Missing data and cluster graphs: cluster-level missingness vs variable-level missingness

Willow J. Scott
Sorbonne Université, INSERM,
Institut Pierre Louis d'Epidémiologie
et de Santé Publique,
F75012, Paris, France
willow.scott@inserm.fr

Eugenio Valdano
Sorbonne Université, INSERM,
Institut Pierre Louis d'Epidémiologie
et de Santé Publique,
F75012, Paris, France
eugenio.valdano@inserm.fr

Charles K. Assaad
Sorbonne Université, INSERM,
Institut Pierre Louis d'Epidémiologie
et de Santé Publique,
F75012, Paris, France
charles.assaad@inserm.fr

Abstract

Missing data is pervasive in many scientific domains such as public health, environmental science, and the social sciences. Recoverability from missing data is typically studied using fully specified variable-level missingness models despite that, in many applications, only coarse structural information is available, for instance when variables are grouped into clusters due to limited knowledge or interpretability reasons. In this paper, we investigate recoverability from such abstract representations. We introduce two classes of cluster-based missingness graphs: the m-C-DMG, which retains variable-specific missingness indicators, and the cm-C-DMG, which aggregates missingness mechanisms at the cluster level. We formalize the notion of compatibility between these abstract graphs and underlying variable-level missingness models, and study how this abstraction affects the recoverability of probabilistic and causal queries. In particular, we give graphical conditions of recovering the joint distribution as well as graphical conditions of recovering a macro causal effect.

1 Introduction

Missing data are ubiquitous in many scientific domains, including public health, environmental science, and the social sciences. Understanding when meaningful quantities can be recovered from incomplete data is therefore a fundamental problem. Existing methods effectively approximate the true distribution of incomplete data when missingness occurs randomly or can be explained by observed values. However, such methods produce biased results when data is Missing Not At Random (MNAR), which is the reality for most datasets in fields like health and epidemiology [Rubin, 1976, Peek and Rodrigues, 2018]. Methods for imputation with MNAR missingness include autoencoders [Pereira et al., 2024], which provide little theoretical guarantees, and pattern-mixture models [Little, 1993], which rely on semi-parametric or parametric assumptions. This work focuses on a non-parametric alternative for dealing with MNAR data. In such settings, graphical models [Pearl, 2009] provide a powerful framework to characterize recoverability of joint, conditional, and causal quantities. However, classical graph-based missing data theory addresses missing data at the level of

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

fully specified variable-level models, where the missingness mechanism of each variable is explicitly represented [Mohan et al., 2013, Mohan and Pearl, 2014, 2021].

In many modern applications, however, the underlying data-generating mechanisms are only partially known or are intentionally abstracted [Peters et al., 2013, Inoue et al., 2022, Assaad et al., 2023, Anand and Hripcsak, 2025]. Variables are thus often grouped into clusters, either for interpretability or because only aggregated structural information is available. For example, clusters are necessary to deal with uncertainty around the relationships between sexual practices and use of prevention in studies on predictors of risk of acquiring Human Immunodeficiency Virus (HIV). In a conceptual graph, sexual behavior is treated as a cluster of measures such as condom use and number of sexual partners. Table 1 exhibits missingness in a synthetic dataset of clusters representing Sexual Behavior, Mental Health, and HIV Status. As is the case in most real-world data, missingness may occur in any of the variables and for a variety of reasons. One must thus reason about recoverability from abstract representations of missingness mechanisms rather than from fully specified variable-level graphs.

In this paper, we introduce and study two such abstractions: the *m-C-DMG* and the *cm-C-DMG*. The m-C-DMG retains variable-specific missingness indicators while representing causal relations between variables at the cluster level. The cm-C-DMG is a coarser abstraction that merges missingness indicators within each cluster into a single vertex. While both representations encode missingness mechanisms at a higher level of abstraction, they differ in the amount of information they preserve, raising the question of how this impacts recoverability.

In this paper, we formalize the compatibility between abstract graphs and variable-level missingness models and study when recoverability can be inferred from these abstractions. We provide results for the recovery of both joint distributions and a specific type of causal effects, known as macro causal effects, depending on the level of abstraction. Overall, this work contributes to the theoretical understanding of missing data problems under structural abstraction.

The remainder of the paper is organized as follows. Section 2 presents the background and preliminaries. Section 3 introduces and compares the two proposed abstractions. Section 4 provides conditions for recovering joint distributions from these abstractions. Section 5 gives conditions for recovering macro causal effects. Section 6 uses simulated data to exemplify the recovery results from the proposed abstractions. Finally, Section 7 concludes the paper with a discussion. All proofs are deferred to the Appendix.

Table 1: Example of synthetic individual-level public-health data with missing values. Only the cluster-level graph between HIV status, mental health, and sexual behavior is assumed to be known, since the relationships within each cluster are unspecified.

ID	`Anxiety scale`	`Depression scale`	`Condom use`	`# of partners`	`HIV status`
1	Moderate	None	NA	3	NA
2	None	Minimal	Never	1	NA
3	None	Severe	NA	5	Negative
4	Minimal	None	Almost always	21	Acute
5	Severe	Minimal	Sometimes	NA	Negative
6	None	None	Always	1	Chronic

2 Background and preliminaries

Consider a structural causal model [Pearl, 2009] that induces a causal graph, also known as an acyclic directed mixed graph (ADMG), and a distribution of interest $\Pr(\mathbb{V})$ over a set of variables $\mathbb{V}$. Let $\mathbb{V}$ be partitioned into $\mathbb{V}_o$ and $\mathbb{V}_m$, where $\mathbb{V}_o \subseteq \mathbb{V}$ denotes the set of variables observed for all records in the population, and $\mathbb{V}_m \subseteq \mathbb{V}$ denotes the set of variables missing in at least one record. We refer to $\mathbb{V}_m \cup \mathbb{V}_o$ as the set of substantive variables. A variable V_i is said to be *fully observed* if $V_i \in \mathbb{V}_o$, and *partially observed* if $V_i \in \mathbb{V}_m$. For each partially observed variable $V_i \in \mathbb{V}_m$, we associate two additional variables: a missingness indicator R_{V_i} and a proxy variable V_i^*. The intended semantics is that R_{V_i} indicates the missingness status of V_i^* in the dataset, and is what we actually observe of V_i in the available dataset, *i.e.*,

$$V_i^* = \begin{cases} V_i & \text{if } R_{V_i} = 0, \\ NA & \text{if } R_{V_i} = 1. \end{cases} \tag{1}$$

In what follows we call $\Pr(\mathbb{V}^o, \mathbb{V}^*, R)$ the manifest distribution. It represents the distribution of what we observe in addition to the missingness indicator variables.

Missingness is typically categorized depending on the process that causes it [Rubin, 1976]:

- Missing Completely at Random (MCAR): the probability of missingness in V_i is completely independent of both the value of V_i and all other variables in the dataset: $R_{V_i} \perp\!\!\!\perp (V_o \cup V_m)$. For example, a folder from a file cabinet organized in alphabetical order is lost.
- Missing at Random (MAR): the probability of missingness may depend on observed variables, but not on the missing value itself: $R_{V_i} \perp\!\!\!\perp V_m \mid V_o$. For example, a person may be less likely to disclose their condom use if they have a sexual partner living with HIV.
- Missing Not at Random (MNAR): the probability of missingness depends on unobserved data. For example, not using condoms may decrease the likelihood of being willing to share one's own test results for sexually transmitted infections, and condom usage is not observed.

Given a manifest distribution $\Pr(\mathbb{V}^*, \mathbb{V}^o, \mathbb{R})$, a query Q is said to be *recoverable* if we can compute a consistent estimate of Q as if no data were missing. *i.e.*, if it is decomposable into terms that contain the missingness mechanism $R_{V_i} = 0$ of every partially observed variable V_i in Q [Mohan et al., 2013].

For MCAR data, all $V_i \in \mathbb{V}_m$, $R_{V_i} \perp\!\!\!\perp (\mathbb{V}_o \cup \mathbb{V}_m)$. Therefore, the joint probability is recoverable trivially: $\Pr(\mathbb{V}) = \Pr(\mathbb{V} \mid R_{\mathbb{V}_m} = 0) = \Pr(\mathbb{V}_o, \mathbb{V}^*_m \mid R_{\mathbb{V}_m} = 0)$. Since, conditional on $R_{V_i} = 0$, V_i^* is observed, considering only complete cases (listwise deletion) allows recovery of the joint probability $\Pr(V_i)$. For MAR data, all $V_i \in \mathbb{V}_m$, $R_{V_i} \perp\!\!\!\perp \mathbb{V}_m \mid \mathbb{V}_o$. Thus, $\Pr(\mathbb{V}) = \Pr(\mathbb{V}_m \mid \mathbb{V}_o) \Pr(\mathbb{V}_o) = \Pr(\mathbb{V}_m \mid \mathbb{V}_o, R_{\mathbb{V}_m} = 0) \Pr(\mathbb{V}_o) = \Pr(\mathbb{V}^*_m \mid \mathbb{V}_o, R_{\mathbb{V}_m} = 0) \Pr(\mathbb{V}_o)$ [Mohan et al., 2013], making it possible to estimate the distribution of $\Pr(\mathbb{V})$.

Recoverability is more complicated under MNAR, leading many to make the assumption that data is MAR or to use listwise deletion, leading to biased results [Wothke, 2000]. Going back to the synthetic data in Table 1, suppose that mental health impacts willingness to report sexual behaviors. HIV status is also missing for some, due to an unobserved variable. Thus, MAR would allow recovery of information on sexual behavior, but such an assumption would bias estimation of the distribution of HIV status.

There are existing methods of recovery under MNAR missingness mechanisms. m-ADMGs (otherwise called *Missingness graphs* or *m-graphs*) employ understanding of the causal structure behind missingness to recover queries [Mohan et al., 2013, Mohan and Pearl, 2014, 2021]. An m-ADMG is a graph $\mathcal{G} = (\mathbb{V}, \mathbb{R}, \mathbb{V}^*, \mathbb{E}^{\rightarrow}, \mathbb{E}^{\dashleftarrow\dashrightarrow})$, where $\mathbb{V}$ is a finite set of substantive variables. For each $X \in \mathbb{V}$, $R_X \in \mathbb{R}$ denotes the missingness indicator associated with X, with $\mathbb{R} = \{R_X : X \in \mathbb{V}\}$. Similarly, $X^* \in \mathbb{V}^*$ denotes the observed-data proxy of X, with $\mathbb{V}^* = \{X^* : X \in \mathbb{V}\}$. The set $\mathbb{E}^{\rightarrow}$ contains directed edges, where an edge $X \to Y$ represents a direct causal effect of X on Y. The set $\mathbb{E}^{\dashleftarrow\dashrightarrow}$ contains bidirected edges, where an edge $X \dashleftarrow\dashrightarrow Y$ represents the presence of an unobserved common cause of X and Y. Proxy vertices have no children and no bidirected edges. For each $X \in \mathbb{V}$, the proxy variable X^* is determined by X and R_X: $X^* = X$ when $R_X = 0$, and $X^* = \mathrm{NA}$ when $R_X = 1$. Importantly, an m-ADMG provides a qualitative representation of the causal relations in an underlying structural causal model (like an ADMG), as well as missingness mechanisms and proxies of partially observed variables. Like an ADMG, it is acyclic and may contain both directed and bidirected edges. It is also important to note that a single m-ADMG can be compatible with many manifest distributions. Compatibility means that the manifest distribution satisfies the conditional independence and missingness constraints encoded by the m-ADMG. We can now redefine the notion of recoverability with respect to m-ADMGs.

Definition 1 (Recoverability in m-ADMG, Mohan et al. [2013])**.** *Given a $\mathcal{G}^m$, and a target query Q defined on the variables in $\mathbb{V}$, Q is said to be recoverable in $\mathcal{G}^m$ if there exists an algorithm that produces a consistent estimate of Q for every manifest distribution $\Pr(\mathbb{V}^*, \mathbb{V}^o, \mathbb{R})$ that is (1) compatible with $\mathcal{G}^m$ and (2) strictly positive over complete cases, i.e., $\Pr(\mathbb{V}^*, \mathbb{V}^o, \mathbb{R}) > 0$.*

Fully specified causal graphs are frequently unobtainable in epidemiological settings, where the interplay of biological, social, and behavioral factors lead to densely populated graphs with uncertainty around the exact relationships between variables. Partially specified graphs allow researchers to employ graphical causal methods, despite ambiguity in parts of these systems. In these cases, experts can sometimes specify a cluster-directed mixed graph (C-DMG), also known as a cluster graph [Tikka et al., 2023, Anand et al., 2023, Ferreira and Assaad, 2025], which is an abstract causal graph whose

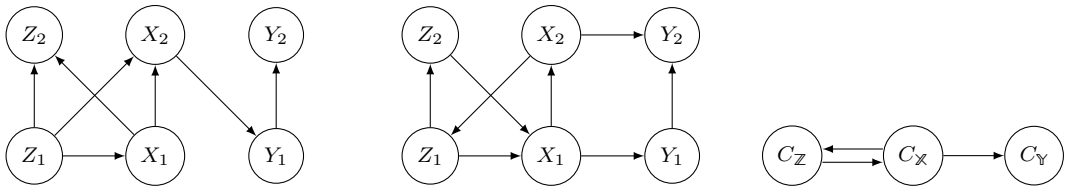


Figure 1: Two ADMGs in (a) and (b) compatible with the same C-DMG in (c).

vertices represent clusters of variables rather than individual variables. A C-DMG may contain directed edges and bidirected edges between clusters. A directed edge $C_i \to C_j$ indicates that at least one variable in cluster C_i has a direct causal effect on at least one variable in cluster C_j. A bidirected edge $C_i \dashleftarrow\!\dashrightarrow C_j$ indicates the presence of latent confounding between variables in the two clusters. Unlike ADMGs and m-ADMGs, C-DMGs may contain directed cycles, since feedback can arise at the cluster level even when the underlying variable-level graph is acyclic. For illustration, Figure 1 presents a C-DMG with two of its compatible ADMGs. However, missingness indicators have not yet been incorporated into cluster graphs, limiting their applicability in incomplete data settings.

