

Theory-independent limits on correlations from generalized Bayesian networks

Joe Henson^{1,2}, Raymond Lal^{3,4} and Matthew F Pusey⁵

¹ University of Bristol, Department of Physics, HH Wills Physics Laboratory, University of Bristol, Bristol, BS8 1TL, UK

² Imperial College London, Department of Physics, South Kensington Campus, London SW7 2AZ, UK

³ Department of Computer Science, University of Oxford, OX1 3QD, UK

⁴ Faculty of Philosophy, University of Cambridge, CB3 9DA, UK

⁵ Perimeter Institute for Theoretical Physics, 31 Caroline Street North, Waterloo, ON N2L 2Y5, Canada

E-mail: j.henson@bristol.ac.uk, rayl@cs.ox.ac.uk and m@physics.org

Received 15 August 2014

Accepted for publication 25 September 2014

Published 20 November 2014

New Journal of Physics **16** (2014) 113043

doi:[10.1088/1367-2630/16/11/113043](https://doi.org/10.1088/1367-2630/16/11/113043)

Abstract

Bayesian networks provide a powerful tool for reasoning about probabilistic causation, used in many areas of science. They are, however, intrinsically classical. In particular, Bayesian networks naturally yield the Bell inequalities. Inspired by this connection, we generalize the formalism of classical Bayesian networks in order to investigate non-classical correlations in arbitrary causal structures. Our framework of ‘generalized Bayesian networks’ replaces latent variables with the resources of any generalized probabilistic theory, most importantly quantum theory, but also, for example, Popescu–Rohrlich boxes. We obtain three main sets of results. Firstly, we prove that all of the observable conditional independences required by the classical theory also hold in our generalization; to obtain this, we extend the classical d -separation theorem to our setting. Secondly, we find that the theory-independent constraints on probabilities can go beyond these conditional independences. For example we find that no probabilistic theory predicts perfect correlation between three parties using only bipartite common causes. Finally, we begin a classification of those



Content from this work may be used under the terms of the [Creative Commons Attribution 3.0 licence](https://creativecommons.org/licenses/by/3.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

causal structures, such as the Bell scenario, that may yield a separation between classical, quantum and general-probabilistic correlations.

Keywords: quantum nonlocality, Bayesian networks, causal structure

1. Introduction

Bell's theorem [5] is a central result in the foundations of quantum mechanics. It reveals that certain quantum correlations are stronger than those obtainable in any locally causal model as defined by Bell. Recently, new results have been obtained by using variations of the scenario that Bell originally considered. For example, Popescu [31] found that sequences of measurements can reveal nonclassicality in more states than the single measurements considered in a Bell scenario. Branciard, Gisin and Pironio [7] found that including the independence of multiple sources could lead to more robust experiments than the single source assumption of a Bell scenario. Using this idea, Fritz [16] showed that the 'free will' assumption of Bell's theorem can be replaced with an assumption about independence of sources, by replacing the measurement settings of the Bell scenario with additional sources. Finally, Bancal *et al* [3] used an elaborate quadripartite scenario to show that explanations of the violation of Bell inequalities using superluminal but finite speed influences are in conflict with the no-signalling principle.

The common theme in these results is the consideration of more complicated causal structures than the one usually assumed in the Bell scenario. This leads to new insights into how quantum theory deviates from classical physics: by considering arbitrary causal structures, these examples expose a rich structure to quantum correlations. However, to clarify and unify these results, it would be helpful to have a *general* framework that formalises the connection between causal structure and observable correlations. There are two desirable features that a general framework of this kind should have. Firstly, it should describe constraints on locally causal models (i.e. defined using classical random variables), for arbitrary causal structures, e.g. it should generalize Bell inequalities. Secondly, it should also allow for non-classical resources—not only of quantum mechanics, but also those of *generalized probabilistic theories* (GPTs). The development of GPTs originates in the fact that, in the Bell scenario, quantum theory cannot achieve the strongest correlations that are consistent with the no-signalling principle [12, 32]. It would be interesting to understand the consequences of different types of causal structure for the separation between classical, quantum, and more general correlations. In particular, this would allow us to pose the question of what is special about quantum correlations in a wider framework than has so far been used.

A framework that achieves the first objective is that of *Bayesian networks*, based on directed acyclic graphs (DAGs). This has been an active area of research by statisticians and computer scientists for several decades, pioneered in particular by Pearl [27, 29]. When this framework is applied to a Bell-type experiment, and the causal structure implied by special relativity and independence of settings is assumed, one obtains exactly Bell's notion of local causality [38]. The significance of this is two-fold: firstly, Bayesian networks are the natural setting for generalising Bell scenarios; secondly, a new formalism—but structurally similar to Bayesian networks—will be needed to describe the behaviour of quantum theory and other GPTs on arbitrary causal structures.

Our contribution. In this paper, we propose a generalization of Bayesian networks which incorporates the framework of GPTs. In particular, we generalize the latent nodes of standard Bayesian networks to allow for resources from an arbitrary GPTs. We then investigate the extent to which results from the causality literature generalize to our approach. We have three main results.

Our first result shows that all the observable conditional independences that follow from a classical Bayesian network still follow in our generalization. The conditional independences mandated by a DAG are characterized graphically by the ‘*d*-separation criterion’. Technically our result is that this criterion is still sound in our generalization. Since our framework goes beyond classical probability theory, we do not have enough structure to even define conditional independences involving latent nodes; hence we require a proof that is very different to the classical case.

Secondly, we also explore what constraints further than the observable conditional independences can be derived for a given causal structure, even in the most general theories. In the case of classical Bayesian networks, all constraints on probability distributions implied by the causal structure are (by definition) conditional independences. However, these conditional independences may involve ‘latent’ variables, which are unobserved. Hence not all of the constraints on observable variables need to take the form of *observable* conditional independences. For example, in the Bell setup, Bell inequalities are constraints on the observable variables that arise from the existence of latent variables. But Bell inequalities are stronger constraints than the observable conditional independences, i.e. the no-signalling conditions.

Since our approach will be to allow arbitrary GPTs, the Bell inequalities in the Bell scenario will not constrain the observable probabilities in a general theory. However, we examine two other quantitative limits on classical correlations that apply to different causal structures. As in the Bell inequality case, these limits do not follow from the observable conditional independences. Nevertheless, we find that both limits do carry over to arbitrary GPTs. Specifically, we show that perfect correlation between three parties cannot be explained by bipartite common causes alone, regardless of which physical theory is used. We also show that any GPT obeys the ‘instrumental inequality’, a close cousin of the Bell inequalities that applies to a simple four-node DAG.

Finally we identify an important classification problem: which are the causal structures that, even classically, have no observable consequences beyond conditional independences? Structures not in this class will certainly be the focus of attention in quantum foundations, but we believe this classification will be of interest in other applications of even the classical causality framework. We make progress on this problem by providing a sufficient condition for our generalized DAG to imply only the observable conditional independences.

Related work. Our work extends Pearl’s research programme [27] to the study of nonlocality. In this respect we build upon the work of Wood and Spekkens [38], who showed that such a connection can be made. Part of our work also builds upon the circuit framework developed by Chiribella, D’Ariano and Perinotti (CDP) [11]. There are several other lines of investigation with similar but distinct aims to ours. Leifer and Spekkens have the ambitious aim of an inherently quantum theory of Bayesian networks [23]. However the Leifer–Spekkens approach is work in progress, and is unlikely to allow for other general probabilistic theories. Fritz has generalized the definitions of classical, quantum, and GPTs correlations beyond the Bell scenario, and provided many interesting examples [16, 17]. But he does not aim to

generalize the standard theory of Bayesian networks directly, and so not all of our results can be translated to his definitions. In appendix A we discuss the connections to these works in more detail. Related work has meanwhile appeared in [9, 30], the latter including extensions of some of our results.

Plan of paper. In section 2 we introduce the background on classical Bayesian networks, in particular the classical d -separation theorem. In section 3 we discuss parts of the CDP circuit framework, which we then build upon to define ‘generalized Bayesian networks’. We then prove the d -separation theorem for our framework. In section 4 we investigate bounds on correlations for the triangle and instrumental inequality scenarios. Finally, in section 5 we provide a sufficient condition on a causal structure for all sets of correlations to be equal.

2. Classical Bayesian networks

We often have reasons to assume a given set of causal relations between random variables. A basic example is the Bell scenario [5], in which we consider probability distributions $P(a, b|x, y)$. The underlying spatio-temporal relations are assumed to constrain these distributions by conditional independences known as the ‘no-signalling’ conditions, e.g. $P(a|x, y) = P(a|x)$. Bell’s locality condition places a further restriction on the possible correlations

$$P(a, b|x, y) = \sum_{\lambda} P(a|x, \lambda)P(b|y, \lambda)P(\lambda). \quad (1)$$

Now, the locality condition can be understood as a condition on the background causal structure, stating that the correlations in $P(a, b|x, y)$ arise through a common cause—a classical random variable λ —that is in the past of both Alice and Bob. Bell inequalities then characterize the correlations that are compatible with this causal structure⁶.

In general, how do we characterize the set of allowed probability distributions given a certain causal structure? In the case where we only consider causal relations between classical random variables, this question is answered by the theory of Bayesian networks. This theory provides a way to describe causal structures, along with rules to determine which probability assignments are consistent with them. Here we provide a brief introduction to this aspect of Bayesian networks, with a view to its generalization in subsequent sections. We largely follow Pearl’s terminology and notation [27].

2.1. Probabilities on graphs

Recall that a *directed graph* G is a pair (V, E) , where V is a set of nodes, and $E \subseteq V \times V$ is a set of directed edges. It is often useful to label the nodes with an index, so that we can write $V = \{X^{(i)}\}_i$. A directed graph may have a *directed cycle*, viz. a sequence of edges $X^{(1)} \rightarrow X^{(2)} \rightarrow \dots \rightarrow X^{(n)} \rightarrow X^{(1)}$. A *DAG* is a directed graph which has no directed cycles.

⁶ At this point one might question the physical motivations for assuming a particular causal structure, especially with regard to the spatio-temporal causal order that is so crucial to the discussion on Bell’s theorem and its consequences. While much could be said on this issue, the main intention here is to discuss the consequences of assuming a causal structure, rather than the many possible motivations for doing so.

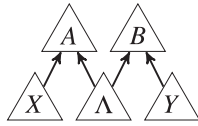


Figure 1. The Bell scenario depicted as a DAG, with hidden variable Λ .

In our work, DAGs will represent causal structure: more specifically, an edge $X \rightarrow Y$ will represent the possibility of direct causal influence from X to Y , where ‘direct causal influence’ will be defined in terms of probabilistic conditional dependence. The nodes that can directly influence Y are all nodes X for which there is an edge $X \rightarrow Y$; these are the *parents* of Y , and the set of all parents of Y is denoted $\text{PA } Y$. Similarly, if $X \rightarrow Y$ then Y is a *child* of X . A *directed path* is a sequence of nodes $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ such that $X^{(i)} \rightarrow X^{(i+1)}$ for $1 \leq i \leq n-1$. In keeping with familial terminology, we say that Y is a *descendant* of X , and X is an *ancestor* of Y , if there is a directed path from X to Y . We also define the following two useful functions on sets of nodes:

- (i) we define $m(U)$ to be the union of the set of nodes U with all the children of each of the nodes in U ;
- (ii) we define $J^-(U)$ to be the union of U with the set of all ancestors of nodes in U (the entire ‘past’ of U).

Now, consider the Bell scenario in which a common cause is assumed to exist. The DAG for this scenario is shown in figure 1, where the A and B nodes represent experiment outcomes in the two wings of the experiment, X and Y are the respective settings, and Λ is the common cause (i.e. the ‘hidden variable’). Writing down such a DAG incorporates various causal assumptions, for example: (i) that the settings are ‘free’, e.g. there are no edges $\Lambda \rightarrow X$ or $\Lambda \rightarrow Y$; and (ii) that the two wings are causally disconnected from each other (which could arise from spacelike separation between Alice and Bob), e.g. there is no edge $X \rightarrow B$.

Let us now consider random variables associated to the nodes of the DAG. Only certain probability distributions will be consistent with the causal structure, if it is to have the intended meaning. As in other treatments, $X^{(1)}$ will denote a random variable, while $x^{(1)}$ denotes the value of the random variable, and the same label will also be used for the node in the graph associated to this variable (it will be clear from the context which is meant). Sometimes capital letters will also be used to signify sets of random variables, and the lowercase letter a value for each of these variables. The basic objects of interest will be probability distributions over all the nodes, $P(g)$. It is convenient to extend this notation to the parents in the following way:

- $\text{PA } X^{(i)}$ is the set of random variables associated to the parents of the node $X^{(i)}$;
- $\text{pa } x^{(i)}$ denotes a values of the random variables $\text{PA } X^{(i)}$.

The notion of causality that we now apply has several equivalent forms [27]. Perhaps the most intuitive is that given a random variable X , once direct causal influence of the parents has been taken into account by conditioning, then X should be independent of every other node, except for its descendants. For our purposes the following form is the most suggestive:

Definition 1. (Markov condition). Let G be a DAG. A probability distribution P is *Markov relative to G* if P satisfies

$$P(x^{(1)}, \dots, x^{(n)}) = \prod_i P(x^{(i)} | \text{pa } x^{(i)}).$$

A simple example is given by a probability distribution P that is Markov with respect to the chain $X \rightarrow Y \rightarrow Z$: this means that Y ‘screens off’ the influence of X from Z , i.e. $P(z|x, y) = P(z|y)$.

Definition 2. A (classical) Bayesian network is a pair (P, G) , where G is a DAG, and P is a probability distribution that is Markov relative to G .

Often, only a subset of the nodes in a Bayesian network represent observable outcomes. These are called *observed* nodes, whereas the other nodes are referred to as *latent* or *hidden* nodes. Latent nodes are usually added by hypothesis in an attempt to explain observed correlations.

