Supplement - Establishing Markov Equivalence in Cyclic Directed Graphs

Tom Claassen[1] **Joris M. Mooij**[2]

[1]Institute for Computing and Information Sciences, Radboud University, Nijmegen, Netherlands
[2]Korteweg-deVries Institute, University of Amsterdam, Amsterdam, Netherlands

Abstract

This supplement contains additional results and proof details related to the UAI2023 submission 'Establishing Markov Equivalence in Cyclic Directed Graphs'. Numbering and notations follow the main article.

1 ADDITIONAL EXPERIMENTAL RESULTS

This section elaborates on the random cyclic graph generating process, and a result that offers some added insight into the inner workings of the two CPAG algorithms.

1.1 GENERATING RANDOM CYCLIC GRAPHS

In contrast to the familiar acyclic graphs, in cyclic graphs there can be *two* edges between each pair of nodes, corresponding to a total of $N(N-1)$ possible directed edges for graphs over N nodes. However, in both the Erdos-Renyi model (all graphs with n edges equally likely) and the Gilbert model (all edges appear with equal probability p), as density or size of the graph increases, the resulting graph is overwhelmingly likely to contain just one, big strongly connected component, with only a few other nodes on its periphery. As a key part of the CET is about invariant edges *between* components in rule (iv) (see e.g. Figure 3 in the main article), just evaluating on arbitrary random graphs would likely lead to an incomplete or biased perspective. In addition, a number of challenges in finding the correct CPAG are related to sequences of connected two-cycles (see main, Figure 2), which in larger fully random graphs are also exceedingly unlikely to appear.

Therefore we tweak the random graph generating process to allow some control over the number and size of the strongly connected components. We introduce a 3-stage process parameterized by size N and density d, as well as parameters p_{two} for the proportion of two-cycles, and p_{acy} and p_{cyc} for the proportion of recursive resp. nonrecursive edges that remain:

1. randomly sample the required number of two-cycles,
2. add random arcs from lower to higher numbered nodes,
3. add completely random arcs for the remaining edges.

Afterwards a random permutation of the nodes is applied to ensure there is no implicit bias in the ordering.

With this procedure, setting $[p_{two}, p_{acy}, p_{cyc}] = [0, 1, 0]$ would lead to a random acyclic graph, whereas setting $[0.1, 0.9, 0]$ would lead to a random acyclic graph with some edges turned into two-cycles. Setting $[0, 0, 1]$ would lead to a fully random cyclic graph in the Erdos-Renyi model. In practice setting e.g. $[p_{two}, p_{acy}, p_{cyc}] = [0.1, 0.82, 0.08]$ leads to a varied number and size of the strongly connected components for graphs of up to $N = 200$ nodes with density $d = 3.0$. For $N = 200$ this leads on average to about 11 nontrivial strongly connected components with average largest component size of about 17 vertices.

For larger/higher density graphs the p_{cyc} proportion should be reduced to avoid collapsing into the 'one big cycle' trap. In our experiments for $d = 5.0$ we used $[p_{two}, p_{acy}, p_{cyc}] = [0.05, 0.93, 0.02]$, which, for $N = 200$ resulted on average in about 5 nontrivial strongly connected components, with an average largest size of about 70 vertices.

Additional implementation details will be published with the accompanying source code.

1.2 RELATIVE TIME SPENT PER STAGE

To take a closer look at the relative contribution of each stage in the two different CPAG procedures to the overall time complexity we also timed each stage separately. Average results are depicted below.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

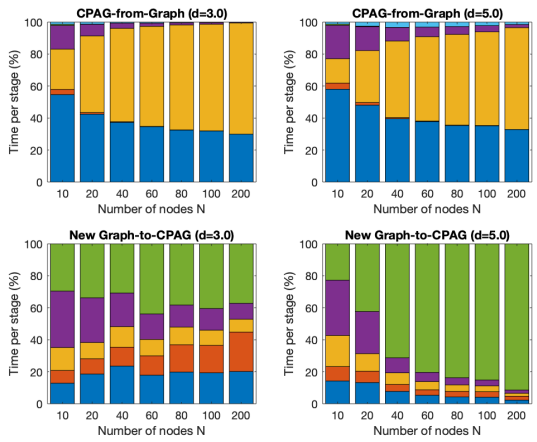

Figure 5: Plots depicting the relative proportion each algorithm spends on average in the different stages, as a function of the size of the graph N, for two different densities $d \in \{3.0, 5.0\}$. Stages are ordered bottom up, i.e. first stage on the x-axis, second stage on top of that etc.

We see that the original CPAG-from-Graph procedure spends the vast majority of its time in the expensive d-separation searches in stage (a) (blue) and (c) (yellow), whereas for sparse graphs the new Graph-to-CPAG version spends roughly equal amounts in each phase. For denser graphs, in the latter starts the final stage (green) that aims to orient invariant edges within and between cycles, starts to dominate, as expected from the complexity analysis in section 4.3.