We end this section by introducing some graphical terminology. Two vertices are said to be *neighbors* if they are adjacent, that is, if they are connected by an edge, regardless of its type or orientation. We use the standard notions of parents, children, ancestors, and descendants induced by directed paths. A non-endpoint vertex on a path is called a *collider* on that path if the two adjacent edges on the path both have arrowheads pointing into it. For example, in the path $A \to B \leftarrow C$, the vertex B is a collider. In contrast, B is a non-collider in paths of the form $A \to B \to C, A \leftarrow B \to C, A \leftarrow B \leftarrow C$.

3 Missingness in cluster graphs

A central objective of this paper is to investigate how missingness mechanisms can be incorporated into cluster graphs. To this end, we introduce two abstract missingness representations, which differ in the level of detail they retain about missingness indicators:

- **missingness-C-DMG (m-C-DMG)**: an abstract graph in which each variable inside a cluster is associated with its own missingness indicator;
- **clustered-missingness-C-DMG (cm-C-DMG)**: a coarser abstract graph in which a single missingness vertex represents the missingness mechanism of the whole cluster.

This distinction is crucial because the two abstractions do not encode the same amount of information. The *m-C-DMG* preserves variable-specific missingness information as it relates to each cluster, whereas the *cm-C-DMG* merges these distinctions into a single cluster-level missingness vertex. As a consequence, the *cm-C-DMG* is generally less informative than the *m-C-DMG*.

Definition 2 (m-C-DMG)**.** *Let $\mathbb{C} = \{C_1, \ldots, C_K\}$ be a partition of a finite set of substantive variables $\mathbb{V}$. An* m-C-DMG *is a graph $\mathcal{G}^{\mathbb{C},m} = (\mathbb{C}, \mathbb{R}, \mathbb{V}^*, \mathbb{E}^{\to}, \mathbb{E}^{\dashleftarrow\!\dashrightarrow})$, where $\mathbb{C}$ is the set of cluster vertices, $\mathbb{R} = \{R_X : X \in \mathbb{V}\}$ is the set of variable-level missingness indicators, and $\mathbb{V}^* = \{X^* : X \in \mathbb{V}\}$ is the set of proxy variables. The edges are defined as follows:*

- *The set $\mathbb{E}^{\to}$ contains directed edges between elements of $\mathbb{C} \cup \mathbb{R} \cup \mathbb{V}^*$. A directed edge $C_{\mathbb{X}} \to C_{\mathbb{Y}}$ between two cluster vertices means that there exist variables $X_i \in C_{\mathbb{X}}$ and $Y_j \in C_{\mathbb{Y}}$ such that X_i has a direct causal effect on Y_j at the variable level. Similarly, a directed edge $C_{\mathbb{X}} \to R_Y$ means that there exists $X_i \in C_{\mathbb{X}}$ such that X_i has a direct causal effect on the missingness indicator R_Y.*
- *The set $\mathbb{E}^{\dashleftarrow\!\dashrightarrow}$ contains bidirected edges between elements of $\mathbb{C} \cup \mathbb{R}$. A bidirected edge $C_{\mathbb{X}} \dashleftarrow\!\dashrightarrow C_{\mathbb{Y}}$ means that there exist variables $X_i \in C_{\mathbb{X}}$ and $Y_j \in C_{\mathbb{Y}}$ that share an unobserved common cause at the variable level.*

Specifying an m-C-DMG can be difficult because it requires distinguishing the missingness mechanisms of individual variables and their relations to the rest of the graph. It can therefore be useful

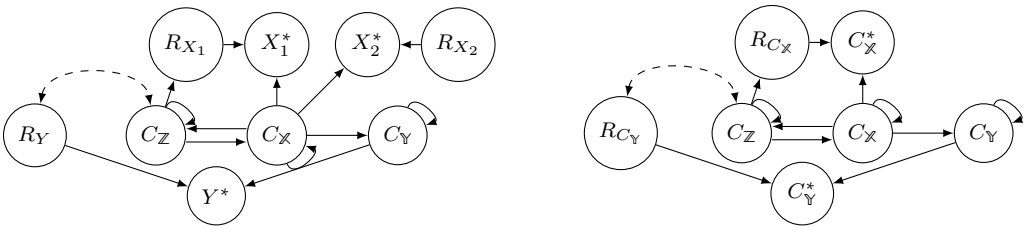


Figure 2: An m-C-DMG (a) and a cm-C-DMG (b), representing conceptual graphs of our running example, with the following clusters: $C_{\mathbb{X}}$: Sexual Behaviors, $C_{\mathbb{Z}}$: Mental Health, $C_{\mathbb{Y}}$: HIV Status. This graph shows a cyclical relationship between mental health and sexual behavior [Alvy et al., Mustanski et al., O'Cleirigh et al.] and the impact of sexual behavior on HIV status [World Health Organization], with missingness on sexual behavior and on HIV status.

to cluster missingness indicators and work with a coarser representation. However, allowing an arbitrary partition of the missingness indicators would make it difficult to track which missingness mechanism corresponds to which substantive variables. For this reason, we focus on a structured form of clustering in which missingness indicators are merged only when their associated variables already belong to the same substantive cluster. We call the m-C-DMG where each missingness indicator $R_{C_{\mathbb{X}}}$ represents the missingness mechanism of an entire cluster $C_{\mathbb{X}}$ a cm-C-DMG.

Definition 3 (cm-C-DMG)**.** *Let $\mathbb{C} = \{C_1, \ldots, C_K\}$ be a partition of a finite set of substantive variables $\mathbb{V}$. A* cm-C-DMG *is a graph $\mathcal{G}^{\mathbb{C},cm} = (\mathbb{C}, \mathbb{R}_{\mathbb{C}}, \mathbb{V}^*, \mathbb{E}^{\rightarrow}, \mathbb{E}^{\dashleftarrow\dashrightarrow})$, where $\mathbb{C}$ is the set of cluster vertices, $\mathbb{R}_{\mathbb{C}} = \{R_C : C \in \mathbb{C}\}$ is the set of cluster-level missingness indicators, and $\mathbb{V}^* = \{X^* : X \in \mathbb{V}\}$ is the set of variable-level proxy vertices. For each cluster $C \in \mathbb{C}$, the node R_C represents the missingness indicators associated with variables in C, namely $\{R_X : X \in C\}$. Missingness indicators associated with variables in different clusters are not merged. The edges are defined as follows:*

- *The set $\mathbb{E}^{\rightarrow}$ contains directed edges between elements of $\mathbb{C} \cup \mathbb{R}_{\mathbb{C}} \cup \mathbb{V}^*$. A directed edge $C_{\mathbb{X}} \to C_{\mathbb{Y}}$ means that there exist variables $X_i \in C_{\mathbb{X}}$ and $Y_j \in C_{\mathbb{Y}}$ such that X_i has a direct causal effect on Y_j at the variable level. A directed edge $C_{\mathbb{X}} \to R_{C_{\mathbb{Y}}}$ means that there exist $X_i \in C_{\mathbb{X}}$ and $Y_j \in C_{\mathbb{Y}}$ such that X_i has a direct causal effect on R_{Y_j} at the variable level.*
- *The set $\mathbb{E}^{\dashleftarrow\dashrightarrow}$ contains bidirected edges between elements of $\mathbb{C} \cup \mathbb{R}_{\mathbb{C}}$. A bidirected edge $C_{\mathbb{X}} \dashleftarrow\dashrightarrow C_{\mathbb{Y}}$ means that there exist variables $X_i \in C_{\mathbb{X}}$ and $Y_j \in C_{\mathbb{Y}}$ that share an unobserved common cause at the variable level.*

Unlike m-ADMGs and similarly to C-DMGs, m-C-DMGs and cm-C-DMGs may contain cycles.

In an m-CDMG or cm-C-DMG, for every $X \in \mathbb{V}$, the proxy vertex X^* has exactly two parents: the cluster vertex $C_{\mathbb{X}}$ containing X, and the missingness indicator. It is also important to highlight that proxy vertices have no children and no bidirected edges. The only difference between the two definitions is that in the former, missingness indicator represents the missingness mechanism of the single variable X, whereas in the latter, a missingness indicator represents that of an entire cluster.

3.1 Compatibility classes

Let $compatible(\mathcal{G}^{\mathbb{C},*m})$ denote the set of variable-level m-ADMGs compatible with an abstract graph $\mathcal{G}_{\mathbb{C}}$. Given an m-C-DMG $\mathcal{G}^{\mathbb{C},m}$, let $\pi(\mathcal{G}^{\mathbb{C},m})$ denote the associated cm-C-DMG obtained by merging, within each cluster, all variable-level missingness indicators into a single cluster-level missingness vertex. Under this natural abstraction map, every variable-level m-ADMG compatible with $\mathcal{G}^{\mathbb{C},m}$ is also compatible with $\pi(\mathcal{G}^{\mathbb{C},m})$. Intuitively, this is because the cm-C-DMG forgets distinctions that are preserved by the m-C-DMG. Therefore, the compatibility class of the m-C-DMG is contained in that of the associated cm-C-DMG:

Proposition 1 (Inclusion of compatibility classes)**.** *Let $\mathcal{G}^{\mathbb{C},m}$ be an m-C-DMG and let $\mathcal{G}^{cm} = \pi(\mathcal{G}^{\mathbb{C},m})$ be the associated cm-C-DMG. Then $compatible(\mathcal{G}^{\mathbb{C},m}) \subseteq compatible(\mathcal{G}^{\mathbb{C},cm})$. Moreover, the inclusion is strict whenever at least one cluster contains two variables with distinct missingness mechanisms.*

The inclusion is strict because different variable-level missingness mechanisms may induce the same cluster-level missingness abstraction. For example, consider a cluster $C = \{A, B\}$ with variable-level missingness indicators R_A and R_B. The m-C-DMG can distinguish between situations where R_A depends on one set of variables and R_B depends on another. By contrast, the cm-C-DMG only records that the cluster C is subject to missingness, without retaining which variable-specific indicator is involved. Hence, several distinct m-ADMGs compatible with the same cm-C-DMG may fail to be compatible with the corresponding m-C-DMG.

3.2 Recoverability semantics

Let Q be a target query. We say that Q is *recoverable* in an abstract graph if Q can be recovered uniformly for every variable-level m-ADMG compatible with the abstract graph, *i.e.*, Q is recoverable in $\mathcal{G}^{\mathbb{C},m} \iff Q$ is recoverable in every $\mathcal{G} \in compatible(\mathcal{G}^{\mathbb{C},m})$ and Q is recoverable in $\mathcal{G}^{\mathbb{C},cm} \iff Q$ is recoverable in every $\mathcal{G} \in compatible(\mathcal{G}^{\mathbb{C},cm})$.

In general, a target quantity Q is less likely to be recoverable from a cm-C-DMG than from an m-C-DMG, since the former is a coarser abstraction. This is formalized in the following proposition.

Proposition 2. *Let $\mathcal{G}^{\mathbb{C},m}$ be an m-C-DMG and let $\mathcal{G}^{\mathbb{C},cm} = \pi(\mathcal{G}^{\mathbb{C},m})$ be the associated cm-C-DMG. Then, for any target query Q, Q recoverable in $\mathcal{G}^{\mathbb{C},cm} \Rightarrow Q$ recoverable in $\mathcal{G}^{\mathbb{C},m}$. The converse does not hold in general.*

The above proposition shows that there may exist a cm-C-DMG $\mathcal{G}^{\mathbb{C},cm}$ and a target query Q such that Q is recoverable in every $\mathcal{G}^m \in compatible(\mathcal{G}^{\mathbb{C},cm})$, while Q is not recoverable from $\mathcal{G}^{\mathbb{C},cm}$ alone. This follows from the fact that the cm-C-DMG suppresses variable-specific missingness distinctions that may be essential for deriving a recoverability formula.

Example 1. *Let $C_{\mathbb{X}} = \{X_1, X_2\}$, and suppose the target query is $Q = \Pr(X_1)$. In an m-C-DMG, assume that only the missingness of X_2 may depend on the cluster: $C_{\mathbb{X}} \to R_{X_2}$, and $C_{\mathbb{X}} \not\to R_{X_1}$. Thus, the missingness of X_1 is not affected by variables in $C_{\mathbb{X}}$ and $\Pr(X_1)$ may be recoverable from the observed cases of X_1. After merging R_{X_1} and R_{X_2} into the cluster-level missingness vertex $R_{C_{\mathbb{X}}}$, the associated cm-C-DMG contains $C_{\mathbb{X}} \to R_{C_{\mathbb{X}}}$. This no longer indicates whether the cluster affects the missingness of X_1, of X_2, or both. Hence the cm-C-DMG is compatible with cases in which the missingness of X_1 depends on $C_{\mathbb{X}}$, including on X_1 itself, making $\Pr(X_1)$ non-recoverable in general. Therefore, $\Pr(X_1)$ may be recoverable from the m-C-DMG but not from the associated cm-C-DMG.*

In the following, we focus on two queries: the joint distribution and macro causal effects. For each, we present recoverability results for both m-C-DMGs and cm-C-DMGs. We denote by $\mathcal{G}^{\mathbb{C},*m}$ a graph that may be either an m-C-DMG or a cm-C-DMG. For simplicity, all illustrative examples are presented using cm-C-DMGs. This allows us to emphasize the main ideas while working with the coarser and more compact abstraction.

