Dealing with cycles in graph-based probabilistic models: the case of Logical Credal Networks

Fabio G. Cozman[1] **Radu Marinescu**[2] **Junkyu Lee**[3] **Alexander Gray**[4] **Denis D. Mauá**[5]

[1] Escola Politénica, Universidade de São Paulo, Brazil
[2]IBM Research, Ireland
[3]IBM Research, Yorktown Heights, US
[4]Centaur AI Institute
[5]Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil

Abstract

We examine the consequences of directed cycles in graph-based representations of joint distributions, investigating the effect of cycles on Markov conditions and on Gibbs factorizations. We focus on Logical Credal Networks, a flexible and general formalism, showing that Koster's theory of Directed-Undirected Mixed Graphs (DUMGs) leads to an interesting Gibbs factorization. We show that inferences with DUMGs lead to multilinear programs. We also study the failure of global Markov conditions in cyclic structural equation models, connecting that failure to probabilistic imprecision under interventions.

Keywords. probabilistic logic, credal networks, Markov conditions, cyclic networks

1. Introduction

In this paper, we examine the representation of probabilistic assessments and independence relations through graphs that allow for cycles. We wish to study Markov conditions and Gibbs factorizations associated with such assessments and graphs. We focus on the language of Logical Credal Networks, a rather flexible formalism that lets us touch on several relevant issues [15, 16].

A Logical Credal Network (LCN) in essence consists of a set of assessments such as

$$\mathbb{P}(\phi_j|\psi_j) \in [\alpha_j, \beta_j], \quad j \in \{1, 2, \dots, m\}, \tag{1}$$

where ϕ_j and ψ_j are logical sentences. The original semantics for LCNs associates a directed graph with assessments, and then extracts independence relations from the graph through a local Markov condition [15]. A more detailed description of LCNs is given in Section 3 (we collect necessary concepts about graphs in Section 2).

It turns out that in many circumstances the original semantics of LCNs can be interpreted through the semantics of chain graphs [5]. However, there are LCNs that cannot be interpreted using chain graphs, and perhaps we should consider semantics that are explicitly based on global Markov conditions, so as to produce useful Gibbs factorizations. This question has been raised recently [5], but it has not been investigated in detail.

We examine an interpretation of assessments where conditioning leads to directed edges while conditioned logical connections lead to undirected edges. We obtain a semantics based on Koster's Directed-Undirected Mixed Graphs (DUMGs) [12] and their associated global Markov condition (under a positivity assumption). Section 4 focuses on these topics.

We also look at the possible failure of global Markov conditions applied to cyclic Structural Equation Models. We show that, when the global Markov condition for DUMGs fails, there must be some intervention that produces interval-valued inferences. This exposes an interesting connection between (failure of) separation in graphs and imprecise probabilities. Section 5 focuses on these topics.

Even though we consider Logical Credal Networks in this paper, our results should be useful to other graph-theoretical formalisms that allow for cycles. We summarize our main results in Section 6, where we also collect a number of suggestions for future work.

2. Background: several kinds of graphs

This section collects a number of concepts and results that are used later; it may be skipped in a first reading, and consulted whenever necessary.

A *Directed-Undirected Mixed Graph* (DUMG) consists of a set of nodes and a set of edges (each edge may be directed or undirected). A *path* is a sequence of edges where each edge may be undirected or directed in the same direction, where all visited nodes must be distinct except for the first and last. A *cycle* is a path with identical first and last nodes. An *undirected cycle* is a cycle containing only undirected edges. A *directed cycle* is a cycle

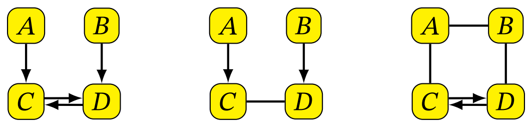

Figure 1. *Left: a directed graph with a cycle. Middle: a chain graph. Right: a DUMG.*

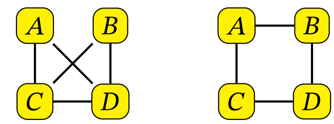

Figure 2. *Left: moral graph of the directed graph in Figure 1 (left): we must connect A and D as they have a common child C, then we must connect B and C as they have a common child D. Right: moral graph of graph in Figure 1 (middle): A and B are connected because they are parents of the same undirected path component.*

that is not undirected, thus containing only undirected edges and directed edges that point in the same direction (note that a directed cycle may have no undirected edges). An *undirected path component* is a maximal set of nodes connected by undirected edges, or a single node without undirected connections to other nodes. For a DUMG G, $U(G)$ denotes the set of undirected path components of G.

A *chain graph* is a DUMG with no directed cycle. A *directed graph* contains only directed edges. An *undirected graph* contains only undirected edges. Figure 1 presents a few relevant examples.

Given a DUMG G and a set $\mathcal{A}$ of its nodes, the subgraph induced by $\mathcal{A}$, denoted by $G_{\mathcal{A}}$, is the graph with all nodes in $\mathcal{A}$ and all edges from G that connect nodes in $\mathcal{A}$.

The set of *parents* of node A, denoted by $\mathsf{pa}(A)$, is the set of all nodes B such that there is a directed edge from B to A. The set of parents of a set of nodes $\mathcal{A}$, denoted by $\mathsf{pa}(\mathcal{A})$, is the union of the parents of each node in $\mathcal{A}$. The *closure* of a set of nodes $\mathcal{A}$, denoted by $\mathsf{cl}(\mathcal{A})$, is the union of $\mathcal{A}$ and the set of nodes that have a path of length 1 to a node in $\mathcal{A}$ (that is, the parents of nodes in $\mathcal{A}$ and the nodes with an undirected edge to a node in $\mathcal{A}$). If there is a path from B to A, B is an *ancestor* of A. The set of ancestors of A is denoted $\mathsf{anc}(A)$. A set $\mathcal{A}$ is an *anterior set* iff for all $A \in \mathcal{A}$ we have $\mathsf{anc}(A) \cup \{A\} \subseteq \mathcal{A}$ (the empty set is an anterior set; the set of all nodes is an anterior set; for each node A, the set $\mathsf{anc}(A) \cup \{A\}$ is an anterior set). Denote by $\mathsf{an}(\mathcal{A})$ the smallest anterior set containing $\mathcal{A}$. (The theory of chain graphs often refers to anterior sets as *ancestral sets* [3].)