We can describe Bell’s theorem in this language [38]. If P is Markov relative to the DAG in figure 1 then

$$P(a, b, x, y, \lambda) = P(a|x, \lambda)P(b|y, \lambda)P(x)P(y)P(\lambda).$$

After marginalizing over λ , and dividing through by $P(x)P(y)$, we obtain Bell’s locality condition, i.e. equation (1). Hence we see that: (i) the idea of a hidden variable λ is identical to the existence of a latent node; (ii) Bell’s locality condition follows from the Markov condition for the Bell DAG. In this way, we can see that Bell applied the same basic account of causality as used in Bayesian networks, albeit applied to a particularly simple and intuitive case. For more complex DAGs, the more general framework is needed.

2.2. A graphical criterion for independence: *d*-separation

A Bayesian network specifies a graph and a probability distribution that decomposes ‘locally’ along the edges of the graph. This means that it encodes certain conditional independences. But in general, further independences will be derivable from those given directly by the fact that P is Markov with respect to G . For example, in the Bell DAG, the Markov condition immediately implies that $P(a|x, \lambda, y) = P(a|x, \lambda)$ (sometimes called ‘parameter-independence’ [33]). But the probability calculus also implies that we can marginalize over λ to obtain $P(a|x, y) = P(a|x)$, i.e. the no-signalling condition. In the theory of Bayesian networks, these additional conditional independences are of paramount importance. Clearly they follow from the structure of the graph alone, but deriving them using probability theory can be impractical, especially in more complicated DAGs. The condition of *d*-separation, developed by Geiger [18] and Verma and Pearl [37], provides a way to ‘read off’ these conditional independences from the structure of the graph.

To gain an intuitive understanding of the *d*-separation condition, let us consider the connected Bayesian networks that have three nodes, X , Y and Z , and two edges. There are three such networks:

- (i) the *chain* $X \rightarrow Z \rightarrow Y$;
- (ii) the *fork* $X \leftarrow Z \rightarrow Y$; and

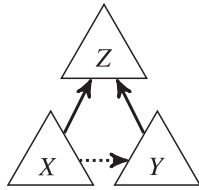


Figure 2. A pseudo-path for the collider.

(iii) the *collider* $X \rightarrow Z \leftarrow Y$.

We can consider whether $P(x, y|z) = P(x|z)P(y|z)$ holds in each of these cases, denoted $X \perp\!\!\!\perp Y \mid Z$. For the chain and fork, it is immediate that X and Y are conditionally independent given Z in any Markov probability distribution (but need not satisfy marginal independence $p(x, y) = p(x)p(y)$). However, in the collider we may not have $X \perp\!\!\!\perp Y \mid Z$, even though X and Y are now marginally independent. For example, Z could hold the value 1 when $x = y$, and 0 otherwise. The same prevention of conditional independence may be caused by conditioning on any node in the mutual future of X and Y in a more general DAG. Roughly speaking, these observations show that, for sets of nodes X , Y and Z , conditional independences $X \perp\!\!\!\perp Y \mid Z$ will follow when Z *contains* the middle node of chains and forks, but *excludes* the middle node of colliders.

We shall use the form of d -separation originally developed by Lauritzen *et al* [22]. Let G be a DAG with disjoint subsets X , Y and Z . Then we define the set $W := G \setminus J^-(X \cup Y \cup Z)$. In words, the set W is every node in G that is not in the inclusive past of any node in X , Y or Z . Now define a *pseudo-path* from node $P^{(1)}$ to node $P^{(p)}$ to be a sequence of nodes $(P^{(1)}, P^{(2)}, \dots, P^{(p)})$ such that, for all $i \in \{1, \dots, p\}$, $P^{(i)} \notin W$, and $m(P^{(i)}) \cap m(P^{(i+1)}) \not\subseteq W$. That is, a pseudo-path does not intersect W , and two sequential elements in a pseudo-path must be adjacent or share a common child that is not in W .

Definition 3. Let G be a DAG G with disjoint subsets X , Y and Z . We say that X and Y are d -separated by Z , written $X \perp\!\!\!\perp Y \mid Z$, if, for all nodes $A \in X$ and $B \in Y$, all pseudo-paths from A to B are non-trivially intersected by Z .

Example 4. (d -separation). As we would expect, for the chain and fork we have $X \perp\!\!\!\perp Y \mid Z$, but this fails for the collider. Consider the dotted line in figure 2. This is a pseudo-path, since W is the empty set in this DAG, and the path has only two sequential elements, with Z as the common child. However this pseudo-path does not intersect Z , and hence $X \perp\!\!\!\perp Y \mid Z$ fails to hold.

The following theorem establishes the link between the d -separation condition and conditional independence.

Theorem 5. (Verma and Pearl [37], Meek [25]). Let G be a DAG with disjoint subsets X, Y and Z . Then:

- (i) If P is Markov with respect to G , then $X \perp\!\!\!\perp Y \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z$.
- (ii) If $X \perp\!\!\!\perp Y \mid Z$ holds for all P which are Markov with respect to G , then $X \perp\!\!\!\perp Y \mid Z$.

Item (i) says that d -separation is a sound criterion for conditional independence, and item (ii) says that d -separation is complete, i.e. *all* robust conditional independences arise through applying the d -separation condition to the underlying DAG. Theorem 5 is of central importance to classical Bayesian networks. For example, many algorithms for causal inference rely exclusively on conditional independences [27].

Example 6. (Conditional independences in the Bell scenario). Consider again the Bell DAG figure 1. We can use the d -separation theorem to derive the usual conditional independences, i.e. the no-signalling conditions. For example, we have $A \perp Y \mid X$, which implies $P(a|x, y) = P(a|x)$. We obtain $A \perp Y \mid X$ as follows. We have $W = \{B\}$. But consider the sequences of nodes between A and Y (the candidate pseudo-paths). For example, $p_1 := (Y, B, \Lambda, A)$ and $p_2 := (Y, \Lambda, A)$ are two such sequences. But p_1 intersects W , and p_2 has a pair sequential elements (Λ, A) that share a common child in W . Similarly, all sequences of nodes between A and Y fail to be pseudo-paths, and hence $A \perp Y \mid \emptyset$ ⁷.

3. Generalized Bayesian networks

We will now extend the definitions of Bayesian networks to go beyond classical theories. This will serve as a framework within which to discuss the differences between the allowed set of probability distributions in classical and quantum systems, and even more general cases. To do this we shall build on the circuit framework for general probabilistic theories that was developed by CDP [11]. This provides a graphical approach which is useful when considering DAGs, and their framework imposes very minimal requirements on the theories it encompasses. We describe this CDP framework in section 3.1. We then introduce our definition of generalized Bayesian networks in section 3.2, after which, in section 3.3, we prove that the d -separation criterion can be extended to generalized Bayesian networks.

3.1. The CDP framework

The CDP framework provides an abstract description of ‘circuits’ consisting of operations (which include preparations, transformations and observations) connected by propagating systems. These will be used to describe sources of general correlations in our generalised Bayesian networks. First, the way in which elements of the circuits compose will be specified (the ‘operational’ part); then the way in which probabilities are attached to circuits will be described. Together these parts constitute what CDP call an *operational-probabilistic theory*.

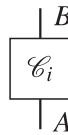
3.1.1. The operational part. To specify the operational part, we consider a collection of named systems A, B, C, \dots , including a *trivial* system I . Systems are the inputs and outputs of *tests* $\{\mathcal{C}_i\}_{i \in X}$, which represent a single use of some physical device, e.g. a Stern–Gerlach device. To prevent the input of a test being its own output, the input and output systems of a test must be

⁷ Note that the Bell DAG here encodes the assumption that the inputs are uncorrelated, i.e. $P(x, y) = P(x)P(y)$. Hence we obtain a ‘no-signalling’ condition that is stronger than the usual one considered for nonlocality in the Bell setup. That is, we obtain $P(a|y) = P(a)$ as well as $P(a|x, y) = P(a|x)$. To allow for the possibility that the inputs are correlated, we would use a different DAG, with extra edges $U \rightarrow X$ and $U \rightarrow Y$, where U represents a correlating variable. With this DAG, we obtain only $P(a|x, y) = P(a|x)$, without $P(a|y) = P(a)$, as expected.

distinct, except when both are trivial. The elements of tests, \mathcal{C}_i , represent operationally distinguishable outcomes of the test. They are referred to as *events*, and are indexed by a finite number of outcomes $i \in X$.

For example, for the test corresponding to the use of a Stern–Gerlach device with a spin-half particle, the outcome set would have two elements, corresponding to the two different spin outcomes. If a test $\{\mathcal{C}_i\}_{i \in X}$ is a singleton, i.e. if there is only one outcome $i = i_0$, then we say that this is a *deterministic test*.

Below we will find it useful to explicitly include the input and output systems in our notation, so an event with input system A and output B will be represented as \mathcal{C}_{iA}^B . The trivial system will not be included explicitly⁸. CDP use a graphical notation that builds upon that of Abramsky and Coecke [1]. A test $\{\mathcal{C}_i\}$ with input system A and output system B is depicted as:



If the input of a test is the trivial system then it is depicted as

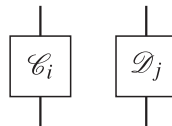


and referred to as a *preparation-test*. *Observation-tests* are the dual notion, for which the output is the trivial system. From now on, we shall omit labelling the systems in the graphical notation.

When the output system of $\{\mathcal{C}_i\}$ is the same as the input system of $\{\mathcal{D}_j\}$, they are composed *in sequence*, depicted as



or symbolically as $\mathcal{C}_{iA}^B \mathcal{D}_{jB}^C$. Otherwise they are composed *in parallel*:



or $\mathcal{C}_{iA}^B \mathcal{D}_{jC}^D$. Each type of composition yields another test, whose outcomes (i, j) are ordered pairs formed by the outcomes i and j of each factor.

If \mathcal{C}_i has input system A and output B , and \mathcal{D}_j has input C and output D , then their parallel composition has the *composite systems*, AC and BD , as inputs and outputs respectively. ‘Composite system’ is a primitive notion for CDP, assumed to satisfy certain basic requirements, and so it is not defined with respect to any other mathematical structure.

⁸ This mimics the use of tensorial notation by Hardy [19, 20].

3.1.2. The probabilistic part. An operational-probabilistic theory is defined as one in which every test from the trivial system to itself (pictorially, a diagram with no input or output wires) is a probability distribution over the outcome set, and where the composition of such tests is given by the corresponding product distribution.

Two tests are called operationally equivalent if substituting one for the other never affects a probability distribution. An operationally equivalent class of observation-events is called an *effect*.

To complete this framework we shall assume the existence of a *unique* deterministic effect \top_A for each system A . Graphically we denote this as:



This assumption is referred to by CDP as *causality*. In particular, ignoring the outcome of any observation-test always corresponds to this unique deterministic effect. This assumption is necessary for the comparison to Bayesian networks below to make sense: CDP show that it is equivalent to the assumption that the probability of an outcome at time t_1 does not depend on which operation is performed at time t_2 , where $t_2 > t_1$. Hence the causality assumption can also be thought of as ‘no-signalling from the future to the past’.

The fact that the deterministic effect is unique trivially implies the following result, which we will use below.

Lemma 7. *The deterministic effect on a composite system AB is equal to the parallel composition of the deterministic effect on A with the deterministic effect on B , or $\top_{AB} = \top_A \top_B$.*

We can now give some examples of causal operational-probabilistic theories.

Example 8. (Quantum theory). Quantum theory will be our main example of an operational-probabilistic theory. Systems A, B, C, \dots are associated to complex Hilbert spaces $\mathcal{H}_A, \mathcal{H}_B, \mathcal{H}_C, \dots$; and in particular, the trivial system is given by the one-dimensional space $\mathcal{H}_I = \mathbb{C}$. Composite systems are given by the vector space tensor product.

Tests are quantum instruments, i.e. sets of completely positive linear maps that sum to a trace preserving map. In particular, deterministic preparation-tests are unit trace positive operators, and observation-tests are of the form $\text{Tr}(E_i \cdot)$, where $\{E_i\}$ is a POVM. Tests compose in sequence by ordinary composition of maps, and in parallel by the vector tensor product. The unique deterministic effect is Tr .

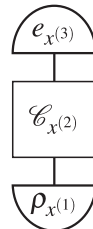
Example 9. (Boxworld). *Boxworld* [4] is a theory defined to produce the maximal violation [32] of the CHSH inequality. The simplest type of system, called a *gbit*, comes with a pair of two-outcome observation-tests $\{e_1, e_2\}$ and $\{f_1, f_2\}$. For any pair of probabilities p_e and p_f there is exactly one deterministic preparation-test ω with $e_1(\omega) = p_e$ and $f_1(\omega) = p_f$. Composite systems get the parallel compositions of these, and there is then a unique deterministic preparation-test for any no-signalling distribution on the outcomes. Subject to these requirements, every other mathematically consistent test is included.

Example 10. (Classical probability theory). We obtain a *classical operational-probabilistic theory* by associating systems A, B, C, \dots with sets $\Lambda_A, \Lambda_B, \Lambda_C$, the trivial system having $\Lambda_I = \{\emptyset\}$. Composite systems are given by the Cartesian product.

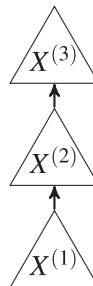
Tests with outcome i from a system A to a system B are given by $p(i, \lambda_B | \lambda_A) \geq 0$, with λ_A and λ_B ranging over Λ_A and Λ_B respectively, and $\sum_{i, \lambda_B} p(i, \lambda_B | \lambda_A) = 1$. Tests compose in sequence by multiplying and the summing over the λ for the intermediate system, and in parallel by multiplying. The unique deterministic effect is $p(\emptyset | \lambda) = 1$.

