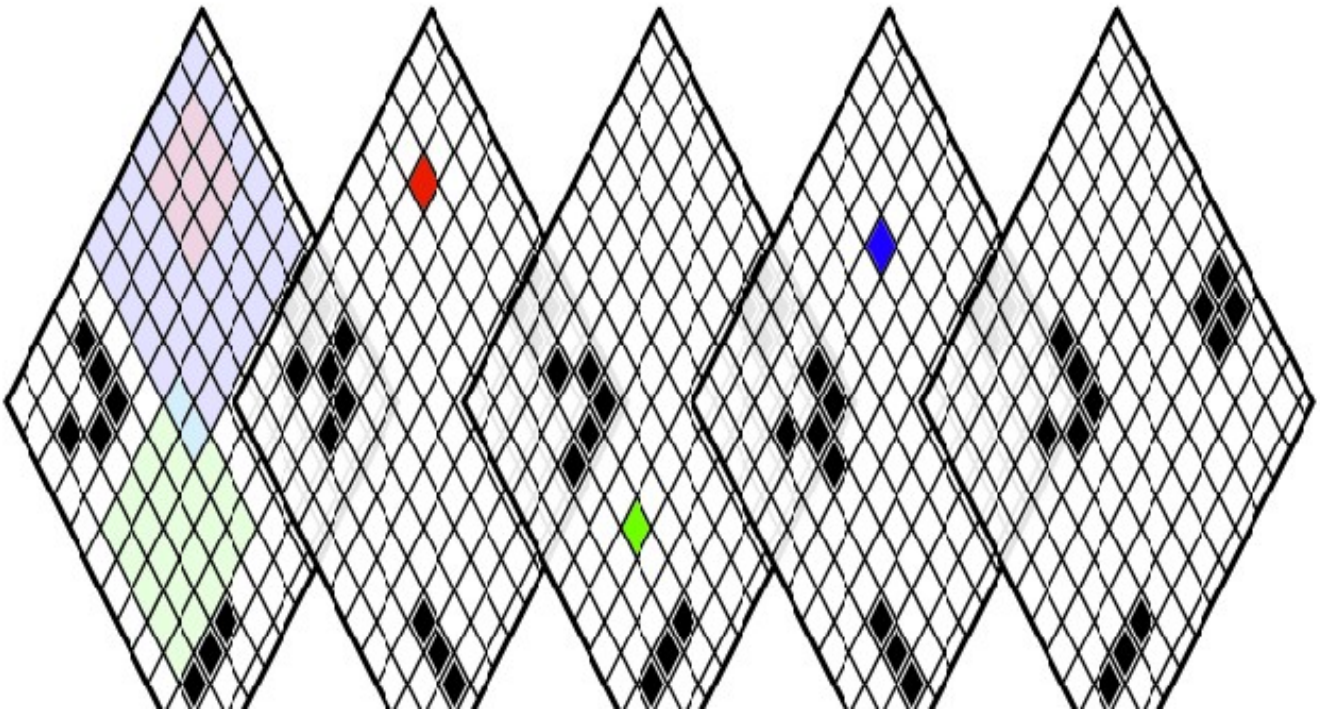


Finite Factored Sets

1. Cartesian Frames and Factored Sets on ArXiv
2. Finite Factored Sets
3. Finite Factored Sets: Introduction and Factorizations
4. Finite Factored Sets: Orthogonality and Time
5. Finite Factored Sets: Conditional Orthogonality
6. Finite Factored Sets: Polynomials and Probability
7. Finite Factored Sets: Inferring Time
8. Finite Factored Sets: Applications
9. Saving Time
10. [AN #163]: Using finite factored sets for causal and temporal inference
11. AXRP Episode 9 - Finite Factored Sets with Scott Garrabrant
12. Finite Factored Sets in Pictures
13. Exploring Finite Factored Sets with some toy examples
14. A simple example of conditional orthogonality in finite factored sets
15. A second example of conditional orthogonality in finite factored sets
16. Counterfactability
17. Countably Factored Spaces

Cartesian Frames and Factored Sets on ArXiv

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Papers on Cartesian frames and factored sets are now on arXiv.