Note that, even though it may seem that for higher densities this final stage in the new Graph-to-CPAG procedure is somehow less efficient, it is still about 4 times as fast as the corresponding stage (f) (light-blue/teal) in the original CPAG-from-Graph procedure. It is just that in the latter stages (a) and (c) are even more expensive.

2 PROOF DETAILS

Lemma 1 *For a directed graph $\mathcal{G}$ and corresponding CMAG $\mathcal{M}$, there is a u-strucuture $\langle X, Z, Z', Y\rangle$ in $\mathcal{M}$ iff there is an uncovered itinerary $\pi = \langle X, Z, U, .., U', Z', Y\rangle$ in $\mathcal{G}$, possibly with $Z = U'$ or $U = U'$, where $\langle X, Z, U\rangle$ and $\langle U', Z', Y\rangle$ are a pair of m.e. conductors w.r.t. the uncovered itinerary π in $\mathcal{G}$.*

Proof By definition 9, a *u*-structure $\langle X, Z, Z', Y\rangle$ implies the existence of an uncovered path $\pi = \langle X, Z, U_1, .., U_k, Z', Y\rangle$ (possibly with $U_1 = U_k$ or $U_1 = Z', U_k = Z$) between nonadjacent X and Y in $\mathcal{M}$, corresponding to an uncovered itinerary in $\mathcal{G}$ where all nodes $\{Z, Z', U_1, .., U_k\}$ are ancestors of each other, but not of X or Y, which implies $\langle X, Z, U_1\rangle$ and $\langle U_k, Z', Y\rangle$ are a pair of m.e. conductors w.r.t. the uncovered itinerary π in $\mathcal{G}$.

Conversely, if $\langle X, Z, U\rangle$ and $\langle U', Z', Y\rangle$ are a pair of m.e. conductors w.r.t. an uncovered itinerary $\pi = \langle X, Z, U, .., U', Z', Y\rangle$ in $\mathcal{G}$, then π is a also an uncovered path $\langle X, Z, .., Z', Y\rangle$ in $\mathcal{M}$, where all intermediate nodes are ancestor of each other, and so $\{Z, U, .., U', Z'\} \subset SCC(Z)$, but not of X or Y, and so $X \longrightarrow Z$ and $Z' \longleftarrow Y$ in $\mathcal{M}$, which by definition 9 implies $\langle X, Z, Z', Y\rangle$ is a *u*-structure. ■

Lemma 2 *In a CMAG $\mathcal{M}$, a pair of nodes $\langle X, Z\rangle$ is part of a u-structure $\langle X, Z, Z', Y\rangle$ with a node $Y \in \mathbf{Y} \subseteq pa(SCC(Z)) \setminus adj(\{X, Z\})$, iff $X \in pa(Z)$, and X and Y are connected in the undirected subgraph over $((SCC(Z) \setminus adj(X)) \cup \{X, Z\} \cup \mathbf{Y}$.*

Proof The given implies the existence of some path from X via adjacent nodes in the undirected subgraph to some node from $\mathbf{Y}$. Let Y be the first node from $\mathbf{Y}$ encountered along this path, then $\langle X, Z_1, .., Z_k, Y\rangle$ is a path over distinct nodes where all $Z_i \in SCC(Z)$ are ancestors of each other, but not of X or Y.

If the path $\langle X, Z_1, .., Z_k, Y\rangle$ is not uncovered, then some subsequence $\langle X, U_1, .., U_m, Y\rangle$ with $\{U_1, .., U_m\} \subset \{Z_1, .., Z_k\}$ can be chosen so that $\langle X, U_1, .., U_m, Y\rangle$ is an uncovered path in the unoriented subgraph (see e.g. Lemma B.1 in (Zhang, 2008)). Furthermore, as all nodes adjacent to X in $\mathcal{M}$ are excluded from this subgraph with the exception of Z, it means that $Z = U_1 = Z_1$. We also know that $m \geq 2$, as all $Y \in \mathbf{Y}$ were taken not to be adjacent to Z, so $Z' = U_m \neq Z$.

Finally, as all $U_i \in SCC(Z)$ are ancestors of each other, but not of X or Y, it also means that each consecutive triple along the path is a noncollider in $\mathcal{M}$, and so in accordance with definition 9 $\langle X, Z, Z', Y\rangle$ is a *u*-structure. ■

For the proof of Theorem 1 we use two helpful results.

Corollary 1 In a CMAG $\mathcal{M}$, a virtual collider triple $\langle A, B, C\rangle$ uniquely corresponds to either:

1. a virtual *v*-structure $\langle A, B, C\rangle$, or
2. a *u*-structures $\langle A, B, B', C\rangle$, or
3. a *u*-structure $\langle A, B', B, C\rangle$,

where for the latter two the complementary triple $\langle A, B', C\rangle$ is also a virtual collider triple.

