APPLICATIONES MATHEMATICAE
51,2 (2024), pp. 109–130

Wojciech Niemiro (Warszawa)

CAUSAL GRAPHS, COMPOSABLE STOCHASTIC PROCESSES AND CONDITIONAL INDEPENDENCE

Abstract. We consider multivariate stochastic processes with causal relations between their components modelled by directed graphs with possible cycles. Our aim is to express conditional independence relations for such processes in terms of separability properties of the underlying graphs. This line of study is quite classical and was initiated in the seminal paper of Pearl (1985), then extended to point processes by Didelez (2007, 2008) and to time series by Eichler (2007) and Eichler and Didelez (2007). In our paper we provide a unifying view and fill in certain gaps. We define a class of models called composable random elements (CRE) which encompasses usual Bayesian networks (BN), dynamic BNs (DBN), continuous time BNs (CTBN) and marked point processes. We show that key results known in the classical setup of directed acyclic graphs (DAG) can be generalised to CREs and remain valid also for graphs containing cycles. For CTBNs, we prove a new theorem that characterises independence between the future of one subprocess and the past of another given the past of a third subprocess. Our paper also tackles causal (interventional) conditional independence relations, strengthening and generalising results of Ay and Polani (2008).

1. Introduction

Background. Probabilistic models of causality were originated in the 1960s by [9] in the context of time series. Granger's definition explicitly and essentially involved ordering of events in time, assuming that a cause always precedes an effect. The pioneering work of [18] dating from the 1980s in-

2020 *Mathematics Subject Classification*: Primary 62H22; Secondary 60J27, 62A01, 62D20.
Key words and phrases: causality, continuous time Bayesian networks, conditioning-by-intervention, separability conditions, directed graphs.
Received 14 December 2023; revised 2 May 2024.
Published online 7 November 2024.

troduced a graph-theoretic representation of causal relations. Pearl's theory did not involve time. In its original form this theory was restricted to directed acyclic graphs (DAG), and thus did not encompass the phenomena of feedback in cause-effect dependences. Later work initiated in [20] generalised graphical models of causality, allowing for cycles in directed graphs. Models based on "structural equations" (structural causal models, SCM) have been intensively examined and extended. The monograph [21] is an excellent overview. Recent developments, with focus on mixed graphs and marginalisation, can be found in [2]. However, this approach leads to certain difficulties and paradoxes (see [15]) if the cycles are present in the underlying graph. A priori it is not clear whether a given set of structural equations has a unique solution and this makes the task of building models rather hard.

On the other hand, a class of "composable Markov processes" (CMP) introduced as early as 1970 by [22] provides natural and flexible examples of causality models which include time and use directed, possibly cyclic, graphs (DG) to represent causal dependence. Research on CMPs was pursued by [17] (who independently introduced these processes under the name of CTBN) and by [3]. Analogous dependence graphs have also been defined and examined for other classes of structured stochastic processes. Discrete time series are considered in [6, 7, 5], multivariate point processes in [4, 14] and diffusion processes in [13].

Outline and the contribution of this paper. We consider both discrete time and continuous time stochastic processes and aim to describe the structure of conditional independence between subprocesses in graph-theoretical terms. In Section 2 we define a class of causal models called "composable random elements" (CRE) which includes as special cases the following models:

- classical Bayesian networks (BN);
- discrete time series and dynamic Bayesian networks (DBN);
- continuous time Bayesian networks (CTBN);
- marked point processes examined in [4].

The definition of a CRE (Assumptions 2.1 and 2.4) allows for cycles in the underlying graph and thus enables us to model the feedback in causal relations. The elementary building blocks of a CRE are *conditional-by-intervention* probability distributions of single nodes given their parents. It turns out that many results on conditional independence and graph separation which are known in the classical setup of DAGs can be generalised to CREs. This approach unifies and simplifies the theory. As a by-product it also reveals certain gaps in existing results, in particular overlooked assumptions about the initial distributions.

Apart from observational independence we also consider interventional (conditional) independence. However, our definition of interventional independence (2.3) is different from and stronger than that in [1], as demonstrated by Example 2.8. Our Theorem 2.6, valid for CREs, generalises and strengthens their results.

In Section 3 we show that certain discrete time processes and CTBNs are special cases of CREs. We highlight the role of assumptions on the initial distributions.

In Section 4 the focus is on the relations between events ordered in time, in the spirit of the Granger definition of causality. This section extends and completes the results of [6, 7, 16]. We consider the structure of conditional independence between the past of one subprocess and the future of another given the past of a third subprocess. This problem is well understood for discrete time processes. In a sense, investigation of a discrete time graphical model can be reduced to considering a space-time graph representing a dynamic Bayesian network (DBN). For CTBNs we need to use completely different methods. Theorem 4.7 is new. Theorem 4.10 provides a more explicit reformulation of the results on δ-separation due to [4] with an independent simple proof.

2. Composable random elements. In this section we present a rather abstract setup. Although our interest is in stochastic processes, the time variable will not be explicitly considered here.

Interventional vs observational conditioning. In probability theory the notion of conditional distribution is defined in terms of a probability measure which describes the joint distribution of all random variables under consideration. In most applications, however, the order is reversed. The joint distribution is usually defined in terms of conditional (and marginal) distributions. In particular, *causal* relations between variables are adequately modelled by *conditional-by-intervention distributions*. They appear (slightly disguised) in the standard definition of classical BN as probabilities of single nodes given the parents, say $p(\mathbf{x}_v \,|\, \mathbf{x}_{\mathrm{pa}(v)})$. In fact, these are also interventional probabilities. In this section we use the conditional-by-intervention probabilities, denoted by $p(\mathbf{x}_v \,\|\, \mathbf{x}_{\mathrm{pa}(v)})$, to define a notion of a *composable random element*, without assuming acyclicity of the underlying graph.

Definitions. Let $\mathcal{V}$ be a finite set. Consider a collection $\mathbf{X} = (\mathbf{X}_v : v \in \mathcal{V})$ of random elements, where $\mathbf{X}_v$ takes values in a measurable space $\mathcal{X}_v$ equipped with a reference measure denoted by $\mathrm{d}\mathbf{x}_v$. A generic element of the Cartesian product $\mathcal{X} = \prod_{v\in\mathcal{V}} \mathcal{X}_v$ is denoted by $\mathbf{x} = (\mathbf{x}_v : v \in \mathcal{V})$. More generally, for a set $\mathcal{A} \subseteq \mathcal{V}$ we write $\mathbf{X}_\mathcal{A} = (\mathbf{X}_v : v \in \mathcal{A})$, $\mathbf{x}_\mathcal{A} = (\mathbf{x}_v : v \in \mathcal{A})$,

and $\mathrm{d}\mathbf{x}_{\mathcal{A}} = \prod_{v\in\mathcal{A}} \mathrm{d}\mathbf{x}_v$. For $\mathcal{A} = \mathcal{V} \setminus \{v\}$ we use the notation $\mathbf{X}_{-v} = (\mathbf{X}_w : w \neq v)$ and analoguosly define $\mathbf{x}_{-v}$.

The elementary building blocks of our causal model are the transition densities $p_v(\mathbf{x}_v \,\|\, \mathbf{x}_{-v})$. From a formal viewpoint, p_v is just a nonnegative function on $\mathcal{X}$ which integrates to 1 with respect to $\mathrm{d}\mathbf{x}_v$ for every $\mathbf{x}_{-v}$. It is interpreted as the conditional-by-intervention distribution of a single component $\mathbf{X}_v$ given the remaining components.

2.1. Assumption. X is a random element with a joint density p with respect to $\mathrm{d}\mathbf{x}$ on $\mathcal{X}$ and this density admits the following factorisation:

$$p(\mathbf{x}) = \prod_{v\in\mathcal{V}} p_v(\mathbf{x}_v \,\|\, \mathbf{x}_{-v}).$$

Moreover, for every $\mathcal{A} \subset \mathcal{V}$ and every $\mathbf{x}_{\mathcal{V}\setminus\mathcal{A}}$ we have

$$\int\limits_{\mathcal{X}_{\mathcal{A}}} \prod_{v\in\mathcal{A}} p_v(\mathbf{x}_v \,\|\, \mathbf{x}_{-v})\, \mathrm{d}\mathbf{x}_{\mathcal{A}} = 1.$$

If Assumption 2.1 holds then we say $\mathbf{X}$ is a *composable random element* (CRE) ([1]).

Under this assumption we can formally define conditional-by-intervention distributions. For arbitrary disjoint subsets $\mathcal{A}$ and $\mathcal{C}$ of $\mathcal{V}$ we let

$$p(\mathbf{x}_{\mathcal{A}} \,\|\, \mathbf{x}_{\mathcal{C}}) = \int\limits_{\mathcal{X}_{\mathcal{V}\setminus(\mathcal{A}\cup\mathcal{C})}} \prod_{v\in\mathcal{V}\setminus\mathcal{C}} p_v(\mathbf{x}_v \,\|\, \mathbf{x}_{-v})\, \mathrm{d}\mathbf{x}_{\mathcal{V}\setminus(\mathcal{A}\cup\mathcal{C})}. \tag{2.2}$$

If $\mathcal{A}\cup\mathcal{C} = \mathcal{V}$ then we simply have $p(\mathbf{x}_{\mathcal{A}} \,\|\, \mathbf{x}_{\mathcal{C}}) = \prod_{v\in\mathcal{A}} p_v(\mathbf{x}_v \,\|\, \mathbf{x}_{-v})$. The LHS of (2.2) can be interpreted as the probability density of $\mathbf{X}_{\mathcal{A}}$ given that $\mathbf{X}_{\mathcal{C}}$ is forced to assume value $\mathbf{x}_{\mathcal{C}}$ ([2]) in a (real or imaginary) controlled experiment. This interpetation has a clear meaning when CRE $\mathbf{X}$ is a random process evolving in time, as explained in Section 3. Of course, $p(\mathbf{x}_{\mathcal{A}} \,\|\, \mathbf{x}_{\mathcal{C}})$ is in general different from the standard, conditional-by-observation probability density $p(\mathbf{x}_{\mathcal{A}} \,|\, \mathbf{x}_{\mathcal{C}})$ defined by

$$p(\mathbf{x}_{\mathcal{A}} \,|\, \mathbf{x}_{\mathcal{C}}) = \frac{p(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{C}})}{p(\mathbf{x}_{\mathcal{C}})}.$$

Note that we use p as a generic notation for densities; in particular, we will write $p_v(\mathbf{x}_v \,\|\, \mathbf{x}_{-v}) = p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v})$ from now on.

