Indistinguishability of causal relations from limited marginals

Costantino Budroni, Nikolai Miklin, and Rafael Chaves

Phys. Rev. A **94**, 042127 — Published 25 October 2016

DOI: 10.1103/PhysRevA.94.042127

Indistinguishability of causal relations from limited marginals

Costantino Budroni,[1] Nikolai Miklin,[1] and Rafael Chaves[2,3]

[1] Naturwissenschaftlich-Technische Fakultät, Universität Siegen, Walter-Flex-Str. 3, 57068 Siegen, Germany
[2] International Institute of Physics, Federal University of Rio Grande do Norte, 59070-405 Natal, Brazil
[3] Institute for Theoretical Physics, University of Cologne, 50937 Cologne, Germany

We investigate the possibility of distinguishing among different causal relations starting from a limited set of marginals. Our main tool is the notion of adhesivity, that is, the extension of probability or entropies defined only on subsets of variables, which provides additional independence constraints among them. Our results provide a criterion for recognizing which causal structures are indistinguishable when only limited marginal information is accessible. Furthermore, the existence of such extensions greatly simplify the characterization of a marginal scenario, a result that facilitates the derivation of new Bell inequalities both in the probabilistic and entropic frameworks, and the identification of marginal scenarios where classical, quantum, and postquantum probabilities coincide.

I. INTRODUCTION

Deciding global properties of a given object from partial information is a problem often encountered in the most diverse fields. Just to cite a few examples, one can mention: knowledge integration of expert systems [1], database theory [2], causal discovery [3, 4] and the phenomenon of quantum nonlocality [5, 6]. More formally, in all such applications we are facing the so-called marginal problem: deciding whether a given set of marginal probability distributions for some random variables arises from a joint distribution of all these variables. Naturally, extensions to the quantum realm are also known: the quantum marginal problem [7] asks whether marginal density operators describing physical systems have a global extension, a fundamental question in quantum information with the most diverse applications [8, 9].

The reasons for our partial/marginal knowledge of a given system will intrinsically depend on the context and may have a variety of fundamental or practical reasons. For instance, in quantum mechanics incompatible observables cannot be perfectly and jointly measured. In turn, artificial intelligence systems gather information from several sources, each providing partial (typically overlapping) information about the global system [10].

In this paper, we will be mainly interested in the marginal problem arising in causal discovery [3, 4] where the aim is to decide, based on empirical observations, which underlying causal structures can explain our data. In particular, we provide a general result connecting the limited information available, i.e., the marginals, and the set of causal structures that can be distinguished on the basis of such information.

Importantly, causal discovery includes the test of local hidden variable (LHV) models that play a fundamental role in the foundations of quantum mechanics, in particular in Bell's theorem [5] and its practical applications in the processing of information [6]. Not surprisingly, given the importance and wide breadth of applications of the marginal problem within causal inference [3, 4] and related fields [11–14], several approaches have been formulated in order to solve it. In the following, we briefly describe some of the leading approaches.

In the absence of hidden variables –that is, all variables composing a given causal structure are empirically available– causal discovery can be faithfully performed based on the set of conditional independencies (CIs) implied by the model under test [3, 4]. However, causal models with hidden variables imply highly non-trivial constraints on the level of the observed distributions [15–18], constituting still a very active field of research [19–22]. In fact, from the causal perspective, Bell's theorem can be understood as a particular class of a causal discovery problem, where the underlying causal assumptions are those of local causality and measurement independence [23, 24]. Furthermore, as realized by Pitowsky [25, 26], the set of probability distributions compatible with such causal assumptions defines a convex set, more precisely a polytope, which facets are exactly the famous Bell inequalities.

The problem with this probabilistic approach, the most used in the study of Bell nonlocality, is that its useful linear convexity property (allowing the derivation of Bell inequalities via efficient linear programs) does not generalize to more complicated causal structures [18]. In the general case, one has to resort to algebraic geometry tools [27] that, at least in principle, are able to solve the marginal problem for arbitrary causal structures [16]. However, in practice, because of its high computational complexity, its use in causal inference is restricted to very few cases of interest [18, 20]. Thus, alternative approaches have also been pursued.

Instead of looking at the constraints imposed on the level of probabilities, Braunstein and Caves [28] asked what are the LHV constraints on the level of the Shannon entropies of these probabilities, which directly lead to the concept of entropic Bell inequalities. Apart from its fundamental relevance from a information-theoretic point of view [29–35], the entropic approach stands as a meaningful alternative for the fact that it provides a much more convenient route for the study of complex causal network beyond LHV models [36–38] and for extensions of causal discovery to the case of quantum causal structures [39–44].

The entropic approach to the marginal problem and causal discovery has received growing attention [45–56], but still suffers from two main drawbacks. The first stands from a computational complexity issue. The region of compatible entropies characterizing a given causal structure form a convex set, thus enourmously simplying the problem as compared to the highly non-trivial non-convex sets appearing in the algebraic geometry approach. In spite of that, the derivation of entropic Bell inequalities mostly rely on the elimination of variables from a system of linear inequalities via the so-called Fourier-Motzkin elimination algorithim [57] that has a double-exponential computational complexity, which limits its use to very few cases of interest. The second issue arises from the fact that this approach relies on an outer approximation of the region defining valid entropies for a collection of variables [58]. Finding better approximations to the entropy cone is a very active field of research in both classical [58] and quantum information theory [59, 60], but to our knowledge very scarce results [43] are known about the implications of it for marginal problems, that is, the projection of the entropy cone on the subspace defined by the empirically observable variables.

Within this context, the goal of this paper is to propose a new way for characterizing the correlations compatible with a given causal structure. With that aim, we work out the consequences of the the so-called adhesivity property [61–63] in the context of marginal scenarios, and what are its implications for the distinguishability between generic causal structures. Furthermore, by applying the general algorithm to particular cases of interest, we show two by-product advantages of our approach: i) it provides a faster computational algorithm and ii) in some cases provides a better approximation to the true marginal set of correlations characterizing a given causal model.

II. SUMMARY OF THE RESULTS

Given the amount of preliminary notions needed to understand our main results and the length of the associated sections, it is convenient to first give an informal summary of such results. The main problem we address is the possibility of distinguishing different causal structures associated with a set of random variables starting from limited information, i.e., limited marginals of their probability distribution. In Sect. IV, using the notion of adhesivity we identify which causal structures are always consistent with a given marginal scenario (Th. 2). These causal structures are defined as Markov random fields, starting from a graph-theoretic procedure (triangulation) applied to the marginal scenario (hyper)graph.

This result is then used in Sec. VI to prove the main theorem (Th. 4) on the relation between causal structures and marginal scenarios. In simple terms, Th. 4 identifies which causal structures can be falsified, i.e., proven to be inconsistent with a given set of marginals. This identification is done on the basis of the independence relations associated with a causal structure and the corresponding Markov random fields associated with the marginal scenario.

An immediate application of these results is in the characterization of which causal structures and marginals scenarios can lead to a different set of allowed correlations depending on the classical, quantum or even post-quantum nature of the underlying process; an important step in the generalization of Bell's theorem to more complex cases [41, 44] and in the understanding of quantum correlations via informational principles such as information causality [29]. For instance, as a consequence of (Th. 2) it follows that if the marginal scenario corresponds to an acyclic hypergraph, every probability distribution is compatible with a classical description, thus precluding the possibility of observing quantum nonlocality in such cases. A similar conclusion holds for scenarios satisfying the condition in case i) of Th. 4.

In addition, Th. 2 is also used to improve the characterization and approximation of entropy cones and correlation polytopes associated with a marginal scenario, under no assumption of a causal structure (Obs. 1), thus facilitating the derivation of new Bell inequalities. Th. 4 also takes into account these methods and explains when they can be applied to causal structures.

The paper is organized as follows. In Sect. III, we provide a detailed survey of all concepts and tools required for the understanding of the paper. More precisely, in Sect. III A, we describe hypergraphs and their properties; in III B, we have a brief account of causal structures; in III C, we introduce the notion of a marginal scenario and finally in Sects. III D and III E, we review the tools provided by correlation polytopes and entropic cones, respectively. In Sect. IV, we start describing some of the original results in this paper, namely, which causal relations are always consistent with a given marginal scenario. In Sect. V, we show how the notion of adhesivity can be used to compute the Bell inequalities of a given marginal scenario using the associated minimal hypergraph, which in some cases can greatly reduce the computational complexity of the problem. In Sect. VI, we prove general results concerning the indistinguishability of causal structures while in Sect. VII, we put our general approach to analyze a few cases of interest. In Sect. VIII, we summarize our findings and discuss interesting future directions of research.

III. PRELIMINARY NOTIONS

A. Graphs and hypergraphs

A hypergraph $\mathcal{H} = (\mathcal{N}, \mathcal{E})$ is defined by a finite set of nodes $\mathcal{N} = \{1, \ldots, n\}$ and a set of (hyper)edges corresponding to subsets of $\mathcal{N}$, i.e., $\mathcal{E} \subset 2^{\mathcal{N}}$. A graph $\mathcal{G}$ is a special case of an hypergraph where edges have cardinality 2, i.e., $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, with $|E| = 2$ for all $E \in \mathcal{E}$. A graph can also be directed, i.e., have directed edges corresponding to ordered pairs $(i, j) \in \mathcal{E}$, denoted by an arrow from i to j. In the following, by graph we will always mean a undirected graph, unless stated otherwise. See Fig. 1 for examples of graphs, directed graphs, hypergraphs, and additional notions discussed below.

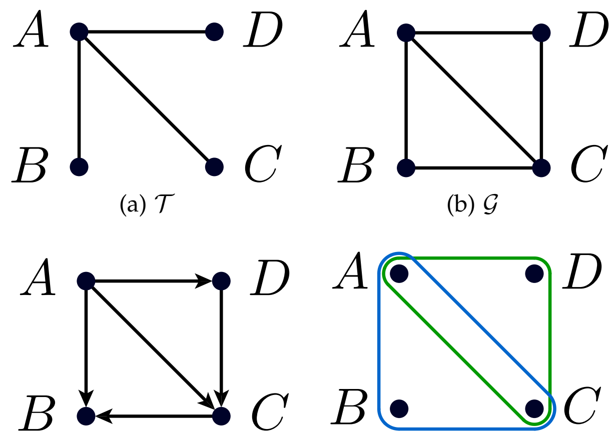
Since we will be interested only in hypergraphs without isolated nodes, we will assume that $\mathcal{N} = \cup_{E \in \mathcal{E}} E$, when not stated otherwise, and we will sometimes denote the hypergraph simply by the set of edges $\mathcal{E}$.

Paths, cycles, and acyclicity are fundamental notions in graph theory. A *path* is a sequence of distinct nodes $v_0, \ldots, v_n$ (except possibly the first and last) connected by edges (v_k, v_{k+1}) $k = 0, \ldots, n$, and a *closed path* or a *loop* is a path with first and last node coinciding, i.e., $v_0 = v_n$. For directed graphs, the definition is analogous with (v_k, v_{k+1}) representing a directed edge. *Acyclic graphs*, also called *tree graphs*, are graphs not containing loops. A graph is *connected* if, for every pair of nodes, there is a path connecting them.