A natural question is whether a classical operational-probabilistic theory is, in fact, a Bayesian network. However, there are two reasons why this is not the case:

1. There is no classical conditioning in an operational-probabilistic theory. That is, a test $\{\mathcal{C}_i\}_i$ should be thought of as a device with an output indicating which classical outcome, e.g. a light that flashes red or green depending on whether spin up or down is detected. However, in general a physical device will have ‘dials’, which can be used to control which operation will take place. This corresponds to allowing a test $\{\mathcal{C}_i\}_i$ to be a function of a classical input. Indeed, this is how the Bell setup is usually conceived, since Alice and Bob each have two possible measurements (observation-tests), and these measurements are chosen based on their input choice, which can be represented as a binary classical random variable.
2. An operational-probabilistic theory carries *two* types of information in each circuit element: the systems that ‘travel’ along the wires, and the classical outputs. The outcome of a test need not tell us everything about the test’s output state, even when the relevant system is classical. Hence a direct interpretation as a DAG, with the outputs of a test translated as the random variables on a node, can easily violate the Markov condition by failing to condition on all the relevant classical information carried by the system. For example, consider the following sequence of tests where each system is classical



For example, suppose that ρ is the preparation of a coin, which can have either heads or tails facing up, and can be black or white. The test \mathcal{C} could change the colour of the coin, but for simplicity let us suppose that each outcome leaves the state of the coin unchanged. The classical outputs are as follows: $x^{(1)}$ is a bit representing the colour of the coin at t_1 , $x^{(2)}$ is a bit representing the face of the coin at t_2 , and $x^{(3)}$ is a bit representing the colour of the coin at t_3 . This yields a classical probability distribution $P(x^{(1)}, x^{(2)}, x^{(3)})$. Now, suppose that we try to interpret this circuit as a classical Bayesian network



The Markov condition implies that $X^{(3)} \perp\!\!\!\perp X^{(1)} \mid X^{(2)}$. But if ρ is the preparation of a coin

with either side facing up, and in each colour with uniform probability, then $X^{(3)}$ is perfectly correlated with $X^{(1)}$, even conditioning on $X^{(2)}$. Hence the Markov condition fails to hold.

In the next subsection we shall connect DAGs with GPTs more carefully, overcoming these two problems.

3.2. Definition of generalized Bayesian networks and examples

Our aim in this subsection will be to generalize Bayesian networks in a way that can allow non-classical resources. We begin by splitting the nodes into two types.

Definition 11. Let G be a DAG with nodes $V = \{X^{(1)}, X^{(2)}, \dots, X^{(m)}\}$. We shall say that G is a *generalized DAG (GDAG)* if V can be partitioned into two sets of nodes:

1. the *observed nodes* $\{X^{(1)}, X^{(2)}, \dots, X^{(n)}\}$ (drawn as triangles), and
2. the *unobserved nodes* $\{X^{(n+1)}, \dots, X^{(m)}\}$ (drawn as circles).

We choose this terminology because all classical data, e.g. the outcomes of measurements, will be associated to observed nodes. On the other hand, the unobserved nodes will replace ‘latent’ random variables with ‘general resources’, e.g. replacing the source λ in the Bell DAG with a general node will allow Alice and Bob to share a quantum state or the state corresponding to a PR box.

We will often apply DAG terminology (parents, children, d -separation, etc) to GDAGs. Unless specified otherwise, the relevant definition should simply be applied to the underlying DAG (i.e. ignoring the distinction between observed and unobserved nodes).

We shall assign CDP tests to each node, and hence we shall use the CDP framework. However, in the previous subsection, we discussed that a circuit element in the CDP framework carries *two* types of data: the classical data associated with an outcome, and the system. In example 10 we noted that this makes it problematic to interpret a CDP circuit as a Bayesian network. Our framework will address this problem by using generalized DAGs. In particular we shall define the *outputs* of observed and unobserved nodes in distinct ways:

1. **Observed nodes:** each observed node will map to a test with no outgoing wires, but will have a classical random variable X assigned to it. In the CDP language, an observed node’s test has the trivial system as output. Where there is an outgoing edge from an observed node, this means there is a choice of test to be performed at the child node, which depends on the value of the classical variable at the parent. CDP call this a ‘conditioned test’ and show that causality is equivalent to them being well defined.
2. **Unobserved nodes:** on the other hand, each unobserved node will output *only* systems, and will not have any non-trivial outcomes assigned to it. For convenience of notation we shall associate a classical random variable with every node⁹ $X^{(i)}$. However, the random variable associated with unobserved nodes will be trivial, taking only one value with probability one.

⁹ As is the case for classical Bayesian networks, we shall use the same symbol $X^{(i)}$ to denote both the node and the random variable associated with the node; context will determine which is being referred to.

Accordingly, we shall associate a non-trivial probability distribution $P(x^{(1)}, x^{(2)}, \dots, x^{(n)})$ only with the observed nodes.

More formally, we have:

Definition 12. Let G be a generalized DAG. Call an edge of G *observed* if it begins on an observed node, and *unobserved* if it begins on an unobserved node.

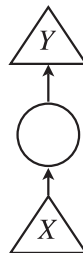
Definition 13. Let G be a generalized DAG with m nodes, of which the first n are observed. A probability distribution P over the observed nodes is *generalized Markov* with respect to G if there exists:

1. a causal operational-probabilistic theory;
2. for every unobserved edge, a distinct system in the theory; and
3. for every node $X^{(i)}$, and every value $\text{opa } x^{(i)}$ of its observed parents, a test $\mathcal{T}_{x^{(i)}}(\text{opa } x^{(i)})_{\text{incU } X^{(i)}}^{\text{outU } X^{(i)}}$ from the composite system $\text{incU } X^{(i)}$ formed by the systems on $X^{(i)}$'s incoming unobserved edges to the composite system $\text{outU } X^{(i)}$ formed by the systems on its outgoing unobserved edges, with
 - (a) an outcome set matching $X^{(i)}$ in the case of an observed node, but
 - (b) a 1-element outcome set in the case of an unobserved node¹⁰ such that

$$P(x^{(1)}, x^{(2)}, \dots, x^{(n)}) = \prod_{i=1}^m \mathcal{T}_{x^{(i)}}(\text{opa } x^{(i)})_{\text{incU } X^{(i)}}^{\text{outU } X^{(i)}}.$$

We say that the *generalized Markov condition (GMC)* is satisfied by a probability distribution P if it is generalized Markov with respect to a given GDAG G .

Example 14. (Prepare and measure). A randomly chosen preparation followed by a fixed measurement can be depicted as



The first node, X , has no incoming edges. Since it is observed, the corresponding test also has no outgoing systems. Hence it corresponds to a test from the trivial system to itself, i.e. a probability distribution p_x . The unobserved node has an incoming edge from X and hence the corresponding test will depend on x . It has one outgoing edge and so the test has a single

¹⁰ Although it is not required for our results, it would be nice if P was independent of how the GDAG is described, in particular of how the incoming and outgoing edges of a node are ordered. See [17] for a sketch proof that should carry over to our setting.

outgoing system, i.e. it is a preparation-test ρ_x for a single system. Finally, the last node corresponds to a test that receives the system from ρ_x and has no outgoing systems, i.e. it is an observation-test $\{e_y\}$. Overall we have

$$P(x, y) = \frac{\text{Diagram: a circle with a semi-circle on top labeled } e_y \text{ and a semi-circle on the bottom labeled } \rho(x)}{p_x}$$

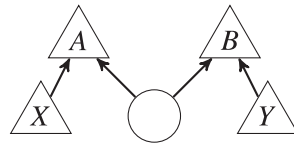
To interpret this diagram it is useful to recall that the composition of two tests from the trivial system to itself is simply multiplication of probability distributions. Hence $P(x, y) = P(y|x)p_x$ where

$$P(y|x) = \frac{\text{Diagram: a circle with a semi-circle on top labeled } e_y \text{ and a semi-circle on the bottom labeled } \rho(x)}{p_x}$$

Definition 15. A *generalized Bayesian network* is a pair (P, G) , such that G is a generalized DAG, and P is generalized Markov with respect to G .

The definition of a generalized Bayesian network is therefore exactly analogous to that of a classical Bayesian network.

Example 16. (Bell setup). We can define a generalized Bayesian network corresponding to the Bell scenario as follows



A probability distribution P that is Markov for this generalized DAG is given by

$$P(a, b, x, y) = \frac{\text{Diagram: a circle with two semi-circles on top labeled } e_a(x) \text{ and } f_b(y), \text{ and a semi-circle on the bottom labeled } \rho}{p_x p_y}$$

In the special case that the operational theory under consideration is quantum theory, this gives

$$P(a, b, x, y) = \text{Tr}((E_a(x) \otimes E_b(y))\rho)p_x p_y,$$

where ρ is a bipartite state and $\{E_a(x)\}$ and $\{E_b(y)\}$ are POVMs for each x, y . This is indeed the standard quantum model of a Bell experiment. This example also illustrates that our formalism describes the classical control of tests as a parameterized family of CDP circuits.

A generalized Bayesian network will allow us to explore the consequences of using non-classical resources in place of classical latent variables. However, we recover classical Bayesian networks if we do not include any unobserved nodes.

Proposition 17. *If all nodes are observed, then a generalized Bayesian network is a classical Bayesian network.*

Proof. If all nodes are observed, then for every node X , we have $\text{incU } X = \text{outU } X = I$. That is, the incoming and outgoing systems of every test are trivial. Then definition 15 becomes

$$P(x^{(1)}, x^{(2)}, \dots, x^{(n)}) = \prod_{i=1}^m \mathcal{T}_{x^{(i)}}(\text{opa } x^{(i)}).$$

For every value of the parents $\text{pa } x^{(i)} = \text{opa } x^{(i)}$, the event $\mathcal{T}_{x^{(i)}}(\text{opa } x^{(i)})$ is a test from I to I and hence a probability distribution on $x^{(i)}$. We can then define

$$P(x^{(i)} | \text{pa } x^{(i)}) := \mathcal{T}_{x^{(i)}}(\text{opa } x^{(i)}),$$

giving a set of conditional probabilities, which, since the composition of tests from I to I is just multiplication, satisfies definition 1. \square

For a given GDAG G , we can identify the following sets of probabilities that are generalized Markov with respect to G :

1. The set \mathcal{G} of probabilities that are generalized Markov for any operational theory.
2. The set \mathcal{Q} of probabilities that are generalized Markov for quantum theory.
3. The set \mathcal{C} of probabilities that are generalized Markov for classical probability theory.

Since classical probability theory can be embedded into quantum theory by using diagonal operators, we have $\mathcal{C} \subseteq \mathcal{Q} \subseteq \mathcal{G}$ for all GDAGs.

\mathcal{C} is closely related to the standard Markov condition on DAGs, with our distinction between observed and unobserved nodes becoming the distinction between observed and latent variables. This is a second sense in which the GMC generalizes the usual Markov condition:

Lemma 18. *Let (P, G) be a generalized Bayesian network with $P \in \mathcal{C}$. Then there exists a classical Bayesian network (P', G') , where G' is the underlying DAG for G , and P and P' agree on the observed nodes defined by G .*

Proof. From definition 15 and example 10, if a generalized Bayesian network (P, G) has $P \in \mathcal{C}$, then each node $X^{(i)}$ has associated to it a probability distribution $p(x^{(i)}, \lambda_{\text{outU } X^{(i)}} | \lambda_{\text{incU } X^{(i)}}, \text{opa } x^{(i)})$, where $x^{(i)}$ is the output, $\lambda_{\text{incU } X^{(i)}}$ is the classical state associated to the incoming edges from unobserved nodes, $\lambda_{\text{outU } X^{(i)}}$ is the classical state associated to the outgoing edges if the node is unobserved (and is trivial otherwise), and $\text{opa } x^{(i)}$ is the output of the observed parents. In this case, we can define a classical random variable $Y^{(i)}$, with values referred to as $y^{(i)}$, for each node: for observed nodes this is simply the output random variable, so that $y^{(i)} := x^{(i)}$, whereas for unobserved nodes it ranges over the classical states on the set of all the outgoing edges, so that $y^{(i)} := \lambda_{\text{outU } X^{(i)}}$. We can now define a probability distribution $P'(y^{(i)} | \text{pa } y^{(i)})$ from $p(x^{(i)}, \lambda_{\text{outU } X^{(i)}} | \lambda_{\text{incU } X^{(i)}}, \text{opa } x^{(i)})$ in the obvious way. This implies that P' is Markov with respect to the underlying DAG of G . Hence we obtain a classical Bayesian network (P', G') , and P' agrees with P on the observed nodes of G by construction. \square

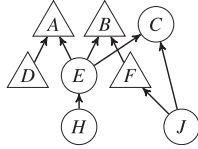


Figure 3. An example for d -separation.

Since classical operational-probabilistic theory is defined using a canonical observation-test, and we only consider tests with a finite number of outcomes, C corresponds to classical probability distributions where all variables, including latent ones, are finite. The results of [17] would suggest that this gives observable probability distributions that are dense in the set that includes infinite-valued latent variables. However, it is very much an open question whether or not these sets are in fact equal, although this is known to be the case in the Bell scenario [14].

Finally, we will use \mathcal{I} to denote the set of probabilities that satisfy all of the observable conditional independences that follow from d -separation. In this notation, the first part of theorem 5 (along with lemma 18) gives $\mathcal{C} \subseteq \mathcal{I}$ for all GDAGs. We will now strengthen this to $\mathcal{G} \subseteq \mathcal{I}$.

3.3. Extending d -separation to generalized Bayesian networks

In GPTs, no-signalling is still valid. Therefore it is to be expected that a generalization of theorem 5, when applied to three disjoint subsets of observed nodes, should obtain. However, the standard proofs of the soundness part of theorem 5 (i.e., item (i)) make use of conditioning on latent variables, the analogue of which is unclear in the general case¹¹. However, by reformulating d -separation before proving the generalization, an alternative proof can be found that does not rely on conditioning on latent variables, and as a result can be more easily generalized.

Lemma 19. (Proof in appendix B). Let G be a DAG with disjoint subsets X , Y and Z , and let $W = G \setminus J^-(X \cup Y \cup Z)$. Then X and Y are d -separated by Z if and only if there exist sets of nodes U and V such that $\{U, V, Z, W\}$ is a partition of G , and

$$X \subseteq U, Y \subseteq V, \quad (2)$$

$$m(U) \cap m(V) \subseteq W. \quad (3)$$

3.4. An example

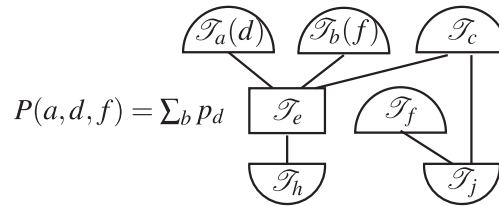
We seek a generalization of theorem 5 from classical to generalized Bayesian networks. The following example is intended to clarify why this is reasonable, and also to elucidate the proof.