We now turn to a definition of *causal conditional independence*. Let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be three disjoint subsets of $\mathcal{V}$, with $\mathcal{C}$ possibly empty. We say that $\mathbf{X}_{\mathcal{B}}$ is *causally independent of* $\mathbf{X}_{\mathcal{A}}$ *imposing* $\mathbf{X}_{\mathcal{C}}$, symbolically $\mathbf{X}_{\mathcal{A}} \not\rightarrow \mathbf{X}_{\mathcal{B}} \,\|\, \mathbf{X}_{\mathcal{C}}$, if

$$p(\mathbf{x}_{\mathcal{B}} \,\|\, \mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{C}}) = p(\mathbf{x}_{\mathcal{B}} \,\|\, \mathbf{x}_{\mathcal{C}}), \tag{2.3}$$

([1]) This term is borrowed from [22]. In fact, composable finite Markov processes introduced by Schweder in *op. cit.* satisfy Assumption 2.1 under an additional condition on the initial distribution; see Proposition 3.9 in the next section.

([2]) The notation $p(\mathbf{x}_{\mathcal{A}} \,\|\, \mathbf{x}_{\mathcal{C}})$ is used instead of Pearl's $p(\mathbf{x}_{\mathcal{A}} \,|\, \mathrm{do}(\mathbf{x}_{\mathcal{C}}))$.

where $(\mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C}) = \mathbf{x}_{\mathcal{A}\cup\mathcal{C}}$ and the symbol $p(\cdot \,\|\, \cdot)$ on both sides of this equation is defined by (2.2). The symbol $\rhd$ will mean the negation of $\not\triangleright$. Formula (2.3) is an exact counterpart of the analogous property of observational independence: the relation $\mathbf{X}_\mathcal{A} \perp\!\!\!\perp \mathbf{X}_\mathcal{B} \,|\, \mathbf{X}_\mathcal{C}$ is equivalent to

$$p(\mathbf{x}_\mathcal{B} \,|\, \mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C}) = p(\mathbf{x}_\mathcal{B} \,|\, \mathbf{x}_\mathcal{C}).$$

Our definition of causal conditional independence via (2.3) is very natural and properly reflects the intuitive meaning of this notion. Later in this section and in the next section we discuss relations to another definition appearing in the literature.

We are interested in the situation when the structure of causal dependence between components of $\mathbf{X}$ is described by a graph. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a *directed graph with possible cycles*, but without self-loops. If $(v, w) \in \mathcal{E}$ then we write $v \to w$. The set of *parents* of a node v is $\mathrm{pa}(v) = \{w : w \to v\}$.

2.4. Assumption. $\mathbf{X}$ is a random element satisfying Assumption 2.1 and $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a directed graph. Moreover, for all $v \in \mathcal{V}$, $p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v})$ does not depend on $\mathbf{x}_{\mathcal{V}\setminus(\mathrm{pa}(v)\cup\{v\})}$ so that

$$p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v}) = p(\mathbf{x}_v \,\|\, \mathbf{x}_{\mathrm{pa}(v)}).$$

If Assumptions 2.1 and 2.4 hold then we say $\mathbf{X}$ is a $\mathcal{G}$-CRE. Combining these two assumptions we arrive at the following basic factorisation formula for $\mathcal{G}$-CREs:

$$p(\mathbf{x}) = \prod_{v\in\mathcal{V}} p(\mathbf{x}_v \,\|\, \mathbf{x}_{\mathrm{pa}(v)}). \tag{2.5}$$

If $\mathcal{G}$ is a DAG (has no cycles) then $p(\mathbf{x}_v \,\|\, \mathbf{x}_{\mathrm{pa}(v)}) = p(\mathbf{x}_v \,|\, \mathbf{x}_{\mathrm{pa}(v)})$ and (2.5) reduces to the standard definition of joint probability of a Bayesian network (BN). In our setup, formula (2.5) can still be used to define joint probability even for cyclic graphs, provided that Assumption 2.1 holds.

The rationale behind Assumptions 2.1 and 2.4 is that our definition of $\mathcal{G}$-CREs encompasses several important classes of models, as will be shown in Section 3.

In the remaining part of this section we show that results known in the classical setup of DAGs can be generalised to CREs and thus they remain valid for graphs with cycles. We give graph-theoretic characterisations of observational and causal conditional independence. We assume the setup defined above, so the time variable will not directly appear. However, we should bear in mind that if $\mathbf{X}_v$ is a stochastic process then $\mathbf{x}_v$ stands for its entire trajectory, and analogously $\mathbf{x}$ denotes the entire trajectory of $\mathbf{X} = \mathbf{X}_\mathcal{V} = (\mathbf{X}_v : v \in \mathcal{V})$; a similar remark applies to subprocesses indexed by subsets of $\mathcal{V}$.

Graph terminology. Throughout, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a directed graph (DG). We define a *trail* between $u \in \mathcal{V}$ and $w \in \mathcal{V}$ as a sequence

$$u = v_0, e_1, v_1, e_2, v_2, e_3, \ldots, e_{k-1}, v_{k-1}, e_k, v_k = w,$$

where $v_0, v_1, \ldots, v_k$ are *distinct* nodes and $e_1, \ldots, e_k$ are edges, $e_i = (v_{i-1} \to v_i)$ or $e_i = (v_{i-1} \leftarrow v_i)$ ([3]). Let v_i be a nonend node in the trail, that is, $i \neq 0$ and $i \neq k$. We say that

- there is a *chain* connexion at v_i if we have $v_{i-1} \to v_i \to v_{i+1}$ or $v_{i-1} \leftarrow v_i \leftarrow v_{i+1}$;
- there is a *fork* connexion at v_i if we have $v_{i-1} \leftarrow v_i \to v_{i+1}$;
- there is a *collider* connexion at v_i if we have $v_{i-1} \to v_i \leftarrow v_{i+1}$.

A *directed path* from u to w is a trail such that all arrows are directed to the right, i.e. $e_i = (v_{i-1} \to v_i)$. We define the set of *ancestors* as follows:

$$\mathrm{an}(v) = \{v\} \cup \{w : \text{there exists a directed path from } w \text{ to } v\}.$$

Moreover, for $\mathcal{A} \subseteq \mathcal{V}$ we put $\mathrm{an}(\mathcal{A}) = \bigcup_{v \in \mathcal{A}} \mathrm{an}(v)$ and $\mathrm{pa}(\mathcal{A}) = \bigcup_{v \in \mathcal{A}} \mathrm{pa}(v)$.

Below, $\mathcal{A}, \mathcal{B}, \mathcal{C}$ stand for any three disjoint subsets of $\mathcal{V}$ with $\mathcal{C}$ possibly empty. A trail (directed path) from $\mathcal{A}$ to $\mathcal{B}$ is a trail (directed path) from an $a \in \mathcal{A}$ to a $b \in \mathcal{B}$.

Following [1, Def. 2] we say that $\mathcal{B}$ is *u-separated* (unidirectionally separated) from $\mathcal{A}$ by $\mathcal{C}$ if every directed path from $\mathcal{A}$ to $\mathcal{B}$ has a node belonging to $\mathcal{C}$. We write $\mathcal{A} \not\longrightarrow_u \mathcal{B} \,|\, \mathcal{C}$ (and $\mathcal{A} \longrightarrow_u \mathcal{B} \,|\, \mathcal{C}$ if u-separation does not hold).

Next we recall the classical Pearl's definition [18]. We say that $\mathcal{B}$ is *d-separated* from $\mathcal{A}$ by $\mathcal{C}$ if every trail from $\mathcal{A}$ to $\mathcal{B}$ contains a chain $\leftarrow c \leftarrow$ or a chain $\to c \to$ or a fork $\leftarrow c \to$ with $c \in \mathcal{C}$ or a collider $\to v \leftarrow$ with $v \notin \mathrm{an}(\mathcal{C})$. We will then write $\mathcal{A} \perp_d \mathcal{B} \,|\, \mathcal{C}$ (and $\mathcal{A} —_d \mathcal{B} \,|\, \mathcal{C}$ if d-separability does not hold).

Note that the relation of d-separation is symmetric (with respect to $\mathcal{A}$ and $\mathcal{B}$), whereas u-separation is clearly not.

Characterisations of conditional independences. We consider a random element $\mathbf{X} = \mathbf{X}_{\mathcal{V}}$ which satisfies Assumptions 2.1 and 2.4.

A characterisation of causal conditional independence was given in [1, Theorem 2]. We provide two improvements of that result. Firstly, it is generalised from DAGs to general directed graphs. Secondly, the conclusion is strengthened because our definition of causal independence is more restrictive than that used in the cited work.

2.6. Theorem. *Let* $\mathbf{X}$ *be a random element which satisfies Assumptions* 2.1 *and* 2.4. *If* $\mathcal{B}$ *is u-separated from* $\mathcal{A}$ *by* $\mathcal{C}$ *then* $\mathbf{X}_{\mathcal{B}}$ *is causally independent*

([3]) The definition of a trail requires some caution, because there are subtle differences between several definitions appearing in the literature.

of $\mathbf{X}_\mathcal{A}$ *imposing* $\mathbf{X}_\mathcal{C}$*:*

$$\mathcal{A} \not\longrightarrow_u \mathcal{B} \,|\, \mathcal{C} \quad \textit{implies} \quad \mathbf{X}_\mathcal{A} \not\perp \mathbf{X}_\mathcal{B} \,\|\, \mathbf{X}_\mathcal{C}.$$

Proof. Let

$$\bar{\mathcal{A}} = \{v : \mathcal{A} \longrightarrow_u \{v\} \,|\, \mathcal{C}\} \cup \mathcal{A}, \qquad \bar{\mathcal{B}} = \mathcal{V} \setminus (\bar{\mathcal{A}} \cup \mathcal{C}).$$

Obviously, $\bar{\mathcal{B}} \supseteq \mathcal{B}$ and there are no arrows from $\bar{\mathcal{A}}$ to $\bar{\mathcal{B}}$. Consequently, if $v \in \bar{\mathcal{B}}$ then $p(\mathbf{x}_v \,\|\, \mathbf{x}_{\mathrm{pa}(v)})$ does not depend on $\mathbf{x}_{\bar{\mathcal{A}}}$. To lighten notation, in this proof we write $p_v = p(\mathbf{x}_v \,\|\, \mathbf{x}_{\mathrm{pa}(v)})$.