Proof If virtual collider triple $\langle A, B, C\rangle$ corresponds to a virtual *v*-structure, then it cannot be part of a *u*-structure $\langle A, B, B', C\rangle$ or $\langle A, B', B, C\rangle$, as that would imply the path from A to C via B is not uncovered, contrary Definition 9. Similarly, if virtual collider triple $\langle A, B, C\rangle$ corresponds to a *u*-structure $\langle A, B, B', C\rangle$, then it cannot

also correspond to a *u*-structure $\langle A, B', B, C\rangle$, as the combination would imply the presence of edges $A \longrightarrow B$ and $B \longleftarrow C$ in $\mathcal{M}$, which again would contradict the fact that the path $\langle A, B, .., B', C\rangle$ in $\mathcal{M}$ is uncovered. By definition 10, in both cases the *u*-structure would imply that the complementary $\langle A, B', C\rangle$ also satisfies the definition of a virtual collider triple. ■

The second is a well-known result that connects nodes that make or break a *d*-separation to ancestral relations, where we use square brackets $[\mathbf{Z}]$ to indicate *minimal* sets, i.e. sets $\mathbf{Z}$ for which there is no strict subset of $\mathbf{Z}$ that preserves the *d*-separation relation:

Lemma 4 In a directed graph $\mathcal{G}$, if adding a node Z to a conditioning set changes a *d*-separation relation between two nodes X and Y relative to a set $\mathbf{Z}$, then:

1. if $X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, \mathbf{Z} \cup [Z]$, then $Z \in an_{\mathcal{G}}(\{X, Y\} \cup \mathbf{Z})$,
2. if $X \not\perp\!\!\!\perp_{\mathcal{G}} Y \,|\, \mathbf{Z} \cup [Z]$, then $Z \notin an_{\mathcal{G}}(\{X, Y\} \cup \mathbf{Z})$,

with special case:

3. if $X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, [\mathbf{Z}]$, then $\mathbf{Z} \subseteq an_{\mathcal{G}}(\{X, Y\})$.

Proof This result was originally derived in (Claassen and Heskes, 2011, Lemma 2) for acyclic graphs in the possible presence of unobserved confounders and selection bias, but the proof for the first two rules carries directly over to the cyclic directed case considered here (i.e. without confounders and/or selection bias). The proof for the special case, where there is no subset of $\mathbf{Z}$ that can *d*-separate X and Y, did use acyclicity, but also follows similar to the proof of rule 1 in the cyclic case:

By contradiction: let $Z \in \mathbf{Z}$ be a node that is not in $an_{\mathcal{G}}(\{X, Y\})$. Let $\mathbf{Z}' = de(Z) \cap \mathbf{Z}$, i.e. all descendants of Z in $\mathbf{Z}$ (including Z itself). Then none of the nodes in $\mathbf{Z}'$ can be ancestor of X or Y, otherwise Z would be ancestor of X or Y as well, contrary to the assumed. We now show that X and Y are also *d*-separated relative to $\mathbf{Z}^* = \mathbf{Z} \setminus \mathbf{Z}'$. Suppose there is an unblocked path π between X and Y relative to $\mathbf{Z}^*$. Then π cannot contain any noncolliders in $\mathbf{Z}^*$, otherwise it would be blocked. But π must contain at least one node $Z' \in \mathbf{Z}'$ that is a noncollider along π, otherwise the path could not be blocked by adding $\mathbf{Z}'$. Therefore Z' must have at least one outgoing arc along π. Follow π in this direction until either a collider is encountered or the end of π is reached. But if a collider is reached, then there must be a node $Z^* \in \mathbf{Z}^*$ that is a descendant of that collider, otherwise π could not be unblocked. But then this node was a descendant of Z', and so also a descendant of Z, which implies it was included in $\mathbf{Z}'$, and therefore not in $\mathbf{Z}^*$. And if the end of the path is reached then there is a directed path from Z' to X or Y, and so also from Z to X or Y, contrary the assumed. Therefore there can be no unblocked path in $\mathcal{G}$ between X and Y relative to $\mathbf{Z}^*$, which in turn implies they are *d*-separated by $\mathbf{Z}^* \subsetneq \mathbf{Z}$, which implies that $\mathbf{Z}$ was not a minimal separating set, contrary the assumed in rule 3. QED. ■

We are now ready to prove the new ancestral CET:

Theorem 1 *Two CMAGs $\mathcal{M}_1$ and $\mathcal{M}_2$ corresponding to cyclic directed graphs $\mathcal{G}_1$ resp. $\mathcal{G}_2$ are Markov equivalent iff*

(i) they have the same skeleton,

(ii) they have the same v-structures,

(iii) they have the same virtual collider triples,

(iv) if $\langle A, B, C\rangle$ and $\langle A, D, C\rangle$ are virtual collider triples, then B is an ancestor of D in $\mathcal{M}_1$ iff B is an ancestor of D in $\mathcal{M}_2$.