Consider the GDAG depicted in figure 3. This is the Bell GDAG with three extra unobserved nodes, C , J and H , added. Intuitively, the addition of these nodes does nothing to alter the possible GMC probability distributions on the outcomes of the observed nodes. For example, the standard no-signalling conditions should still be satisfied. To investigate this, let $X := \{A, D\}$ and $Y := \{F\}$, and let Z be empty. These sets satisfy the conditions of lemma 19 with

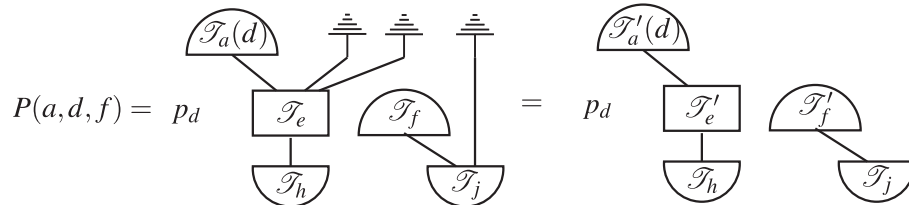
¹¹ See [23] for progress towards such a concept in the case of quantum theory.

$U := \{A, D, E, H\}$, $V := \{F, J\}$ and $W := \{B, C\}$. Hence X and Y are d -separated by the empty set. Therefore, to prove the soundness of the d -separation criterion in our setting, we need to show that X and Y are independent in any GMC probability distribution on this graph.

To establish this, we only need to consider $P(x, y) = P(a, d, f)$, which will be the marginal of a probability distribution that satisfies the GMC with respect to the whole GDAG. $P(a, d, f)$ can therefore be represented graphically as



To be consistent with our motivation, it should not be necessary to mention the nodes in B and C when defining this probability distribution, because they are to the future of all of A , D and F . This is indeed the case: this probability distribution still satisfies the GMC with respect to the graph with these two nodes removed. To see this, note that the outcome b only appears in the effect $\mathcal{T}_b(f)_{E \rightarrow B}$ above, and so summing over all possible outcomes in this factor gives the unique deterministic effect. The test $\mathcal{T}_c_{E \rightarrow C, J \rightarrow C}$ is also a deterministic effect, on $(E \rightarrow C)(J \rightarrow C)$. We use lemma 7, which states that the deterministic effect on a product of systems is the product of the deterministic effect on the systems separately. This gives



where in the last diagram we define the primed tests as the product of the unprimed tests with any following deterministic effects, for example in the case of E

$$\mathcal{T}'_{e H \rightarrow E} = \mathcal{T}_{e H \rightarrow E}^{E \rightarrow A, E \rightarrow B, E \rightarrow C} \top_{E \rightarrow B} \top_{E \rightarrow C}. \quad (4)$$

This result is equivalent to the statement that $P(a, d, f)$ fulfils the GMC for the original GDAG with B and C removed. Once this is done, we only need to note that the circuit has divided into two pieces, one referring to ad but not f , and one referring to f but not ad . Recalling that the definition of operational-probabilistic theories requires that tests from I to I compose by multiplication, this establishes that $P(a, d, f) = P(a, d)P(f)$.

There are two main steps in this example, which are both relevant to the general case. The first was to see that all nodes in W (that is, B and C) can be removed from the GDAG, in the following sense: if the probability distribution $P(x, y)$ fulfils the GMC on the whole graph G then its restriction to $G' = G \setminus W$ fulfils the GMC on G' . Above this is symbolized by absorbing the deterministic effects corresponding to outcomes of nodes in W into the preceding test. Secondly, after this step, the circuit separates into two parts, and hence the probability distribution can be seen to factorize in the required way.

3.5. The d -separation condition: general case

We now seek to show that, as in the above example, d -separation in a GDAG G implies conditional independence for all probability distributions that are GMC with respect to G .

Lemma 20. (*Proof in appendix B*). Consider a GDAG G and a subset $W \subseteq G$ that contains all of its own descendants. If probability distribution $P(g)$ fulfils the GMC on G then the probability distribution $P(g')$ (derived from $P(g)$ by marginalizing over outcomes in W) fulfils the GMC on $G' = G \setminus W$.

This lemma can be applied to eliminate the set W in the reformulation of d -separation given above, simplifying our task to proving the following.

Lemma 21. (*Proof in appendix B*). Let X , Y and Z be disjoint sets of observed nodes in a GDAG G' . Suppose G' can be partitioned into $\{U, V, Z\}$ such that

$$X \subseteq U, Y \subseteq V \quad (5)$$

$$m(U) \cap m(V) = \emptyset. \quad (6)$$

then $X \perp\!\!\!\perp Y \mid Z$ in any GMC probability distribution on G' .

Finally, we can prove our d -separation theorem.

Theorem 22. Let G be a generalized DAG with disjoint observed subsets X , Y and Z . Then

- (i) If P is generalized Markov with respect to G , then $X \perp\!\!\!\perp Y \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z$.
- (ii) If $X \perp\!\!\!\perp Y \mid Z$ holds for all P which are generalized Markov with respect to G , then $X \perp\!\!\!\perp Y \mid Z$.

Proof. To prove item (i), we combine lemmas 20 and 21. Item (ii) is a consequence of $C \subseteq \mathcal{G}$ and the classical theorem 5 part (ii). \square

4. Beyond conditional independence: quantitative bounds on correlations

In the Bell scenario, Bell inequalities limit the classical correlations (establishing $C \subsetneq \mathcal{Q}$), and Tsirelson inequalities limit the quantum correlations (establishing $\mathcal{Q} \subsetneq \mathcal{G}$). What limits the correlations in a general probabilistic theory? In the Bell scenario, a general probabilistic theory is limited *only* by the no-signalling principle (see for example [4]). In our notation, this means that $\mathcal{G} = \mathcal{I}$ for Bell GDAGs. Here we show that this fact does not extend to every scenario, i.e. we provide examples for which $\mathcal{G} \subsetneq \mathcal{I}$. In other words, causal structure can impose quantitative limits *beyond* the conditional independences between observed nodes, *independently* of the precise physical theory under consideration.

4.1. The triangle

The triangle scenario, shown in figure 4, has already received some interest in quantum foundations [8, 10, 16] and the causality literature [34]. Branciard *et al* initially introduced the scenario with definitions matching our C and \mathcal{Q} [8]. It was noted that understanding the classical

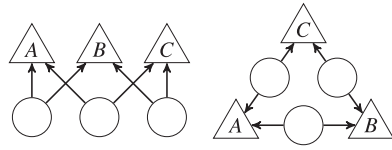


Figure 4. The ‘triangle’ GDAG drawn in two different ways.

correlations C in this scenario is much more mathematically challenging than in the Bell scenario. Nevertheless, Fritz showed that there exist quantum correlations for this scenario which cannot be reproduced using classical sources, i.e. $C \not\subseteq \mathcal{Q}$ [16]. A key part of this proof was showing that any $P \in C$ satisfies a ‘monogamy’ inequality

$$I(A: B) + I(B: C) \leq H(B). \quad (7)$$

In other words, the stronger the correlations between A and B , the weaker must be the correlations between B and C .

This has some interesting consequences. For example, note that there are no independences between observed nodes for this GDAG. Hence the ‘perfectly correlated bits’ distribution $P(0, 0, 0) = P(1, 1, 1) = \frac{1}{2}$ is in \mathcal{I} . However, this perfect correlation violates equation (7), and hence cannot be produced using classical sources.

Here we show that equation (7) this holds for any $P \in \mathcal{G}$, and hence perfect correlation cannot be produced in this GDAG using any generalized probabilistic theory. In other words, $\mathcal{G} \subsetneq \mathcal{I}$. We do this by first proving an important fact about \mathcal{G} in this scenario:

Theorem 23. *Suppose $P \in \mathcal{G}$ for the GDAG in figure 4. Then there exists another probability distribution P' , such that:*

1. $P'(a, c) = P(a)P(c)$,
2. $P'(a, b) = P(a, b)$, and
3. $P'(b, c) = P(b, c)$.

For a given P , the existence of P' is then a linear feasibility problem (studied in [2, 15]), and hence an efficiently checkable necessary condition for $P \in \mathcal{G}$ (and thus also for \mathcal{Q} and C).

Proof. By the definition of \mathcal{G} , there exists a causal operational-probabilistic theory with preparations ρ, σ, τ and observation-tests $\{e_a\}, \{f_b\}, \{g_c\}$ such that

$$P(a, b, c) = \begin{array}{c} \text{Diagram showing three preparation nodes } \rho, \sigma, \tau \text{ at the bottom and three observation nodes } e_a, e_b, e_c \text{ at the top. Arrows connect } \rho \text{ to } e_a \text{ and } e_b, \sigma \text{ to } e_b \text{ and } e_c, \text{ and } \tau \text{ to } e_c \text{ and } e_a. \end{array}$$

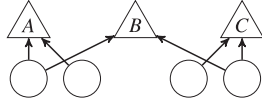


Figure 5. The GDAG for P' .

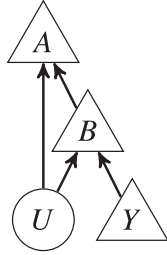


Figure 6. The relevant GDAG for 'instrumental inequalities'.

We can use these, along with the unique deterministic effect, to define

$$P'(a, b, c) = \begin{array}{c} \begin{array}{ccc} e_a & e_b & e_c \\ \hline \rho & \sigma & \sigma & \tau \end{array} \end{array}$$

Notice that $P' \in \mathcal{G}'$ for the GDAG \mathcal{G}' depicted in figure 5. Since A is d -separated from C in this GDAG, theorem 22 gives $P'(a, c) = P'(a)P'(c)$, which, once we have also established items 2 and 3, gives item 1.

Using lemma 7, we find

$$\begin{aligned} P(a, b) &= \sum_c P(a, b, c) = \begin{array}{c} \begin{array}{ccc} e_a & e_b & \text{---} \\ \hline \rho & \sigma & \tau \end{array} \end{array}, \\ P'(a, b) &= \sum_c P'(a, b, c) = \begin{array}{c} \begin{array}{ccc} e_a & e_b & \text{---} \\ \hline \rho & \sigma & \sigma & \tau \end{array} \end{array}. \\ &\quad \begin{array}{c} \text{---} \\ \hline \sigma \end{array} = 1 \end{aligned}$$

giving item 2. Item 3 follows similarly. □

Corollary 24. Equation (7) holds whenever $P \in \mathcal{G}$ for the GDAG in figure 4.

Proof. For any probability distribution $I(A: C|B) \geq 0$ and $H(B|AC) \geq 0$ and so

$$I(A: B) + I(B: C) = H(B) + I(A: C) - I(A: C|B) - H(B|AC) \leq H(B) + I(A: C). \quad (8)$$

Applying theorem 23 we obtain a P' with $I(A: C) = 0$ so that

$$I(A: B) + I(B: C) \leq H(B). \quad (9)$$

But this inequality only involves $P'(a, b)$ and $P'(b, c)$, which equal $P(a, b)$ and $P(b, c)$ respectively, and so this inequality holds for P as well. □

4.2. The instrumental GDAG

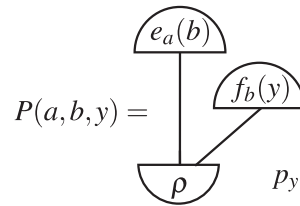
The fact that $\mathcal{C} \subsetneq \mathcal{I}$ for the GDAG in figure 6 has already been noted in the causality literature [28]. The original interest in this DAG arose in the study of cases of imperfect compliance in a controlled trial. For example Y might be a randomly assigned treatment, B the treatment the patient actually follows, and A recovery. There could be factors U that influence both the chance of recovery under each treatment, and also the chance of compliance with a particular treatment. This model does not imply any conditional independences on $\{A, B, Y\}$, but in [28], it is shown that it can still be tested because for any $P \in \mathcal{C}$

$$\max_b \sum_a \max_y P(a, b|y) \leq 1. \quad (10)$$

This is known as the *instrumental inequality*. Here we strengthen this result to

Theorem 25. Equation (10) holds for any $P \in \mathcal{G}$.

Proof. Since $P \in \mathcal{G}$, there is a bipartite preparation-test at U and choices of observation-test at A and B



These can be used to define a no-signalling distribution $P'(a, b|x, y)$ such that $P(a, b|y) = P'(a, b|x = b, y)$. Using no-signalling from y to a we can write $P'(a, b|x, y) = P'(a|x)P'(b|a, x, y)$. We can now adapt the proof in [28] as follows.

For each (a, b) , define $y(a, b)$ as the choice of y the maximizes $P(a, b|y)$. Then

$$\begin{aligned} \sum_a P(a, b|y(a, b)) &= \sum_a P'(a, b|x = b, y(a, b)) \\ &= \sum_a P'(a|x = b)P'(b|a, x = b, y(a, b)). \end{aligned} \quad (11)$$

Certainly $P'(b|a, x, y) \leq 1$, and the final term above is a convex combination of such, and so

$$\sum_a P(a, b|y(a, b)) \leq 1. \quad (12)$$

Recalling the definition of $y(a, b)$ this is exactly

$$\sum_a \max_y P(a, b|y) \leq 1. \quad (13)$$

Since this holds for all b we have equation (10). \square

Since there are no observable independences for this GDAG, \mathcal{I} is just the set of all probability distributions. Hence this result establishes that $\mathcal{G} \subsetneq \mathcal{I}$.

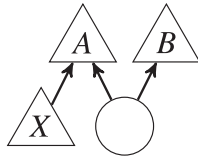


Figure 7. A bipartite Bell scenario where only Alice has a choice of measurement.