We are to show (2.3), i.e.

$$p(\mathbf{x}_\mathcal{B} \,\|\, \mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C}) = p(\mathbf{x}_\mathcal{B} \,\|\, \mathbf{x}_\mathcal{C}).$$

In accordance with our definition (2.2), the RHS of (2.3) is given by

$$\begin{aligned}\mathrm{RHS} &= \int_{\mathcal{X}_{\bar{\mathcal{B}}\setminus\mathcal{B}}} \int_{\mathcal{X}_{\bar{\mathcal{A}}}} \prod_{v\in\bar{\mathcal{B}}} p_v \prod_{v\in\bar{\mathcal{A}}} p_v \,\mathrm{d}\mathbf{x}_{\bar{\mathcal{A}}} \,\mathrm{d}\mathbf{x}_{\bar{\mathcal{B}}\setminus\mathcal{B}} \\ &= \int_{\mathcal{X}_{\bar{\mathcal{B}}\setminus\mathcal{B}}} \prod_{v\in\bar{\mathcal{B}}} p_v \Big(\int_{\mathcal{X}_{\bar{\mathcal{A}}}} \prod_{v\in\bar{\mathcal{A}}} p_v \,\mathrm{d}\mathbf{x}_{\bar{\mathcal{A}}} \Big) \mathrm{d}\mathbf{x}_{\bar{\mathcal{B}}\setminus\mathcal{B}} = \int_{\mathcal{X}_{\bar{\mathcal{B}}\setminus\mathcal{B}}} \prod_{v\in\bar{\mathcal{B}}} p_v \,\mathrm{d}\mathbf{x}_{\bar{\mathcal{B}}\setminus\mathcal{B}}\end{aligned}$$

in view of Assumption 2.1, because $\prod_{v\in\bar{\mathcal{B}}} p_v = \prod_{v\in\bar{\mathcal{B}}} p(\mathbf{x}_v \,\|\, \mathbf{x}_{\mathrm{pa}(v)})$ does not depend on $\mathbf{x}_{\bar{\mathcal{A}}}$. Similarly we compute the LHS of (2.3):

$$\begin{aligned}\mathrm{LHS} &= \int_{\mathcal{X}_{\bar{\mathcal{B}}\setminus\mathcal{B}}} \int_{\mathcal{X}_{\bar{\mathcal{A}}\setminus\mathcal{A}}} \prod_{v\in\bar{\mathcal{B}}} p_v \prod_{v\in\bar{\mathcal{A}}\setminus\mathcal{A}} p_v \,\mathrm{d}\mathbf{x}_{\bar{\mathcal{A}}\setminus\mathcal{A}} \,\mathrm{d}\mathbf{x}_{\bar{\mathcal{B}}\setminus\mathcal{B}} \\ &= \int_{\mathcal{X}_{\bar{\mathcal{B}}\setminus\mathcal{B}}} \prod_{v\in\bar{\mathcal{B}}} p_v \Big(\int_{\mathcal{X}_{\bar{\mathcal{A}}\setminus\mathcal{A}}} \prod_{v\in\bar{\mathcal{A}}\setminus\mathcal{A}} p_v \,\mathrm{d}\mathbf{x}_{\bar{\mathcal{A}}\setminus\mathcal{A}} \Big) \mathrm{d}\mathbf{x}_{\bar{\mathcal{B}}\setminus\mathcal{B}} = \int_{\mathcal{X}_{\bar{\mathcal{B}}\setminus\mathcal{B}}} \prod_{v\in\bar{\mathcal{B}}} p_v \,\mathrm{d}\mathbf{x}_{\bar{\mathcal{B}}\setminus\mathcal{B}}.\end{aligned}$$

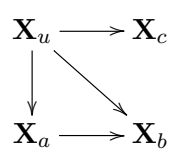
We see that RHS = LHS, which completes the proof. ∎

The above proof is essentially the same as in [1]. However, our definition of causal independence differs from that in the cited paper. In [1, (2)] the authors define causal conditional independence via the following condition (rewritten in our notation):

$$p(\mathbf{x}_\mathcal{B} \,\|\, \mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{C}) = \int_{\mathcal{X}_\mathcal{A}} p(\mathbf{x}_\mathcal{B} \,\|\, \mathbf{x}'_\mathcal{A}, \mathbf{x}_\mathcal{C}) p(\mathbf{x}'_\mathcal{A} \,\|\, \mathbf{x}_\mathcal{C}) \,\mathrm{d}\mathbf{x}'_\mathcal{A}. \tag{2.7}$$

Clearly (2.3) implies (2.7). The converse is not true, as illustrated by the following example.

2.8. Example. Consider four binary random variables $\mathbf{X}_a, \mathbf{X}_b, \mathbf{X}_c, \mathbf{X}_u$ and the following graph of causal relations:



Let $\mathbb{P}(\mathbf{X}_u = 0) = \mathbb{P}(\mathbf{X}_u = 1) = 1/2$, $\mathbf{X}_a = 1 - \mathbf{X}_u$ and $\mathbf{X}_b = \mathbb{1}(\mathbf{X}_a = \mathbf{X}_u)$. The values of $\mathbb{P}(\mathbf{X}_c \,\|\, \mathbf{X}_u)$ are irrelevant (as in fact is the very presence of node c). Put $\mathcal{A} = \{a\}$, $\mathcal{B} = \{b\}$, $\mathcal{C} = \{c\}$. Clearly we have

$$\mathbb{P}(\mathbf{X}_b = 1 \,\|\, \mathbf{X}_a = 0, \mathbf{X}_c) = \mathbb{P}(\mathbf{X}_b = 1 \,\|\, \mathbf{X}_a = 1, \mathbf{X}_c) = 1/2,$$

so (2.7) is true. On the other hand,

$$\mathbb{P}(\mathbf{X}_b = 1 \,\|\, \mathbf{X}_c) = 0,$$

so (2.3) is false.

The difference between (2.7) and (2.3) is by no means technical and pertains to the fundamental question: "what do we mean by causality?". Imagine the following story standing behind our example. There are two railway tracks (marked "1" and "0") between Undover and Andover. A train from Undover to Andover departs just a minute before a train from Andover to Undover. $\mathbf{X}_u = 1$ means that the train from Undover chooses track '1', and $\mathbf{X}_u = 0$ if it chooses track '0'. Analogously, $\mathbf{X}_a$ denotes the choice of the track by the train from Andover. Now, $\mathbf{X}_b = 1$ denotes the event of crash (both trains have chosen the same track). In the "observational regime" we have $\mathbb{P}(\mathbf{X}_b = 1) = 0$, because Undover lets Andover know which track is free. On the other hand, "conditioning-by-intervention" $\mathbb{P}(\cdot \,\|\, X_a)$ actually blocks the flow of information from Undover to Andover and thus causes a railway crash (with probability $1/2$). Note that the choice of the imposed value $\mathbf{x}_a$ is unimportant but the mere *fact of imposing* a value is important. We claim that, intuitively, we should say that $\mathbf{X}_a$ *does* have causal effect on $\mathbf{X}_b$ (imposing $\mathbf{X}_c$). Our definition is consistent with the intuitive sense of causal independence whereas the definition proposed by Ay and Polani is not. $\triangle$

The following proposition shows that Theorem 2.6 gives in some sense a full characterisation of causal conditional independence.

2.9. Proposition. *If $\mathcal{B}$ is not u-separated from $\mathcal{A}$ by $\mathcal{C}$ in a directed graph $\mathcal{G}$ then there exists a probability density $p(\mathbf{x})$ which satisfies Assumptions* 2.1 *and* 2.4 *with respect to $\mathcal{G}$ such that $\mathbf{X}_\mathcal{B}$ is not causally independent of $\mathbf{X}_\mathcal{A}$ imposing $\mathbf{X}_\mathcal{C}$:*

$$\mathcal{A} \longrightarrow_u \mathcal{B} \,|\, \mathcal{C} \quad \textit{implies} \quad \mathbf{X}_\mathcal{A} \rhd \mathbf{X}_\mathcal{B} \,\|\, \mathbf{X}_\mathcal{C} \textit{ for some } \mathbf{X} \textit{ satisfying } 2.1 \textit{ and } 2.4.$$

Proof. Let $\tau = (a = v_0 \to v_1, \to \cdots \to v_{n-1} \to v_n = b)$ be a directed path with $a \in \mathcal{A}$, $b \in \mathcal{B}$ which does not intersect $\mathcal{C}$. Consider a graph with all the arrows removed except those which are present in τ. With a slight abuse of notation we put $\mathcal{G}_\tau = (\mathcal{V}, \mathcal{E}_\tau)$ where $\mathcal{E}_\tau = \mathcal{E} \cap \tau$. The graph $\mathcal{G}_\tau$ is a DAG and any probability density which factorises along $\mathcal{G}_\tau$ satisfies Assumptions 2.1 and 2.4. Such a density makes all nodes in $\mathcal{V} \setminus \tau$ mutually independent and independent of $\mathcal{V} \cap \tau$, so $p(\mathbf{x}_b \,\|\, \mathbf{x}_a, \mathbf{x}_\mathcal{C}) = p(\mathbf{x}_b \,|\, \mathbf{x}_a)$ and $p(\mathbf{x}_b \,\|\, \mathbf{x}_\mathcal{C}) = p(\mathbf{x}_b)$.

It is therefore enough to choose a density on nodes belonging to τ such that $\mathbf{X}_a$ and $\mathbf{X}_b$ are dependent variables. ∎

Now we turn to the characterisation of observational conditional independence, denoted $\mathbf{X}_\mathcal{A} \perp\!\!\!\perp \mathbf{X}_\mathcal{B} \,|\, \mathbf{X}_C$. Theorem 2.10 below generalises the classical result on d-separation [18] to possibly cyclic graphs and composable random elements which satisfy Assumption 2.4. Several results similar to Theorem 2.10 can be found in the literature, e.g. [20, 15] and [3, Proposition 5]. These results are obtained in different setups under different sets of assumptions.

