Under consideration for publication in Math. Struct. in Comp. Science

Causal Inference via String Diagram Surgery

A Diagrammatic Approach to Interventions and Counterfactuals

Bart Jacobs, Aleks Kissinger, and Fabio Zanasi

Received 8 December 2021

Extracting causal relationships from observed correlations is a growing area in probabilistic reasoning, originating with the seminal work of Pearl and others from the early 1990s. This paper develops a new, categorically oriented view based on a clear distinction between syntax (string diagrams) and semantics (stochastic matrices), connected via interpretations as structure-preserving functors.

A key notion in the identification of causal effects is that of an intervention, whereby a variable is forcefully set to a particular value independent of any prior propensities. We represent the effect of such an intervention as an endofunctor which performs 'string diagram surgery' within the syntactic category of string diagrams. This diagram surgery in turn yields a new, interventional distribution via the interpretation functor. While in general there is no way to compute interventional distributions purely from observed data, we show that this is possible in certain special cases using a calculational tool called comb disintegration.

We demonstrate the use of this technique on two well-known toy examples: one where we predict the causal effect of smoking on cancer in the presence of a confounding common cause and where we show that this technique provides simple sufficient conditions for computing interventions which apply to a wide variety of situations considered in the causal inference literature; the other one is an illustration of counterfactual reasoning where the same interventional techniques are used, but now in a 'twinned' set-up, with two version of the world — one factual and one counterfactual — joined together via exogenous variables that capture the uncertainties at hand.

1. Introduction

Causality is about understanding the mechanics of the world around us. This world presents itself in the form of streams of observations, in which statistical (in)dependencies can be recognised. A big question, both in science and in daily life, is: how to distinguish correlation from causation and recognise genuine causal relationships?

An important conceptual tool for distinguishing correlation from causation is the possibility of *intervention*, i.e. forcing some variable to take a specific value in a way that is independent of the other variables being considered. For example, a randomised drug trial attempts to destroy any confounding 'common cause' explanation for correlations between drug use and recovery by randomly assigning a patient to the control or treatment group, independent of any background factors. In an ideal setting, the observed correlations of such a trial will reflect genuine causal influence. Unfortunately, it is not always possible (or ethical) to ascertain causal effects by means of actual interventions. For instance, one is unlikely to get ethical approval to run a clinical trial on whether smoking causes cancer by randomly assigning 50% of the patients to smoke, and waiting a bit to see who gets cancer. However, in certain situations it is possible to predict the effect of such a hypothetical intervention from purely observational data, together with an assumed underlying graph structure.

In this paper, we will focus on the problem of *causal identifiability*. For this problem, we are given observational data as a joint distribution on a set of variables and we are furthermore provided with a *causal structure* associated with those variables. This structure, which typically takes the form of a directed acyclic graph or some variation thereof, tells us which variables can in principle have a causal influence on others. The problem then becomes whether we can measure how strong those causal influences are, by means of computing an *interventional* distribution. That is, can we ascertain what would have happened if a (hypothetical) intervention had occurred?

Note that this is distinct from the related problem of *causal discovery*, where a causal structure is not given from the start, but must be discovered purely from the observational data, subject to certain well-behavedness assumptions [28]. Causal identifiability assumes that the causal structure is already provided, either from previously doing causal discovery or by making use of some additional knowledge about the problem at hand. In particular, this means that causal identifiability is trivial when all variables are observed. However, it becomes a difficult and important problem in the presence of confounding variables (unobserved common causes) or selection bias (conditioning on common effects). In this paper, we will focus on the former.

Over the past 3 decades, a great deal of work has been done in identifying necessary and sufficient conditions for causal identifiability in the presence of confounding variables, starting with very specific special cases such as the *back-door* and *front-door* criteria [28] and progressing to more general necessary and sufficient conditions for causal identifiability based on the **do**-calculus [17], or combinatoric concepts such as confounded components in semi-Makovian models [34, 35].

This style of causal reasoning relies crucially on a delicate interplay between syntax and semantics, which is often not made explicit in the literature. The syntactic object of interest is the causal structure (e.g. a causal graph), which captures something about our understanding of the world, and the mechanisms which gave rise to some observed phenomena. The semantic object of interest is the data: joint and conditional probability distributions on some variables. Fixing a causal structure entails certain constraints on which probability distributions can arise, hence it is natural to see distributions satisfying those constraints as models of the syntax.

In this paper, we make this interplay precise using functorial semantics in the spirit of Lawvere [23], and develop basic syntactic and semantic tools for causal reasoning in this setting. We take as our starting point a functorial presentation of Bayesian networks similar to the one appearing in [12]. The syntactic role is played by string diagrams, which give an intuitive way to represent morphisms of a monoidal category as boxes plugged together by wires. Given a directed acyclic graph (dag) G, we can form a free category Syn_G whose arrows are (formal) string diagrams which represent the causal structure syntactically. Structure-preserving functors from Syn_G to Stoch, the category of stochastic matrices, then correspond exactly to Bayesian networks based on the dag G.

Within this framework, we develop the notion of intervention as an operation of 'string diagram surgery'. Intuitively, this cuts a string diagram at a certain variable, severing its link to the past. Formally, this is represented as an endofunctor on the syntactic category $\mathsf{cut}_X \colon \mathsf{Syn}_G \to \mathsf{Syn}_G$, which propagates through a model $\mathcal{F} \colon \mathsf{Syn}_G \to \mathsf{Stoch}$ to send observational probabilities $\mathcal{F}(\omega)$ to interventional probabilities $\mathcal{F}(\mathsf{cut}_X(\omega))$.

The cut_X endofunctor gives us a diagrammatic means of computing interventional distributions given complete knowledge of $\mathcal{F}$. However, more interestingly, we can sometimes compute interventionals given only partial knowledge of $\mathcal{F}$, namely some observational data. We show that this can also be done via a technique we call *comb disintegration*, which is a string diagrammatic version of a technique called *c-factorisation* introduced by Tian and Pearl [35]. Our approach generalises disintegration, a calculational tool whereby a joint state on two variables is factored into a single-variable state and a channel, representing the marginal and conditional parts of the distribution, respectively. Disintegration has recently been formulated categorically in [7] and using string diagrams in [6]. We take the latter as a starting point, but instead consider a factorisation of a three-variable state into a channel and a *comb*. The latter is a special kind of map which allows inputs and outputs to be interleaved. They were originally studied in the context of quantum communication protocols, seen as games [14], but have recently been used extensively in the study of causally-ordered quantum [5, 30] and generalised [22] processes. While originally imagined for quantum processes, the categorical formulation given in [22] makes sense in both the classical case (Stoch) and the quantum. Much like Tian and Pearl's technique, comb factorisation allows one to characterise when the confounding parts of a causal structure are suitably isolated from each other, then exploit that isolation to perform the concrete calculation of interventional distributions.

However, unlike in the traditional formulation, the syntactic and semantic aspects of causal identifiability within our framework are connected. Namely, we can give conditions for causal identifiability in terms of factorisation of a morphism in Syn_G, whereas the actual concrete computation of the interventional distribution involves factorisation of its interpretation in Stoch. Thanks to the functorial semantics, the former immediately implies the latter.

The interventional techniques in terms of string diagrams and their interpretations can also be applied to counterfactual queries. There, we use two copies of the relevant string diagram, one for the 'actual' and one for the 'counterfactual' world. These two copies are connected via some shared states, which capture the sense in which random variables should take the same value in the real and counterfactual world. This approach, which is also described in functional form [2, 29], is elaborated here in terms of factorisation of stochastic matrices into a deterministic part and an exogenous state. By sharing these states between the two copies of the world, we are able to more effectively make use of our knowledge of what *did* happen in order to predict what *would have* happened, had some past detail been different. One standard example from the literature is elaborated as illustration of how this works.

To introduce our framework itself, we make use of a running example taken from Pearl's book [28]: identifying the causal effect of smoking on cancer with the help of an auxiliary variable (the presence of tar in the lungs). After providing some preliminaries on stochastic matrices and the functorial presentation of Bayesian networks in Sections 2 and 3, we introduce the smoking example in Section 4. In Section 5 we formalise the notion of intervention as string diagram surgery, and in Section 6 we introduce the combs and prove our main calculational result: the existence and uniqueness of comb factorisations. In Section 7, we show how to apply this theorem in computing the interventional distribution in the smoking example. In Section 8, we provide a more general version of the theorem, which captures (and slightly generalises) the conditions given in [35]. In Section 9, we focus on counterfactual reasoning, illustrating how it can also be modelled with string diagram surgery. Finally, in Section 10, we summarise our results and describe several avenues of future work.

This work is an extended version of the conference paper [19], which includes all the missing proofs and a completely new part on counterfactuals (Section 9).

2. Stochastic Matrices and Conditional Probabilities

Symmetric monoidal categories (SMCs) give a very general setting for studying processes which can be composed in sequence (via the usual categorical composition $\circ$) and in parallel (via the monoidal composition $\otimes$). Throughout this paper, we will use *string diagram* notation [33] for depicting composition of morphisms in an SMC. In this notation, morphisms are depicted as boxes with labelled input and output wires, composition $\circ$ as 'plugging' boxes together, and the monoidal product $\otimes$ as placing boxes side-by-side. Identity morphisms are depicted simply as a wire and the unit I of $\otimes$ as the empty diagram. The 'symmetric' part of the structure consists of symmetry morphisms, which enable us to permute inputs and outputs arbitrarily. We depict these as wire-crossings: . Morphisms whose domain is I are called *states*, and they will play a special role throughout this paper.

