Supplementary Materials for: Full Law Identification in Graphical Models of Missing Data: Completeness Results

For clearer presentation of materials in this supplement, we switch to a single-column format. In **Appendix A**, we provide an overview of the nested Markov model. We summarize the necessary concepts required in order to explain our proof of completeness for identification of the full law in missing data acyclic directed mixed graphs (ADMGs). These concepts draw on the binary parameterization of nested Markov models of an ADMG. In **Appendix B**, we provide a concrete example of the odds ratio parameterization. In **Appendix C**, we present proofs that were omitted from the main body of the paper for brevity.

A. Background: Fixing and Nested Markov Models of an ADMG

Given a DAG $\mathcal{G}(V \cup U)$ where U contains variables that are unobserved, the *latent projection operator* onto the observed margin produces an acyclic directed mixed graph $\mathcal{G}(V)$ that consists of directed and bidirected edges (Verma & Pearl, 1990). The bidirected connected components of an ADMG $\mathcal{G}(V)$, partition the vertices V into distinct sets known as districts. The district membership of a vertex V_i in $\mathcal{G}$ is denoted $\mathrm{dis}_{\mathcal{G}}(V_i)$, and the set of all districts in $\mathcal{G}$ is denoted $\mathcal{D}(\mathcal{G})$.

(Evans, 2018) showed that the nested Markov model (Richardson et al., 2017) of an ADMG $\mathcal{G}(V)$ is a smooth super model with fixed dimension, of the underlying latent variable model, that captures all equality constraints and avoids non-regular asymptotics arising from singularities in the parameter space (Drton, 2009; Evans, 2018). We use this fact in order to justify the use of nested Markov models of a missing data ADMG in order to describe full laws that are Markov relative to a missing data DAG with hidden variables. That is, the nested Markov model of a missing data ADMG $\mathcal{G}(V)$, where $V = \{O, X^{(1)}, R, X\}$, is a smooth super model of the missing data DAG model $\mathcal{G}(V \cup U)$. We also utilize nested Markov models of an ADMG $\mathcal{G}(V \setminus X^{(1)})$, corresponding to projection of the missing data ADMG $\mathcal{G}(V)$ onto variables that are fully observable. While such a model does not capture all equality constraints in the true observed law, it is still a smooth super model of it, thus providing an *upper bound* on the model dimension of the observed law.

CADMGs and Kernels

The nested Markov factorization of $p(V)$ relative to an ADMG $\mathcal{G}(V)$ is defined with the use of conditional distributions known as *kernels* and their associated *conditional ADMGs* (CADMGs) that are derived from $p(V)$ and $\mathcal{G}(V)$ respectively, via repeated applications of the *fixing operator* (Richardson et al., 2017). A CADMG $\mathcal{G}(V, W)$, is an ADMG whose nodes can be partitioned into random variables V and *fixed* variables W, with the restriction that only outgoing edges may be adjacent to variables in W. A kernel $q_V(V \mid W)$ is a mapping from values in W to normalized densities over V i.e., $\sum_{v \in V} q_V(v \mid w) = 1$ (Lauritzen, 1996). Conditioning and marginalization operations in kernels are defined in the usual way.

Fixing and Fixability

In Section 4 of the main paper, we provided an informal description of fixing as the operation of inverse-weighting by the propensity score of the variable being fixed; we now formalize this notion. A variable $A \in V$ is said to be *fixable* if the paths $A \to \cdots \to X$ and $A \leftrightarrow \cdots \leftrightarrow X$ do not both exist for all $X \in V \setminus \{A\}$. Given a CADMG $\mathcal{G}(V, W)$ where A is fixable, the graphical operator of fixing, denoted $\phi_A(\mathcal{G})$, yields a new CADMG $\mathcal{G}(V \setminus A, W \cup A)$ with all incoming edges into A being removed, and A being set to a fixed value a. Given a kernel $q_V(V \mid W)$, the corresponding probabilistic operation of fixing, denoted $\phi_A(q_V; \mathcal{G})$ yields a new kernel

$$q_{V \setminus A}(V \setminus A \mid W \cup A) \equiv \frac{q_V(V \mid W)}{q_V(A \mid \mathrm{mb}_{\mathcal{G}}(A), W)},$$

where $\mathrm{mb}_{\mathcal{G}}(A)$ is the *Markov blanket* of A, defined as the bidirected connected component (district) of A (excluding A itself) and the parents of the district of A, i.e., $\mathrm{mb}_{\mathcal{G}}(A) \equiv \mathrm{dis}_{\mathcal{G}}(A) \cup \mathrm{pa}_{\mathcal{G}}(\mathrm{dis}_{\mathcal{G}}(A)) \setminus \{A\}$. It is easy to check that when $\mathcal{G}$ is a DAG, i.e., there are no bidirected edges, the denominator in the probabilistic operation of fixing, reduces to the familiar definition of a simple propensity score.

The notion of fixability can be extended to a set of variables $S \subseteq V$ as follows. A set S is said to be fixable if elements in S can be ordered into a sequence $\sigma_S = \langle S_1, S_2, \dots \rangle$ such that S_1 is fixable in $\mathcal{G}$, S_2 is fixable in $\phi_{S_1}(\mathcal{G})$, and so on. This notion of fixability on sets of variables is essential to the description of the nested Markov model that we present in the following section.

Nested Markov Factorization

Given a CADMG $\mathcal{G}$, A set $S \subseteq V$ is said to be *reachable* if there exists a valid sequence of fixing operations on vertices $V \setminus S$. Further, S is said to be *intrinsic* if it is reachable, and forms a single bidirected connected component or district in $\phi_{\sigma_{V \setminus S}}(\mathcal{G})$, i.e., the CADMG obtained upon executing all fixing operations given by a valid fixing sequence $\sigma_{V \setminus S}$.

A distribution $p(V)$ is said to obey the nested Markov factorization relative to an ADMG $\mathcal{G}(V)$ if for every fixable set S, and any valid fixing sequence σ_S,

$$\phi_{\sigma_S}(p(V); \mathcal{G}) = \prod_{D \in \mathcal{D}(\phi_{\sigma_S}(\mathcal{G}))} q_D(D \mid \mathrm{pa}_{\phi_{\sigma_S}(\mathcal{G})}(D)),$$

where all kernels appearing in the product above can be constructed by combining kernels corresponding to intrinsic sets i.e., $\{q_I(I \mid \mathrm{pa}_{\mathcal{G}}(I)) \mid I \text{ is intrinsic in } \mathcal{G}\}$. Such a construction is made possible by the fact that all the sets D quantified in the product are districts in a reachable graph derived from $\mathcal{G}$.