Proof We show that in terms of the CPAG the first 3 rules are equivalent to the first 4 rules in the original CET, and that the last rule is sound and implies the last two rules in the original CET, which means the combined set of rules is sound and sufficient to ensure Markov equivalence.

(i) By lemma 3 (see below), two nodes in a CMAG $\mathcal{M}$ are adjacent iff they are (virtually) adjacent in the underlying cyclic graph $\mathcal{G}$, and so rule (i) is equivalent between the two CETs.

(ii)+(iii) By definitions 4 and 5 and rule (i), an unshielded triple $\langle A, B, C\rangle$ in a CPAG is either a conductor, an unshielded perfect nonconductor, or an unshielded imperfect nonconductor in $\mathcal{G}$. Therefore (ii).a+(ii).b in the original CET are equivalent to 'have the same unshielded perfect and imperfect nonconductors' (as the remaining unshielded triples then all must correspond to conductors). A perfect nonconductor in $\mathcal{G}$ is a *v*-structure in the CMAG $\mathcal{M}$, and by definition 8 the subset of *imperfect* nonconductors is equivalent to virtual *v*-structures. By corollary 1, a virtual collider triple $\langle A, B, C\rangle$ is either a virtual *v*-structure, or part of a *u*-structure $\langle A, B, B', C\rangle$ or $\langle A, B', B, C\rangle$ for which, by definition 10, the complementary $\langle A, B', C\rangle$ is also a virtual collider triple. By lemma 1, that means that, depending on the skeleton from rule (i), either $\langle A, B, U\rangle$ and $\langle U', B', C\rangle$ are a pair of m.e. conductors w.r.t. uncovered itinerary $\langle A, B, U, .., U', B', C\rangle$, or $\langle A, B', U'\rangle$ and $\langle U, B, C\rangle$ are a pair of m.e. conductors w.r.t. uncovered itinerary $\langle A, B', U', .., U, B, C\rangle$. The latter all follow from rule (iii) in the original CET, and therefore rules (ii) + (iii) combined are equivalent to rules (ii).a + (ii).b + (iii) in the original CET.

(iv) Again by corollary 1, the virtual collider triples in rule (iv) either correspond to a virtual *v*-structure (equivalent to an unshielded imperfect nonconductor in $\mathcal{G}$), or are part of a *u*-structure (equivalent to a pair of m.e. conductors on an uncovered itinerary in $\mathcal{G}$). Therefore, for rule (iv)

we can consider three distinct cases: 1) both virtual collider triples $\langle A, B, C\rangle$ and $\langle A, D, C\rangle$ correspond to virtual *v*-structures, 2) one virtual collider triple corresponds to a virtual *v*-structure, and the other is part of a *u*-structure, or 3) both $\langle A, B, C\rangle$ and $\langle A, D, C\rangle$ are part of a *u*-structure. Below we will tackle each of these cases in turn:

Case (iv).1: if both $\langle A, B, C\rangle$ and $\langle A, D, C\rangle$ correspond to virtual *v*-structures, then they satisfy rule (iv) of the original CET, and so imply that B is ancestor of D in $\mathcal{M}_1$ iff and only iff B is ancestor of D in $\mathcal{M}_2$, and v.v. by symmetry.

Case (iv).2: without loss of generality, assume $\langle A, B, C\rangle$ is a virtual *v*-structure, and $\langle A, D, C\rangle$ is part of a *u*-structure $\langle A, D, D', C\rangle$. Note this implies there cannot be an edge between C and D, otherwise $\langle A, D, D', C\rangle$ would not be a *u*-structure.

Therefore, if $D \longrightarrow B$ in $\mathcal{M}_1$, then $D \longrightarrow B \longleftarrow C$ would be a (virtual) *v*-structure, and already be invariant by rule (ii)+(iii), and so also imply $D \longrightarrow B$ in $\mathcal{M}_2$. If $D \longleftarrow B$ in $\mathcal{M}_1$, i.e. B is NOT a descendant of D in $\mathcal{M}_1$, then by rule (iv) of the original CET, B is also not a descendant of D in $\mathcal{M}_2$, and so imply $D \longleftarrow B$ in $\mathcal{M}_2$. The only remaining possibility is $D \multimap\!\!\!\!- B$ in $\mathcal{M}_1$, which by symmetry then must also apply to $\mathcal{M}_2$. Therefore rule (iv) is also sound for case 2.

Case (iv).3: now both virtual collider triples $\langle A, B, C\rangle$ and $\langle A, D, C\rangle$ are part of a *u*-structure, but neither are virtual *v*-structures. Then by definition 10, either $A \longrightarrow B$ or $C \longrightarrow B$ is in $\mathcal{M}_1$, so without loss of generality assume $A \longrightarrow B$. Then C cannot have an edge to B in $\mathcal{M}_1$, otherwise $\langle A, B, C\rangle$ would be a (virtual) *v*-structure. Similarly, virtual collider triple $\langle A, D, C\rangle$ implies either $A \longrightarrow D$, or $C \longrightarrow D$, but not both (or it would be covered by case (iv).2 already).