A *clique* is a set of nodes $v_0, \ldots, v_n$ pairwise connected by an edge, i.e. $(v_i, v_j) \in \mathcal{E}$ for all $i, j = 1, \ldots, n$, $i \neq j$. Given a graph $\mathcal{G}$, we can construct a hypergraph from it, called the *clique hypergraph* $\mathcal{H}^{\rm cl}_{\mathcal{G}}$, with the same nodes and hyperedges in $\mathcal{H}^{\rm cl}_{\mathcal{G}}$ corresponding to cliques in $\mathcal{G}$. Similarly, a hypergraph $\mathcal{H} = (\mathcal{N}, \mathcal{E})$ can be transformed into a graph by constructing the *2-section* $[\mathcal{H}]_2$: we connect by edges in $\mathcal{G}$ all nodes that are connected by at least one hyperedge in $\mathcal{H}$. Notice that given a hypergraph $\mathcal{H}$, the clique graph of its 2-section will have, in general, extra hyperedges with respect to $\mathcal{H}$ (cf. Fig. 2).

A hypergraph $\mathcal{H} = (\mathcal{N}, \mathcal{E})$ is a *partial hypergraph* of $\mathcal{H}' = (\mathcal{N}, \mathcal{E}')$ if for any $E \in \mathcal{E}$ there exist $E' \in \mathcal{E}'$ such that $E \subset E'$. Equivalently, we will stay that $\mathcal{H}'$ *extends*, or is an *extension* of, $\mathcal{H}$ (cf. Fig. 2).

Given two disjoint subsets of nodes A, B they are said to be *separated* by a subset C if for each pair $a \in A, b \in B$, all the paths from a to b pass through C, i.e., if we remove C, A and B are no longer connected. In addition, C is called a minimal separator if $C \setminus \{v_i\}$ is no longer a separator for any $v_i \in C$.



(a) $\mathcal{T}$ (b) $\mathcal{G}$ (c) DAG (d) $\mathcal{H}$

FIG. 1. **Examples of graphs and hypergraphs**. (a) A tree graph $\mathcal{T}$ where $\{B, A, D\}$ is a path. (b) A graph $\mathcal{G}$ where $\{A, C, B, A\}$ is a loop and $\{A, B, C\}$ and $\{A, C, D\}$ are cliques. B and D are separated by $\{A, C\}$. (c) A directed graph where $\{A, C, B, A\}$ is not a directed path, because direction of (A, B) is not respected. This graph does not contain loops i.e. it is a directed acyclic graph (DAG). B and D are as well separated by $\{A, C\}$, but $\{C\}$ is a minimal separator. (d) A hypergraph $\mathcal{H}$ where nodes B and D are separated by $\{A, C\}$.

An important notion is also that of *triangulated*, or *chordal* graphs, namely, graphs for which every cycle $v_0, \ldots, v_n$ of length $n \geq 4$, contains a *chord*, i.e., an edge connecting (v_i, v_{i+2}). Given any graph, $\mathcal{G}$, additional edges can be added such that the obtained graph, $\mathcal{G}'$, is triangulated, and we will refer to $\mathcal{G}'$ as the *triangulation* of $\mathcal{G}$, see, e.g., Fig. 5 (b),(c).

For hypergraphs, the generalization of the notions of acyclicity and tree is not straightforward and several definition have been proposed (cf. Ref. [64]). For reasons that will be clear in Sect. IV, here we will focus on the notion of α-acyclicity, developed in the framework of database theory, which we will simply call acyclicity. There are several equivalent characterizations of this property (cf. Refs. [64, 65]), but we will focus on three of them: a characterization via the so-called *Graham algorithm*, one via the *running intersection property* of hyperedges, and the characterization as a clique hypergraph of a chordal graph.

Graham algorithm is defined as follows. Given a hypergraph described by hyperedges $\mathcal{E} = \{E_1, \ldots, E_n\}$, apply the following operations whenever they are possible

a) Delete a node i if it appears in exactly one hyperedge.

b) Delete a hyperedge E if $E \subset E'$.

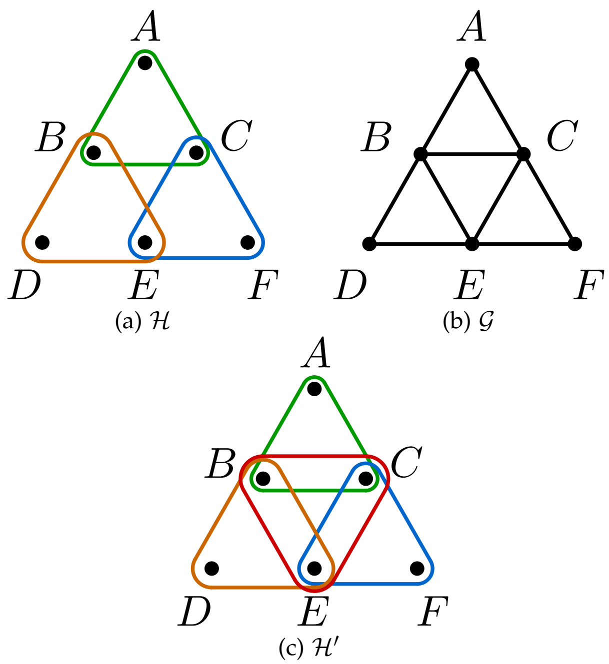


FIG. 2. **Example of graphs and hypergraphs**. (a) A cyclic hypergraph $\mathcal{H}$. (b) A 2-section graph $\mathcal{G} = [\mathcal{H}]_2$ of the hypergraph. (c) A clique hypergraph $\mathcal{H}'$ of $\mathcal{G}$. Notice that $\mathcal{H}' \neq \mathcal{H}$. In fact, $\mathcal{H}'$ is an extension of $\mathcal{H}$.

Acyclic hypergraphs are those for which Graham algorithm returns the empty set. Given a hypergraph $\mathcal{H}$, its *reduced hypergraph* is the hypergraph obtained by applying only operation b) of the Graham algorithm.

A hypergraph has the running intersection property if there exists an ordering of the edges, $E_1, \ldots, E_n$ such that

$$S_i := E_i \cap (E_1 \cup \ldots \cup E_{i-1}) \subset E_j, \text{ with } j < i. \quad (1)$$

In addition, for a connected and reduced hypergraph, the set $\{S_i\}$ corresponds to the *set of minimal separators* of the graph, i.e., S_i separates $R_i := E_i \setminus S_i$ from $(E_1 \cup \ldots \cup E_{i-1}) \setminus S_i$. It can be proven that the running intersection property is equivalent to the empty set output for the Graham algorithm, so it can be used as an alternative definition of an acyclic hypergraph (see, e.g., [65]).

The third equivalent property is defined in terms of graphs: a hypergraph is acyclic iff its hyperedges correspond to the set of cliques of a triangulated graph (see e.g., Ref. [65]).

In order to clarify the above notions, it is instructive to apply them to the simple example depicted in Fig. 2. For instance, we can apply the Graham algorithm to the hypergraph in Fig. 2 (a). By applying operation (a), we remove the nodes A, D, F and we are left with the edges $\{\{B,C\},\{B,E\},\{C,E\}\}$. At this point the algorithm stops, because we cannot remove any edge via operation (b), or any other node with operation (a). The hypergraph $\mathcal{H}$ is thus not acyclic. We can apply the same procedure to $\mathcal{H}'$: by removing the nodes A, D, F and we are left with the edges $\{\{B,C,E\},\{B,C\},\{B,E\},\{C,E\}\}$. We can then continue and remove the edges $\{B,C\},\{B,E\},\{C,E\}$ via operation (b), and finally the nodes B, E, C connected by a single edge $\{B,C,E\}$. The hypergraph $\mathcal{H}'$ is thus acyclic. Equivalently, one can see that the graph $\mathcal{G}$, obtained as the 2-section $[\mathcal{H}]_2 = [\mathcal{H}']_2$ has as cliques exactly the hyperedges of $\mathcal{H}'$, but not those of $\mathcal{H}$. Finally, for the running intersection property of $\mathcal{H}'$, we can choose the ordering $E_1 = \{B,C,E\}$, and for E_2, E_3, E_4 any ordering of the remaining edges. It is clear that any intersection of edges is contained in E_1. One can also straightforwardly check that $\mathcal{H}$ does not have the running intersection property. Similarly, one can easily check the property of separators of the sets $S_i = E_i \cap (E_1 \cup \ldots \cup E_{i-1})$, $i = 2, 3, 4$, for the hypergraph $\mathcal{H}'$.

B. Causal structures

For a set of random variables $X_1, \ldots, X_n$ some additional logical/causal constraints may apply. Usually, these constraints correspond to nonlinear constraints for probabilities (e.g., factorization properties) and to linear constraints for entropies (e.g., vanishing of the mutual information). Two common examples are given below:

- *Deterministic dependence:* The variable X_i is said to be a deterministic function of X_j if their joint probability distribution satisfy $P(x_i, x_j) = \delta_{x_i, F(x_j)} P(x_j)$, where the deterministic dependence is given by $x_i = F(x_j)$. Such type of constraints are usually present in network coding [58]. While not strictly necessary, deterministic constraints also play an important role in the derivation of Bell inequalities [66].

- *Conditional or unconditional independence:* The variable X_i is said to be independent of the variable X_j, if the joint probability distribution satisfy $P(x_i, x_j) = P(x_i)P(x_j)$. Similarly, variable X_i is said to be conditionally independent (CI) of X_j given X_k, if $P(x_i, x_j, x_k) = P(x_i|x_k)P(x_j|x_k)P(x_k)$, that is, X_k screens off the correlations between the two other variables. We will denote the two situations as $(X_i \perp X_j)$ and $(X_i \perp X_j | X_k)$, respectively.

Upper-case letters (e.g., X) denote the random variables, and lower-case letters (e.g., x) the specific value they assume.

As we will see next, conditional and unconditional independence play a crucial role in the study of Bayesian networks and Markov random fields [65], this is why we will focus on them in the remaining of the paper.

1. *Bayesian networks*

A Bayesian network (BN) is a probabilistic model for which conditional dependencies can be represented via a DAG. More precisely, the probability distribution factorizes as

$$P(x_1,\dots,x_n) = \prod_{i=1}^{n} P(x_i|\mathrm{Pa}_i), \tag{2}$$

where Pa_i denotes the *parents* of the node i, i.e., the nodes with arrows pointing at i. The above factorization of the probability distribution gives rise to the *local Markov property*

$$(X_i \perp \mathrm{Nd}_i|\mathrm{Pa}_i), \tag{3}$$

namely that X_i is independent of its *nondescendants* Nd_i, i.e., nodes reachable from X_i via a directed path, given its parents.

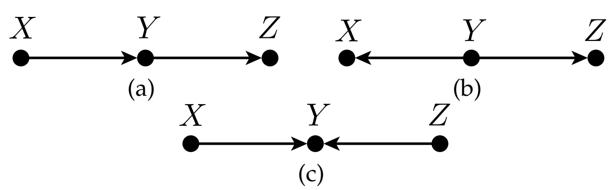


FIG. 3. **Examples of different Bayesian networks**. (a) A DAG representing a Markov chain $X \to Y \to Z$ implying the CI $(X \perp Z|Y)$. (b) A DAG where the variable Y is a common parent of X and Z, once more implying the CI $(X \perp Z|Y)$. (c) A DAG where the variables X and Z have a common child Y. In this case $(X \perp Z)$, but $(X \not\perp Z|Y)$.