(Richardson et al., 2017) noted that when a distribution $p(V)$ is nested Markov relative to an ADMG $\mathcal{G}$, all valid fixing sequences yield the same CADMG and kernel so that recursive applications of the fixing operator on a set $V \setminus S$ can simply be denoted as $\phi_{V \setminus S}(\mathcal{G})$ and $\phi_{V \setminus S}(q_V; \mathcal{G})$ without explicitly specifying any particular valid order. Thus, the construction of the set of kernels corresponding to intrinsic sets can be characterized as $\{q_I(I \mid \mathrm{pa}_{\mathcal{G}}(I)) \mid I \text{ is intrinsic in } \mathcal{G}\} = \{\phi_{V \setminus I}(p(V; \mathcal{G})) \mid I \text{ is intrinsic in } \mathcal{G}\}$, and the nested Markov factorization can be re-stated more simply as, for every fixable set S we have,

$$\phi_S(p(V; \mathcal{G})) = \prod_{D \in \mathcal{D}\left(\phi_S(\mathcal{G})\right)} \phi_{V \setminus D}(p(V); \mathcal{G}),$$

An important result from (Richardson et al., 2017) states that if $p(V \cup U)$ is Markov relative to a DAG $\mathcal{G}(V \cup U)$, then $p(V)$ is nested Markov relative to the ADMG $\mathcal{G}(V)$ obtained by latent projection.

Binary Parameterization of Nested Markov Models

From the above factorization, it is clear that intrinsic sets given their parents form the atomic units of the nested Markov model. Using this observation, a smooth parameterization of discrete nested Markov models was provided by (Evans & Richardson, 2014). We now provide a short description of how to derive the so-called Moebius parameters of a *binary* nested Markov model.

For each district $D \in \mathcal{D}(\mathcal{G})$, consider all possible subsets $S \subseteq D$. If S is intrinsic (that is, reachable and bidirected connected in $\phi_{V \setminus S}(\mathcal{G})$), define the head H of the intrinsic set to be all vertices in S that are childless in $\phi_{V \setminus S}(\mathcal{G})$, and the tail T to be all parents of the head in the CADMG $\phi_{V \setminus S}(\mathcal{G})$, excluding the head itself. More formally, $H \equiv \{V_i \in S \mid \mathrm{ch}_{\phi_{V \setminus S}(\mathcal{G})}(V_i) = \emptyset\}$, and $T \equiv \mathrm{pa}_{\phi_{V \setminus S}(\mathcal{G})}(H) \setminus H$. The corresponding set of Moebius parameters for this intrinsic head and tail pair parameterizes the kernel $q_S(H = 0 \mid T)$, i.e., the kernel where all variables outside the intrinsic set S are fixed, and all elements of the head are set to zero given the tail. Note that these parameters are, in general, *variationally dependent* (in contrast to variationally independent in the case of an ordinary DAG model) as the heads and tails in these parameter sets may overlap. The joint density for any query $p(V = v)$, can be obtained through the Moebius inversion formula; see (Lauritzen, 1996; Evans & Richardson, 2014) for details. For brevity, we will denote $q_S(H = 0 \mid T)$ as simply $q(H = 0 \mid T)$, as it will be clear from the given context what variables are still random in the kernel corresponding to a given intrinsic set.

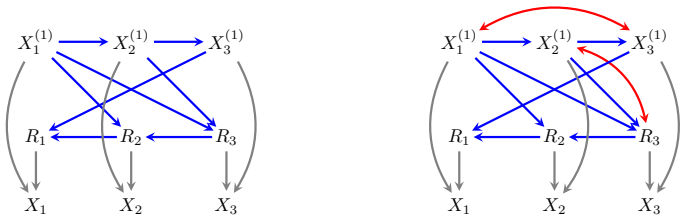
Binary Parameterization of Missing Data ADMGs

We use the parameterization described in the previous section in order to count the number of parameters required to parameterize the full law of a missing data ADMG and its corresponding observed law. We then use this to reason that if the number of parameters in the full law exceeds those in the observed law, it is impossible to establish a map from the observed law to the full law. This in turn implies that such a full law is not identified.

The binary parameterization of the **full law** of a missing data ADMG $\mathcal{G}(X^{(1)}, O, R, X)$ is exactly the same as that of an ordinary ADMG, except that the deterministic factors $p(X_i \mid R_i, X_i^{(1)})$, can be ignored, as $X_i = X_i^{(1)}$ with probability one when $R_i = 1$, and $X_i =?$ with probability one when $R_i = 0$.

The **observed law** is parameterized as follows. First, variables in $X^{(1)}$ are treated as completely unobserved, and an observed law ADMG $\mathcal{G}(X, O, R)$ is obtained by applying the latent projection operator to $\mathcal{G}(X^{(1)}, O, R, X)$. The Moebius parameters are then derived in a similar manner as before, with the additional constraint that if $X_i \in X$ appears in the head of a Moebius parameter, and the corresponding missingness indicator R_i appears in the tail, then the kernel must be restricted to cases where $R_i = 1$. This is because when $R_i = 0$, the probability of the head taking on any value, aside from those where $X_i =?$, is deterministically defined to be 0.

Note that parameterizing the observed law by treating variables in $X^{(1)}$ as fully unobserved does not quite capture all equality constraints that may be detectable in the observed law, as these variables are, in fact, sometimes observable when their corresponding missingness indicators are set to one. Indeed, a smooth parameterization of the observed law of missing data models that captures all constraints implied by the model, is still an open problem. Nevertheless, parameterizing an observed law ADMG, such as the one mentioned earlier, provides an *upper bound* on the number of parameters required to parameterize the true observed law. This suffices for our purposes, as demonstrating that the upper bound on the number of parameters in the observed law is less than the number of parameters in the full law, is sufficient to prove that the full law is not identified.



(a) $\mathcal{G}_a$ (b) $\mathcal{G}_b$

Figure 1. (a) The missing data DAG model used in Scenario 2. (b) the missing data ADMG model used in Scenario 3.