2.10. Theorem. *Consider a random element* $\mathbf{X}$ *which satisfies Assumptions* 2.1 *and* 2.4 *with respect to* $\mathcal{G}$. *If* $\mathcal{B}$ *is d-separated from* $\mathcal{A}$ *by* $\mathcal{C}$ *then* $\mathbf{X}_\mathcal{B}$ *is conditionally independent of* $\mathbf{X}_\mathcal{A}$ *given* $\mathbf{X}_\mathcal{C}$*:*

$$\mathcal{A} \perp_d \mathcal{B} \,|\, \mathcal{C} \quad \textit{implies} \quad \mathbf{X}_\mathcal{A} \perp\!\!\!\perp \mathbf{X}_\mathcal{B} \,|\, \mathbf{X}_\mathcal{C}.$$

We omit the proof, because the proof of an analogous result in [4, Theorem 2] applies without changes to our setup. It is based on the factorisation of densities ([4]) equivalent to our formula (2.5) combined with the results of [10, 11] and therefore goes through under Assumptions 2.1 and 2.4.

2.11. Proposition. *If* $\mathcal{A}$ *and* $\mathcal{B}$ *are not d-separated by* $\mathcal{C}$ *in a directed graph* $\mathcal{G}$ *then there exists a probability density* $p(\mathbf{x})$ *which statisfies Assumptions* 2.1 *and* 2.4 *with respect to* $\mathcal{G}$ *such that* $\mathbf{X}_\mathcal{A}$ *and* $\mathbf{X}_\mathcal{B}$ *are not conditionally independent given* $\mathbf{X}_\mathcal{C}$*:*

$\mathcal{A} \mathrel{—}_d \mathcal{B} \,|\, \mathcal{C}$ *implies* $\mathbf{X}_\mathcal{A} \not\perp\!\!\!\perp \mathbf{X}_\mathcal{B} \,|\, \mathbf{X}_\mathcal{C}$ *for some* $\mathbf{X}$ *satisfying* 2.1 *and* 2.4.

Proof. Let τ be a trail that joins $\mathcal{A}$ to $\mathcal{B}$ and is not d-separated by $\mathcal{C}$. Consider the set $\mathcal{E}_\tau$ of edges which consists of the edges in τ and also the edges leading from colliders of τ to $\mathcal{C}$. Now $\mathcal{G}_\tau = (\mathcal{V}, \mathcal{E}_\tau)$ is a DAG and $\mathcal{A}$ is not d-separated from $\mathcal{B}$ with respect to $\mathcal{G}_\tau$. A classical result [19, Theorem 1.2.4] ensures existence of p which factorises along $\mathcal{G}_\tau$ such that $\mathbf{X}_\mathcal{A} \not\perp\!\!\!\perp \mathbf{X}_\mathcal{B} \,\|\, \mathbf{X}_\mathcal{C}$. It is clear that this p satisfies Assumptions 2.1 and 2.4 with respect to $\mathcal{G}$. ∎

3. Stochastic processes. In this section and in the rest of the paper, $\mathbf{X} = (X_v(t),\ t \in \mathbb{T},\ v \in \mathcal{V})$ is a multivariate stochastic process, with components indexed by $v \in \mathcal{V}$. The time variable t is either discrete ($\mathbb{T} = \{0, 1, \dots, n\}$) or continuous ($\mathbb{T} = [0, u]$). Stochastic processes (and subprocesses) are denoted in bold, for example $\mathbf{X}_v = (X_v(t),\ t \in \mathbb{T})$ for $v \in \mathcal{V}$ and $\mathbf{X}_\mathcal{C} = (X_v(t),\ t \in \mathbb{T}, v \in \mathcal{C})$ for $\mathcal{C} \subseteq \mathcal{V}$. If t is fixed then random variables (or vectors) $X_v(t)$ and $X_\mathcal{C}(t) = (X_v(t),\ v \in \mathcal{C})$ are written in normal

([4]) However, to obtain factorisation of densities for stochastic processes, some assumption on the initial distribution is needed; see comments after Corollary 3.11. In the cited paper this fact seems to have been overlooked.

font. Accordingly, small case letters $x_v, x_{\mathcal{C}}$ denote values of random variables (vectors) $X_v(t), X_{\mathcal{C}}(t)$ while $\mathbf{x}_v, \mathbf{x}_{\mathcal{C}}$ denote trajectories of subprocesses $\mathbf{X}_v, \mathbf{X}_{\mathcal{C}}$.

Discrete time processes and DBNs. Let $\mathbf{X} = (X_v(t))$ be a multivariate stochastic process in discrete time ($t = 0, 1, \dots, n$), with components indexed by $v \in \mathcal{V}$. Assume that $x_v(t) \in \mathcal{S}_v$, where $\mathcal{S}_v$ can be either $\mathbb{R}$ equipped with the Lebesgue measure or a finite set with the counting measure. The joint distribution of $\mathbf{X}$ can be defined in terms of an initial density $p(x(0))$ and subsequent conditional densities $p(x(t) \,|\, x(t-1), \dots, x(0))$. We assume that the following independence conditions hold:

$$p(x(0)) = \prod_{v \in \mathcal{V}} p(x_v(0)), \tag{3.1a}$$

(3.1b)

$$p(x(t) \,|\, x(t-1), \dots, x(0)) = \prod_{v \in \mathcal{V}} p(x_v(t) \,|\, x(t-1), \dots, x(0)) \text{ for } t = 1, \dots, n.$$

If (3.1) is satisfied then we can easily check that the density $p(\mathbf{x})$ of $\mathbf{X}$ satisfies Assumption 2.1 with

$$p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v}) = p(x_v(0)) \prod_{t=1}^{n} p(x_v(t) \,|\, x(t-1), \dots, x(0)).$$

Indeed,

$$\begin{aligned} p(\mathbf{x}) &= p(x(0)) \prod_{t=1}^{n} p(x(t) \,|\, x(t-1), \dots, x(0)) \\ &= \Big(\prod_{v \in \mathcal{V}} p(x_v(0)) \Big) \prod_{t=1}^{n} \prod_{v \in \mathcal{V}} p(x_v(t) \,|\, x(t-1) \dots, x(0)) \\ &= \prod_{v \in \mathcal{V}} \Big(p(x_v(0)) \prod_{t=1}^{n} p(x_v(t) \,|\, x(t-1) \dots, x(0)) \Big) = \prod_{v \in \mathcal{V}} p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v}), \end{aligned}$$

which verifies the first condition of Assumption 2.1. To see that the second condition is also fulfilled, consider the following scenario. Let us fix $\mathbf{x}_{\mathcal{V} \setminus \mathcal{A}}$ and allow $\mathbf{X}_{\mathcal{A}}$ to evolve according to the equations analogous to (3.1), namely

$$p(x_{\mathcal{A}}(0)) = \prod_{v \in \mathcal{A}} p(x_v(0)),$$

$$p(x_{\mathcal{A}}(t) \,|\, x(t-1), \dots, x(0)) = \prod_{v \in \mathcal{A}} p(x_v(t) \,|\, x(t-1), \dots, x(0)) \text{ for } t{=}1, \dots, n$$

(note that $x(s)$ for $s < t$ include both fixed components $x_{\mathcal{V} \setminus \mathcal{A}}(s)$ and values of $x_{\mathcal{A}}(s)$ generated before t). The process $\mathbf{X}_{\mathcal{A}}$ obtained in this way has probability distribution $p(\mathbf{x}_{\mathcal{A}} \,\|\, \mathbf{x}_{\mathcal{V} \setminus \mathcal{A}}) = \prod_{v \in \mathcal{A}} p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v})$. Therefore $p(\mathbf{x}_{\mathcal{A}} \,\|\, \mathbf{x}_{\mathcal{V} \setminus \mathcal{A}})$

can indeed be interpreted as a conditional-by-intervention distribution. In particular, it integrates to 1 with respect to $\mathbf{x}_{\mathcal{A}}$. Moreover, this reasoning also justifies definition (2.2).

Now we consider processes which satisfy Assumption 2.4 with respect to a causal graph $\mathcal{G}$. In the present context of discrete time processes it is convenient to consider graphs with possible self-loops. Let $\mathcal{G}' = (\mathcal{V}, \mathcal{E}')$ be a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with added arrows $v \to v$ for some (not necessarily all) nodes v. Put $\mathrm{pa}'(v) = \{w : (w, v) \in \mathcal{E}'\}$. Assume that the transition probabilities at node v depend on the past configurations of parents of v:

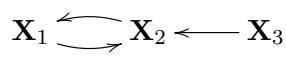
$$p(x_v(t) \,|\, x(t-1), \dots, x(0)) = p(x_v(t) \,|\, x_{\mathrm{pa}'(v)}(t-1), \dots, x_{\mathrm{pa}'(v)}(0)) \tag{3.2}$$

for $t = 1, \dots, n$. If (3.2) is true then Assumption 2.4 is satisfied: $p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v}) = p(\mathbf{x}_v \,\|\, \mathbf{x}_{\mathrm{pa}(v)})$.

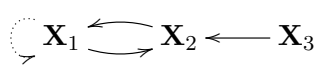
Formula (3.2) implicitly defines an *acyclic* directed graph (DAG) on the set of "space-and-time" nodes (v, t). The arrows lead from (w, s) to (v, t) whenever $w \in \mathrm{pa}'(v)$ and $s < t$. The collection of random variables $(X_v(t))$ forms a dynamic Bayesian network (DBN) ([5]).

Let us illustrate our considerations by a simple example.

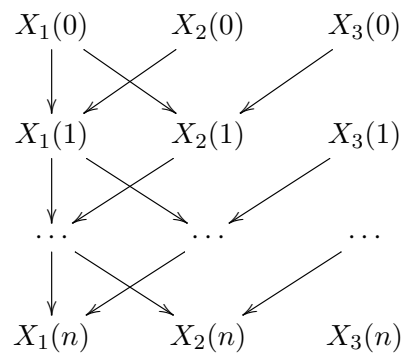
3.3. Example. Let $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = (X_1(t), X_2(t), X_3(t) : t = 0, 1, \dots, n)$. The graph $\mathcal{G}$ is



The augmented graph $\mathcal{G}'$ is obtained by adding the self-loop $\mathbf{X}_1 \to \mathbf{X}_1$, indicated below by a dotted arrow:



For simplicity assume that the process $\mathbf{X}$ is Markov. The structure of dependence of the $3(n+1)$ random variables is given by the following graph:



([5]) The transition densities $p(x_v(t) \,|\, x_{\mathrm{pa}'(v)}(t-1), \dots, x_{\mathrm{pa}'(v)}(0))$ could also be written using double bars, as $p(x_v(t) \,\|\, x_{\mathrm{pa}'(v)}(t-1), \dots, x_{\mathrm{pa}'(v)}(0))$, because they are conditional-by-intervention densities in the model given by the "space-and-time" DAG.