4 Recovering joint distribution

Recovering the joint distribution provides the basis for many downstream probabilistic queries, including marginal and conditional distributions. In the presence of missing data, however, the joint distribution is not directly available from the observed data and may fail to be recoverable. In the following, we provide necessary conditions for recoverability.

Theorem 1 (Necessary conditions for joint recoverability)**.** *Let $\mathcal{G}^{\mathbb{C},*m}$ be an m-C-DMG or a cm-C-DMG. The necessary condition for recovering any joint distribution $\Pr(\mathbb{C})$ compatible with $\mathcal{G}^{\mathbb{C},*m}$ is the absence of any vertex $C_{\mathbb{X}} \in \mathbb{C}^m$ such that:*

- *$C_{\mathbb{X}}$ and its R-vertex are neighbors, or*
- *$C_{\mathbb{X}}$ and its R-vertex are connected by a path in which all intermediate vertices are colliders and elements of $\mathbb{C}^m \cup \mathbb{C}^o$.*

In the following, we provide sufficient conditions for recovering the joint distribution over all clusters. Theorem 2 uses two notions: missingness-indicator blocks and the Markov blanket $\mathbb{R}^{(a)}$, both of which are defined in the Appendix.

Theorem 2 (Sufficient conditions for joint recoverability)**.** *Let $\mathcal{G}^{\mathbb{C},cm} = \left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}, \mathbb{E}^{\mathbb{C}}\right)$ be a cm-C-DMG. Let $\mathbb{R}^{(1)}, \ldots, \mathbb{R}^{(K)}$ be the missingness-indicator blocks, and for any $1 \leq a \leq K$ let $MB^o\left(\mathbb{R}^{(a)}\right), MB^m\left(\mathbb{R}^{(a)}\right)$ denote the fully observed and partially observed parts of the substantive Markov blanket of $\mathbb{R}^{(a)}$. If there are no edges (including self-loops) between missingness-indicator vertices and for every missingness-indicator block $\mathbb{R}^{(a)}$, $\mathbb{R}^{(a)} \cap R_{MB^m\left(\mathbb{R}^{(a)}\right)} = \varnothing$, then the joint distribution $\Pr(\mathbb{c})$ is recoverable using $\mathcal{G}^{\mathbb{C},cm}$ and given by*

$$\Pr(\mathbb{c}) = \frac{\Pr\left(\mathbb{R}_{\mathbb{C}} = 0, \mathbb{c}\right)}{\prod_{a=1}^{K} \Pr\left(\mathbb{R}^{(a)} = 0 \mid MB^o\left(\mathbb{R}^{(a)}\right), MB^m\left(\mathbb{R}^{(a)}\right), R_{MB^m\left(\mathbb{R}^{(a)}\right)} = 0\right)}. \tag{2}$$

It is important to note that the conditions of Theorem 2 imply those of Theorem 1, whereas the converse does not hold. Further details are provided in the Appendix. These results are closely related to the recoverability results established by Mohan and Pearl [2014] for m-ADMGs. They show that analogous graphical conditions can be formulated directly at the cluster level. Moreover, because m-C-DMGs and cm-C-DMGs may contain cycles, the results suggest that the acyclicity assumption imposed in standard ADMGs is not essential for this type of recoverability result.

In the following, we illustrate the above theorems on our running example and show that the joint distribution is recoverable in this case.

Example 2. *Consider the cm-C-DMG in Figure 2 (b). In this graph, $C_{\mathbb{X}}$ is not adjacent to $R_{C_{\mathbb{X}}}$, and there is no path from $C_{\mathbb{X}}$ to $R_{C_{\mathbb{X}}}$ whose intermediate vertices are all colliders belonging to $\mathbb{C}^m \cup \mathbb{C}^o$. Similarly, $C_{\mathbb{Y}}$ is not adjacent to $R_{C_{\mathbb{Y}}}$, and no such collider path connects them. No partially observed cluster violates the conditions of the theorem, thus the joint distribution $\Pr(c_{\mathbb{X}}, c_{\mathbb{Y}}, c_{\mathbb{Z}})$ is recoverable.*

Interestingly, based on the above theorem, we can show that recoverability of the joint distribution in the cm-C-DMG is equivalent to recoverability of the joint distribution in the m-C-DMG.

Proposition 3. *Let $\mathcal{G}^{\mathbb{C},m}$ be an m-C-DMG and let $\mathcal{G}^{\mathbb{C},cm} = \pi(\mathcal{G}^{\mathbb{C},m})$ be the associated cm-C-DMG. Then, Theorem 1 is satisfied in $\mathcal{G}^{\mathbb{C},cm}$ $\iff$ Theorem 1 is satisfied in $\mathcal{G}^{\mathbb{C},m}$. However, while Theorem 2 is satisfied in $\mathcal{G}^{\mathbb{C},cm}$ $\implies$ Theorem 2 is satisfied in $\mathcal{G}^{\mathbb{C},m}$, the converse is not true.*

5 Recovering macro causal effects

So far, we have focused on the recoverability of probabilistic, non-causal queries, namely, joint distributions. We now turn to the recovery of macro causal effects, which is a central objective in many applications. For instance, we may want to assess the effect of sexual behavior on the risk of acquiring HIV, even when data is partially observed. Such questions require going beyond observational distributions and reasoning about interventions at the cluster level.

Before presenting the main results of this section, we introduce several key tools that will be used to derive recoverability conditions for macro causal effects.

We begin by introducing the notion of blocked paths. Let $\mathcal{G}^{\mathbb{C},*m}$ denote either an m-C-DMG or a cm-C-DMG. A path between two vertices is said to be blocked by a set $C_{\mathbb{W}}$ if at least one of the following conditions holds: either the path contains a non-collider that belongs to $C_{\mathbb{W}}$, or the path contains a collider such that neither the collider nor any of its descendants belongs to $C_{\mathbb{W}}$. Otherwise, the path is said to be active given $C_{\mathbb{W}}$.

Two sets of vertices $C_{\mathbb{X}}$ and $C_{\mathbb{Y}}$ are said to be d-separated by $C_{\mathbb{W}}$ in $\mathcal{G}^{\mathbb{C},*m}$ if every path between a vertex in $C_{\mathbb{X}}$ and a vertex in $C_{\mathbb{Y}}$ is blocked by $C_{\mathbb{W}}$. We write this as $\mathbb{X} \perp\!\!\!\perp_{\mathcal{G}^{\mathbb{C},*m}} \mathbb{Y} \mid \mathbb{W}$. The notion of d-separation was originally introduced for ADMGs [Pearl, 2009]. It was recently shown to be sound for C-DMGs, which may contain cycles [Ferreira and Assaad, 2025]. Since m-C-DMGs and cm-C-DMGs are particular forms of C-DMGs, this soundness result applies directly to both graph classes considered here.

To reason about interventions, we also use mutilated graphs. Given a graph $\mathcal{G}^{\mathbb{C},*m}$ and a set of vertices $C_{\mathbb{X}}$, we denote by $\mathcal{G}^{\mathbb{C},*m}_{\overline{C_{\mathbb{X}}}}$ the graph obtained by removing all incoming directed edges into vertices in $C_{\mathbb{X}}$. Similarly, $\mathcal{G}^{\mathbb{C},*m}_{\underline{C_{\mathbb{X}}}}$ denotes the graph obtained by removing all outgoing directed edges

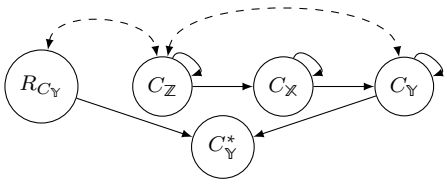


Figure 3: cm-DMG where the macro causal effect of $C_\mathbb{X}$ on $C_\mathbb{Y}$ is recoverable but the joint distribution is not.

from vertices in $C_\mathbb{X}$. $\mathcal{G}^{\mathbb{C},*m}_{\overline{C_\mathbb{X}}\underline{C_\mathbb{Z}}}$ denotes the graph obtained by removing incoming edges into $C_\mathbb{X}$ and outgoing edges from $C_\mathbb{Z}$.

Using d-separation and the notion of mutilated graphs, we can now present the do-calculus, a set of rules that can be used to identify a causal effect. The do-calculus was originally introduced for ADMGs and m-ADMGs and was later shown to be sound for C-DMGs [Ferreira and Assaad, 2025]. In the following, we present the corresponding rules for m-C-DMGs and cm-C-DMGs.

Definition 4 (Rules of the do-calculus in m-C-DMGs and cm-C-DMGs)**.** *The three following rules of the do-calculus are the following:*

Rule 1: $\Pr(\mathbb{c}_\mathbb{y} \mid do(\mathbb{c}_\mathbb{z}), \mathbb{c}_\mathbb{x}, \mathbb{c}_\mathbb{w}) = \Pr(\mathbb{c}_\mathbb{y} \mid do(\mathbb{c}_\mathbb{z}), \mathbb{c}_\mathbb{w})$ *if* $\mathbb{C}_\mathbb{Y} \perp\!\!\!\perp_{\mathcal{G}^{\mathbb{C},*m}_{\overline{\mathbb{C}_\mathbb{Z}}}} \mathbb{C}_\mathbb{X} \mid \mathbb{C}_\mathbb{Z}, \mathbb{C}_\mathbb{W}$

Rule 2: $\Pr(\mathbb{c}_\mathbb{y} \mid do(\mathbb{c}_\mathbb{z}), do(\mathbb{c}_\mathbb{x}), \mathbb{c}_\mathbb{w}) = \Pr(\mathbb{c}_\mathbb{y} \mid do(\mathbb{c}_\mathbb{z}), \mathbb{c}_\mathbb{x}, \mathbb{c}_\mathbb{w})$ *if* $\mathbb{C}_\mathbb{Y} \perp\!\!\!\perp_{\mathcal{G}^{\mathbb{C},*m}_{\overline{\mathbb{C}_\mathbb{Z}}\underline{\mathbb{C}_\mathbb{X}}}} \mathbb{C}_\mathbb{X} \mid \mathbb{C}_\mathbb{Z}, \mathbb{C}_\mathbb{W}$

Rule 3: $\Pr(\mathbb{c}_\mathbb{y} \mid do(\mathbb{c}_\mathbb{z}), do(\mathbb{c}_\mathbb{x}), \mathbb{c}_\mathbb{w}) = \Pr(\mathbb{c}_\mathbb{y} \mid do(\mathbb{c}_\mathbb{z}), \mathbb{c}_\mathbb{w})$ *if* $\mathbb{C}_\mathbb{Y} \perp\!\!\!\perp_{\mathcal{G}^{\mathbb{C},*m}_{\overline{\mathbb{C}_\mathbb{Z}}\overline{\mathbb{C}_\mathbb{X}(\mathbb{C}_\mathbb{W})}}} \mathbb{C}_\mathbb{X} \mid \mathbb{C}_\mathbb{Z}, \mathbb{C}_\mathbb{W}$

where $\mathbb{C}_\mathbb{X}(\mathbb{C}_\mathbb{W})$ *is the set of vertices in* $\mathbb{C}_\mathbb{X}$ *that are non-ancestors of any vertex in* $\mathbb{C}_\mathbb{W}$ *in the mutilated graph* $\mathcal{G}^{\mathbb{C},*m}_{\overline{\mathbb{C}_\mathbb{Z}}}$.

Having introduced all the necessary tools, we are now in a position to show how to recover macro causal effects from m-C-DMGs and cm-C-DMGs. The results presented in this section provide conditions under which such effects can be identified from the observed data using the graphical structure. This constitutes the main result of this section.

Theorem 3 (Recoverability of macro causal effects in m-C-DMGs and cm-C-DMGs)**.** *A causal query Q is recoverable from an m-C-DMG or cm-C-DMG if there exists a sequence of applications of do-calculus, probability manipulations, and Equation 1 that transforms Q into an expression involving only observable quantities (namely proxy variables, missingness indicators, and fully observed variables) and containing no do-operators.*

In the following, we illustrate the theorem with two examples.

Example 3. *Consider the cm-C-DMG in Figure 2 (b). Suppose we are interested in the macro causal effect* $\Pr(c_\mathbb{Y} \mid do(c_\mathbb{X}))$*. The first step is to transform the partially observed outcome into its observed proxy. Since* $R_{C_\mathbb{Y}}$ *is d-separated from* $C_\mathbb{Y}$ *in the mutilated graph where* $C_\mathbb{X}$ *is intervened on, Rule 1 of the do-calculus gives* $\Pr(c_\mathbb{Y} \mid do(c_\mathbb{X})) = \Pr(c_\mathbb{Y} \mid do(c_\mathbb{X}), R_{C_\mathbb{Y}} = 0)$*. Then, by Equation 1,* $\Pr(c_\mathbb{Y} \mid do(c_\mathbb{X}), R_{C_\mathbb{Y}} = 0) = \Pr(c^*_\mathbb{Y} \mid do(c_\mathbb{X}), R_{C_\mathbb{Y}} = 0)$.

The next step is to remove the do-operator. In this graph, there is no open non-causal path from $C_\mathbb{X}$ *to* $C_\mathbb{Y}$ *after conditioning on the observed outcome missingness status. Hence, by Rule 2 of the do-calculus,* $\Pr(c^*_\mathbb{Y} \mid do(c_\mathbb{X}), R_{C_\mathbb{Y}} = 0) = \Pr(c^*_\mathbb{Y} \mid c_\mathbb{X}, R_{C_\mathbb{Y}} = 0)$.