(3a) Assume $C \longrightarrow D$. Then if $B \longrightarrow D$ in $\mathcal{M}_1$, then $B \longrightarrow D \longleftarrow C$ would be an unshielded collider triple and invariant by rules (ii)+(iii), implying $B \longrightarrow D$ in $\mathcal{M}_2$ as well. Similarly, if $B \longleftarrow D$, then $A \longrightarrow B \longleftarrow D$ would be an unshielded collider triple, and so appear in $\mathcal{M}_2$ as well, leaving the only other option $B \multimap\!\!\!\!- D$ in $\mathcal{M}_1$ as invariant *u*-structure $\langle A, B, D, C\rangle$ and therefore $B \multimap\!\!\!\!- D$ in $\mathcal{M}_2$ as well.

(3b) Assume $A \longrightarrow D$ (so an arc from A into both B and D, but C not adjacent to either). Then if $B \longrightarrow D$ in $\mathcal{M}_1$, then D is a descendant of B, but D is not an ancestor of B, i.e. B and D belong to different strongly connected components. However, then we have two nonadjacent nodes B and C in $\mathcal{M}_1$ which means they can be *d*-separated by some minimal set $\mathbf{Z}$ in the underlying $\mathcal{G}_1$. By lemma 4, rule 3 this means the set $\mathbf{Z}$ cannot contain D (as it is not an ancestor of either B or C), but including it in the conditioning set would unblock a path via D (as D is descendant of both B and C), i.e. $B \not\perp_{\mathcal{G}} C \,|\, \mathbf{Z} \cup [D]$, which by Lemma 4 rule 2 implies that D *cannot* be an ancestor of B (or C). Therefore $B \longrightarrow D$ is then invariant and appears in $\mathcal{M}_2$ as well. Same for the case $B \longleftarrow D$, but then with the roles of B and D reversed, leading to $B \longleftarrow D$ in $\mathcal{M}_2$ as well. That leaves the case $B \multimap\!\!\!\!- D$ in $\mathcal{M}_1$ as the only remaining option, which by symmetry means it must appear in $\mathcal{M}_2$ as well.

Therefore, rule (iv) is also sound for case 3, which implies that indeed rule (iv) in Theorem 1 is sound. As it also covers all instances of rules (iv) and (v) in the original CET it means that, taken together, rules (i)-(iv) of the new CET are sound and imply all invariant features from the original CET. Therefore, Theorem 1 suffices to establish *d*-separation equivalence, which in turn, under the assumed global directed Markov property, ensures 'if and only if' Markov equivalence between two CMAGs $\mathcal{M}_1$ and $\mathcal{M}_2$. QED. ■

From section 4.1:

Lemma 3 *In a CMAG $\mathcal{M}$ corresponding to directed graph $\mathcal{G}$, two variables X and Y are adjacent, iff X and Y are (virtually) adjacent in $\mathcal{G}$.*

Proof Follows directly from Lemma 1 in (Richardson, 1997). ■

Theorem 2 *For two different cyclic directed graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, let $\mathcal{P}_1$ and $\mathcal{P}_2$ be the corresponding CPAGs output by (Graph-to-CPAG) algorithm 2. Then $\mathcal{G}_1$ is Markov equivalent to $\mathcal{G}_2$ iff $\mathcal{P}_1 = \mathcal{P}_2$.*

Proof Soundness of the algorithm follows from Theorem 1, in combination with the fact that each orientation has a direct match to an invariant feature in the CET rules and is therefore sound. As the algorithm processes each rule exhaustively, this guarantees the output is a valid CPAG.

Remainder of the proof strategy carries over directly from Theorem 2 in Richardson (1996): if any of the orientations triggers in one graph but not the other, then there must be a difference in one or more *d*-separation statement(s) meaning they are not Markov equivalent. We already showed in the proof of Theorem 1 that CET rules (i)-(ii) were equivalent to the original CET rules (i)-(iv), which (again by the proof of the original Theorem 2) ensures that, for two Markov equivalent graphs, $\mathcal{P}_1$ and $\mathcal{P}_2$ have the same skeleton, *v*-structures, and virtual collider triples.

The final orientation rule (iv), corresponding to original CET rules (iv)+(v), has a slightly stronger implication than the original, but still cannot introduce or destroy a virtual collider triple, and so if it triggers in one graph, then it triggers in the other graph. Therefore, if $\mathcal{P}_1$ and $\mathcal{P}_2$ differ after processing CET-(iv), then $\mathcal{G}_1$ and $\mathcal{G}_2$ must differ on some invariant feature, and so are not Markov equivalent. ■

3 MARKOV PROPERTIES FOR STRUCTURAL CAUSAL MODELS

We state here some of the key definitions and results in the theory of Structural Causal Models (SCMs). These models, also known as Structural Equation Models (SEMs), were introduced a century ago by Wright (1921) and popularized in AI by Pearl (2009). We follow here the treatment of Bongers et al. (2021), as it deals with cycles in a mathematically rigorous way.