More generally, one has the set of conditional independence relations

$$\mathcal{I}(\mathcal{G}) = \{(X_A \perp X_B|X_C) \mid \mathrm{dsep}_{\mathcal{G}}(A:B|C)\}, \tag{4}$$

where $\mathrm{dsep}_{\mathcal{G}}(A:B|C)$ refers to the d-separation properties of nodes in A and B with respect to nodes in C, namely that every path from $a \in A$ to $b \in B$, or vice versa, is *blocked* by a node in C. The path is said to be blocked if it contains one of the following: $x \to c \to y$, or $x \leftarrow c \leftarrow y$, or $x \to z \leftarrow y$, for x, y, z, c in the path, $c \in C$, $z \notin C$ and no descendant of z is in C (cf. Ref. [3]).

Bayesian networks are of particular relevance to formalize causal relations. Within this context, such causal models have been called causal Bayesian networks [3], as opposed to traditional Bayesian networks that formalize conditional independence relations without having necessarily a causal interpretation. To exemplify, consider the three DAGs shown in Fig. 3. DAGs (a) and (b) clearly imply a different set of causal relations between variables X, Y and Z: in both cases the correlations between X and Z are mediated via Y but in (a) X is a parent of Y while in (b) the reverse is true. In spite of their clear causal differences, both causal models imply the same set of CIs, namely that $(X \perp Z|Y)$. That is, every observable probability distribution $p(x,y,z)$ compatible with (a) is also compatible with (b), thus both models are indistinguishable from observations alone [67]. The DAG (c) in Fig. 3 can nonetheless be distinguished from (a) and (b), since it implies that $(X \not\perp Z|Y)$ and the only CI is given by the independence constraint $(X \perp Z)$.

2. *Markov random fields*

Similarly to Bayesian networks, Markov random fields (MRF) correspond to probabilistic models for which conditional dependencies can be represented by a graph $\mathcal{G}$. In this case, the graph is undirected and it may contain cycles. More precisely, independence relations are given by the *global Markov property*

$$(X_A \perp X_B|X_C) \tag{5}$$

if every path from a node in A to a node in B passes through a node in C, i.e., if C is a separator for A and B in $\mathcal{G}$, a fact denoted as $\mathrm{sep}_{\mathcal{G}}(A:B|C)\}$. We will denote the corresponding set of independence relations as

$$\mathcal{I}(\mathcal{G}) = \{(X_A \perp X_B|X_C) \mid \mathrm{sep}_{\mathcal{G}}(A:B|C)\} \tag{6}$$

As opposed to Bayesian networs (cf. Eq. (2)), MRFs do not admit a unique factorization of the probability distribution. However, for the special case of a triangulated graph, denoting with $C_1,\dots,C_k$ the set of maximal cliques with the running intersection property and $S_i := C_i \cap (C_1 \cup \dots \cup C_{i-1})$, as in Eq. (1), one can write

$$P(x_1,\dots,x_n) = \prod_{i=1}^{k} \frac{P(x_{C_i})}{P(x_{S_i})}. \tag{7}$$

Bayesian networks can be described also via MRF, via the so called *moral graph* (cf. Ref. [65]), however this comes at the price of losing some of the original independence constraints described by the DAG.

C. Marginal scenarios

In many relevant situations, one may have only partial information of the distribution of the variables. This may due to practical limitations in collecting data, e.g., latent variables which cannot be measured, or fundamental limitation, e.g., the impossibility of performing

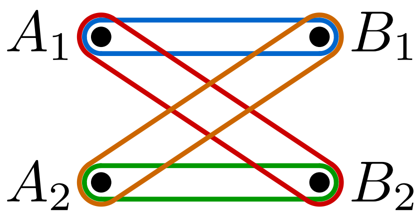


FIG. 4. Hypergraph of the marginal scenario associated with a Bell-CHSH experiment. The observed probabilities correspond to the marginal for $\{A_x, B_y\}$, for $x, y = 1, 2$.

a joint measurement of incompatible quantum observables. This is common in Bell and noncontextuality experiments [6, 68], where one has access only to a limited set of joint probability distributions. Also in purely classical contexts the role of partial information can hardly be overemphasized [3, 13]. For instance, in the so called instrumentality tests modeling randomized experiments [3, 11], the effects from a drug in the recovery of patients is allowed to depend on some unobserved factors that are not under experimental control (social or economical background, etc).

For this reason, we introduce the the notion of marginal scenario. Given a set of random variables $X = \{X_1, \ldots, X_n\}$, a *marginal scenario* is a collection of subsets $\mathcal{M} = \{M_1, \ldots, M_{|\mathcal{M}|}\}$, $M_i \subset X$ of them representing variables that can be jointly measured, i.e., for each M_i, we have access to a probability distribution $P_{M_i}(x_{M_i})$. Moreover, if a set of variables are jointly measurable, we require that the same holds for any subset, i.e., $M \in \mathcal{M}$ and $M' \subset M$ imply $M' \in \mathcal{M}$. Equivalently, one can take only the maximal subsets $S \in \mathcal{M}$.

A marginal scenario can be naturally considered as an hypergraph, with $\mathcal{M}$ the set of hyperedges and $\cup_i M_i$ the set of nodes. We will adopt the maximal subsets convention above and assume that $S \in \mathcal{M}$ are only maximal subsets, i.e., the hypergraph is reduced. We will call such a hypergraph the *marginal scenario hypergraph*, or simply marginal scenario when it is clear we are referring to the hypergraph.

It is again instructive to consider a simple example to fix the above notions. A standard example is given by a Bell experiment [5], in particular by the Clauser-Horne-Shimony-Holt (CHSH) scenario [69]. Two parties, Alice and Bob, perform measurements on their part of a shared quantum system. They can perform one of two measurements, labelled as A_1, A_2 for Alice and B_1, B_2 for Bob. The observed probabilities will then amount to the marginals for $\{A_x, B_y\}$. The corresponding marginal scenario hypergraph is depicted in Fig. 4.

D. Correlation polytopes and non-local correlations

In quantum information, one of the applications of the concept of a marginal scenario is exactly in the study of Bell's theorem and quantum nonlocality [5, 6]. The simplest Bell scenario is given by two distant (ideally space-like separated) parties, Alice and Bob, that at each round of their experiment receive particles produced by a common source. Upon receiving their shares of the physical system, they measure a given observable of it recording the measurement outcome. The measurements choices by Alice and Bob (that are assumed to be independent of how the system has been prepared) are labeled by the variables X and Y while their outcomes are given by A and B, respectively. Under the assumption of local realism, every observable probability distribution that can be obtained in such experiment can be decomposed as the so called local hidden variable (LHV) model

$$p(a_x, b_y) = p(a, b|x, y) = \sum_{\lambda} p(\lambda) p(a|x, \lambda) p(b|y, \lambda), \quad (8)$$

where the variable λ stands for a full description of the source producing the particles and any other mechanism that might affect the measurement outcomes.

As realized by Pitowski [25], the set of probability distributions compatible with the LHV model form a convex set of correlations, the so called correlation polytope. This polytope is characterized by finitely many extremal points, exactly those representing deterministic functions in the LHV decomposition (8). Equivalently, this polytope is characterized by finitely many linear inequalities, the non-trivial ones being exactly the Bell inequalities.

Different methods to characterize such correlation polytopes are available. Here, we will describe a particular method obtained via Fine's theorem [66] and that plays a fundamental role in the derivation of entropic Bell inequalities as we will see in Sect. III E.

Fine's theorem shows that the existence of a LHV model of the form (8) is equivalent to the existence of a joint probability distribution $p(a_1, \ldots, a_m, b_1, \ldots, b_m)$ that marginalizes to the observed distribution $p(a_x, b_y)$ with $x, y = 1, \ldots, m$. The existence of well defined joint distribution over all variables implies that such distribution must respect some constraints, namely, positivity and normalization. That is, $p(a_1, \ldots, a_m, b_1, \ldots, b_m)$ must lie inside a simplex polytope [70]. From this geometric perspective, the correlation polytope is nothing else than the projection of the simplex polytope – characterizing the joint distribution– to a subspace of it that is given by the marginal scenario in question, that is, a projection to the subspace spanned by the observable components $p(a_x, b_y)$ with $x, y = 1, \ldots, m$. Such a projection can be obtained by eliminating, from the corresponding system of linear inequalities describing the simplex, all terms that correspond to non-observables probabilities. This can be achieved, for example, via a standard algorithm known as Fourier-Motzkin elimination [57]. Once removed the redundant inequalities, the remaining set gives the facets of the correlations polytope, that is, tight Bell inequalities.

Unfortunately, such a nice linear convex picture does not hold for more complicated causal structures [19, 21, 39, 71] that now require computationally expensive and highly intractable methods from algebraic geometry [16] in order to deal with the non-linear constraints arising from such models. As mentioned before and explained in more details in the next section, this is one of the reasons why the entropic approach has become a more viable option in the study of complex causal structures. In short, polynomial constraints on the level of probabilities are turned into simpler linear relations in terms of entropies.

E. Entropic cone

Given a collection of n discrete random variables $X_1, \dots, X_n$ with an associated joint probability distribution $P(x_1, \dots, x_n)$, and denoting with X_S the random vector $(X_i)_{i\in S}$, for any subset $S \subset [n] := \{1, \dots, n\}$, the Shannon entropy $H : 2^{[n]} \to \mathbb{R}$ is defined as

$$H(S) := H(X_S) = -\sum_{x_S} P(x_S) \log_2 P(x_S). \tag{9}$$

The above entropies can be arranged in a vector $h = (H(\emptyset), H(X_1), \dots, H(X_1, X_2), \dots, H(X_1, \dots, X_n)) \in \mathbb{R}^{2^n}$. The region

$$\Gamma^*_{[n]} := \overline{\left\{ h \in \mathbb{R}^{2^n} \mid h = (H(S))_{S\subset[n]} \text{ for some entropy } H \right\}}, \tag{10}$$

where $\overline{}$ denotes the closure in $\mathbb{R}^{2^n}$, is known to be a convex cone, also called the *entropy cone* and it has been studied extensively in information theory [58]. A tight and explicit description, however, has not yet been found for $n > 3$, but only some outer approximations of $\Gamma^*_{[n]}$ via polyhedral cones, i.e., cones described by a finite system of linear inequalities $Ax \geq b$, where A is a $m \times n$ real matrix and b a m-dimensional real vector. The most famous outer approximation to the entropic cone is the so-called *Shannon cone* $\Gamma_{[n]}$, defined by

$$h([n] \setminus \{i\}) \leq h([n]), \tag{11a}$$
$$h(S) + h(S \cup \{i, j\}) \leq h(S \cup \{i\}) + h(S \cup \{j\}), \tag{11b}$$
$$h(\emptyset) = 0, \tag{11c}$$

for all $i, j \in [n]$, $i \neq j$, and $S \subset [n] \setminus \{i, j\}$. That is, the Shannon cone associated with n variables is described by $2^{n-2}\binom{n}{2} + n$ inequalities plus one equality constraint (normalization).