Cartesian Frames: https://arxiv.org/abs/2109.10996

Factored Sets: https://arxiv.org/abs/2109.11513

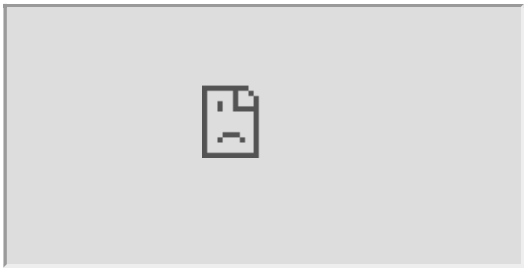
The factored set paper is approximately identical to the sequence here, while the Cartesian frame paper is rewritten by Daniel Hermann and Josiah Lopez-Wild, optimized for an audience of philosophers.

Finite Factored Sets

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the edited transcript of a talk introducing finite factored sets. For most readers, it will probably be the best starting point for learning about factored sets.

Video:



(Lightly edited) slides: https://intelligence.org/files/Factored-Set-Slides.pdf

1. Short Combinatorics Talk

1m. Some Context

Scott: So I want to start with some context. For people who are not already familiar with my work:

- My main motivation is to reduce existential risk.
- I try to do this by trying to figure out how to align advanced artificial intelligence.
- I try to do *this* by trying to become less confused about intelligence and optimization and agency and various things in that cluster.
- My main strategy here is to develop a theory of agents that are embedded in the environment that they're optimizing. I think there are a lot of open hard problems around doing this.
- This leads me to do a bunch of weird math and philosophy. This talk is going to be an example of some weird math and philosophy.

For people who *are* already familiar with my work, I just want to say that according to my personal aesthetics, the subject of this talk is about as exciting as Logical Induction, which is to say I'm really excited about it. And I'm really excited about this audience; I'm excited to give this talk right now.

1t. Factoring the Talk

This talk can be split into 2 parts:

- Part 1: a short pure-math combinatorics talk.

I suspect that if I were better, I would instead be giving a short pure-math category theory talk; but I'm trained as a combinatorialist, so I'm giving a combinatorics talk upfront.

- Part 2: a more applied and philosophical main talk.

This talk can also be split into *4* parts differentiated by color: Motivation, Table of Contents, Main Body, and Examples. Combining these gives us 8 parts (some of which are not contiguous):

	Part 1: Short Talk	**Part 2: The Main Talk**
Motivation	1m. Some Context	2m. The Pearlian Paradigm
ToC	1t. Factoring the Talk	2t. We Can Do Better
Body	1b. Set Partitions, etc.	2b. Time and Orthogonality, etc.
Examples	1e. Enumerating Factorizations	2e. Game of Life, etc.

1b. Set Partitions

All right. Here's some background math:

- A **partition** of a set S is a set X of non-empty subsets of S, called **parts**, such that for each $s \in S$ there exists a unique part in X that contains s.
- Basically, a partition of S is a way to view S as a disjoint union. We have parts that are disjoint from each other, and they union together to form S.
- We'll write Part(S) for the set of all partitions of S.
- We'll say that a partition X is **trivial** if it has exactly one part.
- We'll use bracket notation, $[s]_X$, to denote the unique part in X containing s. So this is like the equivalence class of a given element.
- And we'll use the notation $s \sim_X t$ to say that two elements s and t are in the same part in X.

You can also think of partitions as being like variables on your set S. Viewed in that way, the values of a partition X correspond to which part an element is in.

Or you can think of X as a *question* that you could ask about a generic element of S. If I have an element of S and it's hidden from you and you want to ask a question about it, each possible question corresponds to a partition that splits up S according to the different possible answers.