Given a DUMG and a set of nodes $\mathcal{A}$, the *anterior graph* of $\mathcal{A}$ is the graph induced by $\mathsf{an}(\mathcal{A})$, the smallest anterior set containing $\mathcal{A}$.

A *moralized graph* is obtained from a graph by adding an undirected edge between each two nodes that have children in the same undirected path component and that are not already joined, then turning all edges into undirected ones and merging those pairs of edges that are now identical. Thus a moralized graph is an undirected graph. Figure 2 illustrates this concept.

3. The language: Logical Credal Networks

A quintessential feature of many probabilistic logics is the ability to constrain the probability of a sentence ϕ, perhaps by writing $\mathbb{P}(\phi) \leq \beta$ [11]. Bayesian networks instead rely on precise conditional assessments and directed acyclic graphs to guarantee that a unique joint probability distribution is always specified [7]. While probabilistic logics usually emphasize flexibility, Bayesian networks emphasize guarantees of uniqueness. Credal networks stay somewhere in the middle [4]: a (separately specified) credal network is similar to a Bayesian network but each conditional probability distribution $\mathbb{P}(X_i|\mathsf{pa}(X_i) = \pi)$, for each variable X_i and each possible π, can be selected from a set of distributions (a credal set) $K(X_i|\mathsf{pa}(X_i) = \pi)$.

A *Logical Credal Network* (LCN) consists of a set of *assessments* with syntax given by Expression (1), where $0 \leq \alpha_j \leq \beta_j \leq 1$ and ϕ_j and ψ_j are logical sentences (ψ_j may be omitted). We may use, as syntactic sugar, assessments $\mathbb{P}(\phi_j|\psi_j) = \alpha_j$, $\mathbb{P}(\phi_j) = \alpha_j$, $\mathbb{P}(\phi_j|\psi_j) \leq \beta_j$, etc, whenever appropriate. In addition, each assessment is marked either with a "coupled" flag or an "uncoupled" flag, that leads to different semantics as explained later. If nothing is said about the flag of an assessment, we assume it to be "coupled".[1]

We restrict ourselves in this paper to propositional sentences. Sentences ϕ_j and φ_j in Expression (1) contain propositions from a set $\{A_1, \ldots, A_n\}$. Each one of these propositions is associated with a binary indicator variable X_i (where $X_i = 0$ iff A_i is false and $X_i = 1$ iff A_i is true). Thus there is a correspondence between truth assignments that assign true/false to each proposition, and complete configurations of the corresponding binary variables (a *configuration* for a set of variables assigns a precise value for each one of the variables in the set). We could easily accommodate categorical variables by expanding the syntax, but the restriction to binary variables simplifies the discussion. We freely replace propositions by their corresponding (random) variables, and treat logical sentences as (random) variables whenever needed; whenever possible, we just write ϕ instead of $\phi = 1$ (as done in Expression (1)). For instance,

[1] In the original LCN proposal, each assessment had an associate "true" or "false" flag [15] whose meaning was perhaps more difficult to memorize.

we write logical expressions such as $X_2 \vee X_4$ instead of $A_2 \vee A_4$, and $\mathbb{P}(X_1 \wedge \neg X_3)$ to indicate the probability that $A_1 \wedge \neg A_3$ is true. In addition, whenever possible we use comma instead of the symbol $\wedge$ to indicate a conjunction of assignments to variables; for instance, we write $\mathbb{P}(X_1 = 1, X_2 = 0)$ instead of $\mathbb{P}((X_1 = 1) \wedge (X_2 = 0))$.

The original semantics for LCNs translates sets of assessments to directed graphs, referred to as *primal graphs*, from which their semantics is extracted. A simplified version of primal graphs, suitable for our purposes here, is as follows. Given a LCN, its *reduced primal graph* is built by:

1) Creating a node for each variable X_i;
2) For each sentence ψ_j, creating a directed edge from each variable in ψ_j to each variable in the corresponding ϕ_j;
3) For each ϕ_j in a coupled-assessment, creating directed edges in both directions for each pair of variables in ϕ_j.

In this graph, X is an *lcn-parent* of Y iff there is an edge from X to Y. And X is an *lcn-descendant* of Y iff there is a directed path from Y to X in which no intermediate node is an lcn-parent of Y. While lcn-parents are the usual parents in directed graphs, lcn-descendants are slightly different. The independence relations imposed by the LCN are then given by a *local* Markov condition on the reduced primal graph: a node X is independent, given its lcn-parents, of all nodes that are not X itself nor lcn-descendants of X nor lcn-parents of X. We refer to this condition as LMC(LCN) [5], using the explicit reference to "LCN" to emphasize that lcn-descedats are used rather than the usual descendants.

A Bayesian network with binary variables can be easily written down as a LCN with an acyclic reduced primal graph.

Example 3.1. Suppose we have variables D, U, Y and Z, and assessments $\mathbb{P}(U = u) = \gamma_u$, $\mathbb{P}(Z = z) = \delta_z$, $\mathbb{P}(D = d|U = u, Z = z) = \mu_{duz}$ and $\mathbb{P}(Y = y|D = d, U = u) = \nu_{ydu}$ for all possible d, u, y, z (these assessments may be coupled or uncoupled ones with identical effect). The reduced primal graph is depicted in Figure 3. □

By letting assessments be interval-valued, one can easily build credal networks [4] as long as variables are binary. For instance, in the previous example we might have $\mathbb{P}(Z = 1) \in [0.1, 0.2]$. And the syntax is flexible enough so as to allow some practical settings to be easily conveyed; the next example comes from the literature on interventional reasoning [20].