A monoidal category of prime interest in this paper is Stoch, whose objects are finite sets and morphisms $\boldsymbol{f} : A \to B$ are $|B| \times |A|$ dimensional stochastic matrices. That is, they are matrices of positive numbers (including 0) whose columns each sum to 1:

$$\boldsymbol{f} = \{\boldsymbol{f}_i^j \in \mathbb{R}^+ \mid i \in A, j \in B\} \qquad \text{with} \qquad \textstyle\sum_j \boldsymbol{f}_i^j = 1, \text{ for all } i.$$

Note we adopt the physicists convention of writing row indices as superscripts and column indices as subscripts. Stochastic matrices are of interest for probabilistic reasoning, because they exactly capture the data of a conditional probability distribution. That is, if we take $A := \{1, \ldots, m\}$ and $B := \{1, \ldots, n\}$, conditional probabilities naturally arrange themselves into a stochastic matrix:

$$\boldsymbol{f}_i^j := P(B = j | A = i) \quad \rightsquigarrow \quad \boldsymbol{f} = \begin{pmatrix} P(B=1|A=1) & \cdots & P(B=1|A=m) \\ \vdots & \ddots & \vdots \\ P(B=n|A=1) & \cdots & P(B=n|A=m) \end{pmatrix}$$

States, i.e. stochastic matrices from a trivial input $I := \{*\}$, are (non-conditional) probability distributions, represented as column vectors. There is only one stochastic matrix with trivial output: the row vector consisting only of 1's. The latter, with notation ⫯ as on the right, will play a special role in this paper (see (1) below).

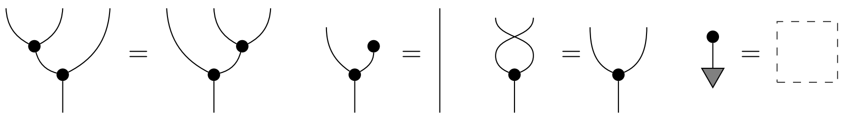
Composition of stochastic matrices is matrix multiplication. In terms of conditional probabilities, this corresponds to multiplication, followed by marginalization over the shared variable: $\sum_B P(C|B)P(B|A)$. Identities are therefore given by identity matrices, which we will often express in terms of the Kronecker delta function $\boldsymbol{\delta}_i^j$.

The monoidal product $\otimes$ in Stoch is the cartesian product on objects, and Kronecker product of matrices: $(\boldsymbol{f} \otimes \boldsymbol{g})_{(i,j)}^{(k,l)} := \boldsymbol{f}_i^k \boldsymbol{g}_j^l$. We will typically omit parentheses and commas in the indices, writing e.g. $\boldsymbol{h}_{ij}^{kl}$ instead of $\boldsymbol{h}_{(i,j)}^{(k,l)}$ for an arbitrary matrix entry of $\boldsymbol{h} \colon A \otimes B \to C \otimes D$. In terms of conditional probabilities, the Kronecker product corresponds to taking product distributions. That is, if $\boldsymbol{f}$ represents the conditional probabilities $P(B|A)$ and $\boldsymbol{g}$ the probabilities $P(D|C)$, then $\boldsymbol{f} \otimes \boldsymbol{g}$ represents $P(B|A)P(D|C)$. Stoch also comes with a natural choice of 'swap' matrices $\boldsymbol{\sigma} : A \otimes B \to B \otimes A$ given by $\boldsymbol{\sigma}_{ij}^{kl} := \boldsymbol{\delta}_i^l \boldsymbol{\delta}_j^k$, making it into a symmetric monoidal category. Every object A in Stoch has three other pieces of structure which will play a key role in our formulation of Bayesian networks and interventions: the *copy* map, the *discarding* map, and the *uniform state*:

$$\left(\text{⋎}\right)_i^{jk} := \boldsymbol{\delta}_i^j \boldsymbol{\delta}_i^k \qquad\qquad \left(\text{⫯}\right)_i := 1 \qquad\qquad \left(\text{⏚}\right)^i := \frac{1}{|A|} \tag{1}$$

Abstractly, this provides Stoch with the structure of a *CDU category.*

Definition 2.1. A *CDU category* (for **c**opy, **d**iscard, **u**niform) is a symmetric monoidal category $(\mathsf{C}, \otimes, I)$ where each object A has a copy map ⋎ : $A \to A \otimes A$, a discarding map ⫯ : $A \to I$, and a uniform state ⏚ : $I \to A$ satisfying the following equations:



(2)

CDU functors are symmetric monoidal functors between CDU categories preserving copy maps, discard maps and uniform states.[†]

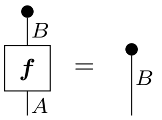
We assume that the CDU structure on I is trivial and the CDU structure on $A \otimes B$ is constructed in the obvious way from the structure on A and B. We also use the first equation in (2) to justify writing 'copy' maps with arbitrarily many output wires: ⋎.

Similar to [3], we can form the free CDU category $\mathsf{FreeCDU}(A, \Sigma)$ over a pair (X, Σ) of a generating set of objects X and a generating set Σ of typed morphisms $f \colon u \to w$, with $u, w \in X^\star$ as follows. The category $\mathsf{FreeCDU}(A, \Sigma)$ has $X^\star$ as set of objects, and

[†] CDU-categories are closely related to *gs-monoidal* categories, structured adopted in the categorical description of resource sensitive algebraic theories [10]. The difference is that CDU-categories also include uniform states ⏚ and the corresponding equation. As we will see, this extra structure is needed in order to account for causal intervention. A CDU category is also very much like a Markov category defined in [13], but Markov categories do not have these uniform states either; additionally in Markov categories the discard equation (3) holds for all maps.

morphisms the string diagrams constructed from the elements of Σ and maps ⋎ $: x \to x \otimes x$, ⫯ $: x \to I$ and ⫰ $: I \to x$ for each $x \in X$, taken modulo the equations (2).

Lemma 2.2. Stoch is a CDU category, with CDU structure defined as in (1).

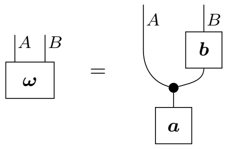


(3)

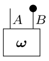
An important feature of Stoch is that $I = \{\star\}$ is the final/terminal object, with ⫯ $: B \to I$ the map provided by the universal property, for any set B. This yields equation (3) on the right, for any $\boldsymbol{f}: A \to B$, justifying the name "discarding map" for ⫯.

We conclude by recording another significant feature of Stoch: *disintegration* [6, 7]. In probability theory, this is the mechanism of factoring a joint probability distribution $P(AB)$ as a product of the first marginal $P(A)$ and a conditional distribution $P(B|A)$. We recall from [6] the string diagrammatic rendition of this process. We say that a morphism $\boldsymbol{f}: X \to Y$ in Stoch has *full support* if, as a stochastic matrix, it has no zero entries. When $\boldsymbol{f}$ is a state, it is a standard result that full support ensures uniqueness of disintegrations of $\boldsymbol{f}$.

Proposition 2.3 (Disintegration). For any state $\boldsymbol{\omega}: I \to A \otimes B$ in Stoch with full support, there exists unique morphisms $\boldsymbol{a}: I \to A, \boldsymbol{b}: A \to B$ such that:



(4)

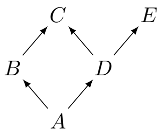
Note that equation (3) and the CDU rules immediately imply that the unique $\boldsymbol{a}: I \to A$ in Proposition 2.3 is the marginal of $\boldsymbol{\omega}$ onto A:  .

3. Bayesian Networks as String Diagrams

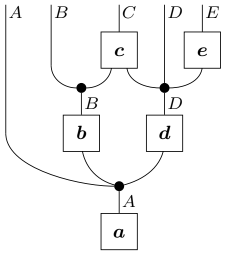
Bayesian networks are a widely-used tool in probabilistic reasoning. They give a succinct representation of conditional (in)dependences between variables as a directed acyclic graph. Traditionally, a Bayesian network on a set of variables $A, B, C, \ldots$ is defined as a directed acyclic graph (dag) G, an assignment of finite sets to each of the nodes $V_G := \{A, B, C, \ldots\}$ of G and a joint probability distribution over those variables which factorises as $P(V_G) = \prod_{A \in V_G} P(A \,|\, \mathrm{Pa}(A))$ where $\mathrm{Pa}(A)$ is the set of parents of A in G. Any joint distribution that factorises this way is said to satisfy the *global Markov property* with respect to the dag G. Alternatively, a Bayesian network can be seen as a dag equipped with a set of conditional probabilities $\{P(A \,|\, \mathrm{Pa}(A)) \mid A \in V_G\}$ which can be combined to form the joint state. Disintegration allows us to switch back and forth between joint states and conditional probabilities and thus yields the equivalence between these two perspectives.

Much like in the case of disintegration in the previous section, Bayesian networks have a categorical description as string diagrams in the category Stoch. This perspective has been widely adopted — with some variations — in recent years [4, 9, 12, 13, 18, 20, 21]. We offer here a presentation inspired by functorial semantics of algebraic theories [24].