Definition 1 (SCM) *A Structural Causal Model (SCM) is a tuple* $M = \langle \mathbf{V}, \mathbf{W}, \mathcal{X}_{\mathbf{V}}, \mathcal{X}_{\mathbf{W}}, \mathbf{f}, P_M \rangle$ *of:*

1. *finite disjoint index sets* $\mathbf{V}, \mathbf{W}$ *for the endogenous and exogenous variables in the model, respectively;*
2. *a product of standard measurable spaces* $\mathcal{X}_{\mathbf{V}} = \prod_{v \in \mathbf{V}} \mathcal{X}_v$*, which define the domains of the endogenous variables;*
3. *a product of standard measurable spaces* $\mathcal{X}_{\mathbf{W}} = \prod_{w \in \mathbf{W}} \mathcal{X}_w$*, which define the domains of the exogenous variables;*
4. *a measurable function* $\mathbf{f} : \mathcal{X}_{\mathbf{V}} \times \mathcal{X}_{\mathbf{W}} \to \mathcal{X}_{\mathbf{V}}$*, the causal mechanism;*
5. *a product probability measure* $P_M = \prod_{w \in \mathbf{W}} P_w$ *on* $\mathcal{X}_{\mathbf{W}}$*, with each* P_w *a probability measure on* $\mathcal{X}_w$*, specifying the exogenous distribution.*

The causal structure of the SCM is encoded by the dependences of the components of $\mathbf{f}$ on the variables in the model. This is formalized by:

Definition 2 (Parent) *Let* M *be an SCM. We call* $i \in \mathbf{V} \cup \mathbf{W}$ *a parent of* $k \in \mathbf{V}$ *if and only if there does not exist a measurable function* $\tilde{f}_k : \mathcal{X}_{\mathbf{V} \setminus \{i\}} \times \mathcal{X}_{\mathbf{W} \setminus \{i\}} \to \mathcal{X}_k$ *such that for* P_M*-almost every* $\mathbf{w} \in \mathcal{X}_{\mathbf{W}}$*, for all* $\mathbf{v} \in \mathcal{X}_{\mathbf{V}}$*,*

$$v_k = \tilde{f}_k(\mathbf{v}_{\setminus i}, \mathbf{w}_{\setminus i}) \iff v_k = f_k(\mathbf{v}, \mathbf{w}).$$

Intuitively, this means that the k'th component of $\mathbf{f}$ does not depend on the i'th variable. This definition allows us to define the directed mixed graph (DMG) associated to an SCM:

Definition 3 (Graph) *Let* M *be an SCM. The graph of* M*, denoted* $\mathcal{G}(M)$*, is defined as the directed mixed graph with nodes* $\mathbf{V}$*, directed edges* $v_1 \to v_2$ *iff* v_1 *is a parent of* v_2 *according to* M*, and bidirected edges* $v_1 \leftrightarrow v_2$ *iff there exists* $w \in \mathbf{W}$ *such that* w *is parent of both* v_1 *and* v_2 *according to* M*.*

If $\mathcal{G}(M)$ is acyclic, we call the SCM M acyclic, otherwise we call the SCM cyclic. If $\mathcal{G}(M)$ contains no bidirected edges, we call the endogenous variables in the SCM M causally sufficient (which is what we assumed in the present work for simplicity).

SCMs provide an implicit description of their solutions.

Definition 4 (Solutions) *A random variable* $\mathfrak{X} = (\mathfrak{X}_{\mathbf{V}}, \mathfrak{X}_{\mathbf{W}})$ *is called a solution of the SCM* M *if* $\mathfrak{X}_{\mathbf{V}} = (\mathfrak{X}_v)_{v \in \mathbf{V}}$ *with* $\mathfrak{X}_v \in \mathcal{X}_v$ *for all* $v \in \mathbf{V}$*,* $\mathfrak{X}_{\mathbf{W}} = (\mathfrak{X}_w)_{w \in \mathbf{W}}$ *with* $\mathfrak{X}_w \in \mathcal{X}_w$ *for all* $w \in \mathbf{W}$*, the distribution* $\mathbb{P}(\mathfrak{X}_{\mathbf{W}})$ *is equal to the exogenous distribution* P_M*, and the structural equations:*

$$\mathfrak{X}_v = f_v(\mathfrak{X}_{\mathbf{V}}, \mathfrak{X}_{\mathbf{W}}) \quad a.s.$$

hold for all $v \in \mathbf{V}$*.*

For acyclic SCMs, solutions exist and have a unique distribution that is determined by the SCM. This is not generally the case in cyclic SCMs, as these could have no solution at all, or could have multiple solutions with different distributions.