Example 3.2. Suppose we have a binary variable Z indicating treatment (yes=1/no=0), and a variable D that indicates whether the treatment was indeed used (yes=1/no=0); finally, we have an observed response Y (positive=1/negative=0), and a latent variable U that is a common cause of D and Y. This is similar to the classic example where Z is an instrumental variable [2]; here we simplify matters by assuming the latent variable U is binary. The intended Bayesian network that carries the relevant information is depicted in Figure 3. Usually one assumes that (infinite) data is available on Z, D and Y; for instance, we may estimate:

$$\begin{array}{ll} \mathbb{P}(Y{=}0,D{=}0,Z{=}0){=}0.288, & \mathbb{P}(Y{=}0,D{=}0,Z{=}1){=}0.002, \\ \mathbb{P}(Y{=}0,D{=}1,Z{=}0){=}0.288, & \mathbb{P}(Y{=}0,D{=}1,Z{=}1){=}0.017, \\ \mathbb{P}(Y{=}1,D{=}0,Z{=}0){=}0.036, & \mathbb{P}(Y{=}1,D{=}0,Z{=}1){=}0.067, \\ \mathbb{P}(Y{=}1,D{=}1,Z{=}0){=}0.288, & \mathbb{P}(Y{=}1,D{=}1,Z{=}1){=}0.014. \end{array}$$

We should take these assessments as uncoupled ones (if they are adopted as coupled-assessments instead, we produce all possible directed edges among Y, D and Z). And to "build" the intended graph, we can just introduce uncoupled-assessments $\mathbb{P}(D = 1, Y = 1|U = 1) \in [0, 1]$, $\mathbb{P}(D = 1|Z = 1) \in [0, 1]$ and $\mathbb{P}(Y = 1|Z = 1) \in [0, 1]$. The reduced primal graph for this LCN is indeed in Figure 3. □

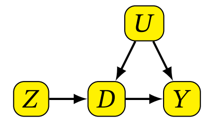

Figure 3. *The reduced primal graph of a Bayesian network / an acyclic reduced primal graph.*

A LCN may have a cyclic primal graph:

Example 3.3. Suppose we have binary variables A, B, C, D, and coupled-assessments $\mathbb{P}(C = 1|A = 1) = \alpha$, $\mathbb{P}(D = 1|B = 1) = \beta$ $\mathbb{P}(C \wedge D) = \gamma$. The reduced primal graph of this LCN is depicted in Figure 1 (left). □

Example 3.4. The same reduced primal graph in Figure 1 (left) is obtained by assessments $\mathbb{P}(A = 1) = \alpha$, $\mathbb{P}(B = 1) = \beta$, $\mathbb{P}(C = 1|A = 1, D = 1) = \gamma$, and $\mathbb{P}(D = 1|B = 1, C = 1) = \delta$. □

If a pair of nodes is connected by two edges in distinct directions, for instance $A \leftrightarrows B$, we say they are connected by *opposing arrows*. If we take the reduced primal graph and replace each pair of opposing arrows by an undirected edge, we obtain a new graph that is referred to as the *structure* of the LCN [5].

A *chain LCN* is a LCN such that its structure is a chain graph. Figure 4 (left) depicts the reduced primal graph for an example chain LCN, and its structure (right).

For a chain LCN, the LMC(LCN) has the same effect as the usual global Markov condition for chain graphs applied to the structure, as long as probabilities are positive [5]. This global Markov condition is a somewhat complex condition: sets of nodes $\mathcal{A}$ and $\mathcal{B}$ are independent given set $\mathcal{C}$ if $\mathcal{A}$ and $\mathcal{B}$ are separated by $\mathcal{C}$ in the moralized (thus undirected) anterior graph of $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$.

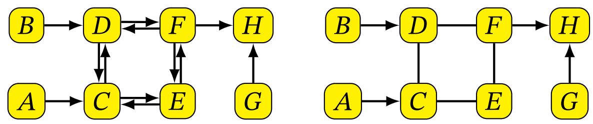

Figure 4. *Left: reduced primal graph for a chain LCN. Right: structure of the chain LCN.*

We refer to this global Markov condition as GMC, and abuse terminology in the remainder of this paper by applying the GMC to undirected graphs, directed graphs, directed-undirected graphs as needed, with the understanding that an appropriate definition of moralized anterior graph applies in each particular case.

For chain LCNs there exists a variant of d-separation that implies independence [22] and we do have a Gibbs factorization that follows from the (chain graph) structure, provided that probabilities are positive [3].

The LCN formalism is not the only one that mixes a graph-based representation with flexible probabilistic assessments. Bayesian Logic [1] advocates the use of a directed graph over variables (to represent independence relations) plus probabilistic logic assessments. A similar strategy is adopted by PPL networks [6], that employ graphs both with assessments and with constraints. However, LCNs are somewhat unique in that a graph is not assumed from the outset; rather, the assessments are assumed as input, and the related graphs are built out of them. This strategy may be a bit disconcerting at times: $\mathbb{P}(A \wedge (B \vee \neg B)) = \alpha$ and $\mathbb{P}(A) = \alpha$ have clearly distinct effects on the primal graph, even though assessments are logically equivalent. On the other hand, these syntactic conventions give the user the flexibility to easily include edges if necessary (see the vacuous assessments at the end of Example 3.2).

In any case, one aspect of LCNs sets them apart from seemingly similar proposals [1, 6]: a primal graph may contain directed cycles. This makes LCNs an excellent case study in this paper.

4. Cycles and their Semantics

We first examine the different kinds of cycles in LCNs, and then propose a semantics based on Koster's theory of DUMGs [12, 13].

4.1. Conditioned sentences: opposing arrows or undirected edges?. We should ask whether it is possible to obtain a Gibbs factorization that goes beyond chain LCNs. However, it seems difficult to do so within the confines of local Markov conditions. A local condition is quite weak in the presence of "long" cycles because in a cycle there are no nondescendants [5]. In addition, most results that yield Gibbs factorizations rely on global Markov conditions so as to use (versions of) the Hammersley-Clifford theorem [9, 12, 13, 21]. Alas, the simple and elegant local Markov conditions we have at our disposal seem too weak for an effective extension to cyclic settings.

We thus assume that a global Markov condition is needed. It does not seem that we can find a global condition that is compatible with the LMC(LCN), as a well-known example illustrates [21].

Example 4.1. Take the graph in Figure 1 (left). If this graph is the reduced primal graph of an LCN, then the LMC(LCN) yields $A \perp\!\!\!\perp B$, $A \perp\!\!\!\perp D|B,C$ and $B \perp\!\!\!\perp C|A,D$, while the GMC applied to the moralized reduced primal graph yields $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B|C,D$ (where we use $\perp\!\!\!\perp$ as usual to mean "independence"). The relevant moral graph appears in Figure 2 (left). □

That is, if we adopt the GMC for (possibly cyclic) primal graphs, we have to discard the LMC(LCN) and in fact give up the Gibbs factorization based on chain graphs for chain LCNs.