Let us start with an example. Here is a Bayesian network in its traditional depiction as a dag with an associated joint distribution over its vertices, and as a string diagram in Stoch:

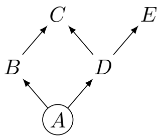


$$P(ABCDE) =$$
$$P(A)P(B|A)P(D|A)P(C|BD)P(E|D)$$

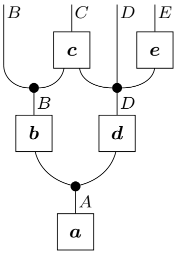


In the string diagram above, the stochastic matrix $\boldsymbol{a}\colon I \to A$ contains the probabilities $P(A)$, $\boldsymbol{b}\colon B \to A$ contains the conditional probabilities $P(B|A)$, $\boldsymbol{c}\colon B \otimes D \to C$ contains $P(C|BD)$, and so on. The entire diagram is then equal to a state $\boldsymbol{\omega}\colon I \to A{\otimes}B{\otimes}C{\otimes}D{\otimes}E$ which represents $P(ABCDE)$.

Note the dag and the diagram above look similar in structure. The main difference is the use of copy maps to make each variable (even those that are not leaves of the dag, A, B and D) an output of the overall diagram. This corresponds to a variable being *observed*. We can also consider Bayesian networks with *latent* variables, which do not appear in the joint distribution due to marginalisation. Continuing the example above, making A into a latent variable yields the following depiction as a string diagram:



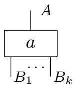
$$P(BCDE) =$$
$$\textstyle\sum_A P(A)P(B|A)P(D|A)P(C|BD)P(E|D)$$



In general, a Bayesian network (with possible latent variables), is a string diagram in Stoch that (1) only has outputs and (2) consists only of copy maps and boxes which each have exactly one output.

By 'a string diagram in Stoch', we mean not only the stochastic matrix itself, but also its decomposition into components. We can formalise exactly what we mean by taking a perspective on Bayesian networks which draws inspiration from functorial semantics of algebraic theories [24]. In this perspective, which mainly elaborates on [12, Ch. 4], we maintain a conceptual distinction between the purely syntactic object (the diagram) and its probabilistic interpretation.

Starting from a dag $G = (V_G, E_G)$, we construct a free CDU category Syn_G which provides the syntax of causal structures labelled by G. The objects of Syn_G are generated by the vertices of G, whereas the morphisms are generated by the following signature:

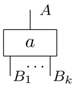
$\Sigma_G = \Bigg\{$ 
$\Bigg|\ A \in V_G$ with parents $B_1, \ldots, B_k \in V_G \Bigg\}$

Then $\mathsf{Syn}_G := \mathsf{FreeCDU}(V_G, \Sigma_G)$. The following result establishes that models (*à la* Lawvere) of Syn_G coincide with G-based Bayesian networks.

Proposition 3.1. There is a 1-1 correspondence between Bayesian networks based on the dag G and CDU functors of type $\mathsf{Syn}_G \to \mathsf{Stoch}$.

Proof. In one direction, consider a Bayesian network consisting of the dag G and, for each node $A \in V_G$, an assignment of a finite set $\tau(A)$ and a conditional probability $P(A|\mathrm{Pa}(A))$. This data yields a CDU functor $\mathcal{F} : \mathsf{Syn}_G \to \mathsf{Stoch}$, defined by the following mappings:

$\mathcal{F} \ :: \ \Bigg\{$ $A \in V_G \ \mapsto\ \tau(A)$

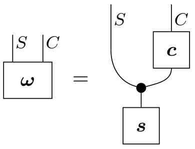

$\mapsto\ \left(\boldsymbol{f}^{j}_{i_1 \ldots i_n} := P(A = j|\mathrm{Pa}(A) = (i_1, \ldots, i_n))\ \right)$

Conversely, let $\mathcal{F} \colon \mathsf{Syn}_G \to \mathsf{Stoch}$ be a CDU functor. This defines a G-based Bayesian network by setting $\tau(A) := \mathcal{F}(A)$ and $P(A = j|\mathrm{Pa}(A) = (i_1, \ldots, i_n)) := \mathcal{F}(a)^{j}_{i_1 \ldots i_n}$. It is immediate that these two mappings are inverse to each other, thus proving the statement. □

This proposition justifies the following definition of a category BN_G of G-based Bayesian networks: objects are CDU functors $\mathsf{Syn}_G \to \mathsf{Stoch}$ and arrows are monoidal natural transformations between them.

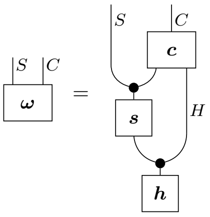
4. Towards Causal Inference: the Smoking Scenario

We will motivate our approach to causal inference via a classic example, inspired by the one given in the Pearl's book [28]. Imagine a dispute between a scientist and a tobacco company. The scientist claims that smoking causes cancer. As a source of evidence, the scientist cites a joint probability distribution ω over variables S for smoking and C for cancer, which disintegrates as in (5) below, with matrix $\boldsymbol{c} = \left(\begin{smallmatrix} 0.9 & 0.7 \\ 0.1 & 0.3 \end{smallmatrix}\right)$. Inspecting this $\boldsymbol{c} : S \to C$, the scientist notes that the probability of getting cancer for smokers (0.3) is three times as high as for non- smokers (0.1). Hence, the scientist claims that smoking has a significant causal effect on cancer.


(5)

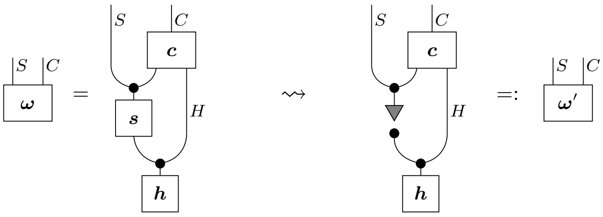
An important thing to stress here is that the scientist draws this conclusion using not only the observational data $\boldsymbol{\omega}$ but also from an assumed *causal structure* which gave rise to that data, as captured in the diagram in equation (5). That is, rather than treating diagram (5) simply as a calculation on the observational data, it can also be treated as an assumption about the actual, physical mechanism that gave rise to that data. Namely, this diagram encompasses the assumption that there is some prior propensity for people to smoke captured by $\boldsymbol{s} : I \to S$, which is both observed and fed into some other process $\boldsymbol{c} : S \to C$ whereby an individual's choice to smoke determines whether or not they get cancer.

The tobacco company, in turn, says that the scientist's assumptions about the provenance of this data are too strong. While they concede that *in principle* it is possible for smoking to have some influence on cancer, the scientist should allow for the possibility that there is some latent common cause (e.g. genetic conditions, stressful work environment, etc.) which leads people both to smoke and get cancer. Hence, says the tobacco company, a 'more honest' causal structure to ascribe to the data ω is (6). This structure then allows for either party to be correct. If the scientist is right, the output of $\boldsymbol{c} : S \otimes H \to C$ depends mostly on its first input, i.e. the causal path from smoking to cancer. If the tabacco company is right, then $\boldsymbol{c}$ depends very little on its first input, and the correlation between S and C can be explained almost entirely from the hidden common cause H.



(6)

So, who is right after all? Just from the observed distribution $\boldsymbol{\omega}$, it is impossible to tell. So, the scientist proposes a clinical trial, in which patients are randomly required to smoke or not to smoke. We can model this situation by replacing $\boldsymbol{s}$ in (6) with a process that ignores its inputs and outputs the uniform state. Graphically, this looks like 'cutting' the link $\boldsymbol{s}$ between H and S:

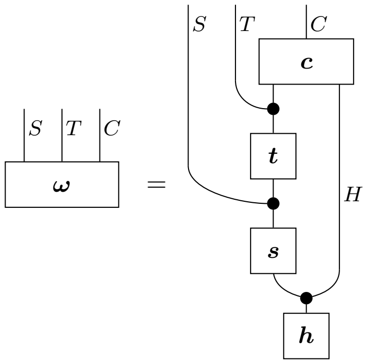


(7)

This captures the fact that variable S is now randomised and no longer dependent on any background factors. This new distribution $\boldsymbol{\omega}'$ represents the data the scientist would have obtained had they run the trial. That is, it gives the results of an *intervention* at $\boldsymbol{s}$. If this $\boldsymbol{\omega}'$ *still* shows a strong correlation between smoking and cancer, one can conclude that smoking indeed causes cancer even when we assume the weaker causal structure (6).

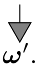
Unsurprisingly, the scientist fails to get ethical approval to run the trial, and hence has only the observational data $\boldsymbol{\omega}$ to work with. Given that the scientist only knows $\boldsymbol{\omega}$ (and not $\boldsymbol{c}$ and $\boldsymbol{h}$), there is no way to compute $\boldsymbol{\omega}'$ in this case. However, a key insight of statistical causal inference is that sometimes it *is* possible to compute interventional distributions from observational ones. Continuing the smoking example, suppose the scientist proposes the following revision to the causal structure: they posit a structure (8) that includes a third observed variable (the presence of T of tar in the lungs), which completely mediates the causal effect of smoking on cancer.

As with our simpler structure, the diagram (8) below contains some assumptions about the origins of the data $\boldsymbol{\omega}$. In particular, by omitting wires, we are asserting there is no *direct* causal link between certain variables.