Definition 5 (Unique solvability) *An SCM* M *is said to be uniquely solvable w.r.t.* $\mathbf{O} \subseteq \mathbf{V}$ *if there exists a measurable mapping* $\mathbf{g}_{\mathbf{O}} : \mathcal{X}_{\mathrm{pa}_{\mathcal{G}(M)}(\mathbf{O}) \setminus \mathbf{O}} \to \mathcal{X}_{\mathbf{O}}$ *such that for* P_M*-almost every* $\mathbf{w} \in \mathcal{X}_{\mathbf{W}}$*, for all* $\mathbf{v} \in \mathcal{X}_{\mathbf{V}}$*:*

$$\mathbf{v}_{\mathbf{O}} = \mathbf{g}_{\mathbf{O}}(\mathbf{v}_{(\mathrm{pa}_{\mathcal{G}(M)}(\mathbf{O}) \setminus \mathbf{O}) \cap \mathbf{V}}, \mathbf{w}_{\mathrm{pa}_{\mathcal{G}(M)}(\mathbf{O}) \cap \mathbf{W}}) \iff \mathbf{v}_{\mathbf{O}} = \mathbf{f}_{\mathbf{O}}(\mathbf{v}, \mathbf{w}).$$

Loosely speaking: the structural equations for $\mathbf{O}$ have an essentially unique solution for $\mathbf{v}_{\mathbf{O}}$ in terms of the other variables appearing in those equations. If M is uniquely solvable with respect to $\mathbf{V}$ (in particular, this holds if M is acyclic), then it induces a unique observational distribution $P_M(\mathfrak{X}_{\mathbf{V}})$, the push-forward of P_M through $\mathbf{g}_{\mathbf{V}}$.

One of the key aspects of SCMs—which we do not discuss here in detail because we do not make use of it in this work—is their causal semantics, which is defined in terms of interventions. Instead, we discuss only their probabilistic properties. In particular, under appropriate assumptions, the graph $\mathcal{G}(M)$ of an SCM M represents conditional independences that its solutions must satisfy. As shown already by Spirtes (1994, 1995), the directed global Markov property does *not* hold in general for cyclic SCMs.

Example 1 (d-separation fails) *Consider the SCM* $M = \langle \{1,2,3,4\}, \{5,6,7,8\}, \mathbb{R}^4, \mathbb{R}^4, \mathbf{f}, P_M \rangle$ *where* P_M *is the standard-normal distribution on* $\mathbb{R}^4$*, and the causal mechanism is given by:*

$$\mathbf{f}(\mathbf{x}) = (x_5, x_6, x_1 x_4 + x_7, x_2 x_3 + x_8)$$

The graph $\mathcal{G}(M)$ *has edges* $1 \longrightarrow 3$*,* $2 \longrightarrow 4$*,* $3 \longrightarrow 4$*,* $4 \longrightarrow 3$*. This SCM is uniquely solvable with respect to its strongly*

connected components $\{1\}$, $\{2\}$, *and* $\{3,4\}$. *One can check that for every solution* $\mathfrak{X}$ *of* M, $\mathfrak{X}_1$ *is not independent of* $\mathfrak{X}_2$ *given* $\{\mathfrak{X}_3, \mathfrak{X}_4\}$. *However, the nodes* 1 *and* 2 *are* d*-separated given* $\{3,4\}$ *in* $\mathcal{G}(M)$. *Hence the global directed Markov property does not hold for* M.

For more concrete examples of cyclic SCMs, we refer the reader to (Bongers et al., 2021). Spirtes (1994) proved a weaker Markov property in terms of a 'collapsed graph', assuming causal sufficiency and densities. Forré and Mooij (2017) found the following formulation in terms of 'σ-separation' that is immediately applicable to the graph of the SCM itself.

Definition 6 (Blockable and unblockable noncolliders) *Let* $\mathcal{G}$ *be a directed mixed graph and* π *a path in* $\mathcal{G}$. *We call a noncollider on* π *unblockable if it is not an end-node and it only has outgoing edges on* π *to nodes in the same strongly connected component of* $\mathcal{G}$*; otherwise, it is called blockable.*

If $\mathcal{G}$ is acyclic then all noncolliders are blockable.

Definition 7 (σ-separation) *For a triple of node sets* $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ *in a graph* $\mathcal{G}$, *we say that* $\mathbf{X}$ *is* σ*-connected to* $\mathbf{Y}$ *given* $\mathbf{Z}$ *iff there is an* $X \in \mathbf{X}$ *and* $Y \in \mathbf{Y}$ *such that there is a path* π *between* X *and* Y *on which every blockable noncollider is not in* $\mathbf{Z}$, *and every collider on* π *is an ancestor of* $\mathbf{Z}$*; otherwise* $\mathbf{X}$ *and* $\mathbf{Y}$ *are said to be* σ*-separated given* $\mathbf{Z}$.