Finally, since $C_\mathbb{X}$ *is partially observed, we need to replace it by its proxy. Notice that in our graph we have* $C^*_\mathbb{Y}$ *d-separated from* $R_{C_\mathbb{X}}$ *given* $C_\mathbb{X}$ *and* $R_{C_\mathbb{Y}}$*. So using Rule 1 of the do-calculus we can add* $R_{C_\mathbb{X}} = 0$ *to the expression. Then Using again Equation 1, we obtain* $\Pr(c^*_\mathbb{Y} \mid c_\mathbb{X}, R_{C_\mathbb{Y}} = 0) = \Pr(c^*_\mathbb{Y} \mid c^*_\mathbb{X}, R_{C_\mathbb{X}} = 0, R_{C_\mathbb{Y}} = 0)$.

In the previous example, the missingness indicator of $C_\mathbb{Y}$ can be d-separated from $C_\mathbb{Y}$. This might suggest that such a separation is always required for recoverability.

Example 4. *Consider the cm-C-DMG in Figure 3. Suppose that we are interested in the macro causal effect* $\Pr(C_\mathbb{Y} \mid do(C_\mathbb{X}))$*. The first step is to replace each partially observed variable that is*

not intervened on by its observed proxy. By Rule 1 of the do-calculus, $\Pr(c_{\mathbb{Y}} \mid do(c_{\mathbb{X}})) = \Pr(c_{\mathbb{Y}} \mid do(c_{\mathbb{X}}), R_{C_{\mathbb{Y}}} = 0)$. *Then, by Equation 1,* $\Pr(c_{\mathbb{Y}} \mid do(c_{\mathbb{X}}), R_{C_{\mathbb{Y}}} = 0) = \Pr(c^*_{\mathbb{Y}} \mid do(c_{\mathbb{X}}), R_{C_{\mathbb{Y}}} = 0)$.

The next step is to remove the do-operator. We first use the law of total probability to introduce $C_{\mathbb{Z}}$*:* $\Pr(c^*_{\mathbb{Y}} \mid do(c_{\mathbb{X}}), R_{C_{\mathbb{Y}}} = 0) = \sum_{c_{\mathbb{Z}}} \Pr(c^*_{\mathbb{Y}} \mid do(c_{\mathbb{X}}), R_{C_{\mathbb{Y}}} = 0, c_{\mathbb{Z}}) \Pr(c_{\mathbb{Z}} \mid do(c_{\mathbb{X}}), R_{C_{\mathbb{Y}}} = 0)$. *By Rule 2 of the do-calculus,* $\Pr(c^*_{\mathbb{Y}} \mid do(c_{\mathbb{X}}), R_{C_{\mathbb{Y}}} = 0) = \sum_{c_{\mathbb{Z}}} \Pr(c^*_{\mathbb{Y}} \mid c_{\mathbb{X}}, R_{C_{\mathbb{Y}}} = 0, c_{\mathbb{Z}}) \Pr(c_{\mathbb{Z}} \mid do(c_{\mathbb{X}}), R_{C_{\mathbb{Y}}} = 0)$. *Finally, by Rule 3,* $\Pr(c^*_{\mathbb{Y}} \mid do(c_{\mathbb{X}}), R_{C_{\mathbb{Y}}} = 0) = \sum_{c_{\mathbb{Z}}} \Pr(c^*_{\mathbb{Y}} \mid c_{\mathbb{X}}, R_{C_{\mathbb{Y}}} = 0, c_{\mathbb{Z}}) \Pr(c_{\mathbb{Z}} \mid R_{C_{\mathbb{Y}}} = 0)$.

Interestingly, the previous example shows that a macro causal effect can be recoverable even when the joint distribution is not. This highlights that recoverability of causal effects does not necessarily require recoverability of the full joint distribution.

6 Simulated data

To exemplify recovery of the joint distribution from a cm-C-DMG, we simulated 100 datasets in two different settings[1]. The two settings, referred to as Setting 1 and Setting 2, are based on two different m-ADMGs compatible with the cm-C-DMG in Figure 2b. All variables are binary and generated according to a structural causal model. Each is sampled from a Bernoulli distribution as functions of their parents, with independent exogenous noise terms: $V_i \sim \text{Bernoulli}(\text{expit}(\alpha_i + \sum_{j \in \text{Pa}(V_i)} \beta_{ij} V_j + \gamma_i U_i))$, where $V_i \in \{0, 1\}$ and random coefficients $\in \{-10, 10\}$. We introduced missingness into the datasets through a similar process. See additional details on data simulation in the Appendix.

We estimated the full joint distributions for both settings under four hypotheses: 1) MCAR, using a complete case analysis, 2) MAR, using the Horvitz-Thompson estimator [Horvitz and Thompson, 1952, Robins et al., 1995], 3) MNAR, using the true m-ADMG underlying the data-generating mechanism with the recovery criterion of Mohan and Pearl [2014], and 4) MNAR, using a cm-C-DMG with Equation 2. Figure 4 compares the estimated joint distribution with the true joint distribution, using the normalized total variation distance (NTVD) [Gibbs and Su, 2002]: $\frac{1}{2} \sum_{\mathbb{v}} |\Pr(\mathbb{v}) - \hat{\Pr}(\mathbb{v})|$ where $\Pr$ refers to the true joint distribution and $\hat{\Pr}$ refers to the estimated joint distribution. Results are visualized with ± 1 standard deviation error bars.

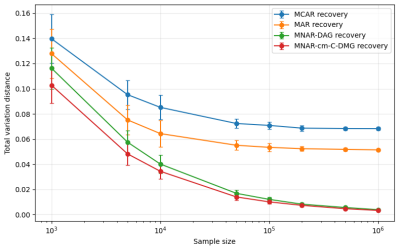
Figure 4 (a,b) shows that in Settings 1 and 2, the NTVD under the MNAR assumption converges to zero for both the fully specified m-ADMG and the cm-C-DMG as the sample size increases. Recovery based on the cm-C-DMG is as accurate as recovery based on the m-ADMG in Setting 2, and slightly more accurate in Setting 1, especially when the number of observations is relatively small. Further investigation is necessary to understand why this is the case. In contrast, under MAR and MCAR assumptions, the NTVD decreases initially but plateaus between approximately 0.05 and 0.45, even for large sample sizes.

Figure 4 (c,d) shows estimation of the macro causal effect of $C_{\mathbb{X}}$ on $C_{\mathbb{Y}}$ based on the estimated joint distributions. Perhaps counter-intuitively, estimation with MCAR recovery, which is less precise for joint distribution estimation, is the most accurate at small sample sizes in both Setting 1 and Setting 2. However, considering the results of Example 3, we note that the recovery of the macro causal effect is ultimately equivalent to assuming MCAR missingness for variables involved in causal effect estimation. As sample size grows, all recovery methods converge to the correct macro causal effect.

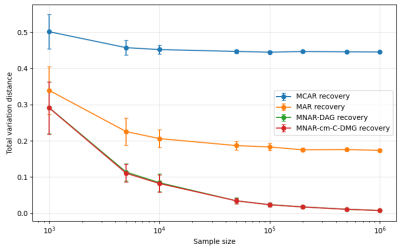
7 Discussion

In this paper, we studied recoverability under abstract missingness representations. We introduced two cluster-based missingness graphs: the m-C-DMG, which preserves variable-specific missingness indicators, and the cm-C-DMG, which further aggregates missingness indicators associated with variables in the same cluster. We formalized compatibility between these abstract graphs and variable-level missingness models and showed how recoverability conclusions can be transferred across abstraction levels. In particular, recoverability from a cm-C-DMG implies recoverability from the corresponding m-C-DMG, while the converse does not hold in general. Moreover, we gave sufficient conditions and necessary conditions for recovering joint distributions and sufficient

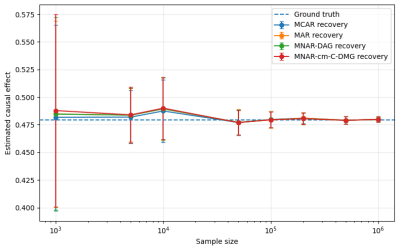
[1]Code simulation study can be accessed at `https://github.com/CIPHOD/pyCIPHOD/tree/main/reproducibility/neurips2026`



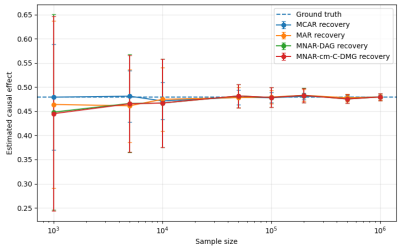
(a) Joint distribution recovery in setting 1.



(b) Joint distribution recovery setting 2.



(c) Macro causal effect estimation in setting 1.



(d) Macro causal effect estimation in setting 2.

Figure 4: Joint distribution recovery (a,b) and macro-level causal effect estimation (c,d) under two settings. Settings 1 and 2 correspond respectively to data-generating processes compatible with two distinct m-ADMGs, both compatible with the cm-C-DMG in Figure 2b.

conditions for recovering macro causal effects. Studying missingness through causal abstractions has several advantages. First, it provides a principled way to analyze settings where only coarse structural knowledge is available. Second, it clarifies how much information is lost when missingness mechanisms are aggregated, and when this loss matters for a specific query.

This work has several limitations. Because m-C-DMGs and cm-C-DMGs are abstractions, they may be conservative: a query that is not recoverable from the abstract graph may still be recoverable in some compatible variable-level graph. Moreover, cm-C-DMGs can hide distinctions between missingness indicators that are essential for recovery. Their usefulness thus depends on whether the chosen clustering preserves the missingness information relevant to the target query. Furthermore, we considered only a special case of cm-C-DMGs, where missingness indicators are clustered only when their associated variables belong to the same substantive cluster. In real applications, if experts cannot specify an m-C-DMG, the partitioning they can provide may not satisfy this simplifying requirement.

Several questions remain open. An important direction is to determine whether our recoverability results are complete. Another direction is to develop algorithms for recoverability directly on these abstract graphs. For non-causal queries, we focused mainly on recovering joint distributions; it would therefore be interesting to further investigate conditional distributions. Furthermore, it would be interesting to study the relative efficiency of recovery based on ADMGs compared with recovery based on cluster graphs in finite-sample settings. Finally, future work should extend the framework to the recovery of micro causal effects under causal abstractions, building on recent work on identifiability from abstract causal graphs [Assaad et al., 2024].

Societal impacts

This paper presents work whose goal is to advance the field of Machine Learning and Epidemiology, and the limits of what can be inferred from data. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgement

We thank anonymous reviewers for their constructive comments. This work was supported by the CIPHOD project (ANR-23-CPJ1-0212-01) as well as by the PP-IMPACT project.

References

Lisa M. Alvy, David J. McKirnan, Gordon Mansergh, Beryl Koblin, Grant N. Colfax, Stephen A. Flores, Sharon Hudson, and Project MIX Study Group. Depression is associated with sexual risk among men who have sex with men, but is mediated by cognitive escape and self-efficacy. 15(6): 1171–1179. ISSN 1573-3254. doi: 10.1007/s10461-010-9678-z.

Tara V. Anand and George Hripcsak. Leveraging cluster causal diagrams for determining causal effects in medicine. *AMIA Annual Symposium Proceedings*, 2025. PMID: 40417561.

Tara V. Anand, Adele H. Ribeiro, Jin Tian, and Elias Bareinboim. Causal effect identification in cluster dags. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i10.26435.

Charles K. Assaad, Imad Ez-Zejjari, and Lei Zan. Root cause identification for collective anomalies in time series given an acyclic summary causal graph with loops. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8395–8404. PMLR, 25–27 Apr 2023.

Charles K. Assaad, Emilie Devijver, Eric Gaussier, Gregor Goessler, and Anouar Meynaoui. Identifiability of total effects from abstractions of time series causal graphs. In Negar Kiyavash and Joris M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 173–185. PMLR, 15–19 Jul 2024.

Simon Ferreira and Charles K. Assaad. Identifying macro conditional independencies and macro total effects in summary causal graphs with latent confounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(25):26787–26795, Apr. 2025. doi: 10.1609/aaai.v39i25.34882.

Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002. doi: 10.1111/j.1751-5823.2002.tb00178.x.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Keisuke Inoue, Beate Ritz, and Onyebuchi A. Arah. Causal effect of chronic pain on mortality through opioid prescriptions: Application of the front-door formula. *Epidemiology*, 33(4):544–552, 2022. doi: 10.1097/EDE.0000000000001477.

Roderick J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993. ISSN 01621459, 1537274X.

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

Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021. doi: 10.1080/01621459.2021.1874961.

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

Brian Mustanski, Robert Garofalo, Amy Herrick, and Geri Donenberg. Psychosocial health problems increase risk for HIV among urban young men who have sex with men: Preliminary evidence of a syndemic in need of attention. 34(1):37–45. ISSN 1532-4796. doi: 10.1007/BF02879919.

Conall O'Cleirigh, Michael E. Newcomb, Kenneth H. Mayer, Margie Skeer, Lara Traeger, and Steven A. Safren. Moderate levels of depression predict sexual transmission risk in HIV-infected MSM: A longitudinal analysis of data from six sites involved in a "prevention for positives" study. 17(5):1764–1769. ISSN 1573-3254. doi: 10.1007/s10461-013-0462-8.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Niels Peek and Pedro Pereira Rodrigues. Three controversies in health data science. *International Journal of Data Science and Analytics*, 6:261–269, 2018. ISSN 2364-4168. doi: https://doi.org/10.1007/s41060-018-0109-y.