We have

$$p(\mathbf{x}_1 \,\|\, \mathbf{x}_2) = p(x_1(0)) \prod_{t=1}^{n} p(x_1(t) \,|\, x_1(t-1), x_2(t-1)),$$

$$p(\mathbf{x}_2 \,\|\, \mathbf{x}_1, \mathbf{x}_3) = p(x_2(0)) \prod_{t=1}^{n} p(x_2(t) \,|\, x_1(t-1), x_3(t-1)),$$

$$p(\mathbf{x}_3) = p(x_3(0)) \prod_{t=1}^{n} p(x_3(t)),$$

and the joint distribution is

$$p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = p(\mathbf{x}_1 \,\|\, \mathbf{x}_2) p(\mathbf{x}_2 \,\|\, \mathbf{x}_1, \mathbf{x}_3) p(\mathbf{x}_3).$$

Moreover, it is easy to verify that, $p(\mathbf{x}_1, \mathbf{x}_2 \,\|\, \mathbf{x}_3) = p(\mathbf{x}_1 \,\|\, \mathbf{x}_2) p(\mathbf{x}_2 \,\|\, \mathbf{x}_1, \mathbf{x}_3)$ can be interpreted as the conditional-by-intervention distribution of $(\mathbf{X}_1, \mathbf{X}_2)$ given $\mathbf{X}_3 = \mathbf{x}_3$. In particular, $\int\int p(\mathbf{x}_1, \mathbf{x}_2 \,\|\, \mathbf{x}_3)\, \mathrm{d}\mathbf{x}_1\, \mathrm{d}\mathbf{x}_2 = 1$ (note that here $\mathrm{d}\mathbf{x}_v = \prod_{t=0}^{n} \mathrm{d}x_v(t)$). $\triangle$

Summarising, we obtain the following conclusions.

3.4. Proposition. *If a discrete-time process* $\mathbf{X}$ *satisfies* (3.1) *and* (3.2) *then* $\mathbf{X}$ *is a random element which satisfies Assumptions* 2.1 *and* 2.4.

3.5. Corollary. *If a discrete-time process* $\mathbf{X}$ *satisfies* (3.1) *and* (3.2) *then Theorems* 2.6 *and* 2.10 *hold true for* $\mathbf{X}$. (*The separation conditions are defined in terms of the causal graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.)

Continuous time Bayesian networks. The same idea as in the previous subsection applies to the case of continuous time processes, but the technical details become more complicated. We restrict ourselves to composable finite Markov processes (CFMP) defined in [22]. (As a matter of fact, we have chosen the term "*composable* random element" to indicate that CFMPs are the basic examples of CREs.)

Let $\mathbf{X} = (X(t),\, 0 \le t \le u)$ be a Markov process with values in a finite state space $\mathcal{S}$. The distribution of $\mathbf{X}$ is characterised by its initial distribution, say ν, and the transition intensities. Assume that there exist bounded functions $Q(t; x, x')$ such that for $x, x' \in \mathcal{S}$,

$$\mathbb{P}(X(t+h) = x' \,|\, X(t) = x) = \begin{cases} Q(t; x, x')h + o(h) & \text{for } x \neq x', \\ 1 + Q(t; x, x)h + o(h) & \text{for } x = x', \end{cases} \tag{3.6}$$

as $h \searrow 0$, where $Q(t; x, x) = -\sum_{x' \neq x} Q(t; x, x')$. We also have $\mathbb{P}(X(0) = x) = \nu(x)$.

The density of the process $\mathbf{X}$ is given by the formula

(3.7)
$$p(\mathbf{x}) = \nu(x(0)) \prod_{t:\, x(t-)\neq x(t)} Q(t;\, x(t-), x(t)) \times \exp\Big[\int_0^u Q(t;\, x(t), x(t))\, \mathrm{d}t\Big].$$

This is a density with respect to a natural measure on the space of trajectories, where each trajectory $\mathbf{x}$ is represented by its "space-time skeleton". For completeness, we give an explicit description of this measure in Appendix A. Integrals with respect to this measure will be simply denoted $\int \cdots \mathrm{d}\mathbf{x}$. Formula (3.7) is well known but it appears in many papers as an expression for *likelihood*, with $\mathbf{x}$ considered to be fixed. Constants are then usually neglected and the initial distribution ν is omitted, as e.g. in [4, 8, 12]. For our purposes ν has to be explicitly taken into consideration.

Now we turn to mutivariate (finite) Markov processes, that is, consider the case when $\mathcal{S} = \prod_{v\in\mathcal{V}} \mathcal{S}_v$ and thus $\mathbf{X} = (\mathbf{X}_v,\, v \in \mathcal{V}) = (X_v(t),\, v \in \mathcal{V},\, 0 \le t \le u)$. The analogue of (3.1) in the present setting is the following "infinitesimal independence" condition (accompanied by the independence condition on the initial distribution):

(3.8a)
$$\nu(x) = \prod_v \nu(x_v),$$

(3.8b)
$$\mathbb{P}(X(t+h) = x' \,|\, X(t) = x) = \prod_v [\mathbb{P}(X_v(t+h) = x'_v \,|\, X(t) = x) + o(h)] \qquad (h \searrow 0) \text{ for } 0 \le t < u.$$

It is clear that the RHS of (3.8b) is $o(h)$ whenever $x_v \neq x'_v$ and $x_w \neq x'_w$ for some $v \neq w$. Thus jumps of $\mathbf{X}$ cannot occur at more than one coordinate simultaneously. Under (3.8b), $\mathbf{X}$ is a composable finite Markov process (CFMP) in the sense of Schweder. The transition intensities are of the following form: for $x \neq x'$,

$$Q(t;\, x, x') = \begin{cases} Q_v(t;\, x_{-v};\, x_v, x'_v) & \text{if } x_v \neq x'_v,\ x_{-v} = x'_{-v}, \\ 0 & \text{if there are } v \neq w \text{ such that} \\ & x_v \neq x'_v,\ x_w \neq x'_w. \end{cases}$$

It is easy to verify that

$$Q(t;\, x, x) = \sum_v Q_v(t;\, x_{-v};\, x_v, x_v),$$

where we put $Q_v(t;\, x_{-v};\, x_v, x_v) = -\sum_{x'_v \neq x_v} Q_v(t;\, x_{-v};\, x_v, x'_v)$. It follows that under (3.8) we have

$$p(\mathbf{x}) = \prod_{v\in\mathcal{V}} p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v}),$$

where

$$p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v}) = \nu(x_v(0)) \prod_{t:\, x_v(t-)\neq x_v(t)} Q_v(t, x_{-v}(t);\, x_v(t-), x_v(t)) \times \exp\Big[\int_0^u Q_v(t;\, x_{-v}(t);\, x_v(t), x_v(t))\,\mathrm{d}t\Big].$$

By comparison of the last equation with (3.7) we can see that for any fixed $\mathbf{x}_{-v} \in \mathcal{X}_{-v}$ the function $\mathbf{x}_v \mapsto p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v})$ is the probability density of the Markov process $\mathbf{X}_v$ with transition intensity defined by $Q_v^{|-v}(t; x_v, x_v') = Q_v(t; x_{-v}(t); x_v, x_v')$ and therefore $\int p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v})\,\mathrm{d}\mathbf{x}_v = 1$ (integration is with respect to the natural measure on the space of trajectories $\mathbf{x}_v$, see Appendix A). Pushing this argument further, for every subset $\mathcal{A}$ of nodes and any fixed $\mathbf{x}_{\mathcal{V}\setminus\mathcal{A}} \in \mathcal{X}_{\mathcal{V}\setminus\mathcal{A}}$, we deduce that the function $\mathbf{x}_\mathcal{A} \mapsto \prod_{v\in\mathcal{A}} p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v})$ is the probability density of the Markov process $\mathbf{X}_\mathcal{A}$ with single-node transition intensities $Q_v^{|\mathcal{V}\setminus\mathcal{A}}(t; x_{\mathcal{A}\setminus\{v\}}(t); x_v, x_v') = Q_v(t; x_{-v}(t); x_v, x_v')$ and hence $\int \prod_{v\in\mathcal{A}} p(\mathbf{x}_v \,\|\, \mathbf{x}_{-v})\,\mathrm{d}\mathbf{x}_\mathcal{A} = 1$. We have thus obtained the following result.

3.9. Proposition. *If a continuous finite Markov process* $\mathbf{X}$ *satisfies* (3.8) *then* $\mathbf{X}$ *is a random element which satisfies Assumption* 2.1.

Put differently, if a CFMP (in the sense of Schweder) has independent initial distribution then it is a CRE in our terminology.

Given a graph $\mathcal{G}$, we say that a CFMP $\mathbf{X}$ is a *continuous time Bayesian network* (CTBN) ([6]) if the transition intensities at v depend on x_{-v} only through $x_{\mathrm{pa}(v)}$, i.e.

$$Q_v(t;\, x_{-v};\, x_v, x_v') = Q_v(t;\, x_{\mathrm{pa}(v)};\, x_v, x_v'). \tag{3.10}$$

Thus $Q_v(t;\, x_{\mathrm{pa}(v)};\, \cdot, \cdot)$ is a matrix of intensities of transitions at node v at time t if the configuration of parent nodes at this time is $x_{\mathrm{pa}(v)}$.

3.11. Corollary. *If a continuous-time Markov process* $\mathbf{X}$ *satisfies* (3.8) *and* (3.10) *then* $\mathbf{X}$ *is a* $\mathcal{G}$*-CRE and consequently Theorems* 2.6 *and* 2.10 *hold true for* $\mathbf{X}$.

Note that condition (3.8a) (independent initial distribution) is needed in Corollary 3.11. To see that (3.8b) and (3.10) *do not imply* 2.10, it is enough to consider the trivial case of $X(t)$ with $t \in \{0\}$ (a process indexed by a degenerate interval of time).

To illustrate Proposition 3.9 and Corrollary 3.11 we provide the following example, which is a continuous time analogue of Example 3.3.

([6]) In fact, it is usually assumed that a CTBN is a time-homogeneous process, i.e. $Q(t; x, x')$ does not depend on t, but we do not need this assumption.