B. Example: Odds Ratio Parameterization

To build up a more concrete intuition for Theorems 1 and 3, we provide an example of the odds ratio parameterization for the missing data models used in Scenarios 2 and 3 of the main paper, reproduced here in Figs. 1(a, b). Utilizing the order R_1, R_2, R_3 on the missingness indicators, the odds ratio parameterization of the missing data process for both models is as follows.

$$\frac{1}{Z} \times \left(\prod_{k=1}^{3} p(R_i \mid R_{-i} = 1, X^{(1)}) \right) \times \text{OR}(R_1, R_2, \mid R_3 = 1, X^{(1)}) \times \text{OR}(R_3, (R_1, R_2) \mid X^{(1)}). \tag{1}$$

We now argue that each piece in Eq. 1 is identified. Note that, in the missing data DAG shown in Fig. 1(a), $R_i \perp\!\!\!\perp X_i^{(1)} \mid R_{-i}, X_{-i}^{(1)}$ by d-separation. The same is true for the missing data ADMG in Fig. 1(b) by m-separation. Thus, in both cases, the product over conditional pieces of each R_i given the remaining variables is not a function $X_i^{(1)}$, and is thus a function of observed data. We now show that $\text{OR}(R_1, R_2 \mid R_3 = 1, X^{(1)})$ is not a function of $X_1^{(1)}, X_2^{(1)}$ by utilizing the symmetry property of the odds ratio.

$$\begin{aligned}
\text{OR}(R_1, R_2 \mid R_3 = 1, X^{(1)}) &= \frac{p(R_1 \mid R_2, R_3 = 1, X_2^{(1)}, X_3^{(1)})}{p(R_1 = 1 \mid R_2, R_3 = 1, X_2^{(1)}, X_3^{(1)})} \times \frac{p(R_1 = 1 \mid R_2 = 1, R_3 = 1, X_2^{(1)}, X_3^{(1)})}{p(R_1 \mid R_2 = 1, R_3 = 1, X_2^{(1)}, X_3^{(1)})} \\
= \text{OR}(R_2, R_1 \mid R_3 = 1, X^{(1)}) &= \frac{p(R_2 \mid R_1, R_3 = 1, X_1^{(1)}, X_3^{(1)})}{p(R_2 = 1 \mid R_1, R_3 = 1, X_1^{(1)}, X_3^{(1)})} \times \frac{p(R_2 = 1 \mid R_1 = 1, R_3 = 1, X_1^{(1)}, X_3^{(1)})}{p(R_2 \mid R_1 = 1, R_3 = 1, X_1^{(1)}, X_3^{(1)})}.
\end{aligned}$$

Thus, from the first equality, the odds ratio is not a function of $X_2^{(1)}$ as $R_1 \perp\!\!\!\perp X_1^{(1)} \mid R_{-1}, X_{-1}^{(1)}$ by d-separation in Fig. 1(a) and by m-separation in Fig. 1(b). A symmetric argument holds for $X_2^{(1)}$ and R_2 as seen in the second and third equalities. Hence, the odds ratio is only a function of $X_3^{(1)}$, which is observable, as the function is evaluated at $R_3 = 1$.

We now utilize an identity from (Chen et al., 2015) in order to simplify the final term in Eq. 1. That is,

$$\begin{aligned}
\text{OR}(R_3, (R_1, R_2) \mid X^{(1)}) &= \text{OR}(R_3, R_2 \mid R_1 = 1, X^{(1)}) \, \text{OR}(R_3, R_1 \mid R_2, X^{(1)}) \\
&= \text{OR}(R_3, R_2 \mid R_1 = 1, X^{(1)}) \, \text{OR}(R_3, R_1 \mid R_2 = 1, X^{(1)}) \underbrace{\frac{\text{OR}(R_3, R_1 \mid R_2, X^{(1)})}{\text{OR}(R_3, R_1 \mid R_2 = 1, X^{(1)})}}_{f(R_1, R_2, R_3 \mid X^{(1)})}.
\end{aligned}$$

The first two pairwise odds ratio terms are functions of observed data using an analogous argument that draws on the symmetry property of the odds ratio and the conditional independence $R_i \perp\!\!\!\perp X_i \mid R_{-i}, X_{-i}^{(1)}$, as before. The final term $f(R_1, R_2, R_3 \mid X^{(1)})$, is a three-way interaction term on the odds ratio scale and can be expressed in three different ways as follows (Chen et al., 2015),

$$\frac{\text{OR}(R_3, R_1 \mid R_2, X^{(1)})}{\text{OR}(R_3, R_1 \mid R_2 = 1, X^{(1)})} = \frac{\text{OR}(R_2, R_3 \mid R_1, X^{(1)})}{\text{OR}(R_2, R_3 \mid R_1 = 1, X^{(1)})} = \frac{\text{OR}(R_1, R_2 \mid R_3, X^{(1)})}{\text{OR}(R_1, R_2 \mid R_3 = 1, X^{(1)})}.$$

From the first equality, we note by symmetry of the odds ratio and conditional independence that f is not a function of $X_1^{(1)}, X_3^{(1)}$. Similarly, from the second equality, we note that f is not a function of $X_2^{(1)}, X_3^{(1)}$. Finally, from the third equality, we note that f is not a function of $X_1^{(1)}, X_2^{(1)}$. Therefore, f is not a function of $X_1^{(1)}, X_2^{(1)}, X_3^{(1)}$ and is identified.

The normalizing function Z, is a function of all the pieces that we have already shown to be identified, and is therefore also identified. Thus, the missing data mechanisms $p(R \mid X^{(1)})$, and consequently, the full laws corresponding to the missing data graphs shown in Figs. 1(a,b) are identified by Remark 2.

C. Proofs

We first prove Lemmas 1 and 2 as we use them in the course of proving Theorems 1 and 3. We start with Lemma 2, as the proof for Lemma 1 simplifies to a special case.

Lemma 2 *A missing data model of an ADMG $\mathcal{G}$ that contains no colluding paths is a submodel of the itemwise conditionally independent nonresponse model described in (Shpitser, 2016; Sadinle & Reiter, 2017).*

Proof. The *complete Markov blanket* of a vertex V_i in an ADMG $\mathcal{G}$, denoted $\mathrm{mb}^c_{\mathcal{G}}(V_i)$ is the set of vertices such that $V_i \perp\!\!\!\perp V_{-i} \setminus \mathrm{mb}^c_{\mathcal{G}}(V_i) \mid \mathrm{mb}^c_{\mathcal{G}}(V_i)$ (Pearl, 1988; Richardson, 2003). In ADMGs, this set corresponds to the Markov blanket of V_i, its children, and the Markov blanket of its children. That is,

$$\mathrm{mb}^c_{\mathcal{G}}(V_i) \equiv \mathrm{mb}_{\mathcal{G}}(V_i) \cup \left(\bigcup_{V_j \in \mathrm{ch}_{\mathcal{G}}(V_i)} V_j \cup \mathrm{mb}_{\mathcal{G}}(V_j) \right) \setminus \{V_i\}.$$

Without loss of generality, we ignore the part of the graph involving the deterministic factors $p(X \mid X^{(1)}, R)$ and the corresponding deterministic edges, in the construction of the Markov blanket and complete Markov blanket of variables in a missing data graph $\mathcal{G}(X^{(1)}, O, R)$. We now show that the absence of non-deterministic colluder paths between a pair $X_i^{(1)}$ and R_i in $\mathcal{G}$ implies that $X_i^{(1)} \notin \mathrm{mb}^c_{\mathcal{G}}(R_i)$.