The above is the minimal set of inequalities that implies monotonicity of entropy, i.e., $H(A|B) \geq 0$, and the submodularity (or strong subadditivity), i.e., $I(A : B|C) := H(A, C) + H(B, C) - H(A, B, C) - H(C) \geq 0$, for any disjoint subsets $A, B, C \subset [n]$ (cf. Ref. [58]).

Inequalities in Eq. (11) are known as *Shannon-type inequalities* in information theory or *polymatroidal axioms* in combinatorial optimization [58]. Given a finite set N and real-valued function $f : 2^N \to \mathbb{R}$, the pair (N, f) is called a polymatroid if f satisfies Eqs. (11) above for $[n] = N$ and $S, \{i, j\} \subset N$.

Geometrically, the entropy cone $\Gamma_{\mathcal{M}}$ associated with the marginal scenario $\mathcal{M}$ corresponds to the projection of the entropic cone onto the subspace of the corresponding variables. Similarly to what happens to the projection of a simplex polytope into a correlation polytope, for a polyhedral cone, such a projection can be obtained by eliminating, from the corresponding system of linear inequalities, all terms that correspond to non-observables terms. After removing redundant inequalities, the remaining set gives facets of the entropic cone in the observable subspace.

Conditional independence constraints arising from BN or MRF, i.e., of the form $(X_A \perp X_B|X_C)$, corresponds to linear constraints on the vector of entropies, i.e., hyperplanes defined by the equation $I(A : B|C) = 0$. Linear constraints from causal structures not only reduce the dimension of the problem, but as well, when applied to polymatriods, reduce the number of axioms needed to describe the constrained cone. For instance, it was shown in Ref. [72], that if $I(A : B|C) = 0$, with A, B, C disjoint sets of variables, then from the set of polymatroid axioms in Eq. (11), the following inequalities are redundant

$$I(A : E|BC) \geq 0, \tag{12}$$
$$I(B : E|AC) \geq 0, \tag{13}$$

where $E \in [n] \setminus \{A, B\} \cap C$. Hence, in this case one can generate a compact representation of polymatroid axioms [72].

IV. ADHESIVITY AND INDEPENDENCE CONSTRAINTS ASSOCIATED WITH A MARGINAL SCENARIO

The main result of this section is that when one has only partial information about (i.e., only some marginals of) a probability distribution, such marginals are always consistent with a global distribution where additional independence constraints are imposed. We start introducing the notion of adhesivity and restating in our language a theorem by Vorob'ev [61] (Th. 1), then we connect this result with the notion of marginal scenario to prove which independence constraints are always compatible with a set of marginals (Th. 2).

On the one hand, such additional constraints simplify the characterization of the entropy cone and correlation polytope associated with the marginal scenario. On the other hand, this result allows us to identify which causal structures can be distinguished when we have access only to some restricted set of marginals.

The notion of adhesivity, albeit in different terms, has first been introduced for probability distributions [61, 62] and subsequently extended to entropies [63]. In the framework of Bell and noncontextuality inequalities, similar ideas have been investigated by several authors [36, 49, 73–75], but never in full generality.

A. Adhesivity of probabilities

Adhesivity can be explained in simple terms as follows. Given two sets of variables $X_I = (X_i)_{i\in I}$ and $X_J = (X_j)_{j\in J}$ and two probability distributions $p(x_I)$ and $p'(x_J)$ such that p and p' coincide on the variables $X_{I\cap J}$, we can define a probability distribution on $I \cup J$ as

$$P(x_{I\cup J}) = \begin{cases} 0 & \text{if } p(x_{I\cap J}) = 0, \\ \frac{p(x_I)p'(x_J)}{p(x_{I\cap J})} & \text{otherwise.} \end{cases} \tag{14}$$

One can easily check that this is a valid probability distribution on the set of variables in $I \cup J$.

The construction in Eq. (14) implies that every two marginals of a probability distribution are always consistent with a probability distribution conditionally independent on their intersection, i.e., $(X_{I\setminus J} \perp X_{J\setminus I} \mid X_{I\cap J})$, since $p(x_I)p'(x_J)/p(x_{I\cap J}) = p(x_I|x_{I\cap J})p'(x_J|x_{I\cap J})p(x_{I\cap J})$. We call such an extension of p and p' an *adhesive extension*.

Similarly, two polymatroid (N, h) and (M, g) coinciding on $N \cap M$ are said to *adhere* or to have an *adhesive extension* if there exist a polymatroid $(N \cup M, f)$ extending h, g, i.e., $f(I) = h(I)$ for $I \subset N$, $f(J) = h(J)$ for $J \subset M$, which is also *modular* on N and M, that is, $f(N \cup M) = f(N) + f(M) - f(N \cap M)$ or, equivalently, such that N and M are conditionally independent on the intersection $N \cap M$. As a consequence of the construction in Eq. (14) for probabilities, restrictions of entropies have an adhesive extension, whereas general polymatroids do not (cf. Ref. [63]).

This observation is at the basis of the derivation of several *non-Shannon* information inequalities, i.e., information inequalities that do not follow from Eqs. (11) (cf. Ref. [63]). Starting from the first non-Shannon inequality derived by Zhang and Yeung [76], infinitely many inequalities have been derived by Matúš [77], and several others authors investigated the problem [78–80].

B. Marginal scenarios admitting a global extension

From the adhesivity property of probability distributions, one can extend probabilities defined on a marginal scenario to a joint probability distribution over all variables, which satisfies extra conditional independence constraints that depends on the marginal scenario.

The theorem below has first been stated without proof by Vorob'ev in Ref. [61], and subsequently explicitly proven in Ref. [62], but also independently derived by other authors [81–83]. The original proof, however, used a quite different terminology. It is helpful to restate it in the language of hypergraphs, and to present a sketch of it, in order to understand the role of the adhesivity property.

Theorem 1. [Vorob'ev] *A set of probabilities associated with an acyclic marginal scenario hypergraph $\mathcal{M}$ admits a global extension to a single probability distribution. Moreover, the extension can be chosen as a MRF described by the 2-section graph $[\mathcal{M}]_2$.*

Sketch of the proof.– Let $\mathcal{M}$ be the marginal scenario hypergraph, by definition, it is a reduced hypergraph. If $\mathcal{M}$ is acyclic, we can find an ordering $M_1, \ldots, M_n$ of its hyperedges respecting the running intersection property. The construction of a global probability distribution can then be obtained by induction on n, the number of hyperedges. For $n = 1$, $P(M_1)$ is a valid probability distribution (to simplify the notation, we will use $P(M_i)$ as a shorthand for $P(x_{M_i})$, etc). We then apply the inductive hypothesis. Let us assume that for $n - 1$ $P(M_1 \cup \ldots \cup M_{n-1})$ is a valid probability distribution extending the marginals $P(M_i)$ for $1 \le i \le n - 1$. We want to extend it to $P(M_1 \cup \ldots \cup M_{n-1} \cup M_n)$. By the running intersection property, $M_n \cap (M_1 \cup \ldots \cup M_{n-1}) =: S_n \subset M_j$ for $j < n$. Denoting by P_{M_i} the marginal probability distribution on M_i, we define $R_n := M_n \setminus S_n$ and

$$P(R_n|S_n) := \frac{P_{M_n}(M_n)}{P_{M_j}(S_n)}, \tag{15}$$

defining $0/0$ to be zero as in Eq. (14), and for the joint distribution

$$\begin{aligned} P(M_1 \cup \ldots \cup M_{n-1} \cup M_n) &:= P(R_n|S_n) \\ &\quad P(M_1 \cup \ldots \cup M_{n-1} \setminus S_n|S_n)P(S_n). \end{aligned} \tag{16}$$

By the adhesivity property, this is a valid probability distribution, and its marginals coincide with $P(M_i)$ for $1 \le i \le n$, so it is an extension of the marginal scenario. In addition, it is modular over the intersection, i.e.

$$(R_n \perp (M_1 \cup \ldots \cup M_{n-1}) \setminus S_n | S_n) \tag{17}$$

Since $\mathcal{M}$ is connected and reduced, the set of minimal separators precisely corresponds to the set S_n above.

The modularities of the constructed distribution are thus precisely those implied by the MRF defined by $[\mathcal{M}]_2$. $\square$

In the next section, we will see the application of this result to general marginal scenario, i.e., not necessarily acyclic.

C. Maximal set of independence conditions associated with a marginal scenario

We will now see the implications of Vorob'ev's theorem on general marginal scenarios. More precisely, we will discuss which independence conditions, arising as MRF conditions, are consistent with a given marginal scenario and how to compute maximal sets of such conditions.

The main result is the following.

Theorem 2. *Given a joint probability distribution P on n variables $X_1, \dots, X_n$, and a marginal scenario $\mathcal{M}$, the marginals P_{M_i} for $M_i \in \mathcal{M}$ are consistent with a probability distribution arising from a MRF associated with the 2-section graph $[\mathcal{T}]_2$, where $\mathcal{T}$ is an acyclic hypergraph extending $\mathcal{M}$.*

Proof.– An acyclic hypergraph $\mathcal{T}$ extending $\mathcal{M}$ can always be found. It is the clique hypergraph of a triangulation of the graph $[\mathcal{M}]_2$. The marginals in $\mathcal{M}$ are consistent with the marginals in $\mathcal{T}$ extracted from the same probability distribution P. Since $\mathcal{T}$ is an acyclic hypergraph, we can apply the construction in Th. 1 to obtain a MRF with independence relations described by the 2-section graph $[\mathcal{T}]_2$. $\square$

For any given marginal scenario $\mathcal{M}$, an acyclic hypergraph extending it corresponds to the clique hypergraph of the triangulation of the 2-section $[\mathcal{M}]_2$, hence the maximum set of independence constraint will correspond to the triangulation with the minimum number of edges, also called the *minimum triangulation*. The problem of computing the minimum triangulation is known to be NP-hard [84]. However, it is much easier to calculate *a minimal triangulation*, namely, a triangulation such that by removing any edge the obtained graph is no longer a chordal graph. Notice that such a minimal triangulation is not necessarily the one with the smallest number of edges among all possible triangulations. Several algorithms have been developed to compute a minimal triangulation, which run in $O(n+m)$ steps, n being the number of nodes and m the number of edges of a graph [84].

In the following, we will adopt the above terminology also for hypergraphs, namely, we will speak about the minimum acyclic hypergraph extending $\mathcal{M}$, in the sense of the minimum triangulation, and a minimal hypergraph extending $\mathcal{M}$, in the sense of a minimal triangulation.

V. OPTIMAL CHARACTERIZATION OF THE MARGINAL SCENARIO FOR PROBABILITIES AND ENTROPIES

As a consequence of the above results, the characterization of a given marginal scenario $\mathcal{M}$, in terms of inequalities for the probability vector or entropy vector, can be computed from those associated with a minimal acyclic hypergraph extending $\mathcal{M}$. This approach offers advantages both for the probabilistic and the entropic approach. Here, we will give a brief summary of the two results, but later we mostly discuss the entropic approach.