5. Towards a classification of ‘interesting’ GDAGs

It is known that a Bell scenario where only one party has a choice of measurement figure 7 is not ‘interesting’. What exactly does this mean? Certainly it doesn’t mean that there are no restrictions on the probability distributions: there is still no-signalling from Alice to Bob. However, this is a conditional independence $X \perp\!\!\!\perp B$ which follows from d -separation. Hence, by definition, it is satisfied by all distributions in \mathcal{I} . The reason this scenario is not interesting is that even for classical distributions there are no further restrictions, i.e. $\mathcal{C} = \mathcal{I}$.

Since we have seen that for any GDAG $\mathcal{C} \subseteq \mathcal{Q} \subseteq \mathcal{G} \subseteq \mathcal{I}$, GDAGs in which $\mathcal{C} = \mathcal{I}$ must have $\mathcal{C} = \mathcal{Q} = \mathcal{G} = \mathcal{I}$. Hence there is very little to say about such GDAGs except for listing the observable conditional independences. It is therefore of interest to classify which GDAGs have $\mathcal{C} = \mathcal{I}$ and which do not. The GDAGs that do not are then candidates for quantum advantages in (‘black-box’) information processing, settings to compare quantum theory to more general theories, and so on.

Here we make significant progress towards such a classification by providing a sufficient condition for $\mathcal{C} = \mathcal{I}$ and providing strong evidence that our condition is also necessary, at least for small GDAGs. This classification problem may be of interest even for purely classical causal inference, since if one has a candidate causal structure for which $\mathcal{C} \subsetneq \mathcal{I}$ then it can be ruled out by checks that go beyond observable conditional independences (like Bell inequalities). On the other hand, if a candidate causal structure has $\mathcal{C} = \mathcal{I}$ then checking the observable conditional independences implied by d -separation suffices for the existence of a (classical) model.

5.1. A sufficient condition for $\mathcal{C} = \mathcal{I}$

We begin by observing that certain changes to a GDAG can only make \mathcal{C} smaller. We will use the notation $X \rightsquigarrow Y$ to denote the existence of a directed path from a node X to node Y , where any intermediate nodes are unobserved.

Theorem 26. *Consider the set of classical correlations \mathcal{C}_G for a GDAG G . Suppose that one of the following transformations is performed on G , producing a GDAG H :*

1. *Removal of an edge.*
2. *Removal of an isolated unobserved node.*
3. *Addition of an edge $X \rightarrow Y$ where previously $X \rightsquigarrow Y$.*
4. *Addition of an edge $X \rightarrow Y$ where previously $\text{PA } X \subseteq \text{PA } Y$ and $\text{PA } X$ contained at least one unobserved node.*

Then $\mathcal{C}_H \subseteq \mathcal{C}_G$.

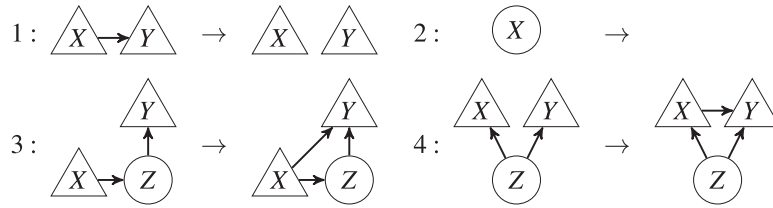


Figure 8. Illustrations of the allowed transformations in theorem 26.

These transformations are illustrated in figure 8.

Proof. We need to prove that if $P \in C_H$ (i.e. P is classical for the new GDAG H) then $P \in C_G$ (i.e. P is classical for the old GDAG G). We shall use the fact that P is classical for a GDAG if and only if there exists a functional causal model for P using the underlying DAG [27]. In a functional causal model, the value of each node Z is given by a function $z = f(\text{pa } z, n_z)$ of its parents and a noise variable, and the noise variables are independently distributed. For each transformation, we shall show that if a functional causal model exists for P defined on H , then a functional causal model exists for P defined on G :

1. In H the argument to a function has been removed, e.g. if a node Z has parents X and Y , then $z = f(x, y, n_z)$ becomes $z = f'(x, n_z)$. We can define a functional causal model for G using the one for H by allowing the function to trivially depend on its new argument, e.g. $f(x, y, n_z) = f'(x, n_z)$. By definition, this gives the same probabilities for all nodes.
2. We can define a model for G by giving the isolated node Z an arbitrary error variable N_Z and making Z an arbitrary function of it. This has no effect on the probabilities for any other variable, which includes all the observable variables.
3. In both G and H we have $X \rightsquigarrow Y$, but in H we also have $X \rightarrow Y$. To define a model for G we must absorb the dependence of Y on X that exists for H . We can do so by using the unobserved nodes $Z^{(i)}$ in the path $X \rightsquigarrow Y$. Specifically, for each of the random variables $Z^{(i)}$ defined for H , we define an ‘enlarged’ variable $W^{(i)}$ that includes a copy of X , when defining a model for G . That is, $z = f(\text{pa } z, n_z)$ becomes $w := (z, x) = (f(\text{pa } z, n_z), x)$. We then replace the dependence of the function at Y on X by its copy in W , i.e. $y = f(z, x, n_y)$ becomes $y = f(z', n_y)$. This procedure does not affect any of the observable probabilities.
4. In H , the variable Y is now a function of X . In turn, X is a function of its parents and an error variable N_X . But since $\text{PA } X \subseteq \text{PA } Y$, to define a model for G we need only ensure the dependence of Y on N_X . Since N_X is independently distributed, we can move this into an unobserved parent of X , say Z , which exists by assumption. Specifically, we define $z' := (z, n_x)$, and then $x = f(z, n_x)$ for H becomes $x = f'(z') := f(z, n_x)$ for G . We let Y be calculated as before, but in place of the direct dependence on X , we use the same function used to calculate x at X , e.g. $y = g(x, z, n_y)$ becomes $y = g'(z', n_y) := g(f(z, n_x), z, n_y)$. The only variable whose probabilities have been changed is Z , which is not observable. \square

The sufficient condition for $\mathcal{I} = \mathcal{C}$ is as follows. If starting with a given GDAG one can apply a sequence of the above transformations and produce a GDAG with:

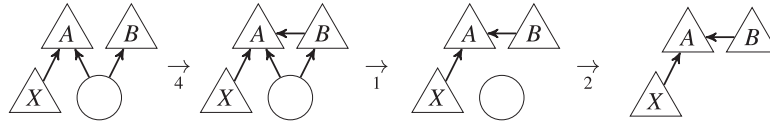


Figure 9. Repeated applications of theorem 26 transform the GDAG of figure 7 into a new GDAG without enlarging C . (The numbers under the arrows indicate the relevant transformation from theorem 26.) Allowed distributions on the final GDAG are constrained only by $X \perp\!\!\!\perp B$, which held for \mathcal{I} in the initial GDAG, and so $C = \mathcal{I}$ in the initial GDAG.

Table 1. The results of our condition for GDAGs of size 1 to 7. It is plausible that the fraction of GDAGs for which $C = \mathcal{I}$ tends to zero as the number of nodes tends to infinity, because larger and larger GDAGs should be more and more likely to contain, for example, a Bell scenario.

Nodes	Number of GDAGs	Number for which our condition holds	Percent
1	2	2	100%
2	7	7	100%
3	40	40	100%
4	420	419	99.8%
5	8628	8532	98.9%
6	357 468	347 287	97.2%
7	299 890 52	283 703 73	94.6%

1. No unobserved nodes, and
2. Requiring no more conditional independences on the observed nodes than the original GDAG did, then $\mathcal{I} = C$ for the original GDAG. To see this, start with some probability distribution in \mathcal{I} . Recalling that the conditional independences are the only restrictions on (G)DAGs with no latent variables, the above two properties ensure that the distribution is classical for the new GDAG. But then by repeated applications of theorem 26 there is a classical model for the original GDAG with the same probabilities for the observed nodes, and we are done.

For example, this condition establishes that the Bell scenario with only one setting, figure 7, indeed has $\mathcal{I} = C$, as shown in figure 9.

5.2. Results for small GDAGs

Using a strategy described in appendix C, and algorithms from [35] to keep track of conditional independences, we have searched for applications of the above condition to all GDAGs with up to seven nodes. The results are shown in table 1. Our condition is powerful enough to show that the overwhelming majority of small GDAGs have $C = \mathcal{I}$. Indeed this is the case for all GDAGs of size three or smaller, and the only GDAG of size four is that of section 4.2 for which it was already known that $C \subsetneq \mathcal{I}$. The 96 GDAGs of size five to which our condition does not apply are mostly trivial modifications of that of section 4.2, for which the proof that $C \subsetneq \mathcal{I}$ will easily carry over. To eliminate such GDAGs from consideration we developed a number of reduction criteria. For completeness these are described in appendix D.

Once these reduction criteria have been applied, there remain 2 GDAGs of size five and 18 of size six. If we can show that these 20 GDAGs have $C \subsetneq \mathcal{I}$ then we will have shown that our necessary condition is also sufficient, at least for GDAGs of size six or less. A full characterization of C in a general scenario is not known, however necessary conditions for membership of C can be derived by searching for ‘Shannon-type entropic inequalities’. These are linear inequalities expressed purely in terms of the Shannon entropy $H(X)$ of subsets of variables. See [10] and references therein for the details of this approach.

For each GDAG we construct the Shannon cone (defined by the positivity of all conditional mutual informations) for all variables, observable and latent. For each node X we add

$$I(X: \text{non-descendants of } X | \text{PA } X) \leq 0$$

to enforce the Markov condition. Finally we use Fourier–Motzkin elimination to project out entropies involving the latent variables. This gives a set of entropic inequalities E_C .

The resulting inequalities are necessary conditions for membership of C . However, we are interested in comparing C with \mathcal{I} . Hence we repeat the process for \mathcal{I} . We start with the Shannon cone on the observable variables, and add $I(X: Y | Z) \leq 0$ whenever X and Y are d -separated by Z . This gives a second set of entropic inequalities $E_{\mathcal{I}}$.

For 19 of the 20 GDAGs we find inequalities in E_C that do not follow from those of $E_{\mathcal{I}}$. Unless the inequality is a non-Shannon-type inequality for \mathcal{I} , this establishes that $C \subsetneq \mathcal{I}$. Since non-Shannon-type inequalities rarely play a role, this is rather good evidence. For most of the GDAGs it is straightforward to find explicit $P \in \mathcal{I}$ that violate one of E_C , thus definitively establishing $C \subsetneq \mathcal{I}$. Curiously, the one GDAG for which $E_{\mathcal{I}} = E_C$ is the bipartite Bell scenario. Fortunately, we already know that $C \subsetneq \mathcal{I}$ for that case! The GDAGs and corresponding inequalities are listed in appendix E.

These results provide excellent evidence that our sufficient condition for $C = \mathcal{I}$ is also necessary for all GDAGs with six or fewer nodes. Perhaps it is in fact necessary for an arbitrary GDAGs.

6. Conclusions

Here we have proposed a way to combine the frameworks of GPTs and causal Bayesian networks. We believe that the results we have obtained suggest that this proposal is worth exploring further, although the two fundamentally distinct types of node mean it is unlikely to be the final word on non-classical causation.

Our first main result was that the graphical d -separation criteria for conditional independence remains sound for generalized networks. This should be useful, since the classical soundness result is very fundamental to the classical theory. For example, the main algorithm for causal inference in the presence of latent variables, IC*, uses only observable conditional independences. Hence it will still operate correctly in our generalization. It would be worth exploring similar generalizations of other fundamental parts of the classical theory, for example the criteria for two causal structures to have the same observable consequences.

We then found that some other constraints on observed probabilities also generalize to this setting. This shows that even in its weakest interpretation, causal structure has more interesting consequences than ‘no signalling’ in the Bell scenario, even extended to include all observable conditional independences. This has interesting foundational consequences. If the violation of

Bell inequalities is to be explained by accepting altered causal structure, one must give up hope of an explanation of observed conditional independences such as no-signalling based on causal structure [38]. We now see that there are other, more intricate, limitations which would also be left unexplained by an altered causal structure. Since our techniques for finding such limits were rather ad hoc, the main open problem here is to obtain a more systematic understanding of these constraints. The entropic inequalities look like a promising place to start: indeed we do not know of any example of such an inequality being violated by any generalized probabilistic theory.

Finally, we have considered the problem of identifying whether or not the only consequences of a GDAG are conditional independences, i.e. $C = \mathcal{I}$. We have presented a sufficient condition. Proving the necessity of this (or any other) condition would shed light on the conceivable forms of ‘device-independent’ non-classicality. If a GDAG has $C \subsetneq \mathcal{I}$, then one could also ask more fine-grained questions: is $C \subsetneq \mathcal{Q}$ (quantum non-classicality), $\mathcal{Q} \subsetneq \mathcal{G}$ (post-quantum correlations), $\mathcal{G} \subsetneq \mathcal{I}$ (theory-independent limits on correlation)? Other interesting classification problems include understanding when the distributions on some nodes, conditioned on some others, form a convex set.

Acknowledgments

We are grateful for useful discussions with Jonathan Barrett, Giulio Chiribella, Tobias Fritz, Anirudh Krishna and Rob Spekkens. Research at Perimeter Institute is supported in part by the Government of Canada through NSERC and by the Province of Ontario through MRI. Work by JH and RL is supported by grants from the John Templeton Foundation. JH also receives support from EPSRC grant *DIQIP* and ERC grant *NLST*.

Appendix A. Comparison with other approaches

Recent work by other authors has also considered correlations on general causal structures. We shall restrict our focus to those approaches which have been specifically used to study classical correlations resulting from quantum processes on general causal structures. Hence we omit works that give a quantum version of Bayesian networks by replacing probabilities with amplitudes (e.g. [36]), or that only apply to states at a single time-step (e.g. [24]), since neither appears to support the causal interpretation which we are interested in. More relevant are the ‘Quantum Causal Networks’ of [21], but these are difficult to compare to our approach since they treat entanglement as a new type of causal relation indicated by an undirected edge, whereas in our approach entanglement requires an analog of a ‘common cause,’ that is, mutual ancestors. Most closely related are two lines of work, based on source-measurement hypergraphs and circuit DAGs respectively. The idea of having two different types of node, and specifically the choice of triangles and circles, comes from a more general project to recast quantum theory as a theory of inference. To aid the reader who has encountered any of these three approaches, here we compare their definitions with ours.