- $X_i^{(1)}$ is not a parent of R_i, as $X_i^{(1)} \to R_i$ is trivially a colluder path.
- $X_i^{(1)}$ is not in the district of R_i, as $X_i^{(1)} \leftrightarrow \cdots \leftrightarrow R_i$ is also a colluder path.

These two points together imply that $X_i^{(1)} \notin \mathrm{mb}_{\mathcal{G}}(R_i)$. We now show that the union over children of R_i and their Markov blankets also exclude $X_i^{(1)}$.

- $X_i^{(1)}$ is not a child of R_i, as directed edges from R_i to variables in $X^{(1)}$ are ruled out by construction in missing data graphs.
- $X_i^{(1)}$ is also not in the district of any children of R_i, as $R_i \to \cdots \leftrightarrow X_i^{(1)}$ is a colluding path.
- $X_i^{(1)}$ is also not a parent of the district of any children of R_i, as $R_i \to \cdots \leftarrow X_i^{(1)}$ is a colluding path.

These three points together rule out the possibility that $X_i^{(1)}$ is present in the union over children and Markov blankets of children of R_i. Thus, we have shown that $X_i^{(1)} \notin \mathrm{mb}^c_{\mathcal{G}}(R_i)$. This implies the following,

$$R_i \perp\!\!\!\perp V \setminus \{R_i, \mathrm{mb}^c_{\mathcal{G}}(R_i)\} \mid \mathrm{mb}^c_{\mathcal{G}}(R_i) \implies R_i \perp\!\!\!\perp X_i^{(1)} \mid \mathrm{mb}^c_{\mathcal{G}}(R_i).$$

By semi-graphoid axioms (see for example, (Lauritzen, 1996; Pearl, 2009)) this yields the conditional independence $R_i \perp\!\!\!\perp X_i^{(1)} \mid R_{-i}, X_{-i}^{(1)}, O$.

The same line of reasoning detailed above can be used for all $R_i \in R$, which then gives us the set of conditional independences implied by the no self-censoring model. That is,

$$R_i \perp\!\!\!\perp X_i^{(1)} \mid R_{-i}, X_{-i}^{(1)}, O, \quad \forall R_i \in R.$$

□

Lemma 1 *A missing data model of a DAG $\mathcal{G}$ that contains no self-censoring edges and no colluders, is a submodel of the itemwise conditionally independent nonresponse model described in (Shpitser, 2016; Sadinle & Reiter, 2017).*

Proof. A DAG is simply a special case of an ADMG with no bidirected edges. Consequently the only two types of colluding paths, are self-censoring edges ($X_i^{(1)} \to R_i$) and colluder structures ($X_i^{(1)} \to R_j \leftarrow R_i$). Thus, the absence of these two structures in a missing data DAG $\mathcal{G}$, rules out all possible colluding paths. The rest of the proof then carries over straightforwardly from Lemma 2. □

Theorem 1 *A full law $p(R, X^{(1)}, O)$ that is Markov relative to a missing data DAG $\mathcal{G}$ is identified if $\mathcal{G}$ does not contain edges of the form $X_i^{(1)} \to R_i$ (no self-censoring) and structures of the form $X_j^{(1)} \to R_i \leftarrow R_j$ (no colluders), and the stated positivity assumption holds. Moreover, the resulting identifying functional for the missingness mechanism $p(R \mid X^{(1)}, O)$ is given by the odds ratio parameterization provided in Eq. 2 of the main draft, and the identifying functionals for the target law and full law are given by Remarks 1 and 2.*

Proof. Given Eq. (2), we know that

$$p(R \mid X^{(1)}, O) = \frac{1}{Z} \times \prod_{k=1}^{K} p(R_k \mid R_{-k} = 1, X^{(1)}, O) \times \prod_{k=2}^{K} \text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X^{(1)}, O),$$

where $R_{-k} = R \setminus R_k, R_{\prec k} = \{R_1, \ldots, R_{k-1}\}, R_{\succ k} = \{R_{k+1}, \ldots, R_K\}$, and

$$\text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X^{(1)}, O) = \frac{p(R_k \mid R_{\succ k} = 1, R_{\prec k}, X^{(1)}, O)}{p(R_k = 1 \mid R_{\succ k} = 1, R_{\prec k}, X^{(1)}, O)} \times \frac{p(R_k = 1 \mid R_{-k} = 1, X^{(1)}, O)}{p(R_k \mid R_{-k} = 1, X^{(1)}, O)},$$

and Z is the normalizing term and is equal to $\sum_r \{\prod_{k=1}^{K} p(r_k \mid R_{-k} = 1, X^{(1)}, O) \times \prod_{k=2}^{K} \text{OR}(r_k, r_{\prec k} \mid R_{\succ k} = 1, X^{(1)}, O)\}$. If we can prove that all the pieces in this factorization are identified, then the missingness process is identified and so is the full law. We provide the proof in two steps. Our proof is similar to the identification proof of the no self-censoring model given in (Malinsky et al., 2019).

Step 1.

We start off by looking at the conditional pieces $p(R_k \mid R_{-k} = 1, X^{(1)}, O)$. Given Lemma. 1, we know that $R_k \perp\!\!\!\perp X_k^{(1)} \mid R_{-k}, X_{-k}^{(1)}, O$. Therefore, $p(R_k \mid R_{-k} = 1, X^{(1)}, O) = p(R_k \mid R_{-k} = 1, X_{-k}^{(1)}, O), \forall k$, is identified for all $R_k \in R$.

Step 2.

For a given $R_k \in R$, in order to prove that the odds ratio is identified, we rewrite the odds ratio in a slightly different way (without loss of generality we drop the fully observed random variables O.)

$$\begin{aligned}
&\text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X^{(1)}) \qquad (2)\\
&\quad = \text{OR}(R_k, R_{k-1} \mid R_{-\{k,k-1\}} = 1, X^{(1)}) \times \text{OR}(R_k, R_{\prec k-1} \mid R_{\succ k} = 1, R_{k-1}, X^{(1)})\\
&\quad = \text{OR}(R_k, R_{k-1} \mid R_{-\{k,k-1\}} = 1, X^{(1)})\\
&\qquad \times \Big\{\text{OR}(R_k, R_{\prec k-1} \mid R_{\succ k} = 1, R_{k-1} = 1, X^{(1)}) \times \underbrace{\frac{\text{OR}(R_k, R_{\prec k-1} \mid R_{\succ k} = 1, R_{k-1}, X^{(1)})}{\text{OR}(R_k, R_{\prec k-1} \mid R_{\succ k} = 1, R_{k-1} = 1, X^{(1)})}}_{f\left(R_{\preceq k} \mid R_{\succ k} = 1, X^{(1)}\right)}\Big\},
\end{aligned}$$

where $R_{\preceq k} = \{R_1, \ldots, R_K\}$. Through recursive applications of this trick, we can rewrite the odds ratio as follows.

$$\text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, X^{(1)}) = \prod_{i=1}^{k-1} \text{OR}(R_k, R_i \mid R_{\{-(k,i)\}} = 1, X^{(1)}) \times f(R_{\preceq i+1} \mid R_{\succ i+1} = 1, X^{(1)}) \qquad (3)$$

Now we need to show that pairwise odds ratios, $\text{OR}(R_k, R_i \mid R_{\{-(k,i)\}} = 1, X^{(1)})$, and terms of the form $f(R_{\preceq i+1} \mid R_{\succ i+1} = 1, X^{(1)})$ are all identified.