Ricardo Cardoso Pereira, Pedro Henriques Abreu, Pedro Pereira Rodrigues, and Mário A.T. Figueiredo. Imputation of data missing not at random: Artificial generation and benchmark analysis. *Expert Systems with Applications*, 249:123654, 2024. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2024.123654.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995. doi: 10.1080/01621459.1995.10476493.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 12 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.581.

Santtu Tikka, Jouni Helske, and Juha Karvanen. Clustering and structural robustness in causal diagrams. *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435.

World Health Organization. Hiv and aids. URL `https://www.who.int/news-room/fact-sheets/detail/hiv-aids`.

Werner Wothke. Longitudinal and multigroup modeling with missing data. In *Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches, and Specific Examples*. Psychology Press, 2000. doi: 10.4324/9781410601940.

A Proofs

A.1 Proof of Proposition 1

Proof. Let $\mathcal{G}^m \in compatible(\mathcal{G}^{\mathbb{C},m})$. We show that $\mathcal{G} \in compatible(\mathcal{G}^{\mathbb{C},cm})$, where $\mathcal{G}^{\mathbb{C},cm} = \pi(\mathcal{G}^{\mathbb{C},m})$.

By definition, π is obtained by merging, for each cluster $C_{\mathbb{X}} \in \mathbb{C}$, all variable-level missingness indicators

$$\{R_X : X \in C_{\mathbb{X}}\}$$

into a single cluster-level missingness indicator $R_{C_{\mathbb{X}}}$. All cluster-level adjacencies and orientations induced by the substantive variables are preserved, while distinctions between missingness indicators inside the same cluster are forgotten.

Since $\mathcal{G}^m$ is compatible with $\mathcal{G}^{\mathbb{C},m}$, every variable-level edge in $\mathcal{G}^m$ induces an edge allowed by $\mathcal{G}^{\mathbb{C},m}$. After applying π, any edge involving a variable-level missingness indicator R_X, with $X \in C_{\mathbb{X}}$, is mapped to the corresponding edge involving $R_{C_{\mathbb{X}}}$. Thus, every edge of $\mathcal{G}^m$ also induces an edge allowed by $\mathcal{G}^{\mathbb{C},cm}$. Therefore,

$$\mathcal{G}^m \in compatible(\mathcal{G}^{\mathbb{C},cm}).$$

As this holds for every $\mathcal{G} \in compatible(\mathcal{G}^m)$, we obtain

$$compatible(\mathcal{G}^{\mathbb{C},m}) \subseteq compatible(\mathcal{G}^{\mathbb{C},cm}).$$

It remains to show that the inclusion is strict in general. Consider a cluster $C_{\mathbb{X}} = \{X_1, X_2\}$. In an m-C-DMG, the missingness mechanisms of X_1 and X_2 are represented by two distinct vertices, R_{X_1} and R_{X_2}. Therefore, the graph can distinguish, for instance, a variable-level m-ADMG in which some variable L points to R_{X_1} but not to R_{X_2}.

After applying π, both R_{X_1} and R_{X_2} are merged into a single vertex $R_{C_{\mathbb{X}}}$. The cm-C-DMG only records that L may affect the missingness mechanism of the cluster $C_{\mathbb{X}}$, without specifying whether this effect is on R_{X_1}, on R_{X_2}, or on both. Hence a variable-level m-ADMG in which L points to R_{X_2} but not to R_{X_1} is compatible with the cm-C-DMG, but not with the original m-C-DMG if the latter allowed only the edge $L \to R_{X_1}$.

Thus, there exist variable-level m-ADMGs compatible with $\mathcal{G}^{cm}$ that are not compatible with $\mathcal{G}^m$. Therefore, the inclusion is strict in general:

$$compatible(\mathcal{G}^{\mathbb{C},m}) \subsetneq compatible(\mathcal{G}^{\mathbb{C},cm}).$$

□

A.2 Proof of Proposition 2

Proof. By Proposition 1, $compatible(\mathcal{G}^m_{\mathbb{C}}) \subseteq compatible(\mathcal{G}^{\mathbb{C},cm})$. Assume that Q is recoverable in $\mathcal{G}^{\mathbb{C},cm}$. By definition, this means that Q is uniquely determined from the observed distribution for all variable-level m-ADMGs in $compatible(\mathcal{G}^{\mathbb{C},cm})$. Since $compatible(\mathcal{G}^m_{\mathbb{C}})$ is a subset of this compatibility class, the same uniqueness holds for all variable-level m-ADMGs compatible with $\mathcal{G}^m_{\mathbb{C}}$. Therefore, Q is recoverable in $\mathcal{G}^m_{\mathbb{C}}$.

We now show that the converse does not hold in general. Let $C = \{A, B\}$ be a cluster and consider the target query

$$Q = P(A).$$

Suppose that, in the m-C-DMG, the cluster C points to the missingness indicator R_B, but not to R_A:

$$C \to R_B, \qquad C \not\to R_A.$$

Thus, the m-C-DMG distinguishes the missingness mechanism of A from that of B. In particular, it records that the missingness of A is not affected by variables in C. Under this abstraction, $P(A)$ is recoverable from the observed distribution.

In the associated cm-C-DMG, however, the indicators R_A and R_B are merged into a single vertex R_C. The edge $C \to R_B$ is therefore abstracted as

$$C \to R_C.$$

This coarser graph no longer specifies whether the missingness mechanism affected by C is that of A, that of B, or both. Hence it is compatible with variable-level m-ADMGs in which the missingness of A depends on A itself. In such cases, $P(A)$ is generally not recoverable because of self-censoring.

Therefore, $P(A)$ can be recoverable in the finer m-C-DMG but not in the associated cm-C-DMG. Hence the converse implication fails in general. □

A.3 Proof of Theorem 1

We first prove the result for cm-C-DMGs.

Definition 5 (Primary path [Ferreira and Assaad, 2025])**.** *Let $\mathcal{G}^* = (\mathbb{V}^*, \mathbb{E}^*)$ be a graph and let $\widetilde{\pi} = \langle V_1^*, \dots, V_n^* \rangle$ be a walk. The primary path of $\widetilde{\pi}$, denoted by $\widetilde{\pi}_p$, is defined iteratively as follows. Set $U_1^* = V_1^*$. For every $k \geq 1$, define $U_{k+1}^* = V^*_{\max i: V_i^* = U_k^* + 1}$, until the last vertex V_n^* is reached. The edge between U_k^* and U_{k+1}^* in $\widetilde{\pi}_p$ has the same endpoint marks as the edge between $V^*_{\max i: V_i^* = U_k^*}$ and $V^*_{\max i: V_i^* = U_k^* + 1}$ in $\widetilde{\pi}$.*

Lemma 1. *Let $\mathcal{G}^{\mathbb{C},cm}$ be a cm-C-DMG. For every $C_{\mathbb{X}} \in \mathbb{C}^m$, there exists a collider path between $C_{\mathbb{X}}$ and $R_{C_{\mathbb{X}}}$ in $\mathcal{G}^{\mathbb{C},cm}$, with every intermediate vertex in $\mathbb{C}^m \cup \mathbb{C}^o$, if and only if there exist a compatible variable-level m-ADMG $\mathcal{G}^m$ and some $X \in C_{\mathbb{X}} \cap \mathbb{V}^m$ such that $\mathcal{G}^m$ contains a path between X and R_X whose intermediate vertices are substantive colliders.*

Proof. Let $\rho(V) = C_V$ denote the clustering map for substantive variables, and let $\rho(R_V) = R_{C_V}$ for their missingness indicators. By compatibility, a variable-level edge between vertices in two distinct clusters induces an edge between the corresponding cluster vertices with the same endpoint marks. Conversely, in every compatible variable-level graph, each cluster-level edge has at least one variable-level witness between its endpoint clusters with the same endpoint marks. We prove the two implications separately.

($\Leftarrow$) Suppose that there exist a variable-level m-ADMG $\mathcal{G}^m$ compatible with $\mathcal{G}^{\mathbb{C},cm}$ and some $X \in C_{\mathbb{X}} \cap \mathbb{V}^m$ such that $\mathcal{G}^m$ contains a path

$$\pi = \langle X, V_1, \dots, V_k, R_X \rangle$$

whose intermediate vertices $V_1, \dots, V_k$ are substantive colliders.

Apply ρ to the vertices of π and contract every maximal consecutive subsequence of identical cluster vertices to one occurrence. Denote the resulting cluster-level sequence by

$$\rho(\pi) = \langle D_0 = C_{\mathbb{X}}, D_1, \dots, D_{m-1}, D_m = R_{C_{\mathbb{X}}} \rangle$$

in $\mathcal{G}^{\mathbb{C},cm}$. This sequence is a walk rather than necessarily a path because distinct variable-level vertices may belong to the same cluster. Let

$$\rho(\pi)_p = \langle U_0, U_1, \dots, U_\ell \rangle$$

be the primary path associated with $\rho(\pi)$. Its endpoints are $U_0 = C_{\mathbb{X}}$ and $U_\ell = R_{C_{\mathbb{X}}}$.

We show that every intermediate vertex U_j, $1 \leq j \leq \ell - 1$, is a collider. Fix such a vertex U_j. Let s_j and t_j denote, respectively, the first and last occurrences of U_j in the portion of $\rho(\pi)$ retained by the primary-path construction. The edge between U_{j-1} and U_j in the primary path is inherited from the edge entering the first occurrence of U_j, whereas the edge between U_j and U_{j+1} is inherited from the edge leaving the last occurrence of U_j. Let V_{s_j} and V_{t_j} be the corresponding variable-level vertices on π. Since U_j is an intermediate vertex of the primary path, V_{s_j} and V_{t_j} are intermediate substantive vertices of π. By assumption, every intermediate vertex of π is a collider. Hence, the edge entering V_{s_j} has an arrowhead at V_{s_j}, and the edge leaving V_{t_j} has an arrowhead at V_{t_j}. Compatibility preserves endpoint marks under clustering. Therefore, the edge entering U_j in $\rho(\pi)_p$ has an arrowhead at U_j, and the edge leaving U_j also has an arrowhead at U_j. Thus, locally,

$$U_{j-1} * \!\!\rightarrow U_j \leftarrow\!\! * U_{j+1},$$

where $*$ denotes either a tail or an arrowhead at the other endpoint. Consequently, U_j is a collider. This argument applies to every intermediate vertex of $\rho(\pi)_p$. Moreover, each such vertex is the

cluster of substantive variables and therefore belongs to $\mathbb{C}^m \cup \mathbb{C}^o$. Hence, $\rho(\pi)_p$ is a path between $C_{\mathbb{X}}$ and $R_{C_{\mathbb{X}}}$ whose intermediate vertices are substantive cluster-level colliders.

($\Rightarrow$) Suppose that $\mathcal{G}^{\mathbb{C},cm}$ contains a path

$$\pi_{\mathbb{C}} = \langle C_0, C_1, \ldots, C_k, C_{k+1} \rangle,$$

where

$$C_0 = C_{\mathbb{X}}, \qquad C_{k+1} = R_{C_{\mathbb{X}}},$$

and every intermediate vertex

$$C_i \in \mathbb{C}^o \cup \mathbb{C}^m, \qquad 1 \le i \le k,$$

is a collider on $\pi_{\mathbb{C}}$.

Choose a partially observed variable $X \in C_{\mathbb{X}} \cap \mathbb{V}^m$, and let $R_X \in R_{C_{\mathbb{X}}}$ be its missingness indicator.

Because $\mathcal{G}^{\mathbb{C},cm}$ admits at least one compatible variable-level m-ADMG, choose $\mathcal{G}_0^m \in compatible(\mathcal{G}^{\mathbb{C},cm})$.

We modify $\mathcal{G}_0^m$ so that it contains a variable-level counterpart of $\pi_{\mathbb{C}}$.

The case $k = 0$ In this case, $C_{\mathbb{X}}$ and $R_{C_{\mathbb{X}}}$ are adjacent. Because a missingness indicator has no substantive children, their edge is either $C_{\mathbb{X}} \to R_{C_{\mathbb{X}}}$ or $C_{\mathbb{X}} \dashleftarrow\dashrightarrow R_{C_{\mathbb{X}}}$. In the first case, add $X \to R_X$ to $\mathcal{G}_0^m$ if it is not already present. This cannot create a directed cycle because no directed path from R_X can reach the substantive variable X, since missingness indicators are not parents of substantive or latent variables.

In the second case, add $X \dashleftarrow\dashrightarrow R_X$ if it is not already present. This does not alter the directed part of the graph.

In either case, the added edge projects to an edge already present in $\mathcal{G}^{\mathbb{C},cm}$. Hence, the resulting graph remains compatible and contains the required length-one path between X and R_X.

The case $k \ge 1$ Because C_1 is a collider on $\pi_{\mathbb{C}}$, the first edge has an arrowhead at C_1. Therefore, it is either $C_0 \to C_1$ or $C_0 \dashleftarrow\dashrightarrow C_1$.

We construct the first variable-level edge separately in these two cases.

Suppose first that $C_0 \to C_1$. Since $\mathcal{G}_0^m$ is compatible with $\mathcal{G}^{\mathbb{C},cm}$, there exist substantive variables

$$A \in C_0 \qquad \text{and} \qquad B \in C_1$$

such that $A \to B$ is an edge of $\mathcal{G}_0^m$.

Let σ be the permutation of the variable-level vertices that exchanges A and X and fixes every other vertex. Let $\mathcal{G}_1^m$ be obtained by applying σ to both endpoints of every edge of $\mathcal{G}_0^m$.