(8)

The absence of an H-labelled input to $\boldsymbol{t}$ says there is no direct causal link from H to T (only mediated by S), and the absence of an S-labelled input wire into $\boldsymbol{c}$ captures that there is no direct causal link from S to C (only mediated by T). In the traditional approach to causal inference, such relationships are typically captured by a graph-theoretic property called *d-separation* (see [28] for a standard and [21] for a categorical account) on the dag associated with the causal structure.

We can again imagine intervening at S by replacing $\boldsymbol{s} : H \to S$ by  $\circ$. Again, this 'cutting' of the diagram will result in a new interventional distribution $\boldsymbol{\omega}'$. However, unlike before, it *is* possible to compute this distribution from the observational distribution $\boldsymbol{\omega}$.

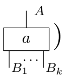
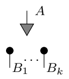
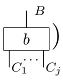
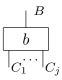
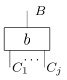
However, in order to do that, we first need to develop the appropriate categorical framework. In Section 5, we will model 'cutting' as a functor. In 6, we will introduce a generalisation of disintegration, which we call *comb disintegration.* These tools will enable us to compute $\boldsymbol{\omega}'$ for $\boldsymbol{\omega}$, in Section 7.

5. Interventional Distributions as Diagram Surgery

The goal of this section is to define the 'cut' operation in (7) as an endofunctor on the category of Bayesian networks. First, we observe that such an operation exclusively concerns the string diagram part of a Bayesian network: following the functorial semantics given in Section 3, it is thus appropriate to define cut as an endofunctor on Syn_G, for a given dag G.

Definition 5.1. For a fixed node $A \in V_G$ in a graph G, let $\mathsf{cut}_A \colon \mathsf{Syn}_G \to \mathsf{Syn}_G$ be the CDU functor freely obtained by the following action on the generators (V_G, Σ_G) of Syn_G:

— For each object $B \in V_G$, $\mathsf{cut}_A(B) = B$.

— $\mathsf{cut}_A($  $) =$  and $\mathsf{cut}_A($  $) =$  for any other  $\in \Sigma_G$.

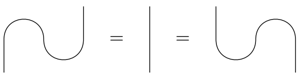
Intuitively, cut_A applied to a string diagram f of Syn_G removes from f each occurrence of a box with output wire of type A.

Proposition 3.1 allows us to "transport" the cutting operation over to Bayesian networks. Given any Bayesian network based on G, let $\mathcal{F} \colon \mathsf{Syn}_G \to \mathsf{Stoch}$ be the corresponding CDU functor given by Proposition 3.1. Then, we can define its A-cutting as the

Bayesian network identified by the CDU functor $\mathcal{F} \circ \mathsf{cut}_A$. This yields an (idempotent) endofunctor $\mathsf{Cut}_A \colon \mathsf{BN}_G \to \mathsf{BN}_G$.

6. The Comb Factorisation

Thanks to the development of Section 5, we can understand the transition from left to right in (7) as the application of the functor Cut_S applied to the 'Smoking' node S. The next step is being able to actually compute the individual Stoch-morphisms appearing in (8), to give an answer to the causality question.

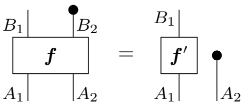


In order to do that, we want to work in a setting where $\boldsymbol{t} \colon S \to T$ can be isolated and 'extracted' from (8). What is left behind is a stochastic matrix with a 'hole' where $\boldsymbol{t}$ has been extracted. To define 'morphisms with holes', it is convenient to pass from SMCs to compact closed categories (see e.g. [33]). Stoch is not itself compact closed, but it embeds into $\mathsf{Mat}(\mathbb{R}^+)$, whose morphisms are *all* matrices over non-negative numbers. $\mathsf{Mat}(\mathbb{R}^+)$ has a (self-dual) compact closed structure; that means, for any set A there is a 'cap' $\cap \colon A \otimes A \to I$ and a 'cup' $\cup \colon I \to A \otimes A$, which satisfy the 'yanking' equations on the right. As matrices, caps and cups are defined by $\cap_{ij} = \cup^{ij} = \delta_i^j$. Intuitively, they amount to 'bent' identity wires. Another aspect of $\mathsf{Mat}(\mathbb{R}^+)$ that is useful to recall is the following handy characterisation of the subcategory Stoch.

Lemma 6.1. A morphism $\boldsymbol{f} \colon A \to B$ in $\mathsf{Mat}(\mathbb{R}^+)$ is a stochastic matrix (thus a morphism of Stoch) if and only if (3) holds, that is, if $\text{⫯} \circ \boldsymbol{f} = \text{⫯}$.

A suitable notion of 'stochastic map with a hole' is provided by a *comb*. These structures originate in the study of certain kinds of quantum channels [5].

Definition 6.2. A 2-comb in Stoch is a morphism $\boldsymbol{f} \colon A_1 \otimes A_2 \to B_1 \otimes B_2$ satisfying, for some other morphism $\boldsymbol{f}' \colon A_1 \to B_1$,



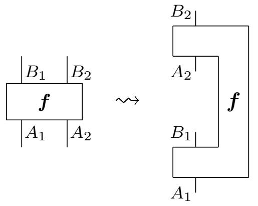
(9)

For this notion of 2-comb it is important to consider the map f with an explicit description of the product $\otimes$ on its domain and codomain.

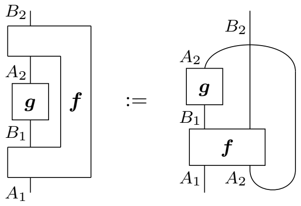
This definition extends inductively to *n-combs*, where we require that discarding the rightmost output yields $\boldsymbol{f}' \otimes \text{⫯}$, for some $(n-1)$-comb $\boldsymbol{f}'$. However, for our purposes, restricting to 2-combs will suffice.

Remark 6.3. Note that Definition 6.2 depends on the specific decomposition of the domain and codomain into a monoidal product of subsystems. For example, any stochastic matrix is trivially a 2-comb with respect to the decompostion $\boldsymbol{f} : A \otimes I \to B \otimes I$, where I is the monoidal unit. Hence, the data associated with a 2-comb is actually a pair of input/output types and a morphism: $((A_1, B_1), (A_2, B_2), \boldsymbol{f} : A_1 \otimes A_2 \to B_1 \otimes B_2)$. We will supress this extra data when it is clear from context.

The intuition behind condition (9) is that the contribution from input A_2 is only visible via output B_2. Thus, if we discard B_2 we may as well discard A_2. In other words, the input/output pair A_2, B_2 happen 'after' the pair A_1, B_1. Hence, it is typical to depict 2-combs in the shape of a (hair) comb, with 2 'teeth', as in (10) below:



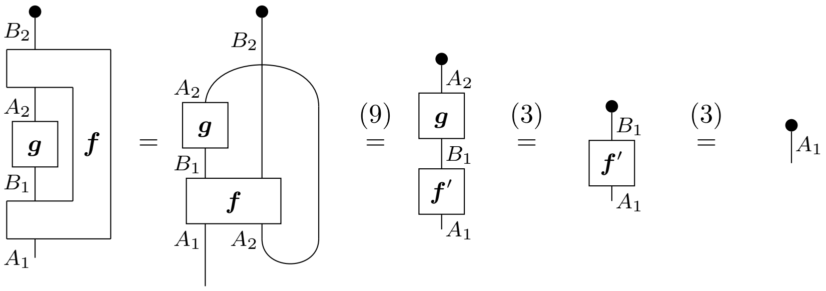
(10)



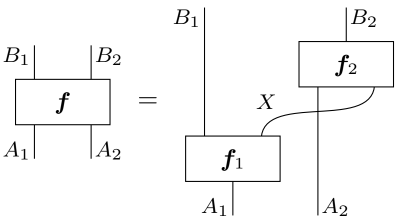
(11)

While combs themselves live in Stoch, $\mathsf{Mat}(\mathbb{R}^+)$ accommodates a second-order reading of the transition $\rightsquigarrow$ in (10): we can treat $\boldsymbol{f}$ as a map which expects as input a map $\boldsymbol{g}\colon B_1 \to A_2$ and produces as output a map of type $A_1 \to B_2$. Plugging $\boldsymbol{g}\colon B_1 \to A_2$ into the 2-comb can be formally defined in $\mathsf{Mat}(\mathbb{R}^+)$ by composing $\boldsymbol{f}$ and $\boldsymbol{g}$ in the usual way, then feeding the output of $\boldsymbol{g}$ into the second input of $\boldsymbol{f}$, using caps and cups, as in (11).