Step 2(a). We start off by proving that for given $R_k, R_i \in R$, the pairwise odds ratio $\text{OR}(R_k, R_i \mid R_{\{-(k,i)\}} = 1, X^{(1)})$ given in Eq. (3) is identified. We know that

$$\text{OR}(R_k, R_i \mid R_{\{-(k,i)\}} = 1, X^{(1)}) = \text{OR}(R_k, R_i \mid R_{\{-(k,i)\}} = 1, X_{\{-(k,i)\}}, X_k^{(1)}, X_i^{(1)}).$$

Consequently, if we can show that the odds ratio is not a function of neither $X_k^{(1)}$ nor $X_i^{(1)}$, then we can safely claim that the odds ratio is only a function of observed data and hence is identified. We get to this conclusion by exploiting the symmetric notion in odds ratios.

$$\begin{aligned}\text{OR}(R_k, R_i \mid R_{\{-(k,i)\}} = 1, X^{(1)}) &= \frac{p(R_k \mid R_i, R_{\{-(k,i)\}} = 1, X^{(1)})}{p(R_k = 1 \mid R_i, R_{\{-(k,i)\}} = 1, X^{(1)})} \times \frac{p(R_k = 1 \mid R_{-k} = 1, X^{(1)})}{p(R_k \mid R_{-k} = 1, X^{(1)})} \\ &= \frac{p(R_i \mid R_k, R_{\{-(k,i)\}} = 1, X^{(1)})}{p(R_i = 1 \mid R_k, R_{\{-(k,i)\}} = 1, X^{(1)})} \times \frac{p(R_i = 1 \mid R_{-i} = 1, X^{(1)})}{p(R_i \mid R_{-i} = 1, X^{(1)})}\end{aligned}$$

In the first equality, we can see that the odds ratio is not a function of $X_k^{(1)}$ since $R_k \perp\!\!\!\perp X_k^{(1)} \mid R_{-k}, X_{-k}^{(1)}$. Similarly, from the second equality, we can see that the odds ratio is not a function of $X_i^{(1)}$ since $R_i \perp\!\!\!\perp X_i^{(1)} \mid R_{-i}, X_{-i}^{(1)}$. Therefore, the pairwise odds ratios are all identified.

Step 2(b). Now we need to show that the second term in Eq. (3) is identified. Since

$$f(R_{\preceq k} \mid R_{\succ k} = 1, X^{(1)}) = f(R_{\preceq k} \mid R_{\succ k} = 1, X_{\succ k}, X_{\preceq k}^{(1)}),$$

we only need to show that this term is not a function of $\{X_1^{(1)}, \ldots, X_k^{(1)}\}$. Because of the way odds ratio is defined, there are k different ways of rewriting the term $f(R_{\preceq k} \mid R_{\succ k} = 1, X^{(1)})$. In the j^{th} specification, f is not a function of $X_j^{(1)}$, because of the fact that $R_j \perp\!\!\!\perp X_j^{(1)} \mid R_{-j}, X_{-j}^{(1)}$. Combining these together, we realize that $f(R_{\preceq k} \mid R_{\succ k} = 1, X^{(1)})$ is not a function of $X_{\preceq k}^{(1)}$, and therefore is identified. For instance, given the triplet $R_i, R_j, R_k \in R$, we can write down $f(R_i, R_j, R_k \mid R_{\succ 3} = 1, X^{(1)})$ in three different ways as follows.

$$\begin{aligned}&f(R_i, R_j, R_k \mid R_{\succ 3} = 1, X^{(1)}) \\ &= \frac{\text{OR}(R_1, R_2 \mid R_{\succ 3} = 1, R_3, X^{(1)})}{\text{OR}(R_1, R_2 \mid R_{\succ 3} = 1, R_3 = 1, X^{(1)})} = \frac{\text{OR}(R_1, R_3 \mid R_{\succ 3} = 1, R_2, X^{(1)})}{\text{OR}(R_1, R_3 \mid R_{\succ 3} = 1, R_2 = 1, X^{(1)})} = \frac{\text{OR}(R_2, R_3 \mid R_{\succ 3} = 1, R_1, X^{(1)})}{\text{OR}(R_2, R_3 \mid R_{\succ 3} = 1, R_1 = 1, X^{(1)})}\end{aligned}$$

From the first equality, we note that f is not a function of $X_1^{(1)}, X_2^{(1)}$. From the second equality, we note that f is not a function of $X_1^{(1)}, X_3^{(1)}$. From the third equality, we note that f is not a function of $X_2^{(1)}, X_3^{(1)}$. Therefore, f is not a function of $X_1^{(1)}, X_2^{(1)}, X_3^{(1)}$ and is identified.

□

Theorem 2 *The graphical condition of no self-censoring and no colluders, put forward in Theorem 1, is sound and complete for the identification of full laws $p(R, O, X^{(1)})$ that are Markov relative to a missing data DAG $\mathcal{G}$.*

Proof. Soundness is a direct consequence of Theorem 1. To prove completeness, it needs to be shown that in the presence of a self-censoring edge, or a colluder structure, the full law is no longer (non-parametrically) identified. A proof by counterexample of both these facts was provided in (Bhattacharya et al., 2019). However, this can also be seen from the fact that self-censoring edges and colluders are special cases of the colluding paths that we prove results in non-identification of the full law in Lemma 3. □

Theorem 3 *A full law $p(R, X^{(1)}, O)$ that is Markov relative to a missing data ADMG $\mathcal{G}$ is identified if $\mathcal{G}$ does not contain any colluding paths and the stated positivity assumption in Section 5 holds. Moreover, the resulting identifying functional for the missingness mechanism $p(R \mid X^{(1)}, O)$ is given by the odds ratio parametrization provided in Eq. 2 of the main draft.*

Proof. The proof strategy is nearly identical to the one utilized in Theorem 1, except the conditional independences $R_k \perp\!\!\!\perp X_k^{(1)} \mid R_{-k}, X_{-k}^{(1)}, O$ come from Lemma 2 instead of Lemma 1. □

Lemma 3 *A full law $p(R, X^{(1)}, O)$ that is Markov relative to a missing data ADMG $\mathcal{G}$ containing a colluding path between any pair $X_i^{(1)} \in X^{(1)}$ and $R_i \in R$, is not identified.*