The graphs $\mathcal{G}_0^m$ and $\mathcal{G}_1^m$ are isomorphic, so the directed part of $\mathcal{G}_1^m$ is acyclic. Moreover, A and X belong to the same cluster. Consequently, applying σ does not change the cluster-level projection of any edge, and therefore $\mathcal{G}_1^m \in compatible(\mathcal{G}^{\mathbb{C},cm})$. The edge $A \to B$ is transformed into $X \to B$. Set $V_1 = B$.

Suppose instead that $C_0 \dashleftarrow\dashrightarrow C_1$. Set $\mathcal{G}_1^m = \mathcal{G}_0^m$, choose any substantive variable $V_1 \in C_1$, and add $X \dashleftarrow\dashrightarrow V_1$ if it is not already present. Since bows are permitted, this addition is valid even if a directed edge already joins X and V_1. The added bidirected edge does not alter the directed part of the graph and projects to the existing cluster-level edge $C_0 \dashleftarrow\dashrightarrow C_1$.

We now construct the remaining edges of the path.

For every $1 \le i \le k-1$, both C_i and C_{i+1} are colliders on $\pi_{\mathbb{C}}$. Hence, the edge between them must have arrowheads at both endpoints and is therefore bidirected:

$$C_i \dashleftarrow\dashrightarrow C_{i+1}.$$

Recursively choose any substantive variable

$$V_{i+1} \in C_{i+1}$$

and add the bidirected edge

$$V_i \dashleftarrow\dashrightarrow V_{i+1}$$

if it is not already present.

Finally, consider the edge between C_k and $R_{C_{\mathbb{X}}}$. Because C_k is a collider, this edge has an arrowhead at C_k. It cannot be of the form

$$R_{C_{\mathbb{X}}} \to C_k,$$

because missingness indicators have no substantive children. It also cannot be of the form

$$C_k \to R_{C_{\mathbb{X}}},$$

because that edge has a tail at C_k. Therefore, the final edge must be

$$C_k \dashleftarrow\dashrightarrow R_{C_{\mathbb{X}}}.$$

Add

$$V_k \dashleftarrow\dashrightarrow R_X$$

if it is not already present.

Let $\mathcal{G}^m$ be the graph obtained from $\mathcal{G}_1^m$ after these additions. Every added edge is bidirected, so the directed part of $\mathcal{G}^m$ is identical to the directed part of $\mathcal{G}_1^m$ and is therefore acyclic. The only added edge incident to a missingness indicator has an arrowhead at that indicator, so no substantive child of a missingness indicator is introduced. Since bows are permitted, the additions preserve the m-ADMG restrictions.

Furthermore, each added edge projects to an edge already present on $\pi_{\mathbb{C}}$. Thus, no new cluster-level edge is introduced. Because $\mathcal{G}_1^m$ was already compatible with $\mathcal{G}^{\mathbb{C},cm}$, it follows that

$$\mathcal{G}^m \in compatible(\mathcal{G}^{\mathbb{C},cm}).$$

By construction, $\mathcal{G}^m$ contains the path

$$\pi = \langle X, V_1, \ldots, V_k, R_X \rangle.$$

The first edge has an arrowhead at V_1, every edge between consecutive intermediate vertices is bidirected, and the final edge has an arrowhead at V_k. Consequently, for every $1 \le i \le k$, both edges incident to V_i on π have arrowheads at V_i. Hence, every intermediate vertex V_i is a substantive collider.

Therefore, there exist a compatible variable-level m-ADMG $\mathcal{G}^m$ and a partially observed variable

$$X \in C_{\mathbb{X}} \cap \mathbb{V}^m$$

such that $\mathcal{G}^m$ contains a path between X and R_X whose intermediate vertices are substantive colliders. □

We now prove the theorem.

Proof. Let $\mathcal{G}^{\mathbb{C},cm} = \left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}, \mathbb{E}^{\mathbb{C}}\right)$ be a cm-C-DMG. Suppose first that there exists $C_{\mathbb{X}} \in \mathbb{C}^m$ such that $C_{\mathbb{X}}$ and $R_{C_{\mathbb{X}}}$ are neighbors.

By the definition of compatibility, we may construct a compatible variable-level m-ADMG $\mathcal{G}^m$ containing some $X \in C_{\mathbb{X}}$ and its missingness indicator $R_X \in R_{C_{\mathbb{X}}}$ such that X and R_X are adjacent.

The variable-level necessary condition of Mohan and Pearl [2014] is violated in this compatible graph. Therefore, the joint distribution is not recoverable in $\mathcal{G}^m$. Since the cm-C-DMG represents a class containing this compatible graph, the joint distribution cannot be recoverable uniformly over the class represented by $\mathcal{G}^{\mathbb{C},cm}$.

Suppose next that there exists $C_{\mathbb{X}} \in \mathbb{C}^m$ such that $C_{\mathbb{X}}$ and $R_{C_{\mathbb{X}}}$ are connected by a path whose intermediate vertices are colliders belonging to $\mathbb{C}^m \cup \mathbb{C}^o$.

By Lemma 1, there exist a compatible variable-level m-ADMG $\mathcal{G}^m$ and some $X \in C_{\mathbb{X}}$ such that X and R_X are connected by a path whose intermediate vertices are substantive colliders.

The second variable-level necessary condition of Mohan and Pearl [2014] is therefore violated. Consequently, the joint distribution is not recoverable in $\mathcal{G}^m$ and cannot be recoverable uniformly over all variable-level m-ADMGs compatible with $\mathcal{G}^{\mathbb{C},cm}$.

Thus, the following conditions are necessary for recoverability:

1. for every $C_{\mathbb{X}} \in \mathbb{C}^m$, $C_{\mathbb{X}}$ and $R_{C_{\mathbb{X}}}$ are not neighbors;
2. for every $C_{\mathbb{X}} \in \mathbb{C}^m$, there is no path between $C_{\mathbb{X}}$ and $R_{C_{\mathbb{X}}}$ whose intermediate vertices are colliders belonging to $\mathbb{C}^m \cup \mathbb{C}^o$.

Extension to m-C-DMGs The proof for m-C-DMGs follows the same argument. It follows by applying the same strategy to the missingness-indicator vertices of the m-C-DMG, Therefore, the stated graphical conditions are necessary for recovering the joint distribution in both cm-C-DMGs and m-C-DMGs. □

A.4 Proof of Theorem 2

Definition 6 (Missingness-indicator blocks)**.** *Let $\mathcal{G}^{\mathbb{C},cm} = \left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}, \mathbb{E}^{\mathbb{C}}\right)$ be a cm-C-DMG. Construct an undirected auxiliary graph $\mathcal{H}_{\mathbb{R}}$ whose vertex set is $\mathbb{R}_{\mathbb{C}}$. Two distinct missingness-indicator vertices R_{C_i} and R_{C_j} are adjacent in $\mathcal{H}_{\mathbb{R}}$ if there exists a collider path between them in $\mathcal{G}^{\mathbb{C},cm}$. Let $\mathbb{R}^{(1)}, \ldots, \mathbb{R}^{(K)}$ be the connected components of $\mathcal{H}_{\mathbb{R}}$. We call these sets the missingness-indicator blocks of $\mathcal{G}^{\mathbb{C},cm}$. For every block $\mathbb{R}^{(a)}$, define*

$$\mathbb{M}_a = \left\{ C \in \mathbb{C}^o \cup \mathbb{C}^m : \begin{array}{l} \textit{there exist } R_{C_i} \in \mathbb{R}^{(a)} \textit{ and a collider path} \\ \textit{between } R_{C_i} \textit{ and } C \textit{ in } \mathcal{G}^{\mathbb{C},cm} \end{array} \right\}.$$

Paths of length one are included. Thus, every substantive cluster adjacent to a member of $\mathbb{R}^{(a)}$ belongs to $\mathbb{M}_a$.

Lemma 2 (Cluster Markov blanket of a missingness block)**.** *Let $\mathbb{R}^{(a)}$ be a missingness-indicator block and let $\mathbb{M}_a$ be defined as in Definition 6. If there are no edges between missingness-indicator vertices then*

$$\mathbb{R}^{(a)} \perp\!\!\!\perp_{\mathcal{G}^{\mathbb{C},cm}} \left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}\right) \setminus \left(\mathbb{R}^{(a)} \cup \mathbb{M}_a\right) \mid \mathbb{M}_a.$$

Consequently, $\mathbb{M}_a$ is a substantive cluster Markov blanket of the missingness-indicator block $\mathbb{R}^{(a)}$.

Proof. Fix arbitrary vertices $R_{C_i} \in \mathbb{R}^{(a)}$ and $Z \in \left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}\right) \setminus \left(\mathbb{R}^{(a)} \cup \mathbb{M}_a\right)$. Consider an arbitrary path

$$\pi = \langle R_{C_i} = Z_0, Z_1, \ldots, Z_q = Z \rangle$$

between R_{C_i} and Z. We show that π is blocked by $\mathbb{M}_a$.

Suppose first that every intermediate vertex $Z_1, \ldots, Z_{q-1}$ is a collider. Then π is a collider path.

If Z is a substantive cluster, then the definition of $\mathbb{M}_a$ implies $Z \in \mathbb{M}_a$, contradicting the choice of Z.

If Z is a missingness-indicator vertex, then R_{C_i} and Z are adjacent in the auxiliary graph $\mathcal{H}_{\mathbb{R}}$. Hence, they belong to the same connected component, and therefore $Z \in \mathbb{R}^{(a)}$, again contradicting the choice of Z.

It follows that π contains at least one intermediate noncollider.

Let Z_j (such that $1 \le j \le q-1$) be the first noncollider encountered when traversing π from R_{C_i} toward Z. Thus,

$$Z_1, \ldots, Z_{j-1}$$

are colliders and Z_j is a noncollider.

We first show that Z_j is a substantive cluster.

Suppose, toward a contradiction, that $Z_j \in \mathbb{R}_{\mathbb{C}}$. Since π is a path and there are no edges between distinct missingness-indicator vertices, both Z_{j-1} and Z_{j+1} are substantive clusters. Moreover, missingness-indicator vertices have no substantive children. Therefore, neither incident edge can have a tail at Z_j. Both edges incident to Z_j on π consequently have arrowheads at Z_j. Thus, Z_j is a collider, contradicting its definition. Hence, $Z_j \in \mathbb{C}^o \cup \mathbb{C}^m$. Since $Z_1, \ldots, Z_{j-1}$ are colliders, the subpath

$$\langle R_{C_i}, Z_1, \ldots, Z_j \rangle$$

is a collider path between a member of $\mathbb{R}^{(a)}$ and the substantive cluster Z_j. By the definition of $\mathbb{M}_a$, $Z_j \in \mathbb{M}_a$.

The vertex Z_j is a noncollider on π and belongs to the conditioning set $\mathbb{M}_a$. Therefore, π is blocked by $\mathbb{M}_a$.

Since the argument applies to every path from every member of $\mathbb{R}^{(a)}$ to every vertex outside $\mathbb{R}^{(a)} \cup \mathbb{M}_a$, we obtain

$$\mathbb{R}^{(a)} \perp\!\!\!\perp_{\mathcal{G}^{\mathbb{C},cm}} \left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}\right) \setminus \left(\mathbb{R}^{(a)} \cup \mathbb{M}_a\right) \mid \mathbb{M}_a.$$

□

Lemma 3 (Transfer of the cluster Markov blanket)**.** *Let $\mathbb{R}^{(a)}$ be a missingness-indicator block and let $\mathbb{M}_a$ be its substantive cluster Markov blanket as defined above. For a collection $\mathbb{A}$ of cluster vertices, let $\mathrm{var}(\mathbb{A})$ denote the union of all variable-level vertices represented by the clusters in $\mathbb{A}$. Define*

$$\mathbf{R}^{(a)} = \mathrm{var}\left(\mathbb{R}^{(a)}\right)$$

and

$$\mathbf{M}_a = \mathrm{var}\left(\mathbb{M}_a\right).$$

Then, in every variable-level m-ADMG $\mathcal{G}^m$ compatible with $\mathcal{G}^{\mathbb{C},cm}$,

$$\mathbf{R}^{(a)} \perp\!\!\!\perp_{\mathcal{G}^m} \mathrm{var}\left(\left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}\right) \setminus \left(\mathbb{R}^{(a)} \cup \mathbb{M}_a\right)\right) \mid \mathbf{M}_a.$$

Furthermore, if P satisfies the global Markov property with respect to $\mathcal{G}^m$, then

$$\mathbf{R}^{(a)} \perp\!\!\!\perp_{P} \mathrm{var}\left(\left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}\right) \setminus \left(\mathbb{R}^{(a)} \cup \mathbb{M}_a\right)\right) \mid \mathbf{M}_a.$$

Consequently, $\mathbf{M}_a$ is a variable-level Markov blanket of the missingness-indicator block $\mathbf{R}^{(a)}$ in every compatible variable-level m-ADMG.

Proof. By Lemma 2,

$$\mathbb{R}^{(a)} \perp\!\!\!\perp_{\mathcal{G}^{\mathbb{C},cm}} \left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}\right) \setminus \left(\mathbb{R}^{(a)} \cup \mathbb{M}_a\right) \mid \mathbb{M}_a.$$

Notice that the sets

$$\mathbb{R}^{(a)}, \qquad \left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}\right) \setminus \left(\mathbb{R}^{(a)} \cup \mathbb{M}_a\right), \qquad \mathbb{M}_a$$

are pairwise disjoint.