A.1. Hypergraphs

Building on the idea of ‘ N -locality’ from [8], in [16] a causal structure is represented by a hypergraph, with vertices representing measurements and edges representing sources. This can

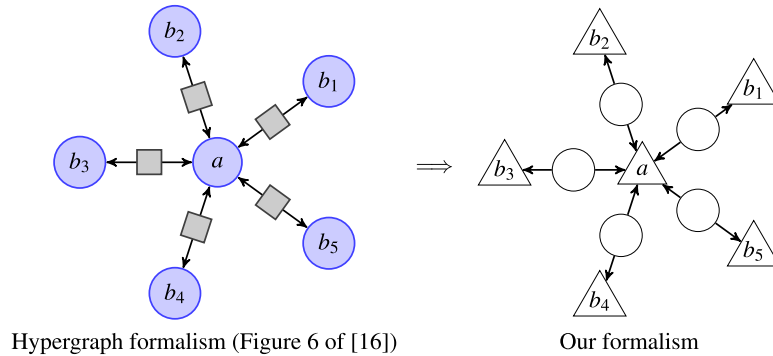


Figure A.1. In the formalism of [16] a causal structure is formally represented by a hypergraph, although the edges are suggestively drawn as squares with arrows to members. To convert to our formalism, an edge becomes an actual (unobserved) node with edges to each member. An application of section 5.1 immediately shows that in this ‘star’ scenario $\mathcal{C} = \mathcal{I}$.

be translated into our formalism by turning each vertex into an observed node, and each hypergraph edge into an unobserved node with an edge going to every member of the hypergraph edge. What is called a ‘correlation’ in [16] then agrees with our definition of a member of \mathcal{I} , and the definitions of classical and quantum correlations map directly to our definitions of \mathcal{C} and \mathcal{Q} .

This close translation means that some of our results touch directly on the results and open problems in [16]. Our triangle result answers the first part of problem 3.4 in [16] in the negative. Our investigation in section 5 seeks to address (a generalization of) problem 3.6 in [16]. For example, the criteria given in section 5.1 enables a graphical proof of the ‘if’ part of theorem 3.8 in [16], see figure A.1.

Many GDAGs in our formalism will not correspond to any hypergraph in the formalism of [16]. For example, the GDAG in section 4.2 cannot be represented as a hypergraph as there is no way to encode the edge from B to A .

A.2. Circuit diagrams

The ubiquitous circuit diagrams used in quantum computing [26] and discussions of GPTs (e.g. [11]) can be viewed as DAGs, and seem to suggest a causal interpretation (see [6] and references therein). Recently this idea has been used specifically for the purpose of exploring Bell-like scenarios [17].

In [17] a causal structure is represented as a DAG. Hence there is only one type of node, which is always associated with a random variable. Any edge can carry ‘hidden variables’ in the classical case or quantum systems in the quantum case. Hence to translate to our formalism, first represent every node as an unobserved node. Then add a supplementary observed node for each of those nodes, and an edge from the unobserved to the supplementary observed node, as in figure A.2. Again the definitions of correlation, classical correlation and quantum correlation appear to coincide with \mathcal{I} , \mathcal{C} and \mathcal{Q} respectively (except that [17] allows infinite-valued latent variables, which as already noted may or may not result in more classical correlations). \mathcal{Q} only

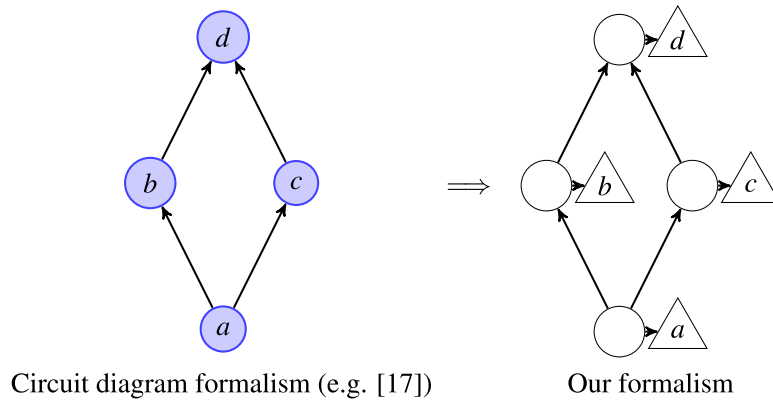


Figure A.2. In the formalism of [17] a causal structure is represented by a DAG. Every edge gets a hidden variable in the classical case and a quantum system in the quantum case, so to represent such a structure in our formalism each node should become two nodes, one of each type, as shown.

matches because every quantum instrument can be replaced by a channel¹² with an additional ‘flag’ system in the output which can later be measured to obtain the result. Finally, [17] considers C-correlations for certain categories \mathcal{C} . This is closely related to the CDP formalism of operational-probabilistic theories and so ranging over all \mathcal{C} should, under the above translation, agree with our \mathcal{G} .

Again many GDAGs in our formalism will not correspond to any DAG in [17]. For example, in the formalism of [17] there is no way to enforce that the edge from B to A in the GDAG of section 4.2 does not carry hidden variables or quantum systems, rather than just the value b as in our formalism.

A.3. Quantum theory as a theory of inference

In [23], Leifer and Spekkens also use GDAGs depicted using circular and triangular nodes. We deliberately use the same notation here, although the approaches are significantly different. The aim in [23] is to generalize the quantum formalism to the point that one can, for example, talk about the joint quantum state of A and B even if A is the input to a channel and B the output. Here we stick to the standard quantum formalism, with tensor products only across space, and limit ourselves to the joint probabilities of the variables on the observed nodes—i.e., the classical variables. In [23], the state of a set of triangular nodes is diagonal in a fixed basis and hence encodes a joint probability distribution. We use the same notation because we expect the possible sets of joint distributions in [23] to match our \mathcal{Q} .

The main reason that the distributions may not be identical is that when an unobserved node has multiple outgoing edges, we associate a Hilbert space to each edge, giving an explicit tensor product structure. In [23], a single Hilbert space is associated with the circular node itself. The meaning of edges is to be in terms of some planned generalization the classical Markov condition to quantum states. Presumably our tensor products will satisfy this condition (see figure A.3 for an example of the likely translation), but there may be quantum states that are

¹² A channel is a quantum instrument with only one outcome, i.e. a completely positive trace preserving map.

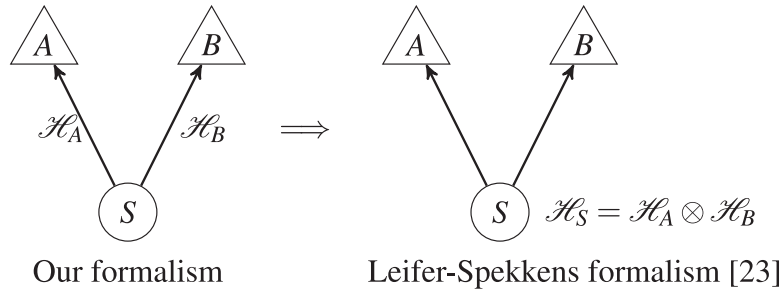


Figure A.3. In our formalism a quantum model for this GDAG consists of two Hilbert spaces, a bipartite quantum state and a POVM on each Hilbert space. In the Leifer–Spekkens formalism there would be a single Hilbert space for \mathcal{H}_S with an associated state, and two POVMs on \mathcal{H}_S satisfying some Markov condition. Translating from the first to the second just involves letting \mathcal{H}_S be the tensor product of the two Hilbert spaces, keeping the state as it is, and tensoring the POVMs with identities so that they act on the whole of \mathcal{H}_S . Until the Leifer–Spekkens formalism has been fully worked out it is difficult to say whether translation in the opposite direction will always be possible.

‘Leifer–Spekkens Markov’ for a GDAG and yet cannot be expressed using our tensor product form.

Appendix B. Proofs of d -separation lemmas

Proof of lemma 19. (‘If.’) We must show that every pseudo-path from X to Y intersects Z . Assume for contradiction that there exists a pseudo-path from X to Y that does not intersect Z . A pseudo-path cannot intersect W by definition. Then, by the assumption that $\{U, V, Z, W\}$ is a partition of G , a pseudopath from X to Y that does not intersect Z can only contain elements in U or V . Such a pseudo-path must at some point contain a pair of sequential elements $a \in U$ and $b \in V$. But we have also assumed that $m(U) \cap m(V) \subseteq W$, i.e. the mutual children of a and b are in W . But this contradicts the definition of a pseudo-path, for which we must have $m(a) \cap m(b) \not\subseteq W$. Hence no such sequential pair in a pseudopath can exist, and therefore there are no pseudopaths from X to Y that do not intersect Z .

(‘Only if.’) Notice that $W = G \setminus J^-(X \cup Y \cup Z)$ is as in the definition of d -separation, and in particular that $W \cap Z = \emptyset$. We obtain the required partition of G as follows. Let U be the union of all pseudo-paths that start at any node in X and finish anywhere in G but without intersecting Z . By the definition of U , we have $X \subseteq U$ and $U \cap Z = \emptyset$. By the definition of a pseudo-path, $U \cap W = \emptyset$. Hence U , W and Z are disjoint. Now define $V := G \setminus (U \cup W \cup Z)$. This defines a partition $\{U, V, Z, W\}$ of G , with $X \subseteq U$. Now, by assumption all pseudopaths from X to Y intersect Z . Therefore $Y \cap U = \emptyset$, by the definition of U . Since we also have $Y \cap Z = \emptyset$ and $Y \cap W = \emptyset$, and since $\{U, V, Z, W\}$ is a partition of G , we therefore have $Y \subseteq V$. Finally, suppose that there exist $a \in U$ and $b \in V$ such that $m(a) \cap m(b) \not\subseteq W$. This defines a pseudo-path from a to b that does not intersect Z . But then by the definition of U , we have $b \in U$ which contradicts the fact that $b \in V$, since $U \cap V = \emptyset$. Hence we have $m(U) \cap m(V) \subseteq W$. \square

Proof of lemma 20. The GMC condition is

$$P(g) = \prod_{i=1}^m \mathcal{T}_{x^{(i)}} \left(\text{opa } x^{(i)} \right)_{\text{incU } X^{(i)}}^{\text{outU } X^{(i)}}, \quad (\text{B.1})$$

and we have

$$P(g') = \sum_w p(g). \quad (\text{B.2})$$

By assumption $W \subseteq G$ contains its own future. A node that is maximal with respect to W is thus maximal with respect to G . Consider such a maximal node $X^{(j)}$, and the following expression

$$\sum_{x^{(j)}} \mathcal{T}_{x^{(j)}} \left(\text{opa } x^{(j)} \right)_{\text{incU } X^{(j)}}^{\text{outU } X^{(j)}}. \quad (\text{B.3})$$

A maximal node has no outgoing systems and so $\text{outU } X^{(j)}$ is in this case empty, so B.3 is an observation test. Furthermore it is either already deterministic (if $X^{(j)}$ is unobserved), or summing over all outcomes $x^{(j)}$ makes it deterministic (if $X^{(j)}$ is observed). For both types of node therefore (B.3) equals the unique deterministic effect on $\text{incU } X^{(j)}$. Applying lemma 7,

$$\sum_{x^{(j)}} \mathcal{T}_{x^{(j)}} \left(\text{opa } x^{(j)} \right)_{\text{incU } X^{(j)}}^{\text{outU } X^{(j)}} = \mathbb{T}_{\text{incU } X^{(j)}} = \prod_{X^{(i) \rightarrow X^{(j)} \in \text{incU } X^{(j)}}} \mathbb{T}_{X^{(i) \rightarrow X^{(j)}}}. \quad (\text{B.4})$$

Summing over $x^{(j)}$ in (B.1), noting that the maximality of $X^{(j)}$ ensures that $x^{(j)}$ appears only in the $i = j$ term, and substituting the above expression for that term we have

$$\sum_{x^{(j)}} P(g) = \prod_{i \in \{1, \dots, m\} \setminus j} \mathcal{T}'_{x^{(i)}} \left(\text{opa } x^{(i)} \right)_{\text{incU } X^{(i)}}^{\text{outU } X^{(i)} \setminus X^{(i) \rightarrow X^{(j)}}}, \quad (\text{B.5})$$

where

$$\mathcal{T}'_{x^{(i)}} \left(\text{opa } x^{(i)} \right)_{\text{incU } X^{(i)}}^{\text{outU } X^{(i)} \setminus X^{(i) \rightarrow X^{(j)}}} = \mathcal{T}_{x^{(i)}} \left(\text{opa } x^{(i)} \right)_{\text{incU } X^{(i)}}^{\text{outU } X^{(i)}} \mathbb{T}_{X^{(i) \rightarrow X^{(j)}}}, \quad (\text{B.6})$$

where $\mathbb{T}_{X^{(i) \rightarrow X^{(j)}}$ is the unique deterministic effect for the system on the edge $X^{(i)}$ to $X^{(j)}$.