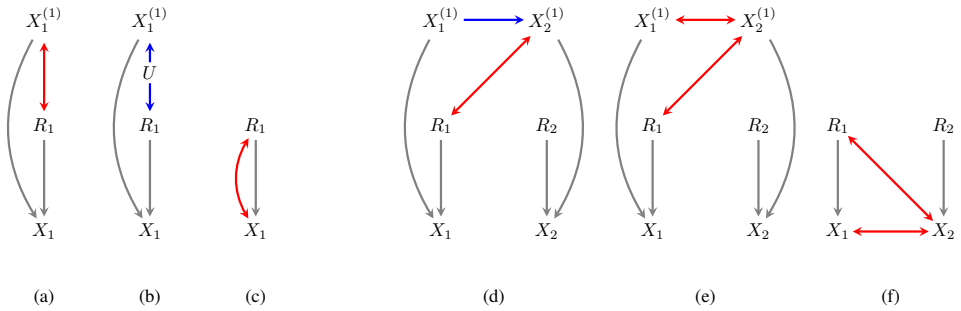


Figure 2. (a, d, e) Examples of colluding paths in missing data models of ADMGs. (b) A DAG with hidden variable U that is Markov equivalent to (a). (c) Projecting out $X_1^{(1)}$ from (a), (f) Projecting out $X_1^{(1)}$ and $X_2^{(1)}$ from (d) and (e).

U	$p(U)$
0	a
1	$1-a$

R_1	U	$p(R_1\|U)$
0	0	b
1	0	$1-b$
0	1	c
1	1	$1-c$

$X_1^{(1)}$	U	$p(X_1^{(1)}\|U)$
0	0	d
1	0	$1-d$
0	1	e
1	1	$1-e$

R_1	$X_1^{(1)}$	U	$p(R_1, X_1^{(1)}, U)$
0	0	0	$a*b*d$
0	0	1	$(1-a)*c*e$
0	1	0	$a*b*(1-d)$
0	1	1	$(1-a)*c*(1-e)$
1	0	0	$a*(1-b)*d$
1	0	1	$(1-a)*(1-c)*e$
1	1	0	$a*(1-b)*(1-d)$
1	1	1	$(1-a)*(1-c)*(1-e)$

R_1	$X_1^{(1)}$	p(Full Law)	X_1	p(Observed Law)
0	0	$a*b*d+(1-a)*c*e$	?	$a*b+(1-a)*c$
	1	$a*b*(1-d)+(1-a)*c*(1-e)$		
1	0	$a*(1-b)*d+(1-a)*(1-c)*e$	0	$a*(1-b)+(1-a)*(1-c)$
	1	$a*(1-b)*(1-d)+(1-a)*(1-c)*(1-e)$	1	

Table 1. Construction of counterexamples for non-identifiablity of the full law in Fig. 2(a) using the DAG with hidden variable U in Fig. 2(b) that is Markov equivalent to (a).

Proof. Proving the non-identifiability of missing data models of an ADMG $\mathcal{G}$ that contains a colluding path can be shown by providing two models $\mathcal{M}_1$ and $\mathcal{M}_2$ that disagree on the full law but agree on the observed law. Coming up with a single example of such a pair of models is sufficient for arguing against non-parametric identification of the full law. Therefore, for simplicity, we restrict our attention to binary random variables. We first provide an example of such a pair of models on the simplest form of a colluding path, a bidirected edge $X_i^{(1)} \leftrightarrow R_i$ as shown in Fig. 2(a). According to Table 1, in order for the observed laws to agree, the only requirement is that the quantity $ab + (1-a)c$ remain equal in both models; hence we can come up with infinitely many counterexamples of full laws that are not the same but map to the same observed law.

Constructing explicit counterexamples are not necessary to prove non-identification as long as it can be shown that there exist at least two distinct functions that map two different full laws onto the exact same observed law. For instance, if the number of parameters in the full law is strictly larger than the number of parameters in the observed law, then there would exist infinitely many such functions. Consequently, we rely on a parameter counting argument to prove the completeness of our results. Since we are considering missing data models of ADMGs, we use the Moebius parameterization of binary nested Markov models of an ADMG described in Appendix A.

The nested Markov model of a missing data ADMG $\mathcal{G}(V)$, where $V = \{O, X^{(1)}, R, X\}$, is a smooth super model of the missing data DAG model $\mathcal{G}(V \cup U)$, and has the same model dimension as the latent variable model (Evans, 2018). We also utilize nested Markov models of an ADMG $\mathcal{G}(V \setminus X^{(1)})$, corresponding to projection of the missing data ADMG $\mathcal{G}(V)$ onto variables that are fully observable. While such a model does not capture all equality constraints in the true observed law, it is still a smooth super model of it, thus providing an *upper bound* on the model dimension of the observed law. This suffices for our purposes, as demonstrating that the upper bound on the number of parameters in the observed law is less than the number of parameters in the full law, is sufficient to prove that the full law is not identified. We first walk the reader through a few examples to demonstrate this proof strategy, and then provide the general argument.

Moebius Parameterization of the Full Law in Fig. 2(d)			
Districts	Intrinsic Head/Tail	Moebius Parameters	Counts
$\{X_1^{(1)}\}$	$\{X_1^{(1)}\}, \{\}$	$q(X_1^{(1)} = 0)$	1
$\{R_2\}$	$\{R_2\}, \{\}$	$q(R_2 = 0)$	1
$\{R_1, X_2^{(1)}\}$	$\{R_1\}, \{\}$	$q(R_1 = 0)$	1
	$\{X_2^{(1)}\}, \{X_1^{(1)}\}$	$q(X_2^{(1)} = 0 \mid X_1^{(1)})$	2
	$\{R_1, X_2^{(1)}\}, \{X_1^{(1)}\}$	$q(R_1 = 0, X_2^{(1)} = 0 \mid X_1^{(1)})$	2
		Total	7
Moebius Parameterization of the Full Law in Fig. 2(e)			
Districts	Intrinsic Head/Tail	Moebius Parameters	Counts
$\{R_2\}$	$\{R_2\}, \{\}$	$q(R_2 = 0)$	1
$\{R_1, X_1^{(1)}, X_2^{(1)}\}$	$\{R_1\}, \{\}$	$q(R_1 = 0)$	1
	$\{X_1^{(1)}\}, \{\}$	$q(X_1^{(1)} = 0)$	1
	$\{X_2^{(1)}\}, \{\}$	$q(X_2^{(1)} = 0)$	1
	$\{R_1, X_2^{(1)}\}, \{\}$	$q(R_1 = 0, X_2^{(1)} = 0)$	1
	$\{X_1^{(1)}, X_2^{(1)}\}, \{\}$	$q(X_1^{(1)} = 0, X_2^{(1)} = 0)$	1
	$\{R_1, X_1^{(1)}, X_2^{(1)}\}, \{\}$	$q(R_1 = 0, X_1^{(1)} = 0, X_2^{(1)} = 0)$	1
		Total	7
Moebius Parameterization of the Observed Law in Fig. 2(f)			
Districts	Intrinsic Head/Tail	Moebius Parameters	Counts
R_2	$\{R_2\}, \{\}$	$q(R_2 = 0)$	1
$\{R_1, X_1, X_2\}$	$\{R_1\}, \{\}$	$q(R_1 = 0)$	1
	$\{X_1\}, \{R_1\}$	$q(X_1 = 0 \mid R_1)$	1
	$\{X_2\}, \{R_2\}$	$q(X_2 = 0 \mid R_2)$	1
	$\{R_1, X_2\}, \{R_2\}$	$q(R_1 = 0, X_2 = 0 \mid R_2)$	1
	$\{X_1, X_2\}, \{R_1, R_2\}$	$q(X_1 = 0, X_2 = 0 \mid R_1, R_2)$	1
		Total	6