A similar approach, albeit with a different terminology, and using a less general version of Th. 2, has been already used in relation with Bell and noncontextuality inequalities. For instance, the decomposition of the CHSH scenario in Fig. 5 was used in the proof of the necessity and sufficiency of Bell inequalities for the existence of a LHV model by Fine [66]. Special cases of Th. 2 have been discussed in Refs. [49, 73, 74], and their application to more general scenarios have been discussed for probabilities [75, 85] and for entropies [36].

A. Triangulation

The first step is to compute a minimal acyclic hypergraph $\mathcal{T}$ extending the marginal scenario $\mathcal{M}$. It can be done as follows:

a1) Compute the 2-section graph $[\mathcal{M}]_2$.

a2) Compute its minimal triangulation.

a3) Take as $\mathcal{T}$ the corresponding clique hypergraph.

In the following, we discuss the above procedure with a simple example. Consider the hypergraph with edges $\{\{A_x, B_y\}\}_{x,y=1,2}$ associated with a Bell experiment and discussed in Sect. III C. One starts with the marginal scenario hypergraph $\mathcal{M}$ of Fig. 5 (a) and computes its 2-section $[\mathcal{M}]_2$ [cf. Fig. 5 (b)]. In this simple case $[\mathcal{M}]_2$ can be triangulated by adding an extra edge connecting A_1 and A_2 [cf. Fig. 5 (c)], or equivalently, connecting B_1 and B_2. Finally, one takes as $\mathcal{T}$ the clique hypergraph of the triangulation [cf. Fig. 5 (d)]. The corresponding MRF independence condition consistent with the marginals in $\mathcal{M}$ is $(B_1 \perp B_2 | A_1, A_2)$.

B. Probabilities

Once $\mathcal{T}$ has been obtained, the probabilistic inequalities describing the marginals consistent with the given scenario $\mathcal{M}$ (describing a correlation polytope, see Sec. III D) can be computed as follows:

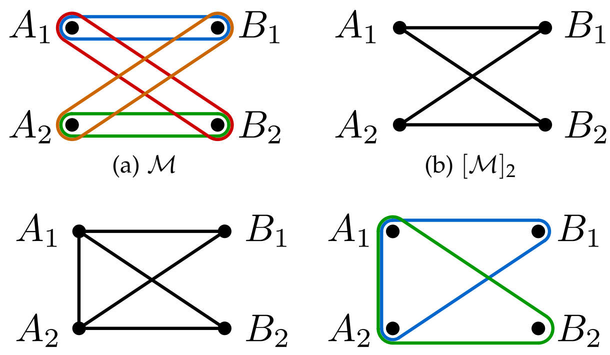


FIG. 5. **Procedure to compute the minimal hypergraph $\mathcal{T}$ extending the marginal scenario $\mathcal{M}$ for the CHSH scenario.** (a) Initial marginal scenario hypergraph corresponding to the CHSH case. (b) 2-section $[\mathcal{M}]_2$ of the original hypergraph. (c) Triangulation of the 2-section graph. (d) Clique hypergraph $\mathcal{T}$ of the triangulation. $\mathcal{T}$ extends $\mathcal{M}$.

b1) Write down the simplex inequalities associated with each maximal clique C_i, i.e., the inequalities corresponding to a classical probability on $|C_i|$ variables (cf. Ref. [26] and Sect. III D for further details on the simplex polytope).

b2) Project such inequalities onto the initial marginal scenario $\mathcal{M}$ (for instance, applying the Fourier-Motzkin elimination, see Sects. III E and III D).

C. Entropies

A similar approach can be applied for deriving entropic inequalities, but it only gives an outer approximation if an exact characterization of the entropy cone is not know (cf. Eq. (19) below).

An alternative approach can be summarized as follows:

c1) Compute the independence constraints $\mathcal{I}(\mathcal{T})$ associated with the MRF graph $[\mathcal{T}]_2$,

c2) Consider the Shannon cone on n variables Γ, with the reduced set of polymatroid axioms associated with $\mathcal{I}(\mathcal{T})$.

c3) Use the linear constraints associated with $\mathcal{I}(\mathcal{T})$, i.e., the vanishing of some conditional mutual information terms, for a partial projection of the full cone onto the marginal scenario. Then, complete the projection with the usual Fourier-Motzkin algorithm.

It is clear that the above approach can be adapted to any type of linear constraints, including those arising from some assumed causal structure, i.e., a BN or MRF, and those arising as deterministic dependence conditions, corresponding to the vanishing of some conditional entropy, discussed in Sect. III B.

In the next section, we will see in details what approaches are possible for characterizing entropic marginals.

D. Outer approximations of the entropy cone

The adhesivity property for restrictions of an entropic polymatroid can be used to obtain outer approximations of the entropy cone as follows.

Theorem 3. *Let $\mathcal{M}$ be the marginal scenario hypergraph and $\mathcal{T}$ an acyclic hypergraph extending it. Let us denote with $\Gamma^*_{\mathcal{T}}$ the entropy cone intersected with the linear constraints defined by $\mathcal{I}(\mathcal{T})$. Then, we have that*

$$\Pi_{\mathcal{M}}(\Gamma^*) = \Pi_{\mathcal{M}}(\Gamma^*_{\mathcal{T}}) \tag{18}$$

where $\Pi_{\mathcal{M}}$ denotes the projection onto the coordinates associated with the marginal scenario $\mathcal{M}$.

Proof.— The result, basically, follows from the fact that the marginals in $\mathcal{M}$ are consistent with the linear constraints $\mathcal{I}(\mathcal{T})$. Given an entropic polymatroid, it is sufficient to take the associated probability distribution and apply Th. 2. The obtained distribution will be consistent with the MRF $[\mathcal{T}]_2$, hence, the associated marginal entropies will be identical to the original ones and inside $\Pi_{\mathcal{M}}(\Gamma^*_{\mathcal{T}})$ by construction. The same construction can be combined with the limits necessary (see, e.g., Refs. [58, 86]) to obtain the closure of entropic polymatroid Γ^*. □

Inspired by the notion of adhesivity, one can also consider the outer approximation

$$\Gamma^*_{\mathcal{T}} \subset \bigcap_k \Gamma^*_{C_k} \tag{19}$$

where C_k are the hyperedges of $\mathcal{T}$ and each $\Gamma^*_{C_k}$ is the entropic cone associated with the variables in C_k embedded in the space of all variables, i.e., the remaining variables are constrained. However, except for the case of entropic polymatroid arising as a restriction of a single polymatroid and few other cases (cf. Ref. [63]), it is not clear whether entropic polymatroid are adhesive, hence, the inclusion in Eq. (19) may be strict.

Now if we take an outer approximation Γ (e.g., the Shannon cone) of the entropic cone Γ^*, and intersect it with the linear subspaces defined by $\mathcal{I}(\mathcal{T})$, we obtain the cone $\Gamma_{\mathcal{T}}$ which satisfies

$$\Pi_{\mathcal{M}}(\Gamma^*) = \Pi_{\mathcal{M}}(\Gamma^*_{\mathcal{T}}) \subset \Pi_{\mathcal{M}}(\Gamma_{\mathcal{T}}) \subset \Pi_{\mathcal{M}}\left(\bigcap_k \Gamma_{C_k}\right). \tag{20}$$

In general, given a marginal scenario $\mathcal{M}$ there exist several minimal acyclic hypergraphs $\{\mathcal{T}_i\}$ extending it. The inclusion in Eq. (20) is valid for each $\Pi_{\mathcal{M}}\left(\Gamma_{\mathcal{T}_i}\right)$, consequently, also for the intersection

$$\Pi_{\mathcal{M}}\left(\Gamma^*\right) \subset \bigcap_i \Pi_{\mathcal{M}}\left(\Gamma_{\mathcal{T}_i}\right). \tag{21}$$

As a conclusion, for each marginal scenario $\mathcal{M}$, we have three different outer approximations for $\Pi_{\mathcal{M}}\left(\Gamma^*\right)$, namely

(i) the intersection of the projections of the full Shannon cones $\Gamma_{[n]}$ with modularity conditions $\{\mathcal{I}(\mathcal{T}_i)\}$, namely, $\bigcap_i \Pi_{\mathcal{M}}\left(\Gamma_{\mathcal{T}_i}\right)$, $\{\mathcal{T}_i\}$ is the set of minimal acyclic hypergraphs extending $\mathcal{M}$

(ii) the projection of the full Shannon cone, namely, $\Pi_{\mathcal{M}}\left(\Gamma\right)$,

(iii) the projection of the intersection of Shannon cones associated with $\{C_k^{(i)}\}_k$ is the set of hyperedges of $\mathcal{T}_i$, i.e., $\bigcap_i \Pi_{\mathcal{M}}\left(\bigcap_k \Gamma_{C_k^{(i)}}\right)$, where each $\Gamma_{C_k^{(i)}}$ is seen as a cone in the space of all variables, and the variables not appearing in $C_k^{(i)}$ are unconstrained (cf. Eq. 19).

We can then summarize the relations among the above cones as follows

Observation 1. *The above approximations satisfy the inclusion relations*

$$\Pi_{\mathcal{M}}\left(\Gamma^*\right) \subset \bigcap_i \Pi_{\mathcal{M}}\left(\Gamma_{\mathcal{T}_i}\right) \subset \Pi_{\mathcal{M}}\left(\Gamma\right) \subset \bigcap_i \Pi_{\mathcal{M}}\left(\bigcap_k \Gamma_{C_k^{(i)}}\right). \tag{22}$$

In general, the inclusion relations from the Observation 1 are proper. In particular, it means that the outer approximation $\bigcap_{i\in\{\mathcal{T}_i\}} \Pi_{\mathcal{M}}\left(\Gamma_{\mathcal{T}_i}\right)$ is tighter than the projection of the Shannon cone –the most widely used method in the literature, see for instance [58, 63, 76, 77]– and thus may contain non-constrained non-Shannon-type inequalities. We will provide some examples of this in Sect. VII.

VI. INDISTINGUISHABILITY OF CAUSAL STRUCTURES

In this section, we will investigate the role of the above results for the case of probability distributions and entropies, where some underlying causal structure is assumed, i.e., some additional conditional independence constraints are present.

The general goal is to characterize the region of probability distributions or entropies compatible with a given causal structure. Via such a characterization, for instance via Bell inequalities, one can check whether some observed data is consistent or inconsistent with the assumed causal structure, thus being of fundamental importance in both quantum information and any other field where causal discovery may play a relevant role. Furthermore, notice that unless one is able to intervene in the physical system under investigation [3], one can never unambiguously prove what is the underlying causal structure. Rather, based on observations alone, one can only prove the compatibility or incompatibility of a given presumed set of causal relations. As expected, the less constraints a given causal structure implies on the distributions compatible with it, the more correlations such models can explain and the smaller is the possibility of falsifying it.