We're also going to use the lattice structure of partitions:

- We'll say that $X \geq_S Y$ (X is finer than Y, and Y is coarser than X) if X makes all of the distinctions that Y makes (and possibly some more distinctions), i.e., if for all $s, t \in S$, $s \sim_X t$ implies $s \sim_Y t$. You can break your set S into parts, Y, and then break it into smaller parts, X.
- $X \vee_S Y$ (the common refinement of X and Y) is the coarsest partition that is finer than both X and Y. This is the unique partition that makes all of the distinctions that either X or Y makes, and no other distinctions. This is well-defined, which I'm not going to show here.

Hopefully this is mostly background. Now I want to show something new.

1b. Set Factorizations

A **factorization** of a set S is a set B of nontrivial partitions of S, called **factors**, such that for each way of choosing one part from each factor in B, there exists a unique element of S in the intersection of those parts.

So this is maybe a little bit dense. My short tagline of this is: "A factorization of S is a way to view S as a product, in the exact same way that a partition was a way to view S as a disjoint union."

If you take one definition away from this first talk, it should be the definition of factorization. I'll try to explain it from a bunch of different angles to help communicate the concept.

If $B = \{b_0, \dots, b_n\}$ is a factorization of S, then there exists a bijection between S and $b_0 \times \cdots \times b_n$ given by $s \mapsto ([s]_{b_0}, \dots, [s]_{b_n})$. This bijection comes from sending an element of S to the tuple consisting only of parts containing that element. And as a consequence of this bijection, $|S| = \prod_{b \in B} |b|$.

So we're really viewing S as a product of these individual factors, with no additional structure.

Although we won't prove this here, something else you can verify about factorizations is that all of the parts in a factor have to be of the same size.

We'll write Fact(S) for the set of all factorizations of S, and we'll say that a **finite factored set** is a pair (S, B), where S is a finite set and $B \in \text{Fact}(S)$.

Note that the relationship between S and B is somewhat loopy. If I want to define a factored set, there are two strategies I could use. I could first introduce the S, and break it into factors. Alternatively, I could first introduce the B. *Any* time I have a finite collection of finite sets B, I can take their product and thereby produce an S, modulo the degenerate case where some of the sets are empty. So S can just be the product of a finite collection of arbitrary finite sets.

To my eye, this notion of factorization is extremely natural. It's basically the multiplicative analog of a set partition. And I really want to push that point, so here's another attempt to push that point:

A **partition** is a set X of non-empty subsets of S such that the obvious function from the disjoint union of the elements of X to S is a bijection.	A **factorization** is a set B of non-trivial partitions of S such that the obvious function to the product of the elements of B from S is a bijection.

I can take a slightly modified version of the partition definition from before and dualize a whole bunch of the words, and get out the set factorization definition.

Hopefully you're now kind of convinced that this is an extremely natural notion.

Andrew Critch: Scott, in one sense, you're treating "subset" as dual to partition, which I think is valid. And then in another sense, you're treating "factorization" as dual to partition. Those are both valid, but maybe it's worth talking about the two kinds of duality.

Scott: Yeah. I think what's going on there is that there are two ways to view a partition. You can view a partition as "that which is dual to a subset," and you can also view a partition as something that is built up out of subsets. These two different views do different things when you dualize.

Ramana Kumar: I was just going to check: You said you can start with an arbitrary B and

then build the S from it. It can be literally any set, and then there's always an S...

Scott: If none of them are empty, yes, you could just take a collection of sets that are kind of arbitrary elements. And you can take their product, and you can identify with each of the elements of a set the subset of the product that projects on to that element.

Ramana Kumar: Ah. So the S in that case will just be tuples.

Scott: That's right.

Brendan Fong: Scott, given a set, I find it very easy to come up with partitions. But I find it less easy to come up with factorizations. Do you have any tricks for...?

Scott: For that, I should probably just go on to the examples.

Joseph Hirsh: Can I ask one more thing before you do that? You allow factors to have one element in them?

Scott: I said "nontrivial," which means it does not have one element.