Table 2. Moebius Parameterization of the Full and Observed Laws of missing data ADMGs

SELF-CENSORING THROUGH UNMEASURED CONFOUNDING:

We start by reanalyzing the colluding path given in Fig. 2(a) and the corresponding projection given in Fig. 2(c). The Moebius parameters associated with the full law are $q(X_1^{(1)} = 0), q(R_1 = 0), q(X_1^{(1)} = 0, R_1 = 1)$, for a total of 3 parameters. The Moebius parameters associated with the observed law in Fig 2(c) are $q(R_1 = 0), q(X_1^{(1)} = 0 \mid R_1 = 0)$, for a total of only 2 parameters. Since $2 < 3$, we can construct infinitely many mappings, as it was shown in Table 1.

SIMPLE COLLUDING PATHS:

Consider the colluding paths given in Fig. 2(d, e) and the corresponding projection (which are identical in both cases) given in Fig. 2(f). The Moebius parameters associated with the full laws and observed law are shown in Table 2. Once again, since the number of parameters in the observed law is less than the number in the full law $(6 < 7)$, we can construct infinitely many mappings.

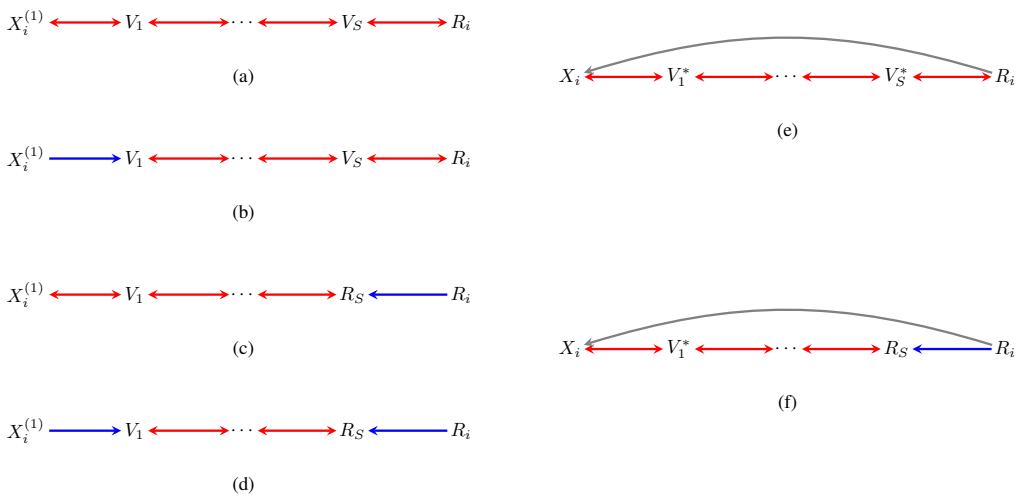


Figure 3. (a) Colluding paths (b) Projecting out $X^{(1)}$

A GENERAL ARGUMENT:

In order to generalize our argument, we first provide a more precise representation (that does not use dashed edges) in Figs. 3(a-d), of all possible colluding paths between $X_i^{(1)}$ and R_i. Without loss of generality, assume that there are K variables in $X^{(1)}$ and there are S variables that lie on the collider path between $X_i^{(1)}$ and R_i, $S \in \{0, 1, \ldots, 2*(K-1)\}$. We denote the sth variable on the collider path by V_s; $V_s \in \{X^{(1)} \setminus X_i^{(1)}, R \setminus R_i\}$. Note that V_S in Figs. 3(c, d) can only belong to $\{R \setminus R_i\}$ by convention. Fig. 3(e) illustrates the corresponding projections of figures (a) and (b), and Fig. 3(f) illustrates the corresponding projections of figures (c) and (d). In the projections shown in Figs. 3(e, f), $V^* \in \{X \setminus X_i^{(1)}, R \setminus R_i\}$.

We now go over each of these colluding paths and their corresponding latent projections, as if they appear in a larger graph that is otherwise completely disconnected. We count the number of Moebius parameters as a function of S, and show that the full law always has one more parameter than the observed law. One can then imagine placing these colluding paths in a larger graph with arbitrary connectivity, and arguing that the full law is still not identified as a consequence of the parameter discrepancy arising from the colluding path alone. That is, if we show a fully disconnected graph containing a single colluding path is not identified, then it is also the case that any edge super graph (super model) is also not identified.

In the following proof we heavily rely on the following fact. Given a bidirected chain of length $V_1 \leftrightarrow, \cdots, \leftrightarrow V_K$, of length K, the number of Moebius parameters required to parameterize this chain is given by the sum of natural numbers 1 to K, i.e., $\frac{K(K+1)}{2}$. This can be seen from the fact that the corresponding Moebius parameters are given by the series,

- $q(V_1 = 0), q(V_1 = 0, V_2 = 0), \ldots, q(V_1 = 0, \ldots, V_K = 0)$ corresponding to K parameters.
- $q(V_2 = 0), q(V_2, V_3 = 0), \ldots, q(V_2 = 0, \ldots, V_K = 0)$ corresponding to $K - 1$ parameters.
- $\ldots$
- $q(V_K = 0)$ corresponding to 1 parameter.

In counting the number of parameters for a disconnected graph (with the exception of the colluding path), we can also exclude the singleton (disconnected) nodes from the counting argument since they account for the same number of parameters in both the full law and observed law. In the full law they are either $q(R_s = 0)$ or $q(X_s^{(1)} = 0)$ and the corresponding parameters in the observed law are $q(R_s = 0)$ or $q(X_s = 0 \mid R_s = 1)$. The Moebius parameter counts for each of the colluding paths in Figs. 3(a-d) and their corresponding latent projections in Figs. 3(e,f) are as follows.