The ideas to be discussed next apply not only to the case of classical causal structures, but also to the quantum [39, 40, 87–90] and even post-quantum [38, 41, 44] generalizations as well. In the following, we will focus on the classical case, that is, all nodes in the associated Bayesian networks or MRFs represent random variables for which a global joint probability distribution can always be assumed to exist. The case of quantum and post-quantum theories will be briefly discussed at the end of this section and presented in full details elsewhere.

Let us consider the causal structure defined by a graph $\mathcal{G}$, which may be either a DAG corresponding to a Bayesian network, or a graph corresponding to a MRF. We will denote the set of independence conditions associated with $\mathcal{G}$ as $\mathcal{I}(\mathcal{G})$, as in Eqs.(4),(6). Let us now assume to have a fixed marginal scenario, with $\mathcal{M}$ the associated hypergraph. Let $\{\mathcal{T}_i\}_i$ be the set of acyclic hypergraph extending $\mathcal{M}$ as in Th. 1. We will denote by $\mathcal{I}(\mathcal{T}_i)$, the set of independence conditions of the corresponding MRF defined by its 2-section $[\mathcal{T}_i]_2$. We have the following result

Theorem 4. *Given $\mathcal{M}, \mathcal{G}$ and $\{\mathcal{T}_i\}_i$, we have three possible cases:*

(i) $\exists i$ such that $\mathcal{I}(\mathcal{G}) \subset \mathcal{I}(\mathcal{T}_i)$.

(ii) $\forall i$ $\mathcal{I}(\mathcal{T}_i) \subset \mathcal{I}(\mathcal{G})$.

(iii) $\forall i$ $\mathcal{I}(\mathcal{G}) \not\subset \mathcal{I}(\mathcal{T}_i)$, and $\exists j$ such that $\mathcal{I}(\mathcal{T}_j) \not\subset \mathcal{I}(\mathcal{G})$.

Then:

In case (i), it is impossible to falsify the causal structure described by $\mathcal{G}$. This follows since for any probability distribution P, its marginals in $\mathcal{M}$ are always consistent with the causal structure described by $\mathcal{G}$. Approach (c1)-(c3) of Sect. V can be used to characterize the marginals associated with $\mathcal{M}$ and $\mathcal{G}$.

In case (ii), marginals associated with $\mathcal{M}$ can still generate correlations that are incompatible with the causal structure associated with $\mathcal{I}(\mathcal{G})$. Approach (c1)-(c3) can be used, *but the obtained constraints are redundant with respect to $\mathcal{I}(\mathcal{G})$. It is, therefore, more convenient to apply approach (c1)-(c3) directly with the conditional independence relations $\mathcal{I}(\mathcal{G})$.*

In case (iii), the marginal correlations associated with $\mathcal{M}$ can again be incompatible with the causal structure associated with $\mathcal{I}(\mathcal{G})$. However, the marginal scenario implies constraints that cannot be combined with those of the causal structure $\mathcal{G}$. Hence, the approach (c1)-(c3) cannot be used with the constraints $\mathcal{I}(\mathcal{T})$.

Proof.— In case (i), the independence constraints of the causal structure are just a subset of the independence constraints consistent with the marginal scenario. As a consequence, given the marginal probabilities $\{P_M\}_{M\in\mathcal{M}}$, one can repeat the construction of Th. 1 and obtain a valid joint probability distribution P that is consistent with the causal structure defined by $\mathcal{G}$.

The above result implies that the approach of Sect. V, both the constructions (b1)-(b3) for probabilities and (c1)-(c3) for entropies, applies also to the case of a causal structure with $\mathcal{I}(\mathcal{G}) \subset \mathcal{I}(\mathcal{T}_i)$.

In case (ii), the marginal scenario implies less constraints than the causal structure $\mathcal{G}$, hence, it is clear that the marginals associated with $\mathcal{M}$ are still able to detect inconsistencies with the causal structure associated with $\mathcal{I}(\mathcal{G})$.

From the point of view of the characterization, one can still use the approach (c1)-(c3) of Sect. V. However, it is more convenient to use in (c3) the linear constraints implied by $\mathcal{I}(\mathcal{G})$, since they also include those associated with $\mathcal{I}(\mathcal{T}_i)$ for all i.

In case (iii), it is again clear that the marginals associated with $\mathcal{M}$ are still able to detect inconsistencies with the causal structure associated with $\mathcal{I}(\mathcal{G})$. However, the approach (c1)-(c3) of Sect. V cannot be used as it generates constraints inconsistent with the causal structure. The situation is clarified by the following inclusion relations among entropy cones in Eqs. (23)-(25). Let us denote by $L_\mathcal{G}$ the subspace of entropy vectors in $\mathbb{R}^{2^n}$ where the linear constraints imposed by $\mathcal{I}(\mathcal{G})$ are satisfied, and similarly for $L_{\mathcal{T}_i}$. The entropy cone associated with a given causal structure, either $\mathcal{G}$ or $\mathcal{T}_i$, will be $\Gamma^* \cap L_{\mathcal{G},\mathcal{T}_i}$. We can write down the relation between the associated entropy cones as

$$\Pi_\mathcal{M}\left(\Gamma^*\right) = \Pi_\mathcal{M}(\Gamma^* \cap L_{\mathcal{T}_i}), \tag{23}$$

$$\Pi_\mathcal{M}\left(\Gamma^* \cap L_\mathcal{G} \cap L_{\mathcal{T}_i}\right) \subset \Pi_\mathcal{M}(\Gamma^* \cap L_\mathcal{G}), \tag{24}$$

$$\begin{aligned}\Pi_\mathcal{M}\left(\Gamma^* \cap L_\mathcal{G}\right) &\subset \Pi_\mathcal{M}(\Gamma^*) \cap \Pi_\mathcal{M}(L_\mathcal{G}) \\ &= \Pi_\mathcal{M}(\Gamma^* \cap L_{\mathcal{T}_i}) \cap \Pi_\mathcal{M}(L_\mathcal{G})\end{aligned} \tag{25}$$

It is then clear that by imposing both the $\mathcal{I}(\mathcal{G})$ and $\mathcal{I}(\mathcal{T}_i)$ conditions one obtains an inner approximation of the entropy cone associated with the causal structure. Vice versa, imposing the causal structure conditions after the projection gives an outer approximation.□

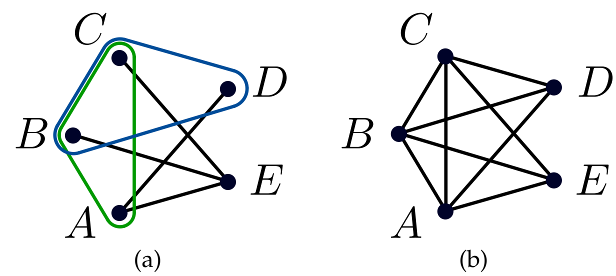


FIG. 6. Marginal scenario hypergraph (a) and (b) its 2-section

To clarify this last part, in particular the impossibility of combining the independence relations $\mathcal{I}(\mathcal{G})$ arising from the causal structure, with the $\{\mathcal{I}(\mathcal{T}_i)\}_i$ arising from the marginal scenario, it is helpful to look at some specific examples. We will discuss them in details in Sect. VII B.

A natural question is, then, how to extend the above results to the case of quantum and post-quantum causal structures. For instance, in the postquantum case if $\mathcal{M}$ is an acyclic hypergraph, then by Th. 1 the observed marginals are always consistent with a classical probability distribution. The same holds, in particular, in the quantum case. However, the fact that different rules for causal inference arise in the quantum and postquantum cases (cf. [38, 40, 41, 44, 88, 90]) together with the different characterization of the associated entropy regions (cf. [38, 40]) makes the above investigation more complex and worth a separate discussion elsewhere [91].

VII. EXAMPLES AND COMPUTATIONAL RESULTS

In order to clarify the results and methods presented in the previous sections, we discuss examples of marginal scenarios and causal structures, together with some computational results.

A. Inclusions in Obs. 1

In the following, we will discuss the possible cases presented in Obs. 1. In particular, we will see examples of strict and non-strict inclusion for the outer approximations of the entropy cone.

1. *Case:* $\bigcap_i \Pi_\mathcal{M}\left(\Gamma_{\mathcal{T}_i}\right) \subsetneq \Pi_\mathcal{M}\left(\Gamma\right) \subsetneq \bigcap_i \Pi_\mathcal{M}\left(\bigcap_k \Gamma_{C_k^{(i)}}\right)$

As already noted by Matúš [63], the proper inclusion $\bigcap_i \Pi_\mathcal{M}\left(\Gamma_{\mathcal{T}_i}\right) \subsetneq \Pi_\mathcal{M}\left(\Gamma\right)$ means existence of non-Shannon-type inequalities in $\Pi_\mathcal{M}\left(\Gamma_{\mathcal{T}_i}\right)$. On the other

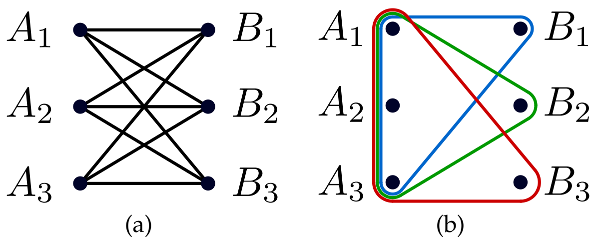


FIG. 7. Marginal scenario (a) and a clique hypergraph (b)

hand, the strict inclusion $\Pi_{\mathcal{M}}(\Gamma) \subsetneq \bigcap_i \Pi_{\mathcal{M}}\left(\bigcap_k \Gamma_{C_k^{(i)}}\right)$ is related to the non-adhesivity of general polymatroids.

To construct an example, is necessary to take at least four variables [63] and our example will consist of five variables. Let us consider the following marginal scenario $\mathcal{M} = \{ABC, BCD, AE, BE, CE, AD\}$, shown on the Fig. 6 (a). One can easily see that 2-section of the hypergraph $\mathcal{M}$ is a triangulated graph shown on the Fig. 6 (b) and thus there is only one corresponding clique hypergraph $\mathcal{T} = \{ABCD, ABCE\}$. The independence constraint $I(D : E|ABC) = 0$ arising from the adhesivity property, gives rise, after projection on the set of entropies given by $\{B, C, D, AD, AE, BD, BE, CD, CE, ABC, BCD\}$, to the following 3 non-redundant non-Shannon-type inequalities

$$H(E|C) + H(C|D) + H(E|B) + H(B,D) + H(A,D) - s_1 H(B,C,D) - s_2 H(A,E) + s_3 H(A,B,C) \geq 0, \quad (26)$$

where the coefficient triplet $(s_1, s_2, s_3) \in \{(1,2,1), (2,1,1), (2,2,2)\}$.

Another interesting aspect of this example is the reduction in the computational time required to compute the projection on a usual desktop computer. More precisely, adding the linear constraint $I(D : E|ABC) = 0$ reduced the time of our computation for the projection from approximately 320 to only 27 seconds.

2. *Case:* $\bigcap_i \Pi_{\mathcal{M}}(\Gamma_{\mathcal{T}_i}) = \Pi_{\mathcal{M}}(\Gamma) = \bigcap_i \Pi_{\mathcal{M}}\left(\bigcap_k \Gamma_{C_k^{(i)}}\right)$

Consider the marginal scenario $\mathcal{M} = \{A_i B_j\}, \forall i, j \in \{1,2,3\}$, shown in Fig. 7 (a), corresponding to a bipartite Bell scenario with three measurement settings per party. The clique hypergraph of one of the triangulations of $\mathcal{M}$ is shown on Fig. 7 (b).