Joseph Hirsh: "Nontrivial" means "not have one element, and not have no elements"?

Scott: No, the empty set has a partition (with no parts), and I will call that nontrivial. But the empty set thing is not that critical.

I'm now going to move on to some examples.

1e. Enumerating Factorizations

Exercise! What are the factorizations of the set $\{0, 1, 2, 3\}$?

Spoiler space:

.

.

.

.

.

.

.

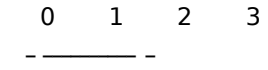
.

.

.

First, we're going to have a kind of trivial factorization:

$$\{ \quad \{ \{ 0 \}, \{ 1 \}, \{ 2 \}, \{ 3 \} \} \quad \}$$



We only have one factor, and that factor is the discrete partition. You can do this for any set, as long as your set has at least two elements.

Recall that in the definition of factorization, we wanted that for each way of choosing one part from each factor, we had a unique element in the intersection of those parts. Since we only have one factor here, satisfying the definition just requires that for each way of choosing one part from the discrete partition, there exists a unique element that is in that part.

And then we want some less trivial factorizations. In order to have a factorization, we're going to need some partitions. And the product of the cardinalities of our partitions are going to have to equal the cardinality of our set S, which is 4.

The only way to express 4 as a nontrivial product is to express it as 2×2. Thus we're looking for factorizations that have 2 factors, where each factor has 2 parts.

We noted earlier that all of the parts in a factor have to be of the same size. So we're looking for 2 partitions that each break our 4-element set into 2 sets of size 2.

So if I'm going to have a factorization of $\{0, 1, 2, 3\}$ that isn't this trivial one, I'm going to have to pick 2 partitions of my 4-element set that each break the set into 2 parts of size 2. And there are 3 partitions of a 4-element sets that break it up into 2 parts of size 2. For each way of choosing a pair of these 3 partitions, I'm going to get a factorization.

$$\{ \quad \{ \{ 0, 1 \}, \{ 2, 3 \} \}, \quad \{ \{ 0, 2 \}, \{ 1, 3 \} \} \quad \}$$

0	1
2	3

{ { { 0 , 1 } , { 2 , 3 } } ,
{ { { 0 , 3 } , { 1 , 2 } } }

0	1
3	2

{ { { 0 , 2 } , { 1 , 3 } } ,
{ { { 0 , 3 } , { 1 , 2 } } }

0	2
3	1

So there will be 4 factorizations of a 4-element set.

In general you can ask, "How many factorizations are there of a finite set of size n?". Here's a little chart showing the answer for n ≤ 25:

S		F a c t (S)
0	1
1	1
2	1
3	1
4	4
5	1
6	61
7	1
8	1681
9	5041
10	15121
11	1
12	13638241
13	1
14	8648641
15	1816214401
16	181880899201
17	1
18	45951781075201
19	1
20	3379365788198401
21	1689515283456001
22	14079294028801
23	1
24	4454857103544668620801
25	538583682060103680001

You'll notice that if n is prime, there will be a single factorization, which hopefully makes sense. This is the factorization that only has one factor.

A very surprising fact to me is that this sequence did not show up on OEIS, which is this database that combinatorialists use to check whether or not their sequence has been studied before, and to see connections to other sequences.

To me, this just feels like the multiplicative version of the Bell numbers. The Bell numbers count how many partitions there are of a set of size n. It's sequence number 110 on OEIS out of over 300,000;

and this sequence just doesn't show up at all, even when I tweak it and delete the degenerate cases and so on.

I am very confused by this fact. To me, factorizations seem like an extremely natural concept, and it seems to me like it hasn't really been studied before.

This is the end of my short combinatorics talk.