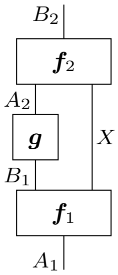
Importantly, for generic $\boldsymbol{f}$ and $\boldsymbol{g}$ of Stoch, there is no guarantee that forming the composite (11) in $\mathsf{Mat}(\mathbb{R}^+)$ yields a valid Stoch-morphism, i.e. a morphism satisfying the finality equation (3). However, if $\boldsymbol{f}$ is a 2-comb and $\boldsymbol{g}$ is a Stoch-morphism, equation (9) enables a discarding map plugged into the output B_2 in (11) to 'fall through' the right side of $\boldsymbol{f}$, which guarantees that the composed map satisfies the finality equation for discarding:



Remark 6.4. An alternative formulation of the comb composition in (11) can be made without leaving the sub-category of stochastic matrices. First, in Stoch, any stochastic matrix satisfying (9) can be *semi-localised*, i.e. factored into two parts $\boldsymbol{f}_1, \boldsymbol{f}_2$ as follows:

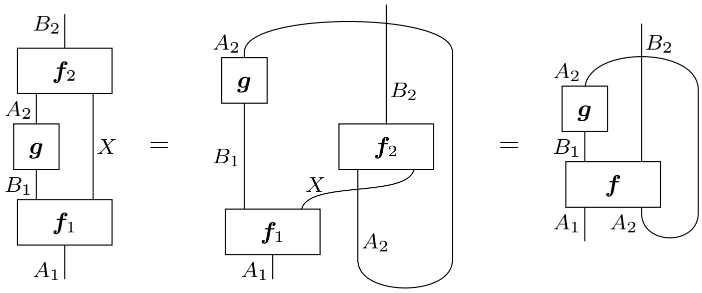


Then, one can show that the composition:



(12)

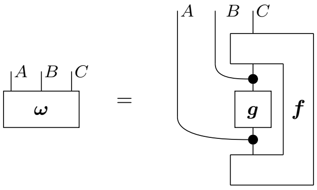
doesn't depend on the particular choice of $\boldsymbol{f}_1, \boldsymbol{f}_2$. The latter can be seen by concrete calculation, or by noting that (12), seen as a diagram in $\mathsf{Mat}(\mathbb{R}^+)$, can be deformed into (11) by re-introducing the feedback loop on A_2:



See e.g. [22] for more details. As it is possible to freely pass between $\boldsymbol{f}$ and $\boldsymbol{f}_1, \boldsymbol{f}_2$, one could equivalently define a comb as an equivalence class of pairs $(\boldsymbol{f}_1, \boldsymbol{f}_2)$ which compose to give $\boldsymbol{f}$, much like those considered in the construction of completely positive maps in [8].

With the concept of 2-combs in hand, we can state our factorisation result.

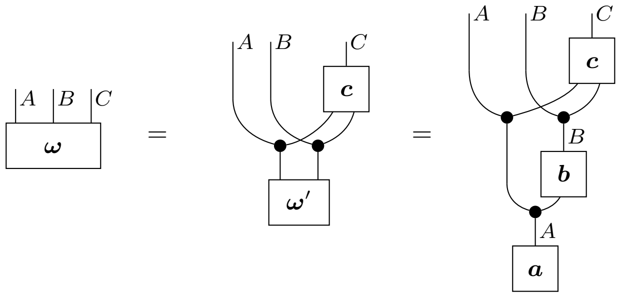
Theorem 6.5. For any state $\boldsymbol{\omega}\colon I \to A \otimes B \otimes C$ of Stoch with full support, there exists a unique 2-comb $\boldsymbol{f} : B \to A \otimes C$ and stochastic matrix $\boldsymbol{g}\colon A \to B$ such that, in $\mathsf{Mat}(\mathbb{R}^+)$:



(13)

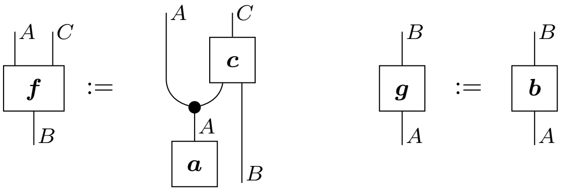
Proof. The construction of $\boldsymbol{f}$ and $\boldsymbol{g}$ mimics the construction of c-factors in [35], using string diagrams and diagrammatic disintegration. Starting with a full-support $\boldsymbol{\omega} : I \to A\otimes B\otimes C$, we apply Theorem 2.3 twice. First we can disintegrate $\boldsymbol{\omega}$ as $(\boldsymbol{\omega}' : I \to A\otimes B, \boldsymbol{c} :$

$A \otimes B \to C$) then further disintegrate $\boldsymbol{\omega}'$ into $(\boldsymbol{a} : I \to A, \boldsymbol{b} : A \to B)$:

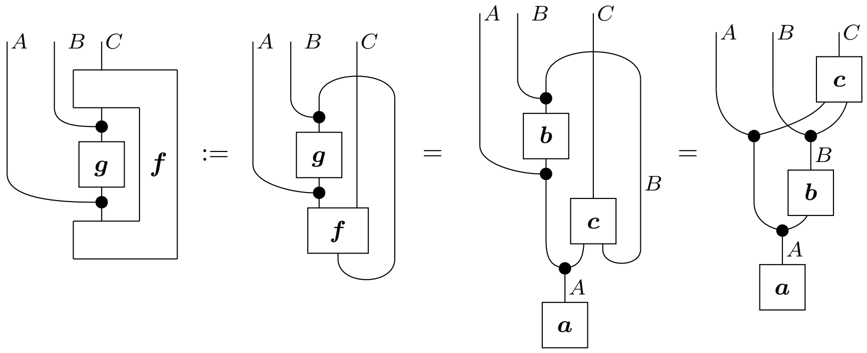


(14)

Now, we let:



Then (13) holds by construction of $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$:



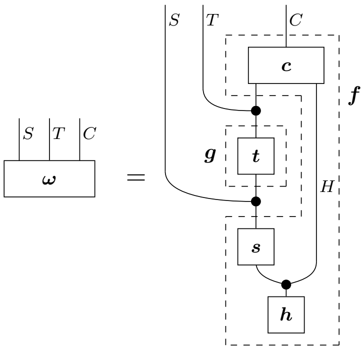
Note the last step above is just diagram deformation and the comonoid laws. The rightmost diagram above is equal to $\boldsymbol{\omega}$ by (14).

For uniqueness, suppose (13) holds for some other $\boldsymbol{f}', \boldsymbol{g}'$. Then by uniqueness of disintegration, it follows that $\boldsymbol{g}' = \boldsymbol{g} = \boldsymbol{b}$. To show that $\boldsymbol{f} = \boldsymbol{f}'$, we expand (13) explicitly in terms of matrices. This equation is equivalent to $\boldsymbol{\omega}^{ijk} = \boldsymbol{f}^{ik}_j \boldsymbol{g}^j_i = (\boldsymbol{f}')^{ik}_j \boldsymbol{g}^j_i$. Note that if $\boldsymbol{g}$ had any zero elements, $\boldsymbol{\omega}$ would not have full support, hence $\boldsymbol{g}^j_i \neq 0$ and therefore $\boldsymbol{f}^{ik}_j = (\boldsymbol{f}')^{ik}_j$ for all i, j, k. □

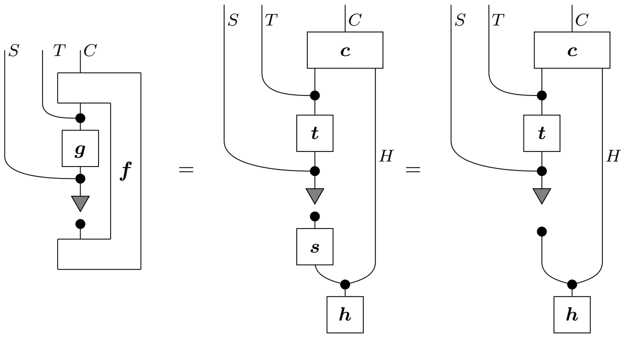
Note that Theorem 6.5 generalises the normal disintegration property given in Theorem 2.3. The latter is recovered by taking $A := I$ (or $C := I$) above.

7. Returning to the Smoking Scenario

We now return to the smoking scenario of Section 4. There, we concluded by claiming that the introduction of an intermediate variable T to the observational distribution $\boldsymbol{\omega} : I \to S \otimes T \otimes C$ would enable us to calculate the interventional distribution. That is, we can calculate $\boldsymbol{\omega'} = \mathcal{F}(\mathsf{cut}_X(\omega))$ from $\boldsymbol{\omega} := \mathcal{F}(\omega)$. Thanks to Theorem 6.5, we are now able to perform that calcuation. We first observe that our assumed causal structure for $\boldsymbol{\omega}$ fits the form of Theorem 6.5, where $\boldsymbol{g}$ is $\boldsymbol{t}$ and $\boldsymbol{f}$ is a 2-comb containing everything else, as in the diagram on the side.



Hence, $\boldsymbol{f}$ and $\boldsymbol{g}$ are computable from $\boldsymbol{\omega}$. If we plug them back together as in (13), we will get $\boldsymbol{\omega}$ back. However, if we insert a 'cut' between $\boldsymbol{f}$ and $\boldsymbol{g}$:



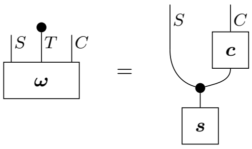
(15)

we obtain $\boldsymbol{\omega'} = \mathcal{F}(\mathsf{cut}_X(\omega))$.