If we consider an intersection of cones $\Gamma_{C_k^{(1)}}$ for cliques $C_k^{(1)} = \{A_1, A_2, A_3, B_k\}$, $k = 1,2,3$, which are the edges of the hypergraph shown on Fig. 7 (b), we

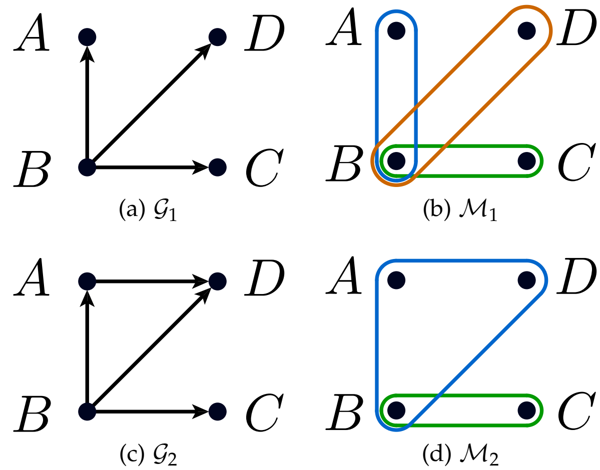


FIG. 8. Example of causal structures (a),(c), and marginal scenarios (b),(d).

find that its projection on the marginal scenario $\mathcal{M}$ differs from the projection of the full cone such that 108 out of 217 rays of the projection $\Pi_{\mathcal{M}}\left(\bigcap_k \Gamma_{C_k^{(1)}}\right)$ are outside of $\Pi_{\mathcal{M}}(\Gamma)$. This can be check via linear programming (cf. Ref. [38] Sect. II of the supplemental material) by simply checking whether such rays are compatible with the basic Shannon inequalities characterizing Γ.

However, if we now consider cliques of the second possible triangulation of $\mathcal{M}$, which are $C_k^{(2)} = \{A_k, B_1, B_2, B_3\}$, $k = 1,2,3$ and compute the intersection $\bigcap_{i=1,2} \Pi_{\mathcal{M}}\left(\bigcap_k \Gamma_{C_k^{(i)}}\right)$ we find, again via linear programming, that all its extremal rays are inside, not only $\Pi_{\mathcal{M}}(\Gamma)$, but also $\Pi_{\mathcal{M}}(\Gamma_{\mathcal{T}_i})$, for all i. Hence we have that all the outer approximations coincide.

The reasons why such an equivalence is interesting is that the calculation of $\bigcap_{i=1,2} \Pi_{\mathcal{M}}\left(\bigcap_k \Gamma_{C_k^{(i)}}\right)$ with a standard Fourier-Motzkin algorithm on a standard desktop takes few minutes, however, a direct computation of $\Pi_{\mathcal{M}}(\Gamma)$ or $\Pi_{\mathcal{M}}(\Gamma_{\mathcal{T}_i})$ seems to be out of computational reach (at least on a usual desktop computer).

B. Three cases in Theorem 4

In this subsection, we will discuss in detail and provide examples for the different cases presented in Theorem 4.

Let us consider four random variables A, B, C, D.

1. Case (i): $\exists i$ such that $\mathcal{I}(\mathcal{G}) \subset \mathcal{I}(\mathcal{T}_i)$

Let us consider the marginal scenario given by $\mathcal{M}_1 = \{AB, BD, BC\}$ and shown on Fig. 8 (b). One can easily see that the clique hypergraph $\mathcal{T}$ of the corresponding triangulation of 2-section graph is unique and coincides with $\mathcal{M}_1$. The independence constraints implied by $\mathcal{T}$ are given by

$$\begin{aligned} &(A \perp C \,|\, B) \\ &(C \perp D \,|\, B) \\ &(D \perp A \,|\, B) \end{aligned} \tag{27}$$

Consider now the two causal structures $\mathcal{G}_1$ and $\mathcal{G}_2$ shown on Fig. 8 (a),(c). The corresponding independence constrains are $\{(A \perp C \,|\, B); (C \perp D \,|\, B); (D \perp A \,|\, B)\}$ for $\mathcal{G}_1$ and $\{(A \perp C \,|\, B); (C \perp D \,|\, B)\}$ for $\mathcal{G}_2$. In both cases $\mathcal{I}(\mathcal{G}_{1,2}) \subset \mathcal{I}(\mathcal{T})$ which means that marginal scenario $\mathcal{M}_1$ from Fig. 8 (b) is insufficient to distinguish causal structures $\mathcal{G}_1$ and $\mathcal{G}_2$. In turn, a marginal scenario, which would be enough to distinguish between these two causal structures is given by $\mathcal{M}_2 = \{ABD, BC\}$ and shown on Fig. 8 (d).

2. Case (ii): $\forall i\ \mathcal{I}(\mathcal{T}_i) \subset \mathcal{I}(\mathcal{G})$

Consider again the causal graph $\mathcal{G}_1$. As we already noted the independence constraints associated with this graph are

$$\begin{aligned} &(A \perp C \,|\, B) \\ &(C \perp D \,|\, B) \\ &(D \perp A \,|\, B) \end{aligned} \tag{28}$$

If we are now interested in the marginal scenario $\mathcal{M}_2 = \{ABD, BC\}$, then one can see that in that case there is again only one possible triangulation of the 2-section graph of $\mathcal{M}_2$ and consequently only one corresponding clique hypergraph. The set of independence constraints, consistent with $\mathcal{M}_2$ is $\{(A \perp C \,|\, B); (C \perp D \,|\, B)\}$, which is a subset of constraints from Eq. (28). In other words, constraints coming from marginal scenario $\mathcal{M}_2$ are redundant to those coming from the causal structure.

3. Case (iii): $\forall i\ \mathcal{I}(\mathcal{G}) \not\subset \mathcal{I}(\mathcal{T}_i)$, and $\exists j$ such that $\mathcal{I}(\mathcal{T}_j) \not\subset \mathcal{I}(\mathcal{G})$

The third one is, arguably, the most interesting case: it shows that the independence constraints arising from the marginal scenario may be "inconsistent" with those associated with the causal structure. An example where this problem arises is the classical case of the information causality scenario [29]: Alice receives two independent inputs X_0, X_1, she creates a message M depending on those inputs that is sent to Bob who provide guesses Y_0, Y_1, respectively of X_0, X_1, on the basis of the message M.

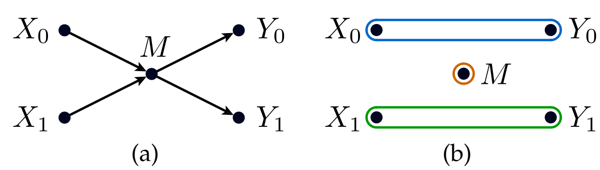


FIG. 9. Causal structure (a) and marginal scenario (b) of classical case of information causality. X_0, X_1 are random inputs for Alice, Y_0, Y_1 are guesses for Bob and M is message which Alice sends to Bob.

The corresponding causal structure is shown in Fig. 9 (a) and the marginal scenario $\mathcal{M}$ in Fig. 9 (b). Once again the clique hypergraph $\mathcal{T}$ coincides with $\mathcal{M}$, hence, it is unique.

We need to show that

$$\mathcal{I}(\mathcal{G}) \not\subset \mathcal{I}(\mathcal{T}), \text{ and } \mathcal{I}(\mathcal{T}) \not\subset \mathcal{I}(\mathcal{G}). \tag{29}$$

For showing $\mathcal{I}(\mathcal{G}) \not\subset \mathcal{I}(\mathcal{T})$, we consider the conditional independence between inputs X_0, X_1 and guesses Y_0, Y_1. I.e.

$$\{X_i \perp Y_j \,|\, M\}_{i,j=0,1} \not\subset \mathcal{I}(\mathcal{T}). \tag{30}$$

An example in the other direction is an independence of message M from the rest of the variables, which is implied by $\mathcal{T}$, and is not consistent with conditional independences $\mathcal{I}(\mathcal{G})$.

The projection $\Pi_{\mathcal{M}}(\Gamma \cap L_{\mathcal{G}})$ gives rise to the following inequalities

$$I(X_0 : Y_0) \geq 0, \quad I(X_1 : Y_1) \geq 0, \tag{31a}$$
$$H(Y_0|X_0) \geq 0, \quad H(X_0|Y_0) \geq 0, \tag{31b}$$
$$H(Y_1|X_1) \geq 0, \quad H(X_1|Y_1) \geq 0, \tag{31c}$$
$$I(X_0 : Y_0) + I(X_1 : Y_1) \leq H(M), \tag{31d}$$

where inequalities (31a, 31b, 31c) are simply polymatroid axioms for the marginals $\{X_0Y_0, X_1Y_1\}$ and one obtains these 6 inequalities, if one computes $\Pi_{\mathcal{M}}(\Gamma \cap L_{\mathcal{T}})$. The last inequality Eq. (31d) is the information causality inequality and is not implied by $\mathcal{I}(\mathcal{T})$.

As a result of the relation from Eq. (29), one cannot combine $\mathcal{I}(\mathcal{T})$ and $\mathcal{I}(\mathcal{G})$. Due to the relation from Eq. (24) and the fact that the Shannon cone is an outer approximation for the case of more than 3 variables, the projection of the Shannon cone with combined constraints $\mathcal{I}(\mathcal{T})$ and $\mathcal{I}(\mathcal{G})$ in this case provides neither an outer nor an inner approximation of $\Pi_{\mathcal{M}}(\Gamma^* \cap L_{\mathcal{G}})$.

VIII. CONCLUSION

Deciding global features of a system of interest with limited information, the so-called marginal problem, is a task often encountered in many fundamental and practical problems. In turn, causal discovery, the inference of causal relations underlying the correlations between observed variables, is yet another basic goal in the most diverse fields. In this paper, we use the notion of adhesivity to investigate marginal problems within causal inference. In particular, we show which causal relations are always compatible with some given marginal information. As a consequence, we are able to identify which causal structures, describing either a Bayesian network or a Markov random field, can be distinguished when only limited marginals are available. In addition, our results provide a method for a faster characterization (in terms of Bell inequalities) of the marginal scenarios associated with a given causal model. This holds true for the both the probabilistic and entropic approaches for Bell inequalities. In particular, in the entropic case our construction allows for a more accurate characterization of allowed regions for entropic marginals, as shown with explicit computational results.

An immediate and interesting open question is the possible generalization of these results to the case where the causal relations between the variables are mediated via quantum or postquantum (nonsignalling) resources. Quantum generalizations of the notion of a causal structure have attracted growing attention [40, 41, 87–90] and we believe that our results could constitute a viable option for the characterization of such quantum structures. Partial results such as, e.g., the fact that classical and postquantum correlations coincide for the case of acyclic marginal scenario hypergraphs (cf. Sect. VI), show that a similar approach can be extended also to the quantum and postquantum case. In particular, this investigation could lead to new insights on which causal structures can demonstrate some sort of non-locality [41, 44]. Finally, another possibility is to try to combine the notion of adhesivity and the algebraic geometry tools [16] required to characterize the set of compatible probabilities associated with complex causal structures [20–22].