> **Ramana Kumar:** If you're willing to do it, I'd appreciate just stepping through one of the examples of the factorizations and the definition, because this is pretty new to me.
>
> **Scott:** Yeah. Let's go through the first nontrivial factorization of $\{0,1,2,3\}$:
>
> $$\{\ \{\{0,1\},\{2,3\}\},\ \{\{0,2\},\{1,3\}\}\ \}$$
>
> | | |
> |---|---|
> | 0 | 1 |
> | 2 | 3 |
>
> In the definition, I said a factorization should be a set of partitions such that for each way of choosing one part from each of the partitions, there will be a unique element in the intersection of those parts.
>
> Here, I have a partition that's separating the small numbers from the large numbers: $\{\{0,1\},\{2,3\}\}$. And I also have a partition that's separating the even numbers from the odd numbers: $\{\{0,2\},\{1,3\}\}$.
>
> And the point is that for each way of choosing either "small" or "large" and also choosing "even" or "odd", there will be a unique element of S that is the conjunction of these two choices.
>
> In the other two nontrivial factorizations, I replace either "small and large" or "even and odd" with an "inner and outer" distinction.
>
> **David Spivak:** For partitions and for many things, if I know the partitions of a set A and the partitions of a set B, then I know some partitions of A + B (the disjoint union) or I know some partitions of A × B. Do you know any facts like that for factorizations?
>
> **Scott:** Yeah. If I have two factored sets, I can get a factored set over their product, which sort of disjoint-unions the two collections of factors. For the additive thing, you're not going to get anything like that because prime sets don't have any nontrivial factorizations.

All right. I think I'm going to move on to the main talk.

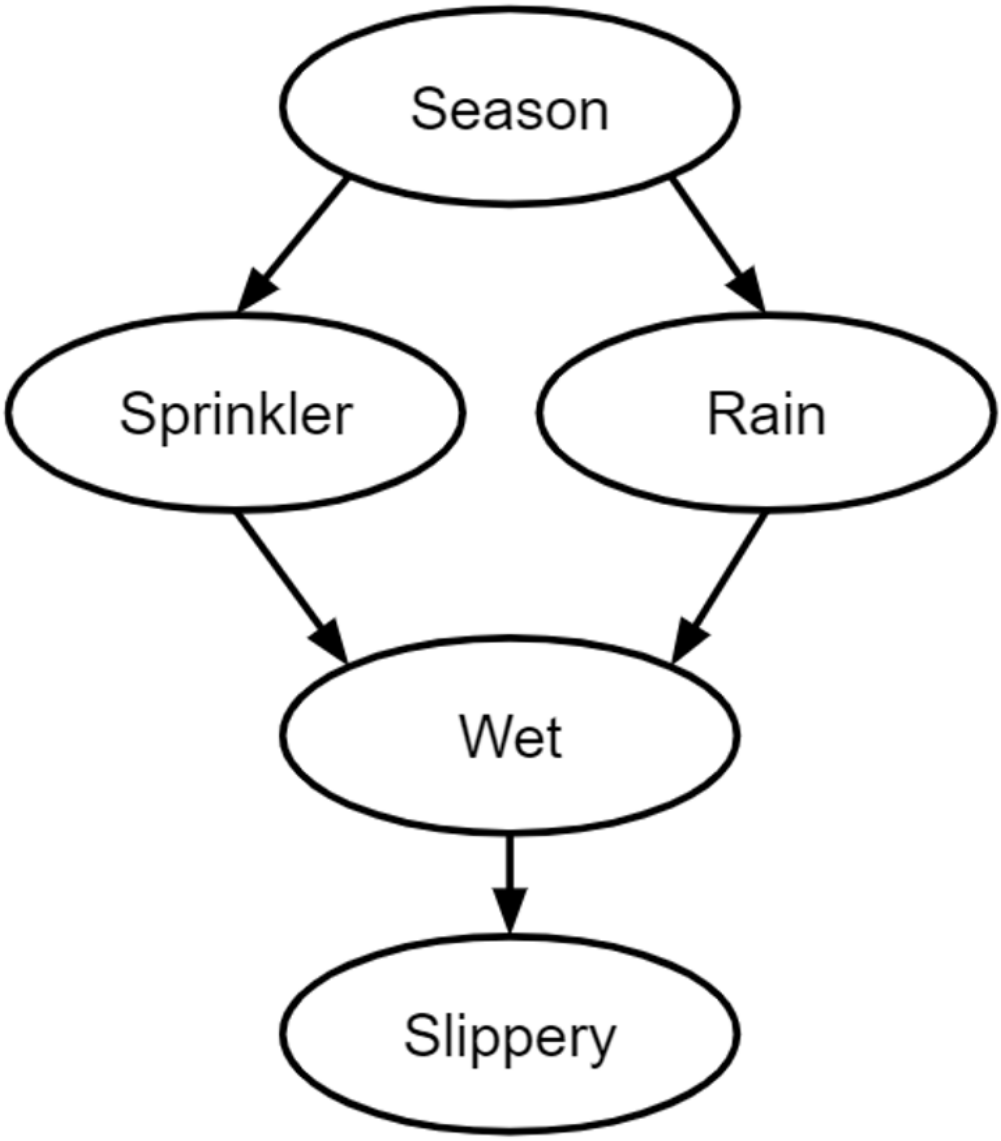
2. The Main Talk (It's About Time)