Let us now consider a concrete example. We fix interpretations for the sets S, T, and C as booleans: $S = T = C = \{0, 1\}$ and let $\boldsymbol{\omega} : I \to S \otimes T \otimes C$ be the stochastic matrix:

$$\boldsymbol{\omega} := \begin{pmatrix} 0.5 \\ 0.1 \\ 0.01 \\ 0.02 \\ 0.1 \\ 0.05 \\ 0.02 \\ 0.2 \end{pmatrix} \begin{array}{l} \leftarrow P(S=0,T=0,C=0) \\ \leftarrow P(S=0,T=0,C=1) \\ \leftarrow P(S=0,T=1,C=0) \\ \leftarrow P(S=0,T=1,C=1) \\ \leftarrow P(S=1,T=0,C=0) \\ \leftarrow P(S=1,T=0,C=1) \\ \leftarrow P(S=1,T=1,C=0) \\ \leftarrow P(S=1,T=1,C=1) \end{array}$$

Now, disintegrating $\boldsymbol{\omega}$:



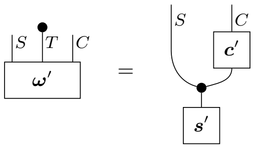
gives $\quad \boldsymbol{c} \approx \begin{pmatrix} 0.81 & 0.32 \\ 0.19 & 0.68 \end{pmatrix}$

The bottom-left element of $\boldsymbol{c}$ is $P(C = 1|S = 0)$, whereas the bottom-right is $P(C = 1|S = 1)$, so this suggests that patients are ≈ 3.5 times as likely to get cancer if they smoke (68% vs. 19%). However, comb-disintegrating $\boldsymbol{\omega}$ using Theorem 6.5 gives $\boldsymbol{g} \colon S \to T$

and a comb $\boldsymbol{f} \colon T \to S \otimes C$ with the following stochastic matrices:

$$\boldsymbol{f} \approx \begin{pmatrix} 0.53 & 0.21 \\ 0.11 & 0.42 \\ 0.25 & 0.03 \\ 0.12 & 0.34 \end{pmatrix} \qquad\qquad \boldsymbol{g} \approx \begin{pmatrix} 0.95 & 0.41 \\ 0.05 & 0.59 \end{pmatrix}$$

Recomposing these with a 'cut' in between, as in the left-hand side of (15), gives the interventional distribution $\boldsymbol{\omega}' \approx (0.38, 0.11, 0.01, 0.02, 0.16, 0.05, 0.07, 0.22)$. Disintegrating:



$$\text{gives} \qquad \boldsymbol{c}' \approx \begin{pmatrix} 0.75 & 0.46 \\ 0.25 & 0.54 \end{pmatrix}.$$

From the interventional distribution, we conclude that, in a (hypothetical) clinical trial, patients are about twice as likely to get cancer if they smoke (54% vs. 25%). So, since $54 < 68$, there was *some* confounding influence between S and C in our observational data, but after removing it via comb disintegration, we see there is still a signficant causal link between smoking and cancer.

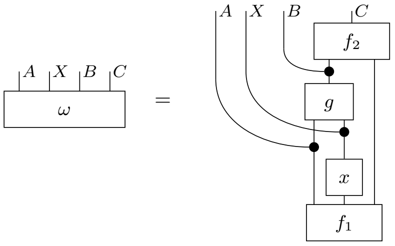
Note this conclusion depends totally on the particular observational data that we picked. For a different interpretation of $\boldsymbol{\omega}$ in Stoch, one might conclude that there is *no* causal connection, or even that smoking *decreases* the chance of getting cancer. Interestingly, all three cases can arise even when a naïve analysis of the data shows a strong direct correlation between S and C. To see and/or experiment with these cases, we have provided the Python code[‡] used to perform these calculations. See also [27] for a pedagocical overview of this example (using traditional Bayesian network language) with some sample calculations.

8. Generic Single Interventions

While we applied the comb decomposition to a particular example, this technique applies essentially unmodified to many examples where we intervene at a single variable (called X below) within an arbitrary causal structure.

[‡] https://gist.github.com/akissinger/aeec1751792a208253bda491ead587b6

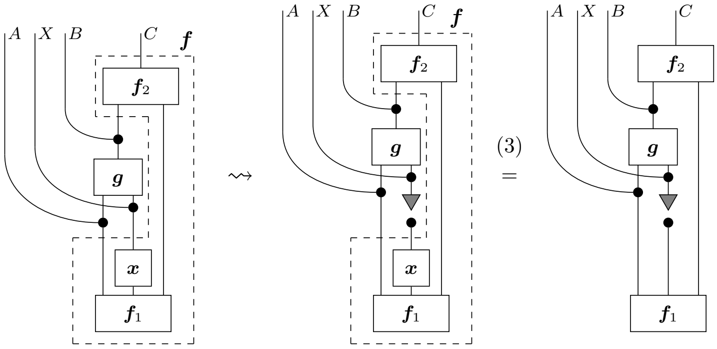
Theorem 8.1. Let G be a dag with a fixed node X that has corresponding generator $x\colon Y_1 \otimes \ldots \otimes Y_n \to X$ in Syn_G. Then, let ω be a morphism in Syn_G of the following form:



(16)

for some morphisms f_1, f_2 and g in Syn_G not containing x as a subdiagram. Then the interventional distribution $\boldsymbol{\omega}' := \mathcal{F}(\mathsf{cut}_X(\omega))$ is computable from any observational distribution $\boldsymbol{\omega} = \mathcal{F}(\omega)$ with full support.

Proof. The proof is very close to the example in the previous section. Interpreting ω into Stoch, we get a diagram of stochastic maps, which we can comb-disintegrate, then recompose with ▼ ∘ ● to produce the interventional distribution:



The right-hand-side above is then $\mathcal{F}(\mathsf{cut}_X(\omega))$. □

This is general enough to cover several well-known sufficient conditions from the causality literature, including single-variable versions of the so-called *front-door* and *back-door* criteria, as well as the sufficient condition based on confounding paths given by Pearl and Tian [35]. As the latter subsumes the other two, we will say a few words about the relationship between the Pearl/Tian condition about confounding paths and Theorem 8.1. In [35], the authors focus on *semi-Markovian* models, where the only latent variables have exactly two observed children and no parents. Suppose we write $A \leftrightarrow B$ if two observed variables are connected by a latent common cause, then one can characterise *confounding paths* as the transitive closure of $\leftrightarrow$. They go on to show that the interventional distribution corresponding to cutting X is computable whenever there are no confounding paths connecting X to one of its children.

We can compare this to the form of expression ω in equation (16). First, note this factorisation implies that all boxes which take X as an input must occur as sub-diagrams of g. Hence, any 'confounding path' connecting X to its children would yield at least one (un-copied) wire from f_1 to g, hence it cannot be factored as (16). Conversely, if there are no confounding paths from X to its children, then we can place the boxes involved in any other confounding path either entirely inside of g or entirely outside of g and obtain factorisation (16). Hence, restricting to semi-Markovian models, the non-confounding-path condition from [35] is equivalent to ours. However, Theorem 8.1 is slightly more general: its formulation doesn't rely on the causal structure ω being semi-Markovian.

9. Counterfactuals

While interventional distributions are a powerful tool for extracting causal information from a probabilistic model, there are certain cases where genuine causal influences remain hidden. In those cases, we can extend string diagram surgery techniques to do *counterfactual reasoning*. This enables us to reason about alternatives for events that have already occurred. For example, we can consider the likelihood of statements like:

> "Had Sally been able to play, the team would have won the game."

Implicit in this statement is the assumption that Sally was, in fact, *not* able to play. We can make this assumption explicit as follows:

> "Given Sally was not able to play, the team would have won, had she been able to play."

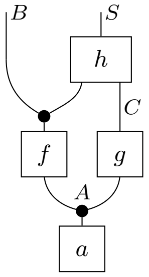
We therefore end up in the seemingly paradoxical situation of needing to condition on a real world event (Sally not being able to play) and considering the outcome of a contradictory event (Sally being able to play).

In order to evaluate the likelihood of such a statement, we need to be able to compare two worlds: the real world, containing events that have already happened, with a hypothetical world which is the same in every way except for a single intervention, namely making Sally able to play.

To see how we could compute probabilities for counterfactual statements using string diagram surgery, we will work out a concrete example of Balke and Pearl [2]. We will do this by giving a diagrammatic version of the 'twin model' technique used there.

Consider three people, Ann, Bob and Carl, who may or may not go to a party. The likelihood that Ann goes to the party is 60%. Bob is very likely (90%) to go to the party if Ann is there, but he will almost surely not go otherwise (97%). For Carl, it is the opposite: he is very likely to go (95%) if Ann does not go, and otherwise he will probably not go (90%). If Bob and Carl both go to the party, they are very likely to have a scuffle (95%). If only one of them is going, it is certain there will be no scuffle (100%), whereas if both are not going there is still a small chance (5%) of them getting in a scuffle somewhere else.