Does the GMC applied to the moralized reduced primal graph (not the moralized structure!) lead to a satisfactory set of independence relations?

To examine this question, consider a setting with zero/one probabilities. Suppose we have coupled-assessments $\mathbb{P}(\phi_j) = 1$. The reduced primal graph has opposing arrows between any two variables that appear in a sentence ϕ_j; the GMC applied to the moralized version of this primal graph introduces many connections through the moralized graph.

Example 4.2. Consider four variables A, B, C, D, and assessments $\mathbb{P}(A \vee B) = \mathbb{P}(A \vee C) = \mathbb{P}(B \vee D) = \mathbb{P}(C \vee D) = 1$. The primal graph is a "square" where sides have opposing arrows, and where *no* independence is extracted from the moralized graph (a fully connected clique). On the other hand, the structure of this LCN is a "square" where sides have undirected edges, and where the GMC produces, for instance, independence of A and D given B and C, a most natural result given the logical constraints that are assigned probability one.

In fact, a LCN whose reduced graph is a long cycle of opposing arrows fails to produce, through the GMC applied to this primal graph, the independence of a variable X, given its parent and child, from the remaining nodes. On the other hand, the structure of this LCN is an undirected cycle where the GMC yields, for any X, given its parent and child, independence from the remaining variables. □

As a digression, note that the *constraint graph* of a set of logical constraints is an undirected graph that captures all connections (and lack thereof) between propositions [8]. Such logical connections are precisely captured by the structure, but not by the primal graph with opposing arrows.

So, it seems to make sense to keep undirected edges when we have propositions in conditioned sentences

ϕ_j; in fact, this comment applies even when probabilities are not just zero/one. Note that some independence relations produced by undirected edges cannot be easily reproduced via directed edges — one instance is the "undirected square" in the previous example.

However, it does not seem desirable to translate every edge in a primal graph directly to an undirected edge: directed graphs lead to unique independence patterns, most notably through colliders, that a translation to undirected edge lacks [20]. Note that the GMC does not refer to a single moralized graph, but rather to a moralized graph for each triple of sets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. So we may have a pattern $A \rightarrow C \leftarrow B$, where A and B are independent despite the fact that the moral graph containing $\{A, B, C\}$ has a connection between A and B.

Indeed, there are two different patterns that lead to opposing arrows in a primal graph, and they should probably be treated differently by a global Markov condition. One pattern comes from logical connections in conditioned sentences ϕ_j, as we have discussed in the previous paragraphs. Another pattern emerges when a set of distinct assessments leads to edges in opposing directions. Consider:

Example 4.3. Suppose we have assessments $\mathbb{P}(C|A \vee B) \in [0.1, 0.9]$, $\mathbb{P}(A|C) \in [0.1, 0.9]$, $\mathbb{P}(B|C) \in [0.1, 0.9]$. Now the primal graph is $A \leftrightarrows C \leftrightarrows B$, and the moralized graph has all variables connected — a sensible outcome, it seems, given that an assessment over $\mathbb{P}(C|A \vee B)$ connects these three variables. On the other hand, the structure is $A - C - B$, imposing independence of A and B given C (exactly what the LMC(LCN) produces).

To emphasize these points, consider assessments $\mathbb{P}(A|B \vee C) \in [0.1, 0.9]$, $\mathbb{P}(B|A \vee D) \in [0.1, 0.9]$, $\mathbb{P}(C|B \vee D) \in [0.1, 0.9]$, $\mathbb{P}(D|B \vee C) \in [0.1, 0.9]$. The primal graph is a "square" where sides are opposing arrows, and all variables are connected in the moralized primal graph. But the GMC applied to the structure produces the independence of A and D given B and C. □

That is, we want the algebraic connection between variables in a collider to be somehow preserved; we should not want these connections to be "erased" by other assessments that generate opposing edges.

Therefore, there are reasons for directed and undirected edges to coexist, even in the presence of cycles. We investigate this possibility in the next subsection.

4.2. Mixed-structures, separation, and factorization. In the previous section we presented arguments in favor of a mix of undirected edges (corresponding to "logical connections") and directed edges (corresponding to "conditioning connections") in the semantics of a LCN.

We adopt a mixed translation from assessments to graph-theoretical constructs; more precisely, from assessments we build a graph as follows:

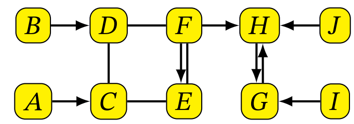

Figure 5. *A mixed-structure that is not a chain graph.*

1) There is a node for each variable X_i;
2) For each assessment $\mathbb{P}(\phi_j|\psi_j) \in [\alpha_j, \beta_j]$ (coupled or uncoupled), there is a directed edge from each variable in ψ_j to each variable in ϕ_j.
3) For each coupled-assessment $\mathbb{P}(\phi_j|\psi_j) \in [\alpha_j, \beta_j]$, there is an undirected edge between any two variables in ϕ_j.

This sort of graph is referred to as the *mixed-structure* of the LCN by Cozman et al. [5], as they wonder whether other semantics for LCNs might be possible so as to extend the factorization obtained for chain LCNs. We now show that such semantics can indeed be specified.

In a mixed-structure, some of the cycles in the primal graph are eliminated, but directed cycles may still be present; therefore, a mixed-structure is a DUMG as previously defined (Section 2). Note that the structure and the mixed-structure of a LCN may differ: the former does not distinguish between opposing arrows produced by distinct assessments from ones produced by a single sentence with several variables, a distinction that mixed-structures do preserve.

Can we find a global Markov condition for mixed-structures that guarantees a Gibbs factorization following from the mixed-structure at least when probabilities are positive, and that corresponds to the GMC when the mixed-structure is a chain graph? We can actually do so by using some relevant, perhaps not as well-known as deserved, results by Koster [12]: he proposed a global Markov condition for DUMGs (using an appropriate separation criterion), and then obtained an interesting Gibbs factorization.