2m. The Pearlian Paradigm

We can't talk about time without talking about Pearlian causal inference. I want to start by saying that I think the Pearlian paradigm is great. This buys me some crackpot points, but I'll say it's the best thing to happen to our understanding of time since Einstein.

I'm not going to go into all the details of Pearl's paradigm here. My talk will not be technically dependent on it; it's here for motivation.

Given a collection of variables and a joint probability distribution over those variables, Pearl can infer causal/temporal relationships between the variables. (In this talk I'm going to use "causal" and "temporal" interchangeably, though there may be more interesting things to say here philosophically.)



Pearl can infer temporal data from statistical data, which is going against the adage that "correlation does not imply causation." It's like Pearl is taking the combinatorial structure of your correlation and

using that to infer causation, which I think is just really great.

> **Ramana Kumar:** I may be wrong, but I think this is false. Or I think that that's not all Pearl needs—just the joint distribution over the variables. Doesn't he also make use of intervention distributions?
>
> **Scott:** In the theory that is described in chapter two of the book *Causality*, he's not really using other stuff. Pearl builds up this bigger theory elsewhere. But you have some strong ability, maybe assuming simplicity or whatever (but not assuming you have access to extra information), to take a collection of variables and a joint distribution over those variables, and infer causation from correlation.
>
> **Andrew Critch:** Ramana, it depends a lot on the structure of the underlying causal graph. For some causal graphs, you can actually recover them uniquely with no interventions. And only assumptions with zero-measure exceptions are needed, which is really strong.
>
> **Ramana Kumar:** Right, but then the information you're using is the graph.
>
> **Andrew Critch:** No, you're not. Just the joint distribution.
>
> **Ramana Kumar:** Oh, okay. Sorry, go ahead.
>
> **Andrew Critch:** There exist causal graphs with the property that if nature is generated by that graph and you don't know it, and then you look at the joint distribution, you will infer with probability 1 that nature was generated by that graph, without having done any interventions.
>
> **Ramana Kumar:** Got it. That makes sense. Thanks.
>
> **Scott:** Cool.

I am going to (a little bit) go against this, though. I'm going to claim that Pearl *is* kind of cheating when making this inference. The thing I want to point out is that in the sentence "Given a collection of variables and a joint probability distribution over those variables, Pearl can infer causal/temporal relationships between the variables.", the words "Given a collection of variables" are actually hiding a lot of the work.

The emphasis is usually put on the joint probability distribution, but Pearl is not inferring temporal data from statistical data alone. He is inferring temporal data from statistical data **and factorization data**: how the world is broken up into these variables.