ACKNOWLEDGMENTS

CB and NM acknowledge financial support from the DFG, the ERC (Consolidator Grant 683107/TempoQ) and the FQXi Fund (Silicon Valley Community Foundation) and the DAAD. RC acknowledges financial support from the Brazilian ministries MEC and MCTIC, the FQXi Fund, the Excellence Initiative of the German Federal and State Governments (Grants ZUK 43 & 81), the US Army Research Office under contracts W911NF-14-1-0098 and W911NF-14-1-0133 (Quantum Characterization, Verification, and Validation), the DFG (GRO 4334 & SPP 1798).

---

[1] S. L. Lauritzen and D. J. Spiegelhalter, Journal of the Royal Statistical Society. Series B (Methodological) , 157 (1988).
[2] M. Studený, in *Advances in Intelligent Computing—IPMU'94* (Springer, 1994) pp. 348–359.
[3] J. Pearl, *Causality* (Cambridge University Press, Cambridge, 2009).
[4] P. Spirtes, N. Glymour, and R. Scheienes, *Causation, Prediction, and Search, 2nd ed.* (The MIT Press, 2001).
[5] J. S. Bell, Physics **1**, 195 (1964).
[6] N. Brunner, D. Cavalcanti, S. Pironio, V. Scarani, and S. Wehner, Rev. Mod. Phys. **86**, 419 (2014).
[7] A. A. Klyachko, Journal of Physics: Conference Series **36**, 72 (2006).
[8] C. Schilling, D. Gross, and M. Christandl, Phys. Rev. Lett. **110**, 040404 (2013).
[9] M. Walter, B. Doran, D. Gross, and M. Christandl, Science **340**, 1205 (2013).
[10] J. Vomlel, *Methods Of Probabilistic Knowledge Integration*, Ph.D. thesis, Czech Technical University, Faculty of Electrical Engineering (1999).
[11] B. Bonet, in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001) pp. 48–55.
[12] N. Friedman, Science **303**, 799 (2004).
[13] G. Ver Steeg and A. Galstyan, in *Proceedings of the 27th conference on Uncertainty in Artificial Intelligence* (2011).
[14] B. Steudel and N. Ay, Entropy **17**, 2304 (2015).
[15] J. Pearl, in *Proceedings of the 11th conference on Uncertainty in Artificial Intelligence* (1995) pp. 435–443.
[16] D. Geiger and C. Meek, in *Proceedings of the 15th conference on Uncertainty in Artificial Intelligence* (1999) pp. 226–235.
[17] J. Tian and J. Pearl, in *Proceedings of the 28th conference on Uncertainty in Artificial Intelligence* (2002) pp. 519–527.
[18] L. D. Garcia, M. Stillman, and B. Sturmfels, Journal of Symbolic Computation **39**, 331 (2005).
[19] C. Kang and J. Tian, in *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence* (2007) pp. 200–208.
[20] C. M. Lee and R. W. Spekkens, arXiv preprint arXiv:1506.03880 (2015).
[21] R. Chaves, Phys. Rev. Lett. **116**, 010402 (2016).
[22] D. Rosset, C. Branciard, T. J. Barnea, G. Pütz, N. Brunner, and N. Gisin, Phys. Rev. Lett. **116**, 010403 (2016).
[23] C. J. Wood and R. W. Spekkens, New J. Phys. **17**, 033002 (2015).
[24] R. Chaves, R. Kueng, J. B. Brask, and D. Gross, Phys. Rev. Lett. **114**, 140403 (2015).

[25] I. Pitowsky, *Quantum probability–quantum logic*, Lecture notes in physics (Springer-Verlag, 1989).
[26] I. Pitowsky, Mathematical Programming **50**, 395 (1991).
[27] R. Hartshorne, *Algebraic geometry*, Vol. 52 (Springer Science & Business Media, 2013).
[28] S. L. Braunstein and C. M. Caves, Phys. Rev. Lett. **61**, 662 (1988).
[29] M. Pawłowski, T. Paterek, D. Kaszlikowski, V. Scarani, A. Winter, and M. Żukowski, Nature **461**, 1101 (2009).
[30] H. Barnum, J. Barrett, L. O. Clark, M. Leifer, R. Spekkens, N. Stepanik, A. Wilce, and R. Wilke, New J. Phys. **12**, 033024 (2010).
[31] A. J. Short and S. Wehner, New J. Phys. **12**, 033023 (2010).
[32] S. W. Al-Safi and A. J. Short, Phys. Rev. A **84**, 042323 (2011).
[33] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf, Ann. Statist. **41**, 2324 (2013).
[34] R. Chaves, J. B. Brask, and N. Brunner, Phys. Rev. Lett. **115**, 110501 (2015).
[35] D. Janzing, R. Chaves, and B. Schoelkopf, arXiv preprint arXiv:1512.02057 (2015).
[36] R. Chaves, L. Luft, and D. Gross, New J. Phys. **16**, 043001 (2014).
[37] R. Chaves, L. Luft, T. O. Maciel, D. Gross, D. Janzing, and B. Schölkopf, Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence , 112 (2014), arXiv:1407.2256.
[38] R. Chaves and C. Budroni, Phys. Rev. Lett. **116**, 240501 (2016).
[39] T. Fritz, New J. Phys. **14**, 103001 (2012).
[40] R. Chaves, C. Majenz, and D. Gross, Nature communications **6**, 5766 (2015).
[41] J. Henson, R. Lal, and M. F. Pusey, New J. Phys. **16**, 113043 (2014).
[42] M. Weilenmann and R. Colbeck, arXiv preprint arXiv:1603.02553 (2016).
[43] M. Weilenmann and R. Colbeck, arXiv preprint arXiv:1605.02078 (2016).
[44] J. Pienaar, arXiv preprint arXiv:1606.07798 (2016).
[45] N. J. Cerf and C. Adami, Phys. Rev. A **55**, 3371 (1997).
[46] R. Chaves and T. Fritz, Phys. Rev. A **85**, 032113 (2012).
[47] T. Fritz and R. Chaves, IEEE Trans. Inform. Theory **59**, 803 (2013).
[48] R. Chaves, Phys. Rev. A **87**, 022102 (2013).
[49] P. Kurzyński, R. Ramanathan, and D. Kaszlikowski, Phys. Rev. Lett. **109**, 020404 (2012).
[50] P. Kurzyński and D. Kaszlikowski, Phys. Rev. A **89**, 012103 (2014).
[51] S. Raeisi, P. Kurzyński, and D. Kaszlikowski, Phys. Rev. Lett. **114**, 200401 (2015).
[52] M. Wajs, P. Kurzyński, and D. Kaszlikowski, Phys. Rev. A **91**, 012114 (2015).
[53] A. R. U. Devi, H. S. Karthik, Sudha, and A. K. Rajagopal, Phys. Rev. A **87**, 052103 (2013).
[54] A. E. Rastegin, Annals of Physics **355**, 241 (2015).
[55] A. E. Rastegin, arXiv preprint arXiv:1403.6945 (2014).
[56] L.-Z. Cao, J.-Q. Zhao, X. Liu, Y. Yang, Y.-D. Li, X.-Q. Wang, Z.-B. Chen, and H.-X. Lu, Scientific reports **6** (2016).
[57] H. P. Williams, Amer. Math. Monthly **93**, 681 (1986).
[58] R. W. Yeung, *Information theory and network coding*, Information technology–transmission, processing, and storage (Springer, 2008).
[59] N. Linden and A. Winter, Communications in Mathematical Physics **259**, 129 (2005).
[60] J. Cadney, N. Linden, and A. Winter, IEEE Transactions on Information Theory **58**, 3657 (2012).
[61] N. N. Vorob'yev, Theory of Probability & Its Applications **12**, 251 (1967).
[62] N. Vorob'ev, Theory of Probability & Its Applications **8**, 420 (1963).
[63] F. Matúš, Discrete Mathematics **307**, 2464 (2007).
[64] C. Beeri, R. Fagin, D. Maier, and M. Yannakakis, J. ACM **30**, 479 (1983).
[65] S. L. Lauritzen, *Graphical models* (Clarendon Press, 1996).
[66] A. Fine, Phys. Rev. Lett. **48**, 291 (1982).
[67] In such cases, in order to distinguish different causal structures, one has to rely on another crucial concept of the mathematical theory of causality, that of an intervention [3].
[68] J. Thompson, P. Kurzynski, S.-Y. Lee, A. Soeda, and D. Kaszlikowski, arXiv preprint arXiv:1304.1292 (2013).
[69] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, Phys. Rev. Lett. **23**, 880 (1969).
[70] S. Boyd and L. Vandenberghe, *Convex optimization* (Cambridge university press, 2009).
[71] C. Branciard, N. Gisin, and S. Pironio, Phys. Rev. Lett. **104**, 170401 (2010).
[72] S. Thakor, A. Grant, and T. Chan, in *Information Theory Workshop (ITW), 2012 IEEE* (2012) pp. 267–271.
[73] C. Budroni and G. Morchio, Journal of Mathematical Physics **51**, 122205 (2010).
[74] C. Budroni and G. Morchio, Foundations of Physics **42**, 544 (2012).
[75] C. Budroni and A. Cabello, Journal of Physics A: Mathematical and Theoretical **45**, 385304 (2012).
[76] Z. Zhang and R. Yeung, IEEE Transactions on Information Theory **44**, 1440 (1998).
[77] F. Matúš, in *Information Theory, 2007. ISIT 2007. IEEE International Symposium on* (IEEE, 2007) pp. 41–44.
[78] K. Makarychev, Y. Makarychev, A. Romashchenko, and N. Vereshchagin, Communications in Information and Systems **2**, 147 (2002).
[79] Z. Zhang, Communications in Information and Systems **3**, 47 (2003).
[80] R. Dougherty, C. Freiling, and K. Zeger, in *2006 IEEE International Symposium on Information Theory* (2006) pp. 233–236.
[81] H. G. Kellerer, Mathematische Annalen **153**, 168 (1964).
[82] H. G. Kellerer, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **3**, 247 (1964).
[83] F. Malvestuto, Discrete Mathematics **69**, 61 (1988).
[84] P. Heggernes, Discrete Mathematics **306**, 297 (2006).
[85] M. Araújo, M. T. Quintino, C. Budroni, M. T. Cunha, and A. Cabello, Phys. Rev. A **88**, 022118 (2013).
[86] F. Matúš, IEEE Transactions on Information Theory **53**, 320 (2007).
[87] M. S. Leifer and R. W. Spekkens, Phys. Rev. A **88**, 052130 (2013).
[88] J. Pienaar and Časlav Brukner, New J. Phys. **17**, 073020 (2015).
[89] T. Fritz, Communications in Mathematical Physics **341**,

391 (2016).
[90] F. Costa and S. Shrapnel, New J. Phys. **18**, 063032 (2016).
[91] N. Miklin *et al.*, *in preparation*.