We now formalize a few concepts, following Koster [12]. For an anterior set $\mathcal{A}$, $\langle\mathcal{A}\rangle$ denotes the union of all anterior sets $\mathcal{B}$ such that $\mathcal{B} \subset \mathcal{A}$. And $J(G)$ denotes, for the DUMG G, the set of all anterior sets $\mathcal{A}$ such that $\langle\mathcal{A}\rangle \subset \mathcal{A}$ (elements of $J(G)$ are called the *join-irreducible elements* of G). The *moral graph* of a DUMG G, denoted by G^m, is obtained by removing all directions of edges, after connecting with an undirected edge each pair of parents of each undirected path component, and then removing multiple parallel undirected edges (this is the exact same definition used to define moralized graphs in Section 2). Figure 6 shows an example. Note that F is a parent of the undirected path component $\{C, D, E, F\}$, and note the different treatment of A and B as parents of an undirected path component (they are connected) and of I and J as parents of nodes connected by opposing arrows (they are not connected).

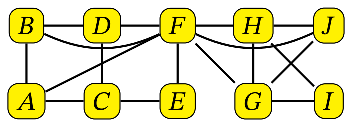

Figure 6. *Moralized graph for the DUMG in Figure 5.*

Koster adopted the *same* global Markov condition used previously [12]: Sets of nodes $\mathcal{A}$ and $\mathcal{B}$ are independent given $\mathcal{C}$ if $\mathcal{A}$ and $\mathcal{B}$ are separated by $\mathcal{C}$ in the moralized anterior graph of $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$.

The key result by Koster [12, Theorem 4.3] can be adapted to our setting, as indicated in more detail later, to obtain the following result: an LCN with a mixed-structure G satisfies the GMC if and only if the joint distribution over all variables satisfies a Gibbs factorization in the cliques of specific moralized anterior graphs, *when* all configurations have positive probability.

This latter positivity assumption is rather inconvenient, but it affects only some parts of Koster's result. Thus it makes sense to present a version of Koster's Theorem 4.3, adapted to the language of LCNs, in parts, where the effect of zero probabilities is isolated.

First, we have:

Theorem 4.1. *Consider a LCN with a mixed-structure G. The following statements are equivalent.*
(a) The LCN satisfies the GMC with respect to G.
(b) The LCN satisfies the GMC with respect to each sub-graph $G_{\mathcal{A}}$ induced by an anterior set $\mathcal{A}$.
(c) The LCN satisfies the GMC with respect to each moralized sub-graph $(G_{\mathcal{A}})^m$ induced by an anterior set $\mathcal{A}$.

A proof of this result is, suitably adapted with respect to definitions, given by Koster [12, Theorem 4.3]. The result does not depend on the values of the joint distribution over variables (those can be zero).

We now look for a Gibbs factorization of the joint distribution. More precisely, consider a LCN with mixed-structure G. We say this LCN satisfies GIBBS(LCN) if and only if any distribution $\mathbb{P}$ that satisfies all constraints implied by assessments (including independence relations) factorize as follows:

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \prod_{\mathcal{A} \in J(G)} \mathbb{P}\big(x_{\mathcal{A}\setminus\langle\mathcal{A}\rangle} | x_{\langle\mathcal{A}\rangle}\big),$$

with

$$\mathbb{P}\big(x_{\mathcal{A}\setminus\langle\mathcal{A}\rangle} | x_{\langle\mathcal{A}\rangle}\big) = \prod_{\mathcal{C} \in C(\mathcal{A})} \rho^{\mathcal{A},\mathcal{C}}(x_{\mathcal{C} \cap \mathsf{cl}(\mathcal{A}\setminus\langle\mathcal{A}\rangle)}),$$

where:

- $C(\mathcal{A})$ denotes the cliques of $(G_{\mathcal{A}})^m$;
- $x_{\mathcal{C} \cap \mathsf{cl}(\mathcal{A}\setminus\langle\mathcal{A}\rangle)}$ denotes the configuration of those values in $\mathcal{C} \cap \mathsf{cl}(\mathcal{A}\setminus\langle\mathcal{A}\rangle)$ as assigned in $\{X_1 = x_1, \dots, X_n = x_n\}$;
- and each $\rho^{\mathcal{A},\mathcal{C}}$ is a non-negative function (a *potential*).

We then have, once again by suitably adapting Koster's proofs:

Theorem 4.2. *Consider an LCN where all configurations have positive probability, and that satisfies GMC with respect to the mixed-structure G. Then GIBBS(LCN) holds.*

In the proof of his Theorem 4.3, Koster employs the Hammersley-Clifford theorem, from which it inherits the assumption of positive probabilities. Another path is to follow results by Moussouris, and assume [18] that any joint distribution associated with an LCN should be the limit of positive distributions on the same underlying moralized mixed-structure, *and* a "barrier property" holds — that is, the probability of a configuration of variables is zero only when some clique of the moralized mixed-structure is assigned probability zero. The condition that any joint distribution is a limit of positive distributions is unfortunately not easy to check; necessary and sufficient conditions for it have been presented by Geiger et al. [10].

Finally, we have:

Theorem 4.3. *Consider a LCN with mixed-structure G and that satisfies GIBBS(LCN). Then the LCN satisfies GMC with respect to G.*

The proof of this result can be obtained by just reproducing the relevant steps in Koster's proof of his Theorem 4.3 [12].

Example 4.4. In the DUMG in Figure 5, the join-irreducible elements are $\{A\}$, $\{B\}$, $\{A,B,C,D,E,F\}$, $\{I\}$, $\{J\}$, and $\{A,B,C,D,E,F,G,H,I,J\}$. Thus we have a factorization that yields the joint distribution as

$$\mathbb{P}(G,H|A,B,C,D,E,F,I,J) \cdot \mathbb{P}(I) \cdot \mathbb{P}(J) \cdot \mathbb{P}(C,D,E,F|A,B) \cdot \mathbb{P}(A) \cdot \mathbb{P}(B).$$

We then have "internal" factorizations that yield

$$\mathbb{P}(C,D,E,F|A,B) = \rho_1(A,B,F) \cdot \rho_2(A,C) \cdot \rho_3(B,D) \cdot \rho_4(C,D) \cdot \rho_5(D,F) \cdot \rho_6(C,E) \cdot \rho_7(E,F),$$

and

$$\mathbb{P}(G,H|A,B,C,D,E,F,I,J) = \rho_8(F,G,H,J) \cdot \rho_9(G,H,I),$$

for suitable positive functions ρ_k. $\Box$.

The reader can consult Koster's work [12] to learn additional important results concerning factorization with respect to DUMGs (and thus with respect to mixed-structures in our case).