3.12. EXAMPLE. Let us consider a Markov process $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = (X_1(t), X_2(t), X_3(t),\, t \in [0,u])$. Assume $\mathbf{X}$ is a CTBN with respect to the following graph $\mathcal{G}$:

$$\mathbf{X}_1 \rightleftarrows \mathbf{X}_2 \longleftarrow \mathbf{X}_3$$

Thus $\mathbf{X}$ is defined via the following conditional intensities: $Q_1(t;\, x_2;\, x_1, x_1')$, $Q_2(t;\, (x_1,x_3);\, x_2, x_2')$ and $Q_3(t;\, x_3, x_3')$. The density of $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$ is given by the formula

$$p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = p(\mathbf{x}_1 \,\|\, \mathbf{x}_2) \cdot p(\mathbf{x}_2 \,\|\, \mathbf{x}_1, \mathbf{x}_3) \cdot p(\mathbf{x}_3),$$

where

$$\begin{aligned}
p(\mathbf{x}_1 \,\|\, \mathbf{x}_2) &= \nu(x_1(0)) \prod_{t:\, x_1(t-)\neq x_1(t)} Q_1(t;\, x_2(t);\, x_1(t-), x_1(t)) \\
&\quad \times \exp\Big[\int_0^u Q_1(t;\, x_2(t);\, x_1(t), x_1(t))\,\mathrm{d}t\Big], \\
p(\mathbf{x}_2 \,\|\, \mathbf{x}_1, \mathbf{x}_3) &= \nu(x_2(0)) \prod_{t:\, x_2(t-)\neq x_2(t)} Q_2(t;\, x_1(t), x_3(t);\, x_2(t-), x_2(t)) \\
&\quad \times \exp\Big[\int_0^u Q_2(t;\, x_1(t), x_3(t);\, x_2(t), x_2(t))\,\mathrm{d}t\Big], \\
p(\mathbf{x}_3) &= \nu(x_3(0)) \prod_{t:\, x_3(t-)\neq x_3(t)} Q_3(t;\, x_3(t-), x_3(t)) \\
&\quad \times \exp\Big[\int_0^u Q_3(t;\, x_3(t), x_3(t))\,\mathrm{d}t\Big].
\end{aligned}$$

If for example we fix $\mathbf{x}_3 = (x_3(t) : 0 \le t \le u)$, then it is easy to notice that $p(\mathbf{x}_1 \,\|\, \mathbf{x}_2)p(\mathbf{x}_2 \,\|\, \mathbf{x}_1, \mathbf{x}_3)$ is the probability density of the process $(\mathbf{X}_1, \mathbf{X}_2)$ with transition intensities

$$\begin{aligned}
Q_1^{|3}(t;\, x_2;\, x_1, x_1') &= Q_1(t;\, x_2;\, x_1, x_1'), \\
Q_2^{|3}(t;\, x_1;\, x_2, x_2') &= Q_2(t;\, x_1, x_3(t);\, x_2, x_2').
\end{aligned}$$

Therefore $\int\int p(\mathbf{x}_1 \,\|\, \mathbf{x}_2)p(\mathbf{x}_2 \,\|\, \mathbf{x}_1, \mathbf{x}_3)\,\mathrm{d}\mathbf{x}_1\,\mathrm{d}\mathbf{x}_2 = 1$ for every $\mathbf{x}_3$. Analogously $\int p(\mathbf{x}_2 \,\|\, \mathbf{x}_1, \mathbf{x}_3)\,\mathrm{d}\mathbf{x}_2 = 1$ for every $(\mathbf{x}_1, \mathbf{x}_3)$ and so on. We can see that the process $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$ is a $\mathcal{G}$-CRE in the sense of Assumptions 2.1 and 2.4. $\triangle$

4. Conditional independence involving time. In this section we tackle independence between the past of one subprocess and the future of another given the past of a third subprocess. More precisely, let $\mathbf{X} = (X_v(t),\, v \in \mathcal{V})$ be a multivariate stochastic process. We want to find necessary and

sufficient conditions for the following independence relation:

$$(X_{\mathcal{B}}(s),\, s > t) \perp\!\!\!\perp (X_{\mathcal{A}}(s),\, s \le t) \,|\, (X_{\mathcal{B}\cup\mathcal{C}}(s),\, s \le t). \tag{4.1}$$

This is a natural problem clearly related to the task of prediction. If we want to predict the *future* of $\mathbf{X}_{\mathcal{B}}$ knowing the *past* of $\mathbf{X}_{\mathcal{B}\cup\mathcal{C}}$ then relation (4.1) tells us that additional information on the past of $\mathbf{X}_{\mathcal{A}}$ is redundant and can be discarded.

We also consider a similar relation with "future" replaced by "immediate future". For a process with discrete time ($t = 0, 1, \ldots, n$) we ask when it is true that

$$X_{\mathcal{B}}(t+1) \perp\!\!\!\perp (X_{\mathcal{A}}(s),\, s \le t) \,|\, (X_{\mathcal{B}\cup\mathcal{C}}(s),\, s \le t). \tag{4.2}$$

For a continuous time process ($t \in [0, u]$) "immediate future" should be understood in some infinitesimal sense. Intuitively, a continuous time analogue of (4.2) is

$$(X_{\mathcal{B}}(t+) \perp\!\!\!\perp (X_{\mathcal{A}}(s),\, s \le t) \,|\, (X_{\mathcal{B}\cup\mathcal{C}}(s),\, s \le t), \tag{4.3}$$

where $X(t+)$ stands for "$(X(s),\, t < s < t + h)$ with $h \searrow 0$". For a Markov process with a finite state space we will give a precise meaning of this informal statement. By convention assume that relation (4.3) is *defined* as follows:

$$\begin{aligned} &\mathbb{P}(X_{\mathcal{B}}(t+h) = x'_{\mathcal{B}} \,|\, X_{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}}(s) = x_{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}}(s),\, s \le t) \\ &\quad = \mathbb{P}(X_{\mathcal{B}}(t+h) = x'_{\mathcal{B}} \,|\, X_{\mathcal{B}\cup\mathcal{C}}(s) = x_{\mathcal{B}\cup\mathcal{C}}(s),\, s \le t) \; + o(h) \qquad (h \searrow 0) \end{aligned} \tag{4.4}$$

for every past trajectory $(x_{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}}(s),\, s \le t)$ of $\mathbf{X}_{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}}$ and every $x'_{\mathcal{B}} \in \mathcal{X}_{\mathcal{B}}$.

In the remaining part of this section we formulate graph-theoretical conditions that ensure the independence relations (4.1), (4.2) and (4.4). We first recall the results for discrete time processes and then show that they remain true also for CTBNs.

We say that $\mathcal{B}$ is *ε-separated* from $\mathcal{A}$ by $\mathcal{C}$ if every trail from $\mathcal{A}$ to $\mathcal{B}$ which ends with an arrow $\to b \in \mathcal{B}$ and has no other nodes in $\mathcal{B}$, satisfies at least one of the following conditions: it contains a chain $\leftarrow c \leftarrow$ or a fork $\leftarrow c \rightarrow$ with $c \in \mathcal{C}$ or a collider $\rightarrow v \leftarrow$ such that $v \notin \mathrm{an}(\mathcal{C})$, or contains a chain $\rightarrow c \rightarrow$ with $c \in \mathcal{C}$ that occurs earlier than some collider. We will then write $\mathcal{A} \perp\!\!\!+_{\varepsilon} \mathcal{B} \,|\, \mathcal{C}$. Negation of this statement is denoted $\mathcal{A} \perp\!\!\!-_{\varepsilon} \mathcal{B} \,|\, \mathcal{C}$.

Remark. The definition of ε-separation given in [16] is incomplete. It does not require that an ε-open trail from $\mathcal{A}$ to $\mathcal{B}$ must not contain nodes belonging to $\mathcal{B}$ other than the end node. However, this requirement is tacitly used in the proof of [16, Theorem 3.9(b)].

The following result was established in [6, Theorem 4.8] (see also [16, Theorem 3.9(a)] for an independent proof).

4.5. THEOREM. *Consider a discrete time process* $\mathbf{X} = (X_v(t),\ t = 0, 1, \ldots, n)$ *satisfying* (3.1) *and* (3.2) *with respect to a graph* $\mathcal{G}$. *If* $\mathcal{B}$ *is* ε*-separated from* $\mathcal{A}$ *by* $\mathcal{C}$ *then* (4.1) *holds for every* $t < n$.

The following proposition proved in [16, Theorem 3.9(b)] is a sort of converse statement.

4.6. PROPOSITION. *If* $\mathcal{B}$ *is not* ε*-separated from* $\mathcal{A}$ *by* $\mathcal{C}$ *then there exists a discrete time process* $\mathbf{X} = (X_v(t),\ t = 0, 1, \ldots, n)$ *satisfying* (3.1) *and* (3.2) *such that* (4.1) *does not hold.*

The proofs of both Theorem 4.5 and Proposition 4.6 given in [16] use d-separation relations in the "space-time" graph (mentioned in the subsection on discrete time processes in Section 3). It turns out that analogous results are true also for CTBNs, but the proofs have to be based on entirely different ideas.

4.7. THEOREM. *Consider a CTBN* $\mathbf{X} = (X_v(t),\ 0 \le t \le u)$ *satisfying* (3.8) *and* (3.10) *with respect to a graph* $\mathcal{G}$. *If* $\mathcal{B}$ *is* ε*-separated from* $\mathcal{A}$ *by* $\mathcal{C}$ *then* (4.1) *holds for every* $t < u$.

Proof. Let

$$\mathcal{D} = \mathrm{an}(\mathcal{B}) \setminus (\mathcal{B} \cup \mathcal{C}), \qquad \mathcal{R} = \mathcal{V} \setminus (\mathcal{A} \cup \mathrm{an}(\mathcal{B}) \cup \mathcal{C}).$$

We assume that $\mathcal{D}$ and $\mathcal{R}$ are nonempty; if $\mathcal{D} = \emptyset$ or $\mathcal{R} = \emptyset$ then the reasoning is similar but simpler, and therefore is omitted.