We can express this situation with the following string diagram, where the wires labelled A, B, and C represent whether Ann, Bob, or Carl go to the party, and S whether there is a scuffle:[§]



(17)

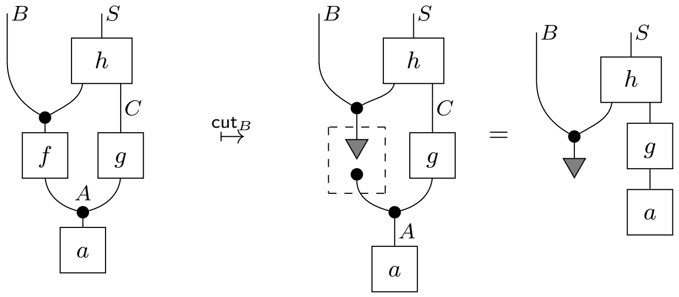
The probabilities given before are then modelled by the following stochastic matrices:

$$\boldsymbol{a} := \begin{pmatrix} 0.40 \\ 0.60 \end{pmatrix} \qquad \boldsymbol{f} := \begin{pmatrix} 0.90 & 0.03 \\ 0.10 & 0.97 \end{pmatrix} \qquad \boldsymbol{g} := \begin{pmatrix} 0.10 & 0.95 \\ 0.90 & 0.05 \end{pmatrix} \qquad \boldsymbol{h} := \begin{pmatrix} 0.95 & 1 & 1 & 0.05 \\ 0.05 & 0 & 0 & 0.95 \end{pmatrix}$$

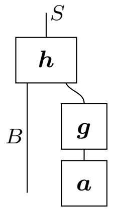
An example of a counterfactual statement is the following:

> "Given Bob did not go to the party, there would have been a scuffle, had he gone." (18)

Since we know all the stochastic matrices already, we can compute the (naïve) interventional distribution to predict how likely a scuffle is if we intervene and *make* Bob go to the party:



Disintegrating on B yields:



$$\approx \begin{pmatrix} 0.97 & 0.58 \\ 0.03 & 0.42 \end{pmatrix}$$

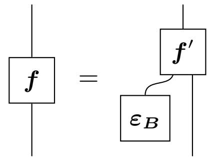
The bottom-right corner of this matrix corresponds to $P(S = 1|B = 1)$, and as we can see, the probability of a scuffle, given the intervention setting $B = 1$, is 42%.

However, we have ignored an important piece of information: Bob actually did *not* go

[§] Note in (17) we treat A and C as latent variables, and choose to observe B and S. This is because B and S are involved in the counterfactual query (18) under consideration.

to the party. Hence, to assign a more accurate probability to statement (18), we should take this into account. As we said before, to do this requires us to consider two copies of the world, one which actually happened (where Bob did not go to the party) and one where we introduce a hypothetical intervention (making Bob go to the party). Crucially, everything else should be kept the same between the two copies of the world.

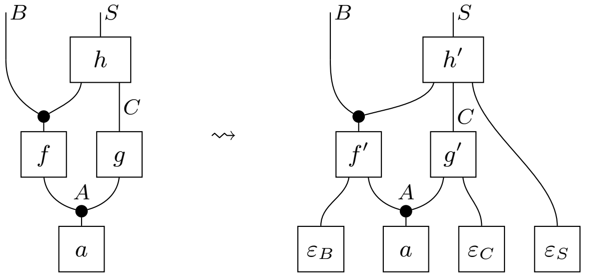
So, what does it mean to "keep everything else the same"? For example, if we look at the stochastic map $\boldsymbol{f}$ relating Ann's attendance to Bob's, the randomness in $\boldsymbol{f}$ can be interpreted as the presence of other, possibly unknown variables affecting Bob's attendance (e.g. the weather or Bob not feeling well). We can isolate those variables by factoring $\boldsymbol{f}$ into a deterministic part $\boldsymbol{f'}$ and an *exogenous state* $\boldsymbol{\varepsilon_B}$. For example, one such factorisation is:



$$\text{where} \qquad \boldsymbol{f'} := \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix} \qquad \boldsymbol{\varepsilon_B} := \begin{pmatrix} 0.027 \\ 0.873 \\ 0.003 \\ 0.097 \end{pmatrix}$$

The deterministic part is a function, which can be expressed as a stochastic matrix whose entries are all only 0 or 1. It is well-known that such a factorisation for stochastic matrices is always possible. For example, it is suggested by Fritz [13] as a candidate axiom for 'well-behaved' probabilistic categories (of which Stoch, which Fritz calls FinStoch, is an example).

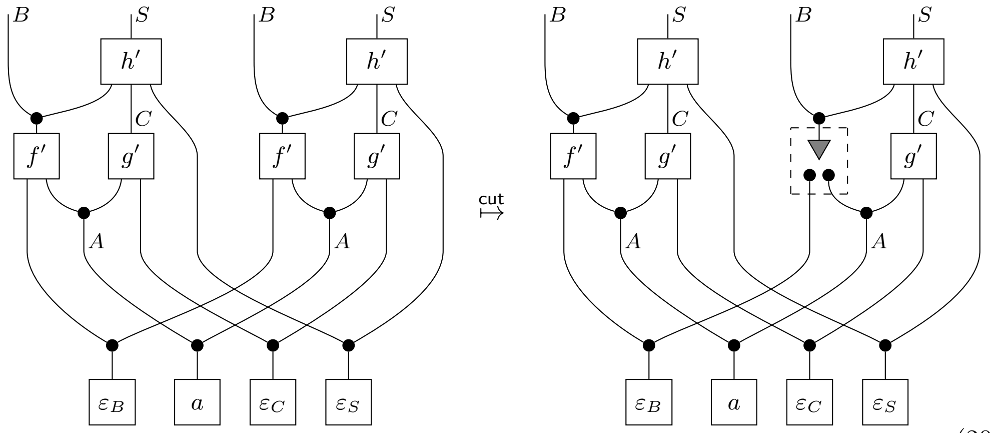
Decomposing f, g, and h into function and exogenous variables, we are able to *exogonize* the diagram, i.e. obtain a new diagram where all of the randomness occurs in boxes with no inputs:



(19)

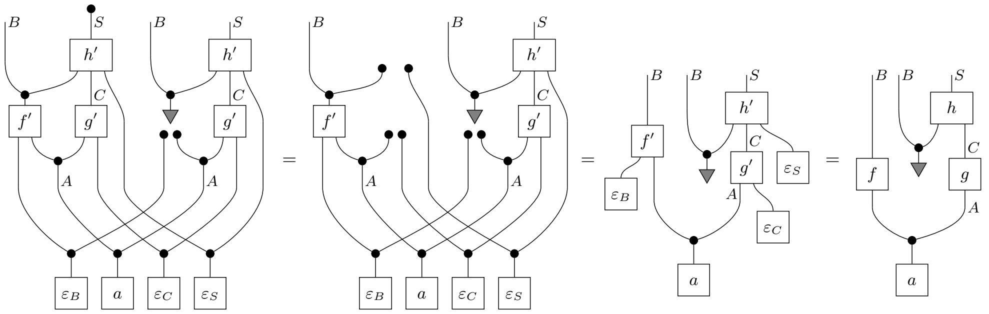
Note that, since a doesn't have any inputs, we can already treat it as an exogenous state.

In order to answer the counterfactual query, we transform our diagram into a "twin" diagram, which contains two copies of all the variables: the first copy represents the "factual world" in which Bob did not go to the party, whereas the second copy represents a counterfactual world, in which Bob went to the party. In order to keep everything else but Bob's attendance the same, the two copies share *the same* exogenous states. We then intervene on the counterfactual copy of B by cutting out the rightmost f':

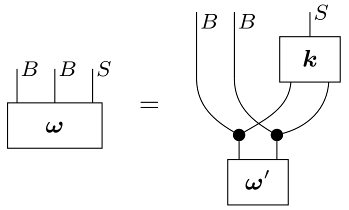


(20)

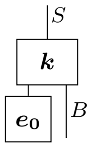
To evaluate the probability of the counterfactual statement (18), we first marginalise out the left copy of S and do some simplification:



The result is a state $\omega : I \to B \otimes B \otimes S$. Disintegrating over the two copies of B, we obtain a stochastic map from 'real' B and 'counterfactual' B to S:



We can condition on the 'real' $B = 0$ by plugging the state $\boldsymbol{e}_0 := (1, 0)$ (meaning "Bob did not go to the party") into the first wire of $\boldsymbol{k}$, to obtain:

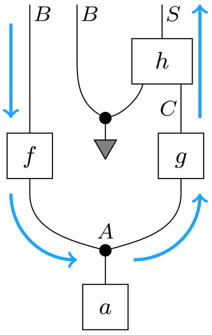


$$\approx \begin{pmatrix} 0.99 & 0.21 \\ 0.01 & 0.79 \end{pmatrix} \tag{21}$$

Again looking at the bottom-right corner, we see the hypothetical probability of a scuffle has gone up to 79%, matching the conclusion from [2].

This is because knowing that Bob actually did not attend the party enables us to draw certain conclusions about what happened at the actual party. Namely, if Bob was not there, it is less likely that Ann was there, which in turn means it was more likely that Carl was. Hence, if we want to think about what would have happened at that *same* party, had Bob gone, he most likely would have seen Carl, and there most likely would have been some fisticuffs.

Crucially, the exogenous states are *shared* between the two copies the diagram, rather than split into two independent copies. This is what enables us to imagine that the real and counterfactual halves are talking about the exact same event, rather than a separate event that happens to be modelled by the same stochastic processes. We can see in that case that the shared $\boldsymbol{a}$ box is what enables information to flow between the two worlds:



That is: when we condition on the counterfactual copy of B in equation (21), this has an effect on the value of S for all values of the 'real' copy of B.

One other thing to note from this example is that it does not actually depend on the decompositions of stochastic maps into functions and exogenous variables. This is not the case in general, and furthermore such decompositions are not unique. Hence counterfactual reasoning is only well-defined when either (i) the decomposition into functions and exogenous states is given in advance, or (ii) the resulting counterfactual query does not depend on the choice of decomposition.

