\tilde{p}_D(0_{R_B})}{\sum_{R_B}\tilde{p}_D(S_A\mid 0_{R_A},R_B)\tilde{p}_D(R_B)}$$

and $p(0_{R_A}\mid 0_{R_B},B,D)$ as $\tilde{p}_D(0_{R_A}\mid 0_{R_B},S_B)$. We then obtain $p(A,B,C,D)$ by dividing the manifest distribution for observed cases $p(0_{R_A},0_{R_B},S_A,S_B,C,D)$ by the above two probabilities.

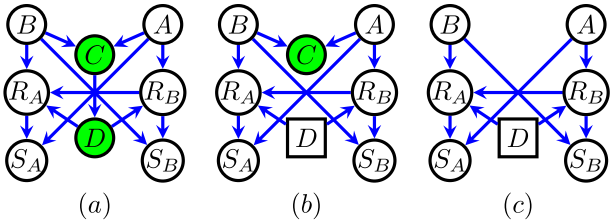

Figure 2: (a) An example where recoverability is possible via **MID**. (b),(c) Graphs corresponding to subproblems considered by **MID** in recovering $p(A, B, C, D)$.

7 NONRECOVERABILITY

The generality of **MID** naturally raises the question of whether it is complete, that is whether whenever it outputs "**cannot recover**" then it is possible to construct two elements of the missingness model that agree on the manifest but disagree on the underlying joint distribution. We leave this difficult question aside in this paper in the interests of space, but note that an approach similar to one used to show completeness for causal effects identification [18] seems promising. That is, use **MID** as a guide for constructing a "zoo" of structures where recoverability does not seem to be possible, and then construct a general method for showing non-recoverability for this "zoo."

Some results on non-recoverability do exist. For example, it can be shown that $p(A)$ is not recoverable in the missingness model with the graph in Fig. 3 (a) [7], and similarly that $p(A, B)$ is not recoverable in the missingness model with the graph in Fig. 3 (b). Characterization of non-recoverability is an open problem.

8 DISCUSSION AND CONCLUSIONS

We have represented missing data as a type of a restricted causal inference problem. Using the machinery of graphical causal models, we have given a general algorithm for recoverability of a joint distribution in MNAR settings. Though we do not require this, our formalism allows the joint distribution we recover to come from a statistical, rather than a causal model – all causal assumptions may be restricted to the missingness model governing the behavior of proxies of missing variables under interventions on indicators. We show that the MCAR, MAR, MNAR taxonomy is not sufficiently granular to classify cases where recoverability is possible. In particular, there are MNAR examples where constraints akin to Verma constraints permit recoverability.

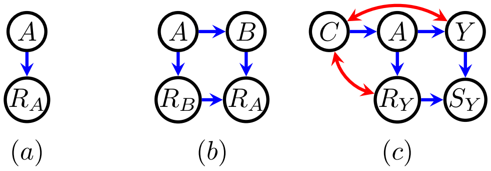

Figure 3: (a) $p(A)$ is not recoverable. (b) $p(A, B)$ is not recoverable. (b) A graph with hidden variables where $p(Y)$ is not recoverable, but $p(Y(a))$ is.

Aside from the algorithm, our formalism allows us to seamlessly integrate issues of identification of causal effects, and recoverability. For instance, it is known that in the graph shown in Fig. 3 (c) (where we treat ↔ edges as indicating the presence of an unobserved parent), $p(Y)$ is not recoverable. However, if the graph on C, A, Y represents a causal model, we can show that $p(Y(a))$ is recoverable. In particular

$$p(Y(a)) = p(S_Y(a, 0_{R_Y})) = \frac{\sum_c p(S_Y, 0_{R_Y} \mid a, c)p(c)}{\sum_c p(0_{R_Y} \mid a, c)p(c)}$$

A similar observation appears in [6], example 3.

By explicitly representing missingness via an intervenable indicator, and a proxy as a response to this intervention, our formalism allows us to reason explicitly about the interpretation of censoring by death using the existing language of interventions. That is if S_X is observed patient history, and 1_{R_X} implies it is missing due to the patient dying, then we may either disallow considering $S_X(0_{R_X})$ (e.g. "resurrecting the patient") for that patient, allow $S_X(0_{R_X})$, but treat it as making statements about exchangeable but different patients who happened to be alive that transfer over to the dead patient in a hypothetical alternative history where the patient never died, and so on.

Note that if we assume a *known* relationship $p(S_M(r_{R_M}) \mid M)$ between M and $S_M(r_{R_M})$ other than direct equality, we can use the approach in this paper to address certain *coarsening* [3] and *measurement error* settings. We do not consider these extensions explicitly here for space reasons, but they are straightforward.

Acknowledgements

This research was supported in parts by grants from NIH R01 AI104459-01A1, NSF #IIS-1302448 and ONR #N00014-10-1-0933 and #N00014-13-1-0153.

References

[1] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[2] Constantine E. Frangakis, Donald B. Rubin, Ming-Wen An, and Ellen MacKenzie. Principal stratification designs to estimate input data missing due to death. *Biometrics*, 63:641–662, 2007.

[3] Daniel F. Heitjan and Donald Rubin. Ignorability and coarse data. *Annals of Statistics*, 19(4):2244–2253, 1991.

[4] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, 1986.

[5] E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[6] Karthika Mohan and Judea Pearl. Graphical models for recovering probabilistic and causal queries from missing data. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1520–1528. Curran Associates, Inc., 2014.

[7] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1277–1285. Curran Associates, Inc., 2013.

[8] J. Neyman. Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. excerpts reprinted (1990) in English. *Statistical Science*, 5:463–472, 1923.

[9] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufmann, San Mateo, 1988.

[10] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[11] Thomas S. Richardson and Jamie M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *preprint:* `http://www.csss.washington.edu/Papers/wp128.pdf`, 2013.

[12] J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

[13] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[14] D. B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[15] D. B. Rubin. Inference and missing data (with discussion). *Biometrika*, 63:581–592, 1976.

[16] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley & Sons, 1987.

[17] Stuart Russell, John Binder, Daphne Koller, and Keiji Kanazawa. Local learning in probabilistic networks with hidden variables. In *Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI-95)*, pages 1146–1152. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1995.

[18] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *National Conference on Artificial Intelligence*, volume 21. AUAI Press, 2006.

[19] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979, 2008.

[20] Ilya Shpitser, Thomas S. Richardson, and James M. Robins. An efficient algorithm for computing interventional distributions in latent variable causal models. In *Uncertainty in Artificial Intelligence*, volume 27. AUAI Press, 2011.

[21] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer Verlag, New York, 1993.

[22] Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Uncertainty in Artificial Intelligence*, volume 18, pages 519–527. AUAI Press, 2002.

[23] T. S. Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University of California, Los Angeles, 1990.