We will use the following notation. For a fixed $t \in [0, u)$ let

$$\begin{aligned}
\mathbf{A} &= (X_{\mathcal{A}}(s),\ 0 \le s \le t), && \text{the past of } \mathbf{X}_{\mathcal{A}},\\
\mathbf{B} &= (X_{\mathcal{B}}(s),\ 0 \le s \le t), && \text{the past of } \mathbf{X}_{\mathcal{B}},\\
\mathbf{C} &= (X_{\mathcal{C}}(s),\ 0 \le s \le t), && \text{the past of } \mathbf{X}_{\mathcal{C}},\\
\mathbf{D} &= (X_{\mathcal{D}}(s),\ 0 \le s \le t), && \text{the past of } \mathbf{X}_{\mathcal{D}},\\
\mathbf{R} &= (X_{\mathcal{R}}(s),\ 0 \le s \le t), && \text{the past of } \mathbf{X}_{\mathcal{R}},\\
\mathbf{F} &= (X_{\mathcal{B}}(s),\ t < s \le u), && \text{the future of } \mathbf{X}_{\mathcal{B}},
\end{aligned}$$

with $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{r}$ and $\mathbf{f}$ denoting the values of these random elements. In these notations the conclusion which we are to prove is the following:

$$p(\mathbf{f} \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}) = p(\mathbf{f} \,|\, \mathbf{b}, \mathbf{c}).$$

We first show that $\mathcal{A} \perp\!\!\!\perp_{\tau\varepsilon} \mathcal{B} \,|\, \mathcal{C}$ implies

$$\mathcal{D} \perp_d \mathcal{A} \,|\, \mathcal{B} \cup \mathcal{C}.$$

Indeed, suppose that there exists a trail τ from $a \in \mathcal{A}$ to $d \in \mathcal{D}$ which is not d-separated by $\mathcal{B} \cup \mathcal{C}$. Let $x \leftarrow y$ be the first entry of τ to $\mathrm{an}(\mathcal{B})$ (the arrow must be directed to a because otherwise x would belong to $\mathrm{an}(\mathcal{B})$). If we had $y \in \mathcal{B} \cup \mathcal{C}$ then τ would be d-separated because y is not a collider. If we had

$y \in \mathcal{D}$ then we would have $\mathcal{A} \rightharpoonup_{\varepsilon} \mathcal{B} \,|\, \mathcal{C}$. To see this, note that $y \in \mathrm{an}(\mathcal{B}) \setminus \mathcal{B}$ implies that there exists $z \in \mathrm{an}(\mathcal{B})$ such that $y \to z$. Let τ_y be the part of τ from a to y. Clearly, τ_y contains neither $\leftarrow v \leftarrow$ nor $\rightarrow v \rightarrow$ nor $\leftarrow v \rightarrow$ with $v \in \mathcal{B} \cup \mathcal{C}$. If τ_y contains $\rightarrow v \leftarrow$ then $v \in \mathrm{an}(\mathcal{B} \cup \mathcal{C})$. Since y is the first entry to $\mathrm{an}(\mathcal{B})$, it follows that $v \in \mathrm{an}(\mathcal{C})$ and thus the collider at v does not ε-block the trail from a to y to z to some $b \in \mathcal{B}$ (by $\mathcal{C}$).

We have obtained a contradiction and therefore $\mathcal{D} \perp_d \mathcal{A} \,|\, \mathcal{B} \cup \mathcal{C}$. By Theorem 2.10 this graph separation statement entails conditional independence $\mathbf{D} \perp\!\!\!\perp \mathbf{A} \,|\, \mathbf{B}, \mathbf{C}$ and thus

$$p(\mathbf{d} \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}) = p(\mathbf{d} \,|\, \mathbf{b}, \mathbf{c}). \tag{4.8}$$

The next step is to show that

$$p(\mathbf{f} \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = p(\mathbf{f} \,|\, \mathbf{b}, \mathbf{c}, \mathbf{d}). \tag{4.9}$$

This follows from $\mathcal{B} \cup \mathcal{C} \cup \mathcal{D} \supseteq \mathrm{an}(\mathcal{B})$. Indeed, since $\mathrm{an}(\mathcal{B})$ is an ancestral set, $\mathbf{X}_{\mathrm{an}(\mathcal{B})}$ is a Markov process and $(X_{\mathrm{an}(\mathcal{B})}(s),\ s > t)$ depends on $(X_{\mathcal{V}}(s),\ s \le t)$ only through $X_{\mathrm{an}(\mathcal{B})}(t)$. Therefore $\mathbf{F}$ is conditionally independent of $\mathbf{A}$ and $\mathbf{R}$ given $(\mathbf{B}, \mathbf{C}, \mathbf{D})$. It follows that $p(\mathbf{f} \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = p(\mathbf{f} \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{r}) = p(\mathbf{f} \,|\, \mathbf{b}, \mathbf{c}, \mathbf{d})$ and we obtain (4.9).

Finally, to obtain the desired conclusion we combine (4.8) with (4.9):

$$\begin{aligned} p(\mathbf{f} \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}) &= \int p(\mathbf{f} \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) p(\mathbf{d} \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}) \, \mathrm{d}\mathbf{d} \\ &= \int p(\mathbf{f} \,|\, \mathbf{b}, \mathbf{c}, \mathbf{d}) p(\mathbf{d} \,|\, \mathbf{b}, \mathbf{c}) \, \mathrm{d}\mathbf{d} = p(\mathbf{f} \,|\, \mathbf{b}, \mathbf{c}), \end{aligned}$$

which completes the proof. ∎

We now turn to a graph-theoretic characterisation of the "immediate future" independence relation (4.4) for CTBNs. The definition of δ-separation and Theorem 4.10 below are due to Vanessa Didelez [4, Theorem 1]. Nevertheless, below we give a more explicit statement of this result with a simple proof to highlight the analogy with ε-separation and Theorem 4.10.

We say that $\mathcal{B}$ is *δ-separated* from $\mathcal{A}$ by $\mathcal{C}$ if every trail from $\mathcal{A}$ to $\mathcal{B}$ which ends with an arrow $\rightarrow b \in \mathcal{B}$ and has no other nodes in $\mathcal{B}$, satisfies at least one of the following conditions: it contains a chain $\leftarrow c \leftarrow$ or a fork $\leftarrow c \rightarrow$ or a chain $\rightarrow c \rightarrow$ with $c \in \mathcal{C}$ or a collider $\rightarrow v \leftarrow$ such that $v \notin \mathrm{an}(\mathcal{C})$. We will then write $\mathcal{A} \not\rightharpoonup_{\delta} \mathcal{B} \,|\, \mathcal{C}$. The negation of this statement is denoted $\mathcal{A} \rightharpoonup_{\delta} \mathcal{B} \,|\, \mathcal{C}$.

4.10. Theorem. *Consider a CTBN $\mathbf{X} = (X_v(t),\ 0 \le t \le u)$ satisfying* (3.8) *and* (3.10) *with respect to a graph $\mathcal{G}$. If $\mathcal{B}$ is δ-separated from $\mathcal{A}$ by $\mathcal{C}$ then* (4.4) *holds for every $t < u$.*

Proof. The proof is similar to that of Theorem 4.7. Let

$$\mathcal{D}' = \mathrm{pa}(\mathcal{B}) \setminus (\mathcal{B} \cup \mathcal{C}), \qquad \mathcal{R}' = \mathcal{V} \setminus (\mathcal{A} \cup \mathrm{pa}(\mathcal{B}) \cup \mathcal{C}).$$

We will use almost the same notation for $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, $\mathbf{R}$ (and $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{r}$) as in the proof of Theorem 4.7, just replacing $\mathcal{D}$ and $\mathcal{R}$ by $\mathcal{D}'$ and $\mathcal{R}'$. We now put

$$\mathbf{D} = (X_{\mathcal{D}'}(s),\, 0 \le s \le t), \qquad \text{the past of } \mathbf{X}_{\mathcal{D}'},$$
$$\mathbf{R} = (X_{\mathcal{R}'}(s),\, 0 \le s \le t), \qquad \text{the past of } \mathbf{X}_{\mathcal{R}'}.$$

The "immediate future" $X_B(t+)$ will be represented by the random event

$$F_h = \{X_{\mathcal{B}}(t+h) = x'_{\mathcal{B}}\}.$$

To verify condition (4.4) it is enough to consider the situation when $x'_{\mathcal{B}}$ differs from $x_{\mathcal{B}}(t)$ at exactly one node, say $v \in \mathcal{B}$. From now on we assume that $x'_v \ne x_v(t)$ and $x'_{\mathcal{B}\setminus\{v\}} = x_{\mathcal{B}\setminus\{v\}}(t)$.

We are to prove the following:

$$\mathbb{P}(F_h \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}) = \mathbb{P}(F_h \,|\, \mathbf{b}, \mathbf{c}) + o(h).$$

We first show that $\mathcal{A} \not\to_\delta \mathcal{B} \,|\, \mathcal{C}$ implies

$$\mathcal{D}' \perp_d \mathcal{A} \,|\, \mathcal{B} \cup \mathcal{C}.$$

Indeed, suppose that there exists a trail τ from $a \in \mathcal{A}$ to $d \in \mathcal{D}'$ which is not d-separated by $\mathcal{B} \cup \mathcal{C}$. By assumption, τ contains neither $\leftarrow v \leftarrow$ nor $\rightarrow v \rightarrow$ nor $\leftarrow v \rightarrow$ with $v \in \mathcal{B} \cup \mathcal{C}$. If all colliders in τ belong to $\mathrm{an}(\mathcal{C})$ (or if there are no colliders) then τ with added $d \rightarrow b \in \mathcal{B}$ is not δ-separated by $\mathcal{C}$, contrary to our assumption. However, τ may contain a collider $\rightarrow v \leftarrow$ with $v \in \mathrm{an}(\mathcal{B})$ and $v \notin \mathrm{an}(\mathcal{C})$. Assume that v is the first such collider. Then we can compose the part of τ ending at v with a directed path from v to the first element of $\mathcal{B}$ to obtain a trail from $\mathcal{A}$ to $\mathcal{B}$ which is not δ-separated by $\mathcal{C}$.