Figures a, b, and e

1. Number of Moebius parameters in Fig. 3(a) is $\frac{(S+2)(S+3)}{2}$

- A bidirected chain $X_i^{(1)} \leftrightarrow, \cdots, \leftrightarrow R_i$ of length $S+2$, i.e., $(S+2)*(S+3)/2$ parameters.

2. Number of Moebius parameters in Fig. 3(b) is $\frac{(S+2)(S+3)}{2}$
   - $q(X_i^{(1)}=0)$, i.e. 1 parameter,
   - A bidirected chain $V_2 \leftrightarrow \cdots \leftrightarrow R_i$ of length S, i.e. $S*(S+1)/2$ parameters,
   - Intrinsic sets involving V_1, i.e., $q(V_1=0 \mid X_i^{(1)}), q(V_1=0, V_2=0 \mid X_i^{(1)}), q(V_1=0,\ldots,R_i=0 \mid X_i^{(1)})$ corresponding to $2*(S+1)$ parameters.
3. Number of Moebius parameters in Fig. 3(e) is $\frac{(S+2)(S+3)}{2}-1$
   - Note that even though each proxy X_s that may appear in the bidirected chain has a directed edge from R_s pointing into it, the corresponding intrinsic head tail pair that involves both variables, will always have $R_i=1$. Hence, we may ignore these deterministic edges and count the parameters as if it were a bidirected chain $V_1^* \leftrightarrow \cdots \leftrightarrow R_i$ of length $S+1$, corresponding to $(S+1)*(S+2)/2$ parameters,
   - When enumerating intrinsic sets involving X_i, we note that $\{X_i, V_1^*, \ldots V_S^*\}$ is not intrinsic as R_i is not fixable (due to the bidirected path between R_i and X_i and the edge $R_i \to X_i$). Thus, as there is one less intrinsic set involving X_i, the number of parameters required to parameterize all intrinsic sets involving X_i is one fewer, i.e., $S+1$ (instead of $S+2$) parameters.

Figures c, d, and f

1. Number of Moebius parameters in Fig. 3(c) is $\frac{(S+2)(S+3)}{2}$
   - $q(R_i=0)$, i.e. 1 parameter,
   - A bidirected chain $X_i^{(1)} \leftrightarrow \cdots \leftrightarrow V_{S-1}$ of length S, i.e. $S*(S+1)/2$ parameters,
   - Intrinsic sets involving R_S, i.e., $q(R_S=0 \mid R_i), q(R_S=0, V_{S-1}=0 \mid R_i), \ldots, q(R_S=0, V_{S-1}=0 \ldots, X_i^{(1)} \mid R_i)$, corresponding to $2*(S+1)$ parameters.
2. Number of Moebius parameters in Fig. 3(d) is $\frac{(S+2)(S+3)}{2}$
   - $q(X_i^{(1)}=0), q(R_i=0)$, i.e. 2 parameters,
   - A bidirected chain $V_2 \leftrightarrow \cdots \leftrightarrow V_{S-2}$ of length $S-2$, i.e. $(S-2)*(S-1)/2$ parameters,
   - Intrinsic sets involving V_1 and not R_S, i.e., $q(V_1=0 \mid X_i^{(1)}), q(V_1=0, V_2=0 \mid X_i^{(1)}), \ldots, q(V_1=0, V_2=0, \ldots, V_{S-1} \mid X_i^{(1)})$, corresponding to $2*(S-1)$ parameters,
   - Intrinsic sets involving R_S and not V_1, i.e., $q(R_S=0 \mid R_i), q(R_S=0, V_{S-1}=0 \mid R_i), \ldots, q(R_S=0, V_{S-1}=0, \ldots, V_2 \mid R_i)$ corresponding to $2*(S-1)$ parameters.
   - The intrinsic set involving both V_1 and R_S, i.e., $q(V_1=0, V_2=0, \ldots, R_S=0 \mid X_i^{(1)}, R_i)$, corresponding to 4 parameters.
3. Number of Moebius parameters in Fig. 3(f) is $\frac{(S+2)(S+3)}{2}-1$
   - $q(R_i=0)$, i.e. 1 parameter,
   - By the same argument as before, deterministic tails can be ignored. Hence, we have a bidirected chain $X_i \leftrightarrow \cdots \leftrightarrow V_{S-1}$ of length S, i.e. $S*(S+1)/2$ parameters,
   - Intrinsic sets involving R_S, i.e., $q(R_S=0 \mid R_i), q(R_S=0, V_{S-1} \mid R_i), \ldots, q(R_S, V_{S-1}, \ldots, V_1 \mid R_i)$, corresponding to $2*S$ parameters, and the special intrinsic set which results in the observed law having one less parameter $q(R_S, V_{S-1}, \ldots, V_1, X_i \mid R_i=1)$ corresponding to just 1 parameter instead of 2 due to the presence of the proxy X_i in the head and the corresponding R_i in the tail.

□

Theorem 4 *The graphical condition of the absence of colluding paths, put forward in Theorem 3, is sound and complete for the identification of full laws $p(R, O, X^{(1)})$ that are Markov relative to a missing data ADMG $\mathcal{G}$.*

Proof. Soundness is a direct consequence of Theorem 3 and completeness is a direct consequence of Lemma. 3. □

References

Bhattacharya, R., Nabi, R., Shpitser, I., and Robins, J. M. Identification in missing data models represented by directed acyclic graphs. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2019.

Chen, H. Y., Rader, D. E., and Li, M. Likelihood inferences on semiparametric odds ratio model. *Journal of the American Statistical Association*, 110(511):1125–1135, 2015.

Drton, M. Discrete chain graph models. *Bernoulli*, 15(3):736–753, 2009.

Evans, R. J. Margins of discrete Bayesian networks. *The Annals of Statistics*, 46(6A):2623–2656, 2018.

Evans, R. J. and Richardson, T. S. Markovian acyclic directed mixed graphs for discrete data. *The Annals of Statistics*, pp. 1452–1482, 2014.

Lauritzen, S. L. *Graphical Models*. Oxford, U.K.: Clarendon, 1996.

Malinsky, D., Shpitser, I., and Tchetgen Tchetgen, E. J. Semiparametric inference for non-monotone missing-not-at-random data: the no self-censoring model. *arXiv preprint arXiv:1909.01848*, 2019.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

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

Richardson, T. S. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. Nested Markov properties for acyclic directed mixed graphs. *arXiv:1701.06686v2*, 2017. Working paper.

Sadinle, M. and Reiter, J. P. Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika*, 104(1):207–220, 2017.

Shpitser, I. Consistent estimation of functions of data missing non-monotonically and not at random. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*. 2016.

Verma, T. and Pearl, J. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, 1990.