4.3. Inference as multilinear programming. We have extended previous results on Gibbs factorization for chain LCNs to a general factorization result, at the cost of discarding the local Markov condition. One advantage of a compact factorization is to reduce the cost of inference, where *inference* refers to the computation of a probability value $\mathbb{P}(\phi_q|\psi_e)$, where ϕ_q is referred to as the *query* and ψ_e is the *evidence*. Because we may have a set of joint distributions that satisfy constraints in a given LCN, we must focus on the computation of probability bounds. Thus we have the *lower* probability

$$\underline{\mathbb{P}}(\phi_q|\psi_e) = \inf \frac{\sum_{x_1,\ldots,x_n \vDash \phi_q \wedge \psi_e} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)}{\sum_{x_1,\ldots,x_n \vDash \psi_e} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)}, \tag{2}$$

where summations go over configurations of variables satisfying appropriate sentences. Similarly, we can compute the *upper probability* $\overline{\mathbb{P}}(\phi_q|\psi_e)$ by using the same Expression (2) except that we must calculate the supremum (instead of the infimum) over all possible joint distributions.

Note that the values to be found in these optimization problems are the values of the potentials $\rho^{\mathcal{A},\mathcal{C}}$ for the various configurations of variables. Note also that the Gibbs factorization discussed in the previous section defines a multilinear expression on those optimization variables.

Now consider the set of constraints that must be satisfied when setting up these optimization problems: for each assessment $\mathbb{P}(\phi_j|\psi_j) \in [\alpha_j, \beta_j]$, we must satisfy

$$\mathbb{P}(\phi_j \wedge \psi_j) - \alpha_j \mathbb{P}(\psi_j) \geq 0$$

and

$$\mathbb{P}(\phi_j \wedge \psi_j) - \beta_j \mathbb{P}(\psi_j) \leq 0,$$

where $\mathbb{P}(\theta)$ for any sentence θ is equal to $\sum_{x_1,\ldots,x_n \vDash \theta} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$, where the summation goes over configurations that satisfy θ, and again $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_1)$ satisfies the factorization. We thus obtain multilinear expressions on the (to be found) values of potentials.

We can then introduce an auxiliary variable t and an auxiliary constraint

$$t \cdot \sum_{x_1,\ldots,x_n \vDash \psi_e} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = 1,$$

and rewrite $\underline{\mathbb{P}}(\phi_q|\psi_e)$ as

$$\inf \sum_{x_1,\ldots,x_n \vDash \phi_q \wedge \psi_e} t\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

(and similarly for the upper probability).

Such maneuvers lead us to a multinear program with a closed feasible region (hence we can replace inf and sup by min and max respectively). We can thus use some of the machinery developed for credal networks [17] when dealing with such programs. In addition, we can resort to simplification results from Bayesian networks; for instance, we can marginalize out variables that cannot affect an inference.

Example 4.5. Consider Example 4.4. Suppose we want $\underline{\mathbb{P}}(A = 1, B = 1, F = 1)$ (that is, $\phi_q = A \wedge B \wedge F$ and ψ_e is a tautology). In the objective function given by Expression (2), we can marginalize out G, H, I, J. Similarly, each constraint generated by assessments that do not refer explicitly to G, H, I, J can have these variables marginalized out of the product that yields the joint distribution. Thus the computation of this lower probability must handle 32 optimization variables that caracterize potentials ρ_1 to ρ_7, plus 2 optimization variables that capture $\mathbb{P}(A)$ and $\mathbb{P}(B)$. □

5. Global Markov condition and imprecise probabilities

As noted in the previous section, we can obtain a satisfying theory of LCNs if we adopt the GMC as defined by Koster for DUMGs. However, things are a bit convoluted in the presence of cycles. We study in this section scenarios where a GMC cannot be consistently imposed, and find an apparently unavoidable connection between those scenarios and probabilistic imprecision. We do so using a class of models that is rather important in practice, namely, structural equation models.

In this section we use the compact notation $\mathbb{P}(X|y_1,\ldots,y_m) = \theta(y_1,\ldots,y_m)$ to mean that $\mathbb{P}(X = 1|Y_1 = y_1,\ldots,Y_m = y_m)$ is equal to one if $\theta(Y_1,\ldots,Y_m)$ is true when $Y_1 = y_1,\ldots,Y_m = y_m$, and equal to zero otherwise.

We start with a well-known example due to Neal [19].

Example 5.1. Take a LCN with the following coupled-assessments:

$$\mathbb{P}(X_1) = 1/2, \qquad \mathbb{P}(X_4) = 1/2, \qquad \mathbb{P}(X_5) = 1/2,$$
$$\mathbb{P}(X_2|x_1,x_3) = x_1 \oplus x_3, \qquad \mathbb{P}(X_3|x_1,x_2) = x_1 \oplus x_2,$$
$$\mathbb{P}(X_6|x_2,x_4,x_5,x_7) = (x_2 \oplus x_4 \oplus x_5) \wedge \neg x_7,$$
$$\mathbb{P}(X_7|x_2,x_4,x_5,x_6) = (x_2 \oplus x_4 \oplus x_5) \wedge x_6,$$

for the possible values of each x_j. The mixed-structure for this LCN is shown in Figure 7 (left). Neal shows that, for each configuration of $\{X_1, X_4, X_5\}$, there is a single configuration for all remaining variables that satisfies all constraints. Consequently, a unique joint distribution is specified by this LCN. Now we must have that $(x_2 \oplus x_4 \oplus x_5)$ must be 0 with probability 1, for otherwise we get logical inconsistency with probability 1. Hence X_4 and X_5 are *not* independent given X_2 (if $X_2 = 0$, then $X_4 \leftrightarrow X_5$; if $X_2 = 0$, then $X_4 \leftrightarrow \neg X_5$), but the global Markov condition says this independence should hold. The moralized anterior graph of $\{X_2, X_4, X_5\}$ is depicted in Figure 7 (right). □

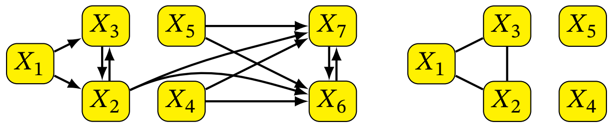

Figure 7. *Left: the mixed-structure for the LCN in Example 5.1. Right: the moralized anterior graph of* $\{X_2, X_4, X_5\}$.