The upshot is that marginalizing over the outcomes for a maximal element $X^{(j)}$ produces a probability distribution that fulfils the GMC for the GDAG with that element removed, $G \setminus X^{(j)}$. Because $W \subseteq G$ contains its own future, this process can be repeated for every element in W , and so marginalizing over every outcome in W results in a distribution satisfying the GMC for $G \setminus W$. \square

Proof of lemma 21. We can write

$$P(x, y, z) = \sum_{u'v'} P(u, v, z), \quad (\text{B.7})$$

where $U' = U \setminus X$, $V' = V \setminus Y$, and the lowercase versions are the associated outcome variables as before. Define $Z_1 = Z \cap m(U) = m(U) \setminus U$ and $Z_2 = Z \setminus Z_1$. Then (using UZ_1 as shorthand for $U \cup Z_1$ and so on)

$$P(y, z) = \sum_u \prod_{P \in UZ_1} \mathcal{T}_p(\text{opa } p)_{\text{incU } P}^{\text{outU } P} \sum_{v'} \prod_{Q \in VZ_2} \mathcal{T}_q(\text{opa } q)_{\text{incU } Q}^{\text{outU } Q} \quad (\text{B.8})$$

$$= \left(\sum_{u'x} \prod_{P \in UZ_1} \mathcal{T}_p(\text{opa } p)_{\text{incU } P}^{\text{outU } P} \right) \left(\sum_{v'} \prod_{Q \in VZ_2} \mathcal{T}_q(\text{opa } q)_{\text{incU } Q}^{\text{outU } Q} \right). \quad (\text{B.9})$$

The factorization above follows because the nodes U , whose outcome variables u are summed over in the first bracket, do not appear in the second bracket. A node in U is never a parent of a node in Z_2 , from the definition of Z_2 and Z_1 above; it is never a parent of a node in V because of condition 6. Conversely, a node in V' is never a parent of a node in Z_1 or of a node in U for the same reasons. It follows trivially that a node in U is not the *child* of a node in V' or vice-versa. This establishes the factorization (and also that the terms in the brackets correspond to closed circuits and are thus probabilities). For the same reasons we also have

$$P(x, y, z) = \left(\sum_{u'} \prod_{P \in UZ_1} \mathcal{T}_p(\text{opa } p)_{\text{incU } P}^{\text{outU } P} \right) \left(\sum_{v'} \prod_{Q \in VZ_2} \mathcal{T}_q(\text{opa } q)_{\text{incU } Q}^{\text{outU } Q} \right). \quad (\text{B.10})$$

Now $P(x|y, z) = P(x, y, z)/P(y, z)$, and the second terms in B.9 and B.10 will cancel. Since $Y \subseteq V$ this means $P(x|y, z)$ is independent of y , establishing the conditional independence of X and Y given Z . \square

Appendix C. A $C = I$ search strategy

It might appear that one has to attempt a potentially unbounded number of transformations to apply the sufficient condition for $C = I$ in section 5.1. Fortunately, if any sequence of transformations exists from a GDAG to one satisfying the criteria given there, then one will be found using the following strategy, as we will show below.

Let T (for ‘tricky’) be the set of all observed nodes that have unobserved parents. Let R (for ‘root’) be the set of all unobserved nodes that have no unobserved parents. Consider every possible ordering of T : T_1, T_2, \dots, T_n , with each element T_i associated with every possible $R_i \in R$, with $R_i \rightsquigarrow T_i$. For each possibility, apply the transformations as follows:

1. Apply transformation 3 to every pair of nodes with $X \rightsquigarrow Y$.
2. For i from 1 to n :
 - (a) Applying transformation 1, remove any edges from T_j (with $j > i$) to T_i , and from any unobserved nodes (except R_i) to T_i .
 - (b) Use transformation 4 to add edges from T_i to T_j (with $j > i$) where possible.
3. Apply transformation 1 to remove any remaining edges incident on unobserved nodes, then use transformation 2 to remove all the unobserved nodes.

It can be seen that transformation 3 can be applied first, as none of the other transformations can increase its applicability. It might as well be applied ‘maximally’ as any unhelpful edges can always be removed later.

It can also be seen that transformation 2 must be applied to all the unobserved nodes at some point, to ensure there are none in the final GDAG, and it can always be applied last, as it cannot increase the applicability of any of the other transformations.

All that remains is to show that the second step makes the best use of transformations 1 and 4. Since removing edges can only add conditional independences, it can only be worth doing if it helps in applying transformations 4. Since we are aiming for a GDAG with no unobserved nodes, the only point in applying transformations 4 to an unobserved node would be if it helped with a future application between observed nodes. Clearly, adding an edge from an observed node to an unobserved node cannot help. Let us consider a situation in which transformations 4 can be used to add an edge from an unobserved node to an observed node. Now, we can (and will) later remove any such edge from unobserved nodes to observed nodes, except if it is required to apply transformations 4. Because of this, the only point of adding the edge would be for it to connect the observed node to the one unobserved parent required to enable this later application of transformations 4. But, from the maximal application of transformations 3, any such role can just as easily be played by the unobserved parent required for the possible application of transformations 4 presently under consideration.

Hence transformations 4 is only worth applying between observed nodes, and of these only the nodes in T are possibilities. After all the transformation we are left with some GDAG, which defines a partial order on T and can be extended to a total order. If we are aiming for a particular order we need to remove any edges from T_j to T_i with $j > i$. The only ultimate use for edges from unobserved nodes is to allow the application of transformations 4, for which only one such edge is needed. If an unobserved node has unobserved parents then by the first step the parent can only have more descendants, making it the same or more useful for the application of transformations 4. Hence the single edge from an unobserved node we keep might as well be from an element of R .

Finally, we need to argue that transformations 4 might as well be applied based on the ordering on T we have defined using the final GDAG. The only point in applying a transformations 4 early is if it helps with a later application of transformations 4. If the later application is to add an edge from X to Y , we can only help by adding an edge from a node in $PA\ X$ to Y . But a node in $PA\ X$ will be before X in the ordering on T , so such an edge will, if possible, be added before when following the above strategy.

Appendix D. GDAG reduction

In order to study whether or not the sufficient condition for $C = \mathcal{I}$ given in section 5.1 might also be necessary for $C = \mathcal{I}$ by checking small GDAGs, it is useful to have a notion of when one GDAG ‘reduces’ to another, such that if the second GDAG has $C \subsetneq \mathcal{I}$ then the first does as well.

D.1. Strong reducibility

We say that a GDAG A is strongly reducible to another GDAG B if the observed nodes in B are a subset of the observed nodes in A , and for any causal operational-probabilistic theory, the set of possible distributions on observable variables in B is equal to the set of distributions obtained by marginalizing distributions on A . We likewise require that \mathcal{I} for B is exactly the marginals of \mathcal{I} for A .

Applying the following transformations to A gives a new graph B to which A is strongly reducible:

1. *Removing a disconnected component.* By the definition of an operational-probabilistic theory, the probabilities for two disconnected components are the products of the probabilities for each. Since marginalising one factor in a product distribution gives the other factor, valid distributions for A marginalize to valid distributions on B . Similarly a valid distribution on B can be taken to a valid distribution on A with the correct marginal by putting an arbitrary model on the removed component. For \mathcal{I} simply note that the d -separation conditions for B are not affected by the presence or absence the disconnected component.
2. *Removing a childless unobserved node.* Such a node represents the unique deterministic effect, which can be factorized into the deterministic effect on each incoming system, which can then be incorporated into the definition of the parent node. To go in the other direction simply use trivial systems for each incoming edge. Such nodes can neither block existing paths nor create a new unblocked path and so do not effect \mathcal{I} either.
3. *Merging an unobserved node with its sole parent, also unobserved.* An unobserved node that has only one parent, which is also unobserved, represents a deterministic test. Its parent can be redefined by applying that test to the relevant output system. To go in the other direction simply use the identity test, whose existence is part of the definition of operational-probabilistic theory. This transformation also does not affect conditional independences among the observed nodes.
4. *Removing an observed node associated with a 1-outcome variable.* Such a node represents the deterministic effect on its inputs, as in the case of a childless unobserved node. Outgoing edges have no effect because they just add a fixed label to children. Finally, removing a node certainly cannot remove conditional independences from the remaining nodes, to ensure it doesn't add any see appendix D.4.
5. *For an observed node X all of whose parents are observed, removing an edge from a parent Y such that all the observable conditional independences from d -separation after the removal already held beforehand.* Such an observed node is specified by a classical conditional probability $p(x|y, z)$. Once the edge from Y is removed we have $X \perp\!\!\!\perp Y \mid Z$ (since if Y is a descendant of X the original graph would have contained a cycle). By assumption $X \perp\!\!\!\perp Y \mid Z$ therefore holds in the original distribution, i.e. $p(x|y, z) = p(x|z)$, and so we can achieve the exact same probability distributions with or without the edge from Y to X . Finally, \mathcal{I} is the same by construction.
6. *Removing an unobserved node whose parents and children are subsets of the parents and children respectively of another unobserved node.* The test at such an unobserved node can simply be incorporated into the other node, with the edges from common parents and children now carrying the systems to/from both. To go in the other direction just add a trivial test to the new node. An unblocked path via the removed node can just as easily go via the other node so \mathcal{I} is unaffected.

D.2. Reducibility

The condition for reducibility is the same as strong reducibility, except that we only consider GPTs that have system types, states, and measurements suitable for perfectly transmitting, encoding, and decoding any finite-valued classical information. This includes classical

probability theory (which defines \mathcal{C}) and quantum theory (which defines \mathcal{Q}). It also includes unspecified theories (which define \mathcal{G}) since any operational-probabilistic theory can always be supplemented with such systems. It does not include, for example, the restriction of quantum theory to operations with a certain amount of noise.

Clearly reducibility is a weaker notion than strong reducibility. In addition to the transformations in the previous subsection, applying the following transformations to A gives a new graph B to which A is reducible:

1. *Merging an unobserved node with its sole child.* To convert a model on the unmerged GDAG to the merged one, simply compose the two tests. To go in the other direction, let the new unobserved node with only one child be the identity test on the edges from unobserved parents, and use classical information encoding states for the incoming edges from observed parents. At the child use the corresponding classical information decoding measurements to recreate the correct dependencies. As for \mathcal{I} , simply note that this change has no effect on the d -separation of observed nodes.
2. *Merging an observed node Y (that has only one sibling, Z) with its unobserved parent X (which is itself parentless).* The pair of nodes X, Y represents a bipartite state at X with a measurement Y on one system. Considered together this is a ‘preparation test’ for the remaining system that goes to Z . But in a causal theory every state is proportional to a deterministic state, so this is equivalent to sampling from the classical probability distribution given by the norms of the states and then preparing the corresponding normalized state. The sampling can be done as the new consolidated node, whilst the preparation can be incorporated into Z . Going in the other direction, we are starting with a single node representing a classical probability distribution. This can be sampled as part of the new unobserved node X , with the resulting classical information transmitted to both children. The copy sent to the observed node Y is simply decoded and output, the copy sent to the other node Z is decoded and then used as the label that previously came from the observed parent. Since an unobserved node cannot be conditioned on, the path from the observed node Y via the unobserved node X operates in exactly the same way as a direct connection as far as d -separation is concerned, so \mathcal{I} is unchanged.

D.3. The implications of reducibility

Suppose we have two GDAGs, and the first is reducible to the second. Suppose the second has $\mathcal{C} \subsetneq \mathcal{I}$, i.e. there exists some $P \in \mathcal{I}$ with $P \notin \mathcal{C}$. Then by reducibility, there exists a $P' \in \mathcal{I}$ for the first GDAG, which marginalizes to P . Suppose $P' \in \mathcal{C}$ for the first GDAG. Then by a second application of reducibility, it marginalizes to a distribution in \mathcal{C} for the second GDAG. But we already said it marginalizes to $P \notin \mathcal{C}$. Hence $P' \notin \mathcal{C}$. We conclude that if a GDAG has $\mathcal{C} \subsetneq \mathcal{I}$ then so does any other GDAG that reduces to it.

Except for items 1, 4 and 5 of appendix D.1, the reduction rules don’t affect the observed nodes and so the marginalization step in the definition of reducibility is irrelevant. For reductions that don’t use those three rules, we therefore have the stronger statement that \mathcal{C} is the same for both GDAGs, and so is \mathcal{I} . In particular $\mathcal{C} \subsetneq \mathcal{I}$ for one GDAG if and only if $\mathcal{C} \subsetneq \mathcal{I}$ for the other.

D.4. *d*-separation without a trivial variable

The following is needed to ensure that transformation item 4 of section D.1 satisfies the part of the definition of strong reducibility relating to \mathcal{I} . Given a GDAG with observed and unobserved nodes, suppose that a distribution P over variables on the observed nodes satisfies all the conditional independences implied by *d*-separation, i.e. $P \in \mathcal{I}$. Suppose further that some variables F always takes a fixed value. Then we claim that P also satisfies all the conditional independences implied by the GDAG with F removed.

Suppose that X and Y are *d*-separated by Z in the new GDAG but not the old. If we imagine removing the edges incident to F one by one, starting with outgoing edges and then moving on to incoming edges, then there must be a ‘critical edge’ wherein X and Y are not *d*-separated by Z before the removal, but are *d*-separated afterwards. Therefore all the unblocked paths before the removal must have passed through the critical edge.

Consider first an outgoing critical edge. Then X and Y are *d*-separated by ZF , because F blocks any otherwise unblocked path from X to Y . That means that $X \perp\!\!\!\perp Y \mid ZF$. But if F takes a fixed value then conditioning on it doesn’t do anything, so $X \perp\!\!\!\perp Y \mid Z$ as required.

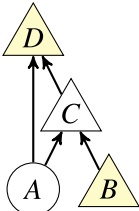
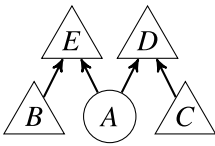
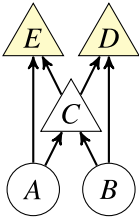
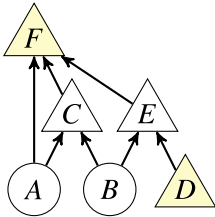
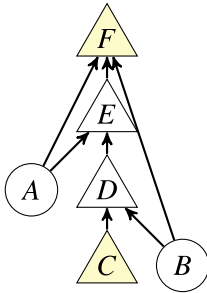
The other case is an incoming critical edge. By construction all the outgoing edges have already been removed, so all the unblocked paths from X to Y are head-to-head at F . If we write $Z = Z_D Z_{ND}$, where Z_D are descendants of F and Z_{ND} are not, then X and Y are *d*-separated by Z_{ND} and so $X \perp\!\!\!\perp Y \mid Z_{ND}$. Furthermore any path from XY to Z_D not blocked by Z_{ND} passes through F , and so $Z_D \perp\!\!\!\perp XY \mid Z_{ND}F$. As before this implies that $Z_D \perp\!\!\!\perp XY \mid Z_{ND}$. By the decomposition property of conditional independences we have $Z_D \perp\!\!\!\perp X \mid YZ_{ND}$ and hence $X \perp\!\!\!\perp Z_D \mid YZ_{ND}$ by the symmetry property. Combining this with $X \perp\!\!\!\perp Y \mid Z_{ND}$ using the contraction property gives $X \perp\!\!\!\perp Y \mid Z_D Z_{ND} = X \perp\!\!\!\perp Y \mid Z$ as required.