Note the small difference with the definition of d-connection: σ-connection only considers the *blockable* noncolliders. The following general result was shown by Forré and Mooij (2017).

Theorem 1 (σ-Separation Markov property) *Let* M *be an SCM that is uniquely solvable w.r.t. each strongly connected component of* $\mathcal{G}(M)$. *Then, the observational distribution of* M *exists and is unique. Furthermore, for a solution* $\mathfrak{X}$ *of* M *and for* $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$*: if* $\mathbf{A}$ *is* σ*-separated from* $\mathbf{B}$ *given* $\mathbf{C}$ *in* $\mathcal{G}(M)$, *then* $\mathfrak{X}_\mathbf{A}$ *is conditionally independent of* $\mathfrak{X}_\mathbf{B}$ *given* $\mathfrak{X}_\mathbf{C}$.

Proof See the proof of Theorem A.21 in Bongers et al. (2021). ■

Under certain additional assumptions, one can show the stronger d-separation criterion (also known as the global directed Markov property).

Theorem 2 (d-Separation Markov property) *Let* M *be an SCM that satisfies one of the following three assumptions:*

1. M *is acyclic;*
2. - *all endogenous domains* $\mathcal{X}_v$ *for* $v \in \mathbf{V}$ *are discrete, and*
   - M *is uniquely solvable w.r.t. each ancestral subset* $A \subseteq \mathbf{V}$ *(that is, each subset* $A \subseteq \mathbf{V}$ *such that* $\mathrm{an}_{\mathcal{G}(M)}(A) = A$*);*
3. - $\mathcal{X}_\mathbf{V} = \mathbb{R}^\mathbf{V}$ *and* $\mathcal{X}_\mathbf{W} = \mathbb{R}^\mathbf{W}$, *and*
   - $\mathbf{f}$ *is a linear mapping, and*
   - *each* $v \in \mathbf{V}$ *has at least one parent in* $\mathbf{W}$ *according to* M, *and*
   - P_M *has a density w.r.t. the Lebesgue measure on* $\mathbb{R}^\mathbf{W}$.

Then, the observational distribution of M *exists and is unique. Furthermore, for a solution* $\mathfrak{X}$ *of* M *and for* $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$*: if* $\mathbf{A}$ *is* d*-separated from* $\mathbf{B}$ *given* $\mathbf{C}$ *in* $\mathcal{G}(M)$, *then* $\mathfrak{X}_\mathbf{A}$ *is conditionally independent of* $\mathfrak{X}_\mathbf{B}$ *given* $\mathfrak{X}_\mathbf{C}$.

Proof See the proof of Theorem A.7 in Bongers et al. (2021). The acyclic case is well known. The discrete case fixes the erroneous theorem by Pearl and Dechter (1996), for which a counterexample was found by Neal (2000), by adding the assumption of unique solvability with respect to each ancestral subset, and extends it to allow for bidirected edges in the graph. The linear case is an extension of existing results for the linear-Gaussian setting without bidirected edges Spirtes (1994, 1995); Koster (1996) to a linear (possibly non-Gaussian) setting with bidirected edges in the graph. ■

For this paper, we assume that the global directed Markov property holds with respect to a graph that contains no bidirected edges. From the above theorem, it follows that this will hold if the data comes from the observational distribution of a causally sufficient SCM that falls into either the acyclic case (1), the discrete case (2), or the linear case (3). Note that these assumptions are sufficient, but not necessary.

References

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. (2021). Foundations of structural causal models with cycles and latent variables. Annals of Statistics, 49(5):2885–2915.

Claassen, T. and Heskes, T. (2011). A logical characterization of constraint-based causal discovery. In Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, pages 135–144.

Forré, P. and Mooij, J. M. (2017). Markov properties for graphical models with cycles and latent variables. arXiv.org preprint, arXiv:1710.08775 [math.ST].

Koster, J. (1996). Markov properties of nonrecursive causal models. The Annals of Statistics, 24(5):2148–2177.

Neal, R. (2000). On deducing conditional independence from d-separation in causal graphs with feedback. Journal of Artificial Intelligence Research, 12:87–91.

Pearl, J. (2009). Causality: Models, Reasoning and Inference. Cambridge University Press.

Pearl, J. and Dechter, R. (1996). Identifying independence in causal graphs with feedback. In Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence (UAI-96), pages 420–426.

Richardson, T. (1996). Discovering cyclic causal structure. Technical Report CMU-PHIL-68, Carnegie Mellon University.

Richardson, T. (1997). A characterization of Markov equivalence for directed cyclic graphs. International Journal of Approximate Reasoning, 17(2-3):107–162.

Spirtes, P. (1994). Conditional independence in directed cyclic graphical models for feedback. Technical Report CMU-PHIL-54, Carnegie Mellon University.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95), pages 499–506.

Wright, S. (1921). Correlation and causation. Journal of Agricultural Research, 20:557–585.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artificial Intelligence, 172(16-17):1873–1896.