That is, if we impose the assessments in Neal's example *and* the GMC, we obtain an inconsistency. This is obviously embarrassing, as this LCN specifies a Structural Equation Model (SEM), a popular language that is often employed in causal reasoning [20]. Neal's example has indeed led to significant discussion in the context of causality; in particular it has led to new variants of d-separation [9].

Conditions that guarantee the consistency of d-separation in recursive SEMs with categorical variables have been examined by Neal [19] and by Forré and Mooij [9]. We need additional concepts to be able to convey such results, based on Forré and Mooij's Definition 3.8.1 [9], as follows.

Suppose G is a directed graph such that every node is associated with a distinct random variable X_i. This graph has the *Structural Equations Property (SEP)* iff
(i) the random variables that have no parents, the *roots*, are all independent and all associated with precise marginal probabilities; and
(ii) the random variables that have parents, the *non-roots*, are all associated with deterministic functions f_i so that $X_i = f_i(\mathsf{pa}(X_i))$.

If, in addition,
(iii) there is a function g_i, for each X_i, such that $X_i = g_i(\mathcal{Y}_i)$, where $\mathcal{Y}_i$ are the roots in the anterior graph of X_i, then the graph has the *Ancestrally-Solvable Structural Equations Property (AS-SEP)*.

And if, in addition,
(iv) for any anterior set, if the root variables in the anterior set are fixed, there is only one possible configuration for the non-root variables (that is, a configuration that satisfies the relevant f_i), and this configuration is exactly given by the relevant g_i (except in a set of probability zero),
then the graph has the *Ancestrally-Uniquely-Solvable Structural Equations Property (AUS-SEP).*[2]

The key result by Forré and Mooij is that the Ancestrally-Uniquely-Solvable Structural Equations Property implies a factorization [9, Theorems 3.8.12 and 3.7.2] and then implies the global Markov condition over the directed graph, assuming probabilities over roots are positive [9, Corollary 3.6.9]. Hence a SEM that satisfies ancestral unique solvability also satisfies the global Markov condition adopted for DUMGs.

This result on AUS-SEP can be explored so as to find an interesting connection between separation and imprecision in probability values. First, suppose that a SEM is such that a single joint probability distribution is specified over all variables (this is the case of Example 5.1). We say the SEM is then a *Bayesian SEM*. As we have seen, a Bayesian SEM may still fail the GMC. But suppose we are interested in interventional reasoning, where we intervene in some variables so as to quantify their causal effect [20]. A usual procedure in interventional reasoning is to assume that the effect of an intervention in variable X can be computed by changing the original SEM, replacing the equation where X is in the left hand side with an assignment $X = x$ (where x is the value fixed by the intervention), and replacing X by its fixed value x in all other equations. When we move to the directed graph representing the intervened SEM, we see that the intervention produces a graph where all edges into X are discarded.

Now if we have a SEM such that every intervention produces another (intervened) SEM such that there is a single joint probability distribution over the variables, we say this SEM is an *interventionally Bayesian SEM*. That is, any intervention leads to point probabilities for events. The terminology is apt: a non-recursive SEM is obviously represented as a Bayesian network, and is interventionally Bayesian.

Now consider a "contrapositive" version of Forré and Mooij's results:

Theorem 5.1. *Suppose G is a directed graph that satisfies SEP with positive probabilities over variables without parents. If probabilities and functions associated with G fail to satisfy the GMC, then there must be an intervention such that the set of possible values of the probability of some variable, under the intervention, is not a singleton.*

Proof. Take a G that satisfies SEP with positive probabilities over the root variables. If the global Markov condition fails, the AUS-SEP must fail given Forré and Mooij's results. That may happen because G satisfies AS-SEP but fails to have a single possible configuration of non-root variables for fixed root variables in some anterior set; it is then enough to intervene so as to "disconnect" the subgraph of this anterior set, and then ask for the probability of a variable that has multiple possible values for a fixed configuration of root variables. It may instead be the case that AS-SEP fails; that is, the value of some non-root variable in an anterior set is not specified by the root variables in the anterior set. But this is only possible if, when we fix the root variables in the anterior set, there is not a unique configuration possible for the non-root variables — otherwise that very function would lead to satisfaction of AS-SEP. Again we can intervene so as to

[2] The definition of AUS-SEP by Forré and Mooij somewhat indirectly asks every possible configuration of non-root variables satisfying the functions f_i to satisfy the (uniquely defined) functions g_i.

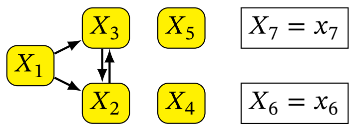

Figure 8. *An intervention on X_6 and X_7 in the SEM of Example 5.1.*

"disconnect" the subgraph of this anterior set, and then ask for the probability of a variable that has multiple possible values for a fixed configuration of root variables. Hence the original SEM is not a Bayesian one. □

That is,

Failure of GMC in SEM (with $\mathbb{P} > 0$)	$\Rightarrow$	Imprecise probability with respect to some intervention.

Example 5.2. Consider Example 5.1. By intervening on X_6 and X_7, we obtain the "sub"-SEM is given by:

$$\mathbb{P}(X_1) = 1/2, \qquad \mathbb{P}(X_4) = 1/2, \qquad \mathbb{P}(X_5) = 1/2,$$
$$\mathbb{P}(X_2|x_1, x_3) = x_1 \oplus x_3, \qquad \mathbb{P}(X_3|x_1, x_2) = x_1 \oplus x_2,$$

plus constraints enforcing $X_6 = x_6$ and $X_7 = x_7$ for the intervened variables. The corresponding graph is depicted in Figure 8. With respect to the intervened model we have, for instance, we have the following tight probability intervals. To obtain $\mathbb{P}(X_2 = 1)$, we note that both values of X_1 let X_2 to be either 0 or 1; hence $\mathbb{P}(X_2 = 1) \in [0, 1]$. By similar reasoning, we have $\mathbb{P}(X_2 = 1, X_3 = 1) \in [0, 1/2]$ and $\mathbb{P}(X_2 = 1|X_3 = 1) \in [0, 1]$. □

6. Conclusion

In this paper we have investigated the consequences of cycles in graph-theoretical models, using the language of Logical Credal Networks as the concrete modeling tool. In this journey, we have proposed a few ideas concerning Logical Credal Networks. In short:

- There are distinct strategies in the translation from conditioned logical sentences to graph-theoretical constructs; a translation based on directed *and* undirected edges (the latter replacing some opposing arrows) may be valuable.
- A Gibbs factorization can be produced out of the resulting undirected/directed graphs produced by a LCN. We have thus answered a question raised by Cozman et al. [5] as they wondered how to extend results for chain LCNs.
- The move to a global Markov condition (that seems necessary to obtain Gibbs factorizations) requires some careful analysis when it comes to structural equation models. When the global Markov condition fails, there is some intervention that leads to probability intervals, thus displaying a veiled connection between separation concepts that are essential to graph-theoretical modeling and probabilistic imprecision.