Appendix E. Small ‘interesting’ GDAGs

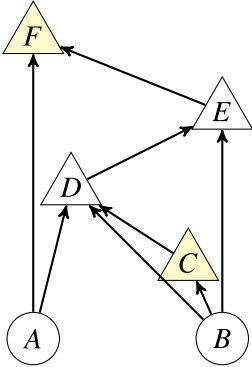
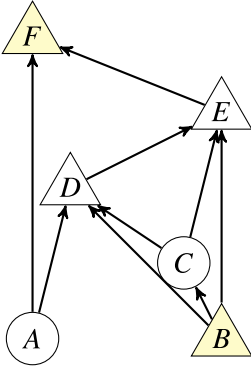
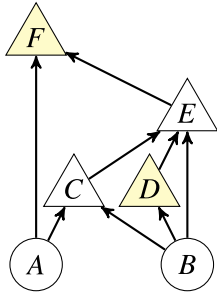
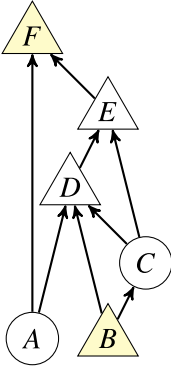
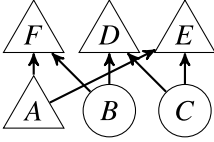
Here we present the all the GDAGs of size at most six which the criteria in section 5 does not identify as having $C = \mathcal{I}$, and the reduction criteria above do not identify as being reducible to a smaller such GDAG. If all these GDAGs have $C \subsetneq \mathcal{I}$ then our criteria is also necessary for $C = \mathcal{I}$, at least for GDAGs of this size.

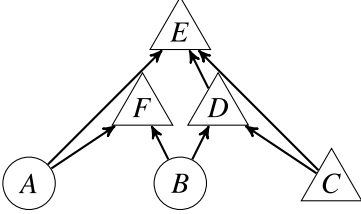
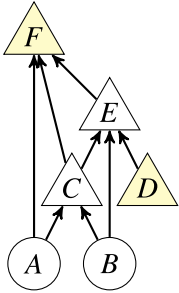
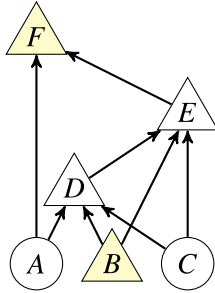
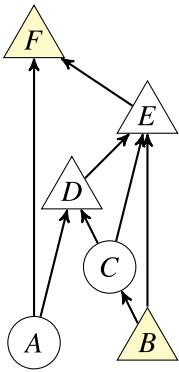
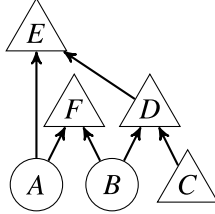
As well as the GDAG itself, we list a generating set of observable independences, which defines \mathcal{I} . We also list a generating set of Shannon-type inequalities for C , excluding those that are Shannon-type inequalities for \mathcal{I} .

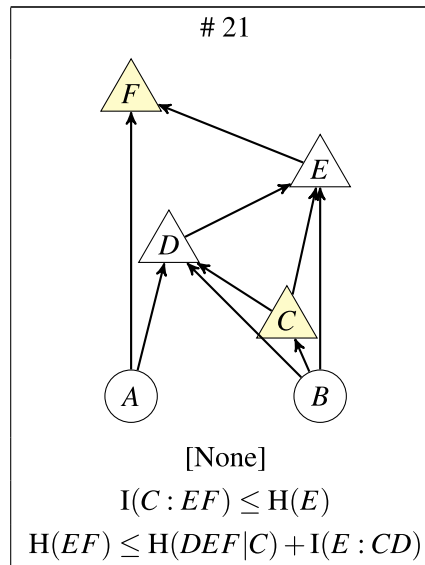
These inequalities provide good evidence that $C \subsetneq \mathcal{I}$. However, technically these inequalities could be non-Shannon inequalities for \mathcal{I} . For most of the GDAGs, we have highlighted a subset of the nodes. The probability distribution defined by perfectly correlated random bits on these nodes, with all other nodes taking a fixed value, is a member of \mathcal{I} yet violates the first entropic inequality listed and is therefore not in C . This closes the non-Shannon ‘loophole’ and establishes that $C \subsetneq \mathcal{I}$ for these GDAGs.

<p>Id</p> <p>GDAG</p> <p>Generating observable independences Further Shannon-type inequalities for \mathcal{C}</p>	<p># 1 (see section 4.2)</p>  <p>[None]</p> <p>$I(B : CD) \leq H(C)$</p>
<p># 2</p>  <p>$B \perp\!\!\!\perp CD, C \perp\!\!\!\perp BE$</p> <p>[None]</p>	<p># 3 (studied in [13])</p>  <p>[None]</p> <p>$I(D : E C) \leq H(C)$</p>
<p># 4</p>  <p>$C \perp\!\!\!\perp D$</p> <p>$I(CEF : D) \leq H(E C)$</p> <p>$H(F CE) \leq H(CF DE)$</p>	<p># 5</p>  <p>$C \perp\!\!\!\perp E \mid D$</p> <p>$I(C : EF) \leq H(E)$</p> <p>$I(C : DEF) \leq H(D)$</p>

<p># 6</p> <p>[None]</p> $I(A : EF) \leq H(E)$ $H(EF) \leq H(DEF A) + I(E : AD)$	<p># 7</p> <p>$C \perp\!\!\!\perp D$</p> $I(CEF : D) \leq H(E C)$ $H(EF) \leq H(CEF D) + I(E : CD)$
<p># 8 (see section 4.1)</p> <p>[None]</p> $I(D : F) + I(E : F) \leq H(F), \text{ and 3 permutations}$ $2(I(D : E : F) + I(D : E) + I(D : F) + I(E : F)) \leq H(D) + H(E) + H(F)$ $I(D : E : F) + I(D : E) + I(D : F) + I(E : F) \leq H(DE), \text{ and 3 permutations}$	
<p># 9</p> <p>[None]</p> $I(E : F D) \leq H(D)$ $I(E : CF D) \leq H(C)$	<p># 10</p> <p>[None]</p> $I(F : DE) \leq H(D)$ $H(DE) \leq H(CDE F) + I(D : CF)$

<p># 11</p>  <p>[None]</p> $I(C : EF) \leq H(E)$ $H(EF) \leq H(DEF C) + I(E : CD)$	<p># 12</p>  <p>[None]</p> $I(B : EF) \leq H(E)$ $H(EF) \leq H(DEF B) + I(E : BD)$
<p># 13</p>  <p>[None]</p> $I(D : EF) \leq H(E)$ $H(EF) \leq H(CEF D) + I(E : CD)$	<p># 14</p>  <p>[None]</p> $I(B : EF) \leq H(E)$ $H(EF) \leq H(DEF B) + I(E : BD)$
<p># 15</p>  <p>$A \perp\!\!\!\perp D, E \perp\!\!\!\perp F \mid A$</p> $I(D : E : F) \leq H(EF AD)$ $2I(D : E : F) + I(AE : D) + I(AF : D) \leq H(DEF A)$ $I(D : E : F) + I(AEF : D) \leq H(DF A), \text{ and } E \leftrightarrow F$	

<p># 16</p>  <p>$C \perp\!\!\!\perp F$</p> $H(C DE) + I(C:D) + I(C:E) + I(F:CDE) \leq H(CDF) + I(D:E F)$ $I(C:D) + I(C:E) + I(F:CDE) \leq H(DE) + I(D:E F)$	
<p># 17</p>  <p>$C \perp\!\!\!\perp D$</p> $I(D:CEF) \leq H(E C)$ $H(E CD) + I(D:CEF) \leq H(CE)$	<p># 18</p>  <p>[None]</p> $I(B:EF) \leq H(E)$ $H(EF) \leq H(DEF B) + I(E:BD)$
<p># 19</p>  <p>[None]</p> $I(B:EF) \leq H(E)$ $H(EF) \leq H(DEF B) + I(E:BD)$	<p># 20</p>  <p>$C \perp\!\!\!\perp F, C \perp\!\!\!\perp E D$</p> $I(C:DEF) \leq H(D F)$



References

- [1] Abramsky S and Coecke B 2004 A categorical semantics of quantum protocols *Proc. 19th Annual IEEE Symp. on Logic in Computer Science* pp 415–25
- [2] Araújo M *et al* 2013 All noncontextuality inequalities for the n -cycle scenario *Phys. Rev. A* **88** 022118
- [3] Bancal J-D *et al* 2012 Quantum non-locality based on finite-speed causal influences leads to superluminal signalling *Nat. Phys.* **8** 867–70
- [4] Barrett J 2007 Information processing in generalized probabilistic theories *Phys. Rev. A* **75** 032304
- [5] Bell J S 1964 On the Einstein–Podolsky–Rosen paradox *Physics* **1** 195–200
- [6] Blute R, Ivanov I and Panangaden P 2003 Discrete quantum causal dynamics *Int. J. Theor. Phys.* **42** 2025–41
- [7] Branciard C, Gisin N and Pironio S 2010 Characterizing the nonlocal correlations created via entanglement swapping *Phys. Rev. Lett.* **104** 170401
- [8] Branciard C *et al* 2012 Bilocal versus nonbilocal correlations in entanglement-swapping experiments *Phys. Rev. A* **85** 032119
- [9] Chaves R, Majenz C and Gross D 2014 Information-theoretic implications of quantum causal structures arXiv:1407.3800
- [10] Chaves R, Luft L and Gross D 2014 Causal structures from entropic information: geometry and novel scenarios *New J. Phys.* **16** 043001
- [11] Chiribella G, D’Ariano G and Perinotti P 2010 Probabilistic theories with purification *Phys. Rev. A* **81** 062348
- [12] Cirelson B S 1980 Quantum generalizations of Bell’s inequality *Lett. Math. Phys.* **4** 93–100
- [13] Evans R 2012 Graphical methods for inequality constraints in marginalized DAGs 2012 *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)* pp 1–6
- [14] Fine A 1982 Hidden variables, joint probability, and the Bell inequalities *Phys. Rev. Lett.* **48** 291–5
- [15] Fritz T and Chaves R 2013 Entropic inequalities and marginal problems *IEEE Trans. Inf. Theory* **59** 803–17
- [16] Fritz T 2012 Beyond Bell’s theorem: correlation scenarios *New J. Phys.* **14** 103001
- [17] Fritz T 2014 Beyond Bell’s theorem II: scenarios with arbitrary causal structure arXiv:1404.4812
- [18] Geiger D 1987 Towards the formalization of informational dependencies *UCLA Computer Science Technical Report 880053* (<http://fmdb.cs.ucla.edu/Treports/880053.pdf>)

- [19] Hardy L 2011 Foliable operational structures for general probabilistic theories *Deep Beauty: Understanding the Quantum World Through Mathematical Innovation* ed H Halvorson (Cambridge: Cambridge University Press)
- [20] Hardy L 2012 The operator tensor formulation of quantum theory *Phil. Trans. R. Soc. A* **370** 3385–417
- [21] Laskey K B 2007 Quantum causal networks *Proc. AAAI Spring Symp. on Quantum Interaction* (Menlo Park, CA: AAAI Press) p 142
- [22] Lauritzen S L *et al* 1990 Independence properties of directed markov fields *Networks* **20** 491–505
- [23] Leifer M S and Spekkens R W 2013 Towards a formulation of quantum theory as a causally neutral theory of bayesian inference *Phys. Rev. A* **88** 052130
- [24] Leifer M and Poulin D 2008 Quantum graphical models and belief propagation *Ann. Phys.* **323** 1899–946
- [25] Meek C 1995 Strong completeness and faithfulness in Bayesian networks *Proc. 11th Annual Conf. on Uncertainty in Artificial Intelligence (UAI-95)* (San Francisco, CA: Morgan Kaufmann) pp 411–8
- [26] Nielsen M A and Chuang I L 2000 *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press)
- [27] Pearl J 2009 *Causality, Models, Reasoning, and Inference* 2nd edn (Cambridge: Cambridge University Press)
- [28] Pearl J 1995 On the testability of causal models with latent and instrumental variables *Proc. 11th Annual Conf. on Uncertainty in Artificial Intelligence (UAI-95)* (San Francisco, CA: Morgan Kaufmann) pp 435–43
- [29] Pearl J 1996 The art and science of cause and effect *UCLA 81st Faculty Research Lecture Series* Available at (http://singapore.cs.ucla.edu/LECTURE/lecture_sec1.htm), and as epilogue of [27]
- [30] Pienaar J and Brukner C 2014 A graph-separation theorem for quantum causal models arXiv:1406.0430
- [31] Popescu S 1995 Bell’s inequalities and density matrices: revealing ‘hidden’ nonlocality *Phys. Rev. Lett.* **74** 2619–22
- [32] Popescu S and Rohrlich D 1994 Quantum nonlocality as an axiom *Found. Phys.* **24** 379–85
- [33] Shimony A 2013 Bell’s theorem *The Stanford Encyclopedia of Philosophy* ed E N Zalta (<http://plato.stanford.edu/archives/win2013/entries/bell-theorem/>)
- [34] Steudel B and Ay N 2010 Information-theoretic inference of common ancestors arXiv:1010.5720
- [35] Studený M 1988 Complexity of structural models *Prague Stochastics ’98: 13th Prague Conf. on Information Theory, Statistical Decision Functions and Random Processes* pp 523–8
- [36] Tucci R R 1995 Quantum Bayesian nets *Int. J. Mod. Phys. B* **9** 295–337
- [37] Verma T and Pearl J 1988 Causal networks: semantics and expressiveness *Proc. 4th Workshop on Uncertainty in Artificial Intelligence (Minneapolis, MN and Mountain View, CA)* pp 352–9
- [38] Wood C J and Spekkens R W 2012 The lesson of causal discovery algorithms for quantum correlations: causal explanations of Bell-inequality violations require fine-tuning arXiv:1208.4119