In [2], the authors use this fact to state that causal Bayesian networks alone are not adequate for counterfactual reasoning, and emphasise instead the utility of structural equational models for this task. However, in certain situations, like the party example, Bayesian networks do suffice.

10. Conclusion and Future Work

This paper takes a fresh, systematic look at the problem of causal identifiability. By clearly distinguishing syntax (string diagram surgery and identification of comb shapes) and semantics (comb-disintegration of joint states) we obtain a clear methodology for computing interventional distributions, and hence causal effects, from observational data. Furthermore, we show that diagram surgery can be used to model counterfactual queries.

A natural next step is moving beyond single-variable interventions to the general case, i.e. situations where we allow interventions on multiple variables which may have some arbitrary causal relationships connecting them. This would mean extending the comb factorisation Theorem 6.5 from a 2-comb and a channel to arbitrary n-combs. This seems to be straightforward, via an inductive extension of the proof in Section 8. A more substantial direction of future work will be the strengthening of Theorem 8.1 from sufficient conditions for causal identifiability to a full characterisation. Indeed, the related condition based on confounding paths from [35] is a necessary and sufficient condition for computing the interventional distribution on a single variable. Hence, it will be interesting to formalise this necessity proof (and more general versions, e.g. [16]) within our framework and investigate, for example, the extent to which it holds beyond the semi-Markovian case.

Throughout this work, we have relied crucially on the fact that observational data has full support. This assumption seems to often be made without comment in the literature on causal identifiability. For example, Tian and Pearl make this assumption implicitly in [35] by dividing by conditional probabilities at will. However, it is a rather strong assumption, and in particular rules out deterministic relationships between variables. There are some tricks one can do to cope with this problem (see e.g. the discussion of determinism in the context of causal discovery in [26]), so it would be interesting to see if they can be adapted to our setting.

While we focus exclusively on the case of taking models in Stoch in this paper, the techniques we gave are posed at an abstract level in terms of composition and factorisation. Hence, we are optimistic about their prospects to generalise to other probabilistic (e.g. infinite discrete and continuous variables) and quantum settings. In the latter case, this could provide insights into the emerging field of *quantum causal structures* [11, 15, 25, 31, 32], which attempts in part to replay some of the results coming from statistical causal reasoning, but where quantum processes play a role analogous to stochastic ones. A key difficulty in applying our framework to a category of quantum processes, rather than Stoch, is the unavailability of canonical (basis-independent) 'copy' morphisms due to the quantum no-cloning theorem [36]. However, a recent proposal for the formulation of 'quantum common causes' [1] suggests a (partially-defined) analogue to the role played by 'copy' in our formulation constructed via multiplication of certain commuting Choi matrices. Hence, it may yet be possible to import results from classical causal reasoning into the quantum case just by changing the category of models.

Acknowledgements Fabio Zanasi is supported by EPSRC grant EP/V002376/1.

References

[1] John-Mark A. Allen, Jonathan Barrett, Dominic C. Horsman, Ciarán M. Lee, and Robert W. Spekkens. Quantum common causes and quantum causal models. *Phys. Rev. X*, 7:031021, Jul 2017.

[2] Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. In B. Hayes-Roth and R. Korf, editors, *Proc. 12th Nat. Conf. on Artificial Intelligence*, pages 230–237. AAAI Press / The MIT Press, 1994.

[3] Filippo Bonchi, Paweł Sobociński, and Fabio Zanasi. Deconstructing Lawvere with distributive laws. *J. Log. Algebr. Meth. Program.*, 95:128–146, 2018.

[4] Benjamin Cabrera, Tobias Heindel, Reiko Heckel, and Barbara König. Updating probabilistic knowledge on condition/event nets using Bayesian networks. In *29th International Conference on Concurrency Theory, CONCUR 2018, September 4-7, 2018, Beijing, China*, pages 27:1–27:17, 2018.

[5] Giulio Chiribella, Giacomo Mauro D'Ariano, and Paolo Perinotti. Quantum circuit architecture. *Phys. Rev. Lett.*, 101:060401, Aug 2008.

[6] Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Math. Struct. in Comp. Sci.*, 29(7):938–971, 2019.

[7] Florence Clerc, Vincent Danos, Fredrik Dahlqvist, and Ilias Garnier. Pointless learning. In J. Esparza and A. Murawski, editors, *Foundations of Software Science and Computation Structures*, number 10203 in Lect. Notes Comp. Sci., pages 355–369. Springer, Berlin, 2017.

[8] Bob Coecke and Chris Heunen. Pictures of complete positivity in arbitrary dimension. *Information and Computation*, 250:50–58, 2016.

[9] Bob Coecke and Robert W. Spekkens. Picturing classical and quantum Bayesian inference. *Synthese*, 186(3):651–696, 2012.

[10] Andrea Corradini and Fabio Gadducci. An algebraic presentation of term graphs, via GS-monoidal categories. *Applied Categorical Structures*, 7(4):299–331, 1999.

[11] Fabio Costa and Sally Shrapnel. Quantum causal modelling. *New Journal of Physics*, 18(6):063032, 2016.

[12] Brendan Fong. Causal theories: A categorical perspective on Bayesian networks. Master's thesis, Univ. of Oxford, 2012. see arxiv.org/abs/1301.6201.

[13] T. Fritz. A synthetic approach to Markov kernels, conditional independence, and theorems on sufficient statistics. *Advances in Math.*, 370:107239, 2020.

[14] Gus Gutoski and John Watrous. Toward a general theory of quantum games. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 565–574. ACM, 2007.

[15] Joe Henson, Raymond Lal, and Matthew F. Pusey. Theory-independent limits on correlations from generalized Bayesian networks. *New Journal of Physics*, 16(11):113043, 2014.

[16] Yimin Huang and Marco Valtorta. On the completeness of an identifiability algorithm for semi-Markovian models. *Annals of Mathematics and Artificial Intelligence*, 54(4):363–408, Dec 2008.

[17] Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. *CoRR*, abs/1206.6831, 2012.

[18] B. Jacobs and F. Zanasi. The logical essentials of Bayesian reasoning. In G. Barthe, J.-P. Katoen, and A. Silva, editors, *Foundations of Probabilistic Programming*, pages 295–331. Cambridge Univ. Press, 2021.

[19] Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. Causal inference by string diagram surgery. In *Foundations of Software Science and Computation Structures - 22nd International Conference, FOSSACS 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6-11, 2019, Proceedings*, pages 313–329, 2019.

[20] Bart Jacobs and Fabio Zanasi. A predicate/state transformer semantics for Bayesian learning. *Electr. Notes Theor. Comput. Sci.*, 325:185–200, 2016.

[21] Bart Jacobs and Fabio Zanasi. A formal semantics of influence in Bayesian reasoning. In K. Larsen, H. Bodlaender, and J.-F. Raskin, editors, *Math. Found. of Computer Science*, volume 83 of *LIPIcs*, pages 21:1–21:14. Schloss Dagstuhl, 2017.

[22] Aleks Kissinger and Sander Uijlen. A categorical semantics for causal structure. In *32nd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2017, Reykjavik, Iceland, June 20-23, 2017*, pages 1–12, 2017.

[23] F. William Lawvere. Functorial semantics of algebraic theories. *Proceedings of the National Academy of Sciences of the United States of America*, 50(5):869, 1963.

[24] F. William Lawvere. Ordinal sums and equational doctrines. In B. Eckmann, editor, *Seminar on Triples and Categorical Homology Theory*, volume 80 of *Lecture Notes in Mathematics*, pages 141–155. Springer-Verlag, 1969.

[25] Matthew S. Leifer and Robert W. Spekkens. Towards a formulation of quantum theory as a causally neutral theory of Bayesian inference. *Phys. Rev. A*, 88:052130, Nov 2013.

[26] Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

[27] Michael Nielsen. If correlation doesn't imply causation, then what does? Available at http://www.michaelnielsen.org/ddi/if-correlation-doesnt-imply-causation-then-what-does, accessed: 2018-11-15.

[28] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[29] Judea Pearl and Thomas S. Verma. *A theory of inferred causation*. Morgan Kaufmann, San Mateo, CA., 1991.

[30] Paolo Perinotti. Causal structures and the classification of higher order quantum computations. 2016.

[31] Jacques Pienaar and Časlav Brukner. A graph-separation theorem for quantum causal models. *New Journal of Physics*, 17(7):073020, 2015.

[32] Katja Ried, Megan Agnew, Lydia Vermeyden, Dominik Janzing, Robert W. Spekkens, and Kevin J. Resch. A quantum advantage for inferring causal structure. *Nat Phys*, 11:1745–2473, 2015.

[33] Peter Selinger. A survey of graphical languages for monoidal categories. In Bob Coecke, editor, *New Structures for Physics*, volume 813 of *Lecture Notes in Physics*, pages 289–355. Springer, 2011.

[34] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21th National Conference on Artificial Intelligence*, pages 1219–1226. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[35] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, July 28 - August 1, 2002, Edmonton, Alberta, Canada.*, pages 567–573, 2002.

[36] William K. Wootters and Wojciech H. Zurek. A single quantum cannot be cloned. *Nature*, 299(5886):802–803, 1982.