The contradiction obtained shows that $\mathcal{D}' \perp_d \mathcal{A} \,|\, \mathcal{B} \cup \mathcal{C}$. By Theorem 2.10 this graph separation statement entails $\mathbf{D} \perp\!\!\!\perp \mathbf{A} \,|\, \mathbf{B}, \mathbf{C}$ and thus (exactly as in the proof of Theorem 4.7)

$$p(\mathbf{d} \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}) = p(\mathbf{d} \,|\, \mathbf{b}, \mathbf{c}). \tag{4.11}$$

Now we are going to show that

$$\mathbb{P}(F_h \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = \mathbb{P}(F_h \,|\, \mathbf{b}, \mathbf{c}, \mathbf{d}) + o(h). \tag{4.12}$$

Indeed, $(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{r}) = (x(s),\, s \le t)$ is the past trajectory of the whole process. Therefore, using the Markov property we obtain

$$\begin{aligned}
&\mathbb{P}(F_h \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{r}) \\
&\qquad = \mathbb{P}(X_{\mathcal{B}}(t+h) = x'_{\mathcal{B}} \,|\, X(s) = x(s),\, s \le t) \\
&\qquad = \mathbb{P}(X_{\mathcal{B}}(t+h) = x'_{\mathcal{B}} \,|\, X(t) = x(t)) \\
&\qquad = \sum\nolimits_{x'_{\mathcal{V}\setminus\mathcal{B}}} \mathbb{P}(X_{\mathcal{B}}(t+h) = x'_{\mathcal{B}},\, X_{\mathcal{V}\setminus\mathcal{B}}(t+h) = x'_{\mathcal{V}\setminus\mathcal{B}} \,|\, X(t) = x(t))
\end{aligned}$$

Since $x'_v \neq x_v(t)$, all the terms with $x'_{\mathcal{V}\setminus\mathcal{B}} \neq x_{\mathcal{V}\setminus\mathcal{B}}(t)$ are $o(h)$, because they correspond to more than one jump of the process. Moreover, $x'_{\mathcal{B}\setminus\{v\}} = x_{\mathcal{B}\setminus\{v\}}(t)$. Therefore

$$\begin{aligned}
&\mathbb{P}(F_h \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{r}) \\
&\quad = \mathbb{P}\big(X_v(t+h) = x'_v,\, X_{-v}(t+h) = x_{-v}(t) \,|\, X(t) = x(t)\big) + o(h) \\
&\quad = Q_v(t; x_{\mathrm{pa}(v)}(t); x_v(t), x'_v)h + o(h) \\
&\quad = q(\mathbf{b}, \mathbf{c}, \mathbf{d})h + o(h)
\end{aligned}$$

for some function q, since $Q_v(t; x_{\mathrm{pa}(v)}(t); x_v(t), x'_v)$ depends only on $x_{\mathrm{pa}(\mathcal{B})}(t)$ and $\mathrm{pa}(\mathcal{B}) \subseteq \mathcal{B} \cup \mathcal{C} \cup \mathcal{D}'$.

Let us make an important remark on the meaning of $o(h)$ terms. All $o(h)$ terms appearing in the last chain of equations can be bounded by a function $r(h)$ *depending only on* h such that $r(h)/h \to 0$ with $h \searrow 0$. This is because in the basic equation (3.6) we have a finite number of $o(h)$ terms (a single such term can depend on h and some x, x'). In view of the above remark we can "integrate $o(h)$ with respect to a probability distribution" and obtain $o(h)$ also bounded by $r(h)$.

We marginalise the equality $\mathbb{P}(F_h \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{r}) = q(\mathbf{b}, \mathbf{c}, \mathbf{d})h + o(h)$ with respect to $\mathbf{r}$ as follows:

$$\begin{aligned}
\mathbb{P}(F_h \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) &= \int \mathbb{P}(F_h \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{r}) p(\mathbf{r} \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})\, \mathrm{d}\mathbf{r} \\
&= \int \big(q(\mathbf{b}, \mathbf{c}, \mathbf{d})h + o(h)\big) p(\mathbf{r} \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})\, \mathrm{d}\mathbf{r} \\
&= q(\mathbf{b}, \mathbf{c}, \mathbf{d})h + o(h).
\end{aligned}$$

Further marginalising with respect to $\mathbf{a}$ we get

$$\begin{aligned}
\mathbb{P}(F_h \,|\, \mathbf{b}, \mathbf{c}, \mathbf{d}) &= \int \mathbb{P}(F_h \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) p(\mathbf{a} \,|\, \mathbf{b}, \mathbf{c}, \mathbf{d})\, \mathrm{d}\mathbf{a} \\
&= \int \big(q(\mathbf{b}, \mathbf{c}, \mathbf{d})h + o(h)\big) p(\mathbf{a} \,|\, \mathbf{b}, \mathbf{c}, \mathbf{d})\, \mathrm{d}\mathbf{a} \\
&= q(\mathbf{b}, \mathbf{c}, \mathbf{d})h + o(h).
\end{aligned}$$

Thus we have proved (4.12).

Finally, to obtain the desired conclusion we combine (4.11) with (4.12):

$$\begin{aligned}
\mathbb{P}(F_h \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}) &= \int \mathbb{P}(F_h \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) p(\mathbf{d} \,|\, \mathbf{a}, \mathbf{b}, \mathbf{c})\, \mathrm{d}\mathbf{d} \\
&= \int \mathbb{P}(F_h \,|\, \mathbf{b}, \mathbf{c}, \mathbf{d}) p(\mathbf{d} \,|\, \mathbf{b}, \mathbf{c})\, \mathrm{d}\mathbf{d} + o(h) = \mathbb{P}(F_h \,|\, \mathbf{b}, \mathbf{c}) + o(h).
\end{aligned}$$

This completes the proof. ∎

Appendix A. Measure on the space of trajectories. Let $\mathbf{X}$ be a continuous time stochastic process with values in a finite space $\mathcal{S}$. A sample path $\mathbf{x}$ of $\mathbf{X}$ in a finite time interval $[0, u]$ is represented by its space-time

skeleton

$$\begin{pmatrix} 0 & t_1 & \cdots & t_i & \cdots & t_n & u \\ x_0 & x_1 & \cdots & x_i & \cdots & x_n & \end{pmatrix},$$

where n is the number of jumps in the interval, $t_1 < \cdots < t_n < u$ are the times of jumps and $x_i = \mathbf{x}(t_i) \in \mathcal{S}$. The space $\mathcal{X}$ of trajectories can thus be identified with

$$\{0\} \times \mathcal{S} \cup \bigcup_{n=1}^{\infty} \{n\} \times \mathcal{S}^{n+1} \times \Delta_n,$$

where $\Delta_n = \{(t_1, \ldots, t_n) : 0 < t_1 < \cdots < t_n < u\}$ (the term $\{0\} \times \mathcal{S}$ corresponds to constant trajectories). We equip the space $\{n\} \times \mathcal{S}^{n+1} \times \Delta_n$ with the product of counting measure on $\mathcal{S}^{n+1}$ and the Lebesgue measure on Δ_n. The resulting measure on $\mathcal{X}$ is denoted $\mathrm{d}\mathbf{x}$.

If we consider a multivariate process $\mathbf{X} = (\mathbf{X}_v,\ v \in \mathcal{V})$ and $\mathcal{A} \subseteq \mathcal{V}$ then we analogously define the measure $\mathrm{d}\mathbf{x}_{\mathcal{A}}$ on the space of trajectories of the subprocess $\mathbf{X}_{\mathcal{A}}$.

References

[1] N. Ay and D. Polani, *Information flows in causal networks*, Adv. Complex Systems 11 (2008), 17–41.

[2] S. Bongers, P. Forré, J. Peters, and J. M. Mooij, *Foundations of structural causal models with cycles and latent variables*, Ann. Statist. 49 (2021), 2885–2915.

[3] V. Didelez, *Graphical models for composable finite Markov processes*, Scand. J. Statist. 34 (2007), 169–185.

[4] V. Didelez, *Graphical models for marked point processes based on local independence*, J. Roy. Statist. Soc. Ser. B 70 (2008), 245–264.

[5] M. Eichler, *Granger causality and path diagrams for multivariate time series*, J. Econometrics 137 (2007), 334–353.

[6] M. Eichler and V. Didelez, *Causal reasoning in graphical time series models*, in: 23rd Annual Conference on Uncertainty in Artificial Intelligence, 2007, 109–116.

[7] M. Eichler and V. Didelez, *On Granger causality and the effect of interventions in time series*, Lifetime Data Analysis 16 (2010), 3–32.

[8] Y. Fan, J. Xu, and C. R. Shelton, *Importance sampling for continuous time Bayesian networks*, J. Machine Learning Res. 11 (2010), 2115–2140.

[9] C. W. J. Granger, *Investigating causal relations by econometric models and cross-spectral methods*, Econometrica 37 (1969), 424–438.

[10] J. Hammersley and P. Clifford, *Markov fields on finite graphs and lattices*, preprint, 1971.

[11] S. Lauritzen, A. P. Dawid, B. Larsen and H. Leimer, *Independence properties of directed Markov fields*, Networks 20 (1990), 491–505.

[12] B. Miasojedow and W. Niemiro, *Geometric ergodicity of Rao and Teh's algorithm for Markov jump processes and CTBNs*, Electron. J. Statist. 11 (2017), 4629–4648.

[13] S. W. Mogensen and N. R. Hansen, *Markov equivalence of marginalized local independence graphs*, Ann. Statist. 48 (2020), 539–559.

[14] S. W. Mogensen and N. R. Hansen, *Graphical modeling of stochastic processes driven by correlated noise*, Bernoulli 28 (2022), 3023–3050.
[15] R. M. Neal, *On deducing conditional independence from d-separation in causal graphs with feedback (research note)*, J. Artificial Intelligence Res. 12 (2000), 87–91.
[16] W. Niemiro and Ł. Rajkowski, *Local dependence graphs for discrete time processes*, in: Proc. 2nd Conference on Causal Learning and Reasoning, Proc. Machine Learning Res. 213 (2023), 772–790.
[17] U. Nodelman, C. R. Shelton, and D. Koller, *Continuous time Bayesian networks*, in: Proc. 18th Conf. on Uncertainty in Artificial Intelligence, 2002, 378–387.
[18] J. Pearl, *Bayesian networks: A model of self-activated memory for evidential reasoning*, in: Proc. 7th Conf. of the Cognitive Science Society, Univ. of California, Irvine, CA, 1985, 15–17.
[19] J. Pearl, *Causality: Models, Reasoning and Inference*, Cambridge Univ. Press, 2009.
[20] J. Pearl and R. Dechter, *Identifying independencies in causal graphs with feedback*, in: Proc. 12th Internat. Conf. on Uncertainty in Artificial Intelligence, 1996, 420–426.
[21] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*, MIT Press, 2017.
[22] T. Schweder, *Composable Markov processes*, J. Appl. Probab. 7 (1970), 400–410.

Wojciech Niemiro
Institute of Applied Mathematics and Mechanics
University of Warsaw
Warszawa, Poland
E-mail: wniem@mimuw.edu.pl