Even though the whole analysis focused on Logical Credal Networks, the connection between separation and imprecision, and the analysis of directed/undirected modeling of sentences, should be of general interest.

It should be noted that there are other possible semantics for mixed-structures that might be studied. One possibility is to use the theory of *MC graphs*, also developed by Koster [14], where one can have undirected, directed *and* bidirected edges. Might it be the case that MC graphs are actually more appropriate?

Assuming DUMGs are the best path, we should investigate their properties — for instance, are they closed under conditioning? Are they useful in causal analysis? And we should also look into inference algorithms and their complexity.

A pressing issue is the positivity assumption that is required by Koster's proofs as he relied on the classic Hammersley-Clifford theorem [14]. We have noted that other conditions, such as the barrier property by Moussouris [18] and the algebraic conditions by Geiger et al. [10], may be adopted. However, these conditions are not simple at all, and it is not clear how they connect with practical scenarios. Given that LCNs can contain logical constraints, more work should be devoted to the effect of zero probabilities.

Moreover, future work should also continue to study the meaning of assessments and separation conditions, and how they interact in practice when there are recursive relations (for instance in structural equation models). As another future path, we note that we have focused only on propositional sentences, but first-order sentences should be investigated — the original work on Logical Credal Networks interprets the semantics of first-order constructs through grounding [15], a strategy that should be better examined.

Perhaps the real conclusion here is that we do not yet have all the tools we need to address cyclic models.

Additional Author Information

Acknowledgements. F.G.C. and D.D.M. thank the support of the Center for Artificial Intelligence (C4AI — c4ai.inova.usp.br) and its funding by IBM Corporation and FAPESP through grant 2019/07665-4. F.G.C. is partially supported by CNPq through grant Pq 305753/2022-3. D.D.M. is partially supported by CNPq through grant Pq 305136/2022-4 and FAPESP Grant 2022/02937-9.

References

[1] Kim Allan Andersen and John N. Hooker. “Bayesian Logic”. In: *Decision Support Systems* 11 (1994), pp. 191–210. DOI: 10.1016/0167-9236(94)90031-0.

[2] A. Balke and J. Pearl. “Counterfactual Probabilities: Computational Methods, Bounds and Applications”. In: *Conference on Uncertainty in Artificial Intelligence* (1994), pp. 46–54.

[3] Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, 1999.

[4] Fabio G. Cozman. “Credal networks”. In: *Artificial Intelligence* 120 (2000), pp. 199–233. DOI: 10.1016/S0004-3702(00)00029-1.

[5] Fabio G. Cozman et al. “Markov conditions and factorization in logical credal networks”. In: *International Journal of Approximate Reasoning* 172, 109237 (2024). DOI: 10.1016/j.ijar.2024.109237.

[6] Fabio Gagliardi Cozman, Cassio Polpo de Campos, and José Carlos Ferreira da Rocha. “Probabilistic Logic with Independence”. In: *International Journal of Approximate Reasoning* 49.1 (2008), pp. 3–17. DOI: 10.1016/j.ijar.2007.08.002.

[7] Adnan Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.

[8] Rina Dechter. “Tractable Structures for Constraint Satisfaction Problems”. In: *Handbook of Constraint Programming*. Elsevier, 2006.

[9] Patrick Forré and Joris M. Mooij. “Markov Properties for Graphical Models with Cycles and Latent Variables”. Oct. 2017. arXiv: 1710.08775 [math.ST].

[10] Dan Geiger, Christopher Meek, and Bernd Sturmfels. “On the toric algebra of graphical models”. In: *The Annals of Statistics* 34.3 (2006), pp. 1463–1492. DOI: 10.1214/009053606000000263.

[11] Joseph Y. Halpern. *Reasoning about Uncertainty*. Cambridge, Massachusetts: MIT Press, 2003.

[12] J. T. Koster. “Gibbs and Markov properties of graphs”. In: *Annals of Mathematics and Artificial Intelligence* 21 (1997), pp. 13–26. DOI: 10.1023/A:1018948915264.

[13] J. T. A. Koster. “Markov properties of nonrecursive causal models”. In: *Annals of Statistics* 24.5 (1996), pp. 2148–2177.

[14] J. T. A. Koster. “Marginalizing and conditioning in graphical models”. In: *Bernoulli* 8.6 (2002), pp. 817–840.

[15] Radu Marinescu et al. “Logical credal networks”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 15325–15337.

[16] Radu Marinescu et al. “Approximate inference in logical credal networks”. In: *Thirty-Second International Joint Conference on Artificial Intelligence*. 2023, pp. 5694–5701.

[17] Denis Deratani Mauá and Fabio Gagliardi Cozman. “Thirty years of credal networks: Specification, algorithms and complexity”. In: *International Journal of Approximate Reasoning* 126 (2020), pp. 133–157. DOI: 10.1016/j.ijar.2020.08.009.

[18] John Moussouris. “Gibbs and Markov random systems with constraints”. In: *Journal of Statistical Physics* 10.1 (1974), pp. 11–33. DOI: 10.1007/BF01011714.

[19] Radford M Neal. “On deducing conditional independence from d-separation in causal graphs with feedback (research note)”. In: *Journal of Artificial Intelligence Research* 12 (2000), pp. 87–91. DOI: 10.1613/jair.689.

[20] Judea Pearl. *Causality: Models, Reasoning, and Inference (2nd edition)*. Cambridge, United Kingdom: Cambridge University Press, 2009.

[21] Peter Spirtes. “Directed cyclic graphical representations of feedback models”. In: *Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 491–498.

[22] Milan Studený. “Bayesian networks from the point of view of chain graphs”. In: *Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 1998, pp. 496–503.