The soundness of cluster d-separation [Ferreira and Assaad, 2025] states that if three disjoint collections $\mathbb{A}, \mathbb{B}, \mathbb{D}$ of cluster vertices are d-separated in the cluster graph, then the unions of the variable-level vertices represented by those collections are d-separated in every compatible variable-level graph. Applying soundness with

$$\mathbb{A} = \mathbb{R}^{(a)},$$

$$\mathbb{B} = \left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}\right) \setminus \left(\mathbb{R}^{(a)} \cup \mathbb{M}_a\right),$$

and

$$\mathbb{D} = \mathbb{M}_a$$

gives, in every compatible variable-level m-ADMG $\mathcal{G}^m$,

$$\mathrm{var}\left(\mathbb{R}^{(a)}\right) \perp\!\!\!\perp_{\mathcal{G}^m} \mathrm{var}\left(\left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}\right) \setminus \left(\mathbb{R}^{(a)} \cup \mathbb{M}_a\right)\right) \mid \mathrm{var}\left(\mathbb{M}_a\right).$$

Using the notation

$$\mathbf{R}^{(a)} = \mathrm{var}\left(\mathbb{R}^{(a)}\right)$$

and

$$\mathbf{M}_a = \mathrm{var}\left(\mathbb{M}_a\right),$$

this becomes

$$\mathbf{R}^{(a)} \perp\!\!\!\perp_{\mathcal{G}^m} \mathrm{var}\left(\left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}\right) \setminus \left(\mathbb{R}^{(a)} \cup \mathbb{M}_a\right)\right) \mid \mathbf{M}_a.$$

Now fix an arbitrary distribution P satisfying the global Markov property with respect to $\mathcal{G}^m$. The preceding graphical d-separation entails the probabilistic conditional independence

$$\mathbf{R}^{(a)} \perp\!\!\!\perp_{P} \mathrm{var}\left(\left(\mathbb{C}^o \cup \mathbb{C}^m \cup \mathbb{R}_{\mathbb{C}}\right) \setminus \left(\mathbb{R}^{(a)} \cup \mathbb{M}_a\right)\right) \mid \mathbf{M}_a.$$

Since $\mathcal{G}^m$ was arbitrary, the conclusion holds in every variable-level m-ADMG compatible with $\mathcal{G}^{\mathbb{C},cm}$. □

We now prove the theorem.

Proof. Let $\mathcal{G}^m$ be an arbitrary variable-level m-ADMG compatible with $\mathcal{G}^{\mathbb{C},cm}$, and let P be an arbitrary distribution satisfying the global Markov property with respect to $\mathcal{G}^m$.

For every missingness-indicator block $\mathbb{R}^{(a)}$, let

$$\mathbf{R}^{(a)} = \text{var}\left(\mathbb{R}^{(a)}\right)$$

and

$$\mathbf{M}_a = \text{var}\left(MB^o\left(\mathbb{R}^{(a)}\right) \cup MB^m\left(\mathbb{R}^{(a)}\right)\right).$$

Write

$$\mathbf{M}_a^o = \text{var}\left(MB^o\left(\mathbb{R}^{(a)}\right)\right)$$

and

$$\mathbf{M}_a^m = \text{var}\left(MB^m\left(\mathbb{R}^{(a)}\right)\right).$$

By Lemma 3,

$$\mathbf{R}^{(a)} \perp\!\!\!\perp_P \left[\left(\mathbb{V}^o \cup \mathbb{V}^m\right) \setminus \mathbf{M}_a\right] \cup \left[\mathbb{R} \setminus \mathbf{R}^{(a)}\right] \mid \mathbf{M}_a. \tag{3}$$

Choose an arbitrary ordering

$$\mathbb{R}^{(1)}, \ldots, \mathbb{R}^{(K)}$$

of the missingness-indicator blocks. By the chain rule,

$$P\left(\mathbb{R} \mid \mathbb{V}^o, \mathbb{V}^m\right) = \prod_{a=1}^{K} P\left(\mathbf{R}^{(a)} \mid \mathbb{V}^o, \mathbb{V}^m, \mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(a-1)}\right). \tag{4}$$

For each a, all variables in the conditioning set of the corresponding factor in (4), except those in $\mathbf{M}_a$, belong to the set on the right-hand side of the conditional independence in (3). Therefore, by decomposition of (3),

$$P\left(\mathbf{R}^{(a)} \mid \mathbb{V}^o, \mathbb{V}^m, \mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(a-1)}\right) = P\left(\mathbf{R}^{(a)} \mid \mathbf{M}_a\right). \tag{5}$$

Substituting (5) into (4) gives

$$P\left(\mathbb{R} \mid \mathbb{V}^o, \mathbb{V}^m\right) = \prod_{a=1}^{K} P\left(\mathbf{R}^{(a)} \mid \mathbf{M}_a\right). \tag{6}$$

Evaluating (6) at $\mathbb{R} = 0$ yields

$$P\left(\mathbb{R} = 0 \mid \mathbb{V}^o, \mathbb{V}^m\right) = \prod_{a=1}^{K} P\left(\mathbf{R}^{(a)} = 0 \mid \mathbf{M}_a^o, \mathbf{M}_a^m\right). \tag{7}$$

By assumption,

$$\mathbb{R}^{(a)} \cap R_{MB^m\left(\mathbb{R}^{(a)}\right)} = \varnothing.$$

Consequently, the missingness indicators associated with the partially observed clusters in $MB^m\left(\mathbb{R}^{(a)}\right)$ belong to $\mathbb{R} \setminus \mathbf{R}^{(a)}$. By the conditional independence in (3),

$$\mathbf{R}^{(a)} \perp\!\!\!\perp_{\cdot} R_{\mathbf{M}_a^m} \mid \mathbf{M}_a^o, \mathbf{M}_a^m$$

It follows that

$$P\left(\mathbf{R}^{(a)} = 0 \mid \mathbf{M}_a^o, \mathbf{M}_a^m\right) = P\left(\mathbf{R}^{(a)} = 0 \mid \mathbf{M}_a^o, \mathbf{M}_a^m, R_{\mathbf{M}_a^m} = 0\right). \tag{8}$$

Substituting (8) into (7) gives

$$P\left(\mathbb{R}=0 \mid \mathbb{V}^o, \mathbb{V}^m\right)=\prod_{a=1}^{K} P\left(\mathbf{R}^{(a)}=0 \mid \mathbf{M}_a^o, \mathbf{M}_a^m, R_{\mathbf{M}_a^m}=0\right). \tag{9}$$

By the definition of conditional probability,

$$P\left(\mathbb{R}=0, \mathbb{V}^o, \mathbb{V}^m\right)=P\left(\mathbb{V}^o, \mathbb{V}^m\right) P\left(\mathbb{R}=0 \mid \mathbb{V}^o, \mathbb{V}^m\right).$$

Using (9) and the assumed positivity condition, we obtain

$$P\left(\mathbb{V}^o, \mathbb{V}^m\right)=\frac{P\left(\mathbb{R}=0, \mathbb{V}^o, \mathbb{V}^m\right)}{\prod_{a=1}^{K} P\left(\mathbf{R}^{(a)}=0 \mid \mathbf{M}_a^o, \mathbf{M}_a^m, R_{\mathbf{M}_a^m}=0\right)}. \tag{10}$$

The numerator in (10) is observable from complete cases. In every denominator factor, the event

$$R_{\mathbf{M}_a^m}=0$$

ensures that all partially observed variables in $\mathbf{M}_a^m$ equal their observed proxies. Hence, each denominator factor is a functional of the observed-data distribution.

The blocks and their substantive Markov blankets are defined entirely from $\mathcal{G}^{\mathbb{C},cm}$. Moreover, the conditional independences used above hold in every compatible variable-level m-ADMG by Lemma 3. Therefore, the resulting observed-data functional is the same for every variable-level m-ADMG compatible with $\mathcal{G}^{\mathbb{C},cm}$.

Rewriting (10) in cluster notation gives

$$\Pr(\mathbb{c})=\frac{\Pr\left(\mathbb{R}_{\mathbb{C}}=0, \mathbb{c}\right)}{\prod_{a=1}^{K} \Pr\left(\mathbb{R}^{(a)}=0 \mid MB^o\left(\mathbb{R}^{(a)}\right), MB^m\left(\mathbb{R}^{(a)}\right), R_{MB^m\left(\mathbb{R}^{(a)}\right)}=0\right)}.$$

Thus, $\Pr(\mathbb{C})$ is recoverable uniformly over all variable-level m-ADMGs compatible with $\mathcal{G}^{\mathbb{C},cm}$.

Extension to m-C-DMGs The proof for m-C-DMGs follows the same argument. It follows by projecting forbidden variable-level paths to the m-C-DMG and applying the soundness of cluster d-separation to its cluster Markov blankets. Therefore, the stated graphical conditions are sufficient for recovering the joint distribution in both cm-C-DMGs and m-C-DMGs. □

A.5 Proof of Proposition 3

Proof. Let $\mathcal{G}_{\mathbb{C}}^m$ be an m-C-DMG and let $\mathcal{G}^{\mathbb{C},cm}=\pi(\mathcal{G}_{\mathbb{C}}^m)$ be the associated cm-C-DMG.

Theorem 1 is satisfied in $\mathcal{G}^{\mathbb{C},cm}$ ⇒ Theorem 1 is satisfied in $\mathcal{G}^{\mathbb{C},m}$

This implication is given by Proposition 2.

Theorem 1 is satisfied in $\mathcal{G}^{\mathbb{C},cm}$ ⇐ Theorem 1 is satisfied in $\mathcal{G}^{\mathbb{C},m}$

Equivalently, we prove the contrapositive. Suppose that the conditions of Theorem 1 are not satisfied in $\mathcal{G}^{\mathbb{C},cm}$. Then, there exists a cluster $C_{\mathbb{X}} \in \mathbb{C}^m$ such that $C_{\mathbb{X}}$ and its cluster-level missingness vertex $R_{C_{\mathbb{X}}}$ are neighbors or are connected by a path whose intermediate vertices are colliders and belong to $\mathbb{C}^m \cup \mathbb{C}^o$.

Suppose $C_{\mathbb{X}}$ is adjacent to $R_{C_{\mathbb{X}}}$ in the cm-C-DMG. Since $R_{C_{\mathbb{X}}}$ is obtained by merging the missingness vertices associated with variables in $C_{\mathbb{X}}$, this adjacency must come from an adjacency in the m-C-DMG between $C_{\mathbb{X}}$ and at least one of its variable-level missingness vertices, say R_i. But then, in $\mathcal{G}_{\mathbb{C}}^m$, the cluster $C_{\mathbb{X}}$ and its missingness vertex R_i are neighbors. Thus the first condition of Theorem 1 is violated in $\mathcal{G}_{\mathbb{C}}^m$.

Suppose that $C_{\not\perp}$ is connected to $R_{C_{\not\perp}}$ via a path in which all intermediate vertices are colliders and elements of $\mathbb{C}^m \cup \mathbb{C}^o$ in the cm-C-DMG. By the definition of a cm-C-DMG, all compatible m-C-DMGs must contain at least one variable $R_{X_i} \in R_{C_{\not\perp}}$ connected to $C_{\not\perp}$ via such a path. Violation of the second condition of Theorem 1 in a cm-C-DMG thus implies violation in all compatible m-C-DMGs.

Therefore,

$$\text{Theorem 1 not satisfied in } \mathcal{G}^{\mathbb{C},cm} \Rightarrow \text{ Theorem 1 not satisfied in } \mathcal{G}^{m}_{\mathbb{C}}.$$

Taking the contrapositive gives

$$\text{Theorem 1 is satisfied in } \mathcal{G}^{m}_{\mathbb{C}} \Rightarrow \text{ Theorem 1 is satisfied in } \mathcal{G}^{\mathbb{C},cm}.$$

Theorem 2 is satisfied in $\mathcal{G}^{\mathbb{C},cm}$ $\Rightarrow$ Theorem 2 is satisfied in $\mathcal{G}^{\mathbb{C},m}$

This implication is given by Proposition 2.

Theorem 2 is satisfied in $\mathcal{G}^{\mathbb{C},cm}$ $\not\Leftarrow$ Theorem 2 is satisfied in $\mathcal{G}^{\mathbb{C},m}$

The converse of the conditions of Theorem 2 fails because merging missingness indicators can merge distinct missingness blocks and thereby create a new violation of the block-intersection condition.

For example, consider an m-C-DMG with missingness-indicator blocks

$$\mathbb{R}^{(1)} = \{R_A, R_X\} \qquad \text{and} \qquad \mathbb{R}^{(2)} = \{R_B, R_Y\},$$

where X and Y belong to the same substantive cluster C. Suppose that the graph contains the collider paths

$$R_A \dashleftarrow\dashrightarrow C_{Z_1} \dashleftarrow\dashrightarrow R_X$$

and

$$R_B \dashleftarrow\dashrightarrow C_{Z_2} \dashleftarrow\dashrightarrow R_Y,$$

together with the edge

$$C_B \longrightarrow R_A,$$

and no other relevant edges.

In the m-C-DMG, C_B belongs to the Markov blanket of $\mathbb{R}^{(1)}$, but its indicator R_B belongs to the distinct block $\mathbb{R}^{(2)}$. Thus, the block-intersection condition can hold for both blocks.

After projection, R_X and R_Y are merged into the single vertex R_C. The projected graph contains the collider path

$$R_A \dashleftarrow\dashrightarrow C_{Z_1} \dashleftarrow\dashrightarrow R_C \dashleftarrow\dashrightarrow C_{Z_2} \dashleftarrow\dashrightarrow R_B.$$

Hence, R_A, R_C, and R_B belong to the same missingness-indicator block. Since $C_B \to R_A$, the cluster C_B belongs to the Markov blanket of this block, while R_B belongs to the block itself. The block-intersection condition therefore fails in the cm-C-DMG.

Thus, the sufficient conditions may hold in an m-C-DMG but fail in its associated cm-C-DMG. □

A.6 Proof of Theorem 3

Lemma 4 (Soundness of do-calculus in cluster missingness graphs)**.** *Let $\mathcal{G}^{\mathbb{C},*m}$ denote either an m-C-DMG or a cm-C-DMG, and let $compatible(\mathcal{G}^{\mathbb{C},*m})$ be the set of variable-level m-ADMGs compatible with it. Then any application of a rule of do-calculus in $\mathcal{G}^{\mathbb{C},*m}$ is valid in every $\mathcal{G} \in compatible(\mathcal{G}^{\mathbb{C},*m})$.*

Proof. Both m-C-DMGs and cm-C-DMGs are C-DMGs whose vertices represent either clusters of substantive variables, missingness indicators, missingness-indicator clusters, or proxy variables. Hence the soundness of do-calculus in C-DMGs established by Ferreira and Assaad [2025] applies directly.

More precisely, if a rule of do-calculus is applicable in $\mathcal{G}^{\mathbb{C},*m}$, then the corresponding d-separation condition holds in the relevant mutilated cluster graph. By the soundness result of Ferreira and Assaad [2025], this d-separation condition entails the corresponding d-separation condition in every

compatible variable-level m-ADMG. The standard soundness of do-calculus at the variable level then implies that the associated do-calculus equality is valid in every $\mathcal{G} \in compatible(\mathcal{G}^{\mathbb{C},*m})$.

Therefore, any do-calculus transformation licensed by $\mathcal{G}^{\mathbb{C},*m}$ is sound for all compatible variable-level missingness graphs. □

Now we proceed to the proof of the theorem.

Proof. Let $compatible(\mathcal{G}^{\mathbb{C},*m})$ denote the set of variable-level m-ADMGs compatible with $\mathcal{G}^{\mathbb{C},*m}$. Suppose that the query Q can be transformed into an expression F using do-calculus, probability manipulations, and Equation 1, where F contains no do-operators and involves only observable quantities.

By Lemma 4, each do-calculus step performed in $\mathcal{G}^{\mathbb{C},*m}$ is sound in every variable-level m-ADMG

$$\mathcal{G} \in compatible(\mathcal{G}^{\mathbb{C},*m}).$$

Moreover, the probability manipulations are algebraic identities and are therefore valid in every compatible model. Equation 1 is the consistency relation defining the proxy variables, and is also valid in every compatible missingness model.

Thus, the whole sequence of transformations from Q to F is valid in every variable-level m-ADMG compatible with $\mathcal{G}^{\mathbb{C},*m}$. Hence, for every such compatible model,

$$Q = F.$$

Since F contains no do-operators and involves only observable quantities, it is a functional of the observed-data distribution. Therefore, Q is uniquely determined from the observed-data distribution for all variable-level m-ADMGs compatible with $\mathcal{G}^{\mathbb{C},*m}$.

By definition, Q is recoverable from $\mathcal{G}^{\mathbb{C},*m}$. □

B Relationship between the necessary and sufficient graphical conditions

The block-intersection condition appearing in the sufficient recoverability result in Tehorem 2 implies the graphical conditions appearing in the necessary recoverability result in Theorem 1. This implication follows directly from the definition of the substantive Markov blanket of a missingness-indicator block. It holds both at the variable level, for m-ADMGs, and at the cluster level, for m-C-DMGs and cm-C-DMGs.

The converse does not hold in general. In particular, the block-intersection condition may fail because a partially observed substantive vertex is connected to its own missingness-indicator block through a collider path containing another missingness-indicator vertex. Such a path is not excluded by the necessary condition, which only considers paths whose intermediate vertices are substantive colliders.

Proposition 4 (The sufficient block condition implies the necessary conditions)**.** *Let $\mathcal{G}$ be either:*

1. *a variable-level m-ADMG, in which case let $\mathbb{A}^m = \mathbb{V}^m$ and associate with each $A = X \in \mathbb{A}^m$ its missingness indicator $R_A = R_X$; or*
2. *an m-C-DMG or a cm-C-DMG, in which case let $\mathbb{A}^m = \mathbb{C}^m$. For an m-C-DMG, associate with each partially observed variable X the pair (C_X, R_X), and for a cm-C-DMG associate with each $A = C \in \mathbb{C}^m$ the cluster-level missingness indicator $R_A = R_C$.*

Let $\mathbb{R}^{(1)}, \ldots, \mathbb{R}^{(K)}$ be the missingness-indicator blocks of $\mathcal{G}$, and let $MB^m(\mathbb{R}^{(i)})$ denote the partially observed substantive part of the Markov blanket of $\mathbb{R}^{(i)}$. Suppose that

$$\mathbb{R}^{(i)} \cap R_{MB^m(\mathbb{R}^{(i)})} = \varnothing \qquad \text{for every } i. \tag{11}$$

Then, for every partially observed substantive vertex A, the following conditions hold:

1. *A and its missingness indicator R_A are not adjacent;*

2. *there is no path between A and R_A whose intermediate vertices are colliders and substantive vertices.*

Proof. We give the proof simultaneously for variable-level and cluster-level graphs.

Fix a partially observed substantive vertex A, and let $\mathbb{R}(A)$ denote the missingness-indicator block containing R_A.

Suppose first that A and R_A are adjacent. Since $R_A \in \mathbb{R}(A)$, the definition of the substantive Markov blanket of a missingness-indicator block implies that $A \in MB^m(\mathbb{R}(A))$. Consequently, $R_A \in R_{MB^m(\mathbb{R}(A))}$. Since $R_A \in \mathbb{R}(A)$ as well, it follows that $R_A \in \mathbb{R}(A) \cap R_{MB^m(\mathbb{R}(A))}$, contradicting (11). Therefore, A and R_A cannot be adjacent.

Suppose next that there exists a path $\pi = \langle A, U_1, \ldots, U_k, R_A \rangle$ such that every intermediate vertex U_j is a substantive collider. The path π is therefore a collider path between the substantive vertex A and the member R_A of the block $\mathbb{R}(A)$. By the definition of the substantive Markov blanket, $A \in MB^m(\mathbb{R}(A))$. As above, this implies $R_A \in \mathbb{R}(A) \cap R_{MB^m(\mathbb{R}(A))}$, again contradicting (11).

Hence neither adjacency nor a substantive-intermediate collider path can exist between a partially observed substantive vertex and its own missingness indicator. □

Remark 1. *The proposition does not require the absence of edges between distinct missingness-indicator vertices. That assumption may be required for the factorization used in a sufficient recoverability theorem, but it is not needed to prove that the block-intersection condition implies the individual necessary graphical conditions.*

It is important to note that the converse of Proposition 4 does not hold. We will show this with a counterexample.

The absence, for every partially observed substantive vertex A, of

Example 5. *First consider a variable-level m-ADMG containing partially observed variables X and Y, a substantive variable Z, and the edges*

$$X \longrightarrow R_Y, \qquad R_Y \dashleftarrow\dashrightarrow Z, \qquad Z \dashleftarrow\dashrightarrow R_X,$$

with no other relevant edges.

There is no edge between the distinct missingness indicators R_X and R_Y. Moreover, X and R_X are not adjacent. The path

$$X \longrightarrow R_Y \dashleftarrow\dashrightarrow Z \dashleftarrow\dashrightarrow R_X$$

does not violate the individual necessary condition because one of its intermediate vertices, namely R_Y, is a missingness indicator rather than a substantive vertex. Assume that Y is otherwise disconnected from R_Y, so the individual necessary conditions also hold for Y.

On the other hand,

$$R_Y \dashleftarrow\dashrightarrow Z \dashleftarrow\dashrightarrow R_X$$

is a collider path between R_Y and R_X. Hence they belong to the same missingness-indicator block, say

$$\mathbb{R}^{(1)} = R_X, R_Y.$$

Since $X \to R_Y$, the substantive vertex X belongs to the partially observed Markov blanket of this block:

$$X \in MB^m(\mathbb{R}^{(1)}).$$

Therefore,

$$R_X \in R_{MB^m(\mathbb{R}^{(1)})}.$$

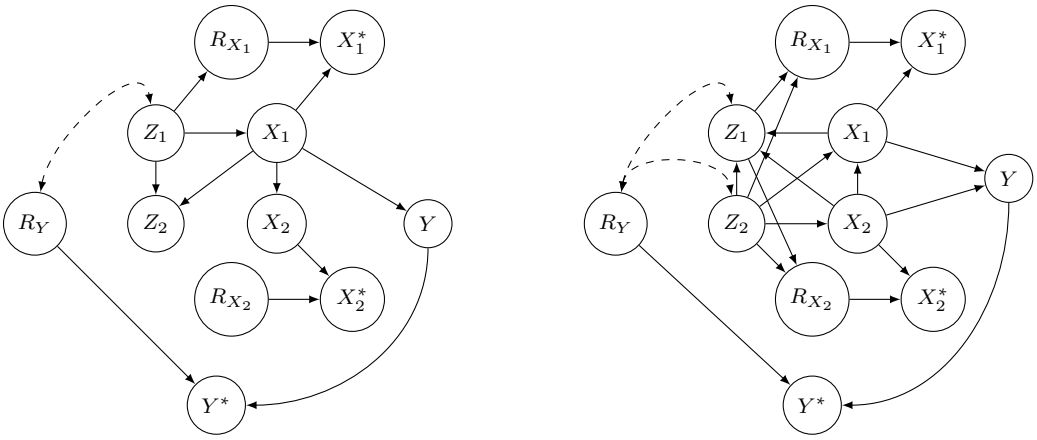
Since $R_X \in \mathbb{R}^{(1)}$, we obtain

$$R_X \in \mathbb{R}^{(1)} \cap R_{MB^m(\mathbb{R}^{(1)})},$$

and the block-intersection condition fails.

The cluster-level counterexample is obtained by replacing the substantive vertices by their corresponding cluster vertices:

$$C_{\mathbb{X}} \longrightarrow R_{C_{\mathbb{Y}}}, \qquad R_{C_{\mathbb{Y}}} \dashleftarrow\dashrightarrow C_{\mathbb{Z}}, \qquad C_{\mathbb{Z}} \dashleftarrow\dashrightarrow R_{C_{\mathbb{X}}}.$$



(a) Setting 1 (b) Setting 2

Figure 5: m-ADMGs compatible with Figure 2b used to generate simulated data.

Let $C_{\mathbb{X}}$ and $C_{\mathbb{Y}}$ be partially observed clusters, and assume that $C_{\mathbb{Y}}$ is otherwise disconnected from $R_{C_{\mathbb{Y}}}$.

There is no edge between the two missingness-indicator vertices. The path from $C_{\mathbb{X}}$ to $R_{C_{\mathbb{X}}}$ contains the intermediate missingness-indicator vertex $R_{C_{\mathbb{Y}}}$, and therefore does not violate the necessary condition requiring all intermediate vertices to belong to $\mathbb{C}^o \cup \mathbb{C}^m$. Nevertheless, $R_{C_{\mathbb{X}}}$ and $R_{C_{\mathbb{Y}}}$ belong to the same missingness-indicator block, while

$$C_{\mathbb{X}} \in MB^m\left(R_{C_{\mathbb{X}}}, R_{C_{\mathbb{Y}}}\right).$$

Thus,

$$R_{C_{\mathbb{X}}} \in R_{C_{\mathbb{X}}}, R_{C_{\mathbb{Y}}} \cap R_{MB^m(R_{C_{\mathbb{X}}}, R_{C_{\mathbb{Y}}})},$$

and the cluster-level block-intersection condition also fails.

The counterexample shows that the block-intersection condition is strictly stronger than the individual necessary graphical conditions. The difference arises because the block condition detects collider connections containing intermediate missingness-indicator vertices, whereas the individual necessary condition only excludes paths whose intermediate vertices are substantive colliders.

C Simulated data

The m-ADMGs in Figure 5, both of which are compatible with the cm-C-DMG in Figure 2b, were used as the true data-generating mechanisms behind the simulated datasets.

Missingness was distributed approximately as follows across simulated datasets. In Setting 1, X_1 was around 35% missing, X_2 was around 20% missing, and Y was around 55% missing. In Setting 2, X_1 was around 25% missing, X_2 was around 65% missing, and Y was around 55% missing.

Data was simulated and analyzed using a MacBook Pro, with negligible execution time (under 1 minute for the whole script).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we introduce two classes of cluster-based missingness graphs, which are relevant in a variety of domains and give the graphical conditions of recoverability for both joint distributions and macro causal effects, all of which are addressed in the abstract.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations to the real-life applicability of the work introduced and the conservativeness of the approach are discussed in the Discussion section.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theorems, formulas and proofs are clearly stated, numbered and referenced in the main paper, with assumptions clearly stated and examples to provide intuition. Proofs are included in the Appendix.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The data generation process, the analysis of experimental data, and its implications for the claims made in our paper are clearly described in the "Simulated data" section, with further details in Appendix C. All code is made available for those who would like to reproduce our experiments.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All code used to simulate and analyze experimental data is open access through the pyCIPHOD package developed by our team.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Choice of accuracy measurements is described in the "Simulated data" section.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The choice of error measurement is specified in the "Simulated data" section.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Computer resources necessary are detailed in Appendix C.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper does not involve real data or human subjects and addresses possible real-world impacts, in line with the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The "Societal impacts" section address impacts of the work.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No data or models are released.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The experiments use the pyCIPHOD package developed by our team. pyCIPHOD is released under the MIT License, and the license terms are respected.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a new reproducibility script within the pyCIPHOD package to reproduce the experiments. The script is documented alongside the package, with instructions describing how to run it and reproduce the reported results.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: the paper does not involve study participants.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is used only for writing, editing, or formatting purposes