I claim that this issue is also entangled with a failure to adequately handle abstraction and determinism. To point at that a little bit, one could do something like say:

"Well, what if I take the variables that I'm given in a Pearlian problem and I just forget that structure? I can just take the product of all of these variables that I'm given, and consider the space of all partitions on that product of variables that I'm given; and each one of those partitions will be its own variable. And then I can try to do Pearlian causal inference on this big set of all the variables that I get by forgetting the structure of variables that were given to me."

And the problem is that when you do that, you have a bunch of things that are deterministic functions of each other, and you can't actually infer stuff using the Pearlian paradigm.

So in my view, this cheating is very entangled with the fact that Pearl's paradigm isn't great for handling abstraction and determinism.

2t. We Can Do Better

The main thing we'll do in this talk is we're going to introduce an alternative to Pearl that does not rely on factorization data, and that therefore works better with abstraction and determinism.

Where Pearl was given a collection of variables, we are going to just consider all partitions of a given set. Where Pearl infers a directed acyclic graph, we're going to infer a finite factored set.

In the Pearlian world, we can look at the graph and read off properties of time and orthogonality/independence. A directed path between nodes corresponds to one node being before the other, and two nodes are independent if they have no common ancestor. Similarly, in our world, we will be able to read time and orthogonality off of a finite factored set.

(Orthogonality and independence are pretty similar. I'll use the word "orthogonality" when I'm talking about a combinatorial notion, and I'll use "independence" when I'm talking about a probabilistic notion.)

In the Pearlian world, *d*-separation, which you can read off of the graph, corresponds to conditional independence in all probability distributions that you can put on the graph. We're going to have a fundamental theorem that will say basically the same thing: conditional orthogonality corresponds to conditional independence in all probability distributions that we can put on our factored set.

In the Pearlian world, *d*-separation will satisfy the compositional graphoid axioms. In our world, we're just going to satisfy the compositional semigraphoid axioms. The fifth graphoid axiom is one that I claim you shouldn't have even wanted in the first place.

Pearl does causal inference. We're going to talk about how to do temporal inference using this new paradigm, and infer some very basic temporal facts that Pearl's approach can't. (Note that Pearl can also sometimes infer temporal relations that *we* can't—but only, from our point of view, because Pearl is making additional factorization assumptions.)

And then we'll talk about a bunch of applications.

Pearl	**This Talk**
A Given Collection of Variables	All Partitions of a Given Set
Directed Acyclic Graph	Finite Factored Set
Directed Path Between Nodes	"Time"
No Common Ancestor	"Orthogonality"
d-Separation	"Conditional Orthogonality"
Compositional Graphoid	Compositional Semigraphoid
d-Separation ↔ Conditional Independence	The Fundamental Theorem
Causal Inference	Temporal Inference
Many Many Applications	Many Many Applications

Excluding the motivation, table of contents, and example sections, this table also serves as an outline of the two talks. We've already talked about set partitions and finite factored sets, so now we're going to talk about time and orthogonality.

2b. Time and Orthogonality

I think that if you capture one definition from this second part of the talk, it should be this one. Given a finite factored set as context, we're going to define the history of a partition.

Let $F = (S, B)$ be a finite factored set. And let $X, Y \in \text{Part}(S)$ be partitions of S.

The **history** of X, written $h^F(X)$, is the smallest set of factors $H \subseteq B$ such that for all $s, t \in S$, if $s \sim_b t$ for all $b \in H$, then $s \sim_X t$.

The history of X, then, is the smallest set of factors H—so, the smallest subset of B—such that if I take an element of S and I hide it from you, and you want to know which part in X it is in, it suffices for me to tell you which part it is in within each of the factors in H.

So the history H is a set of factors of S, and knowing the values of all the factors in H is sufficient to know the value of X, or to know which part in X a given element is going to be in. I'll give an example soon that will maybe make this a little more clear.

We're then going to define **time** from history. We'll say that X is **weakly before** Y, written $X \preceq^F Y$, if $h^F(X) \subseteq h^F(Y)$. And we'll say that X is **strictly before** Y, written $X \prec^F Y$, if $h^F(X) \subset h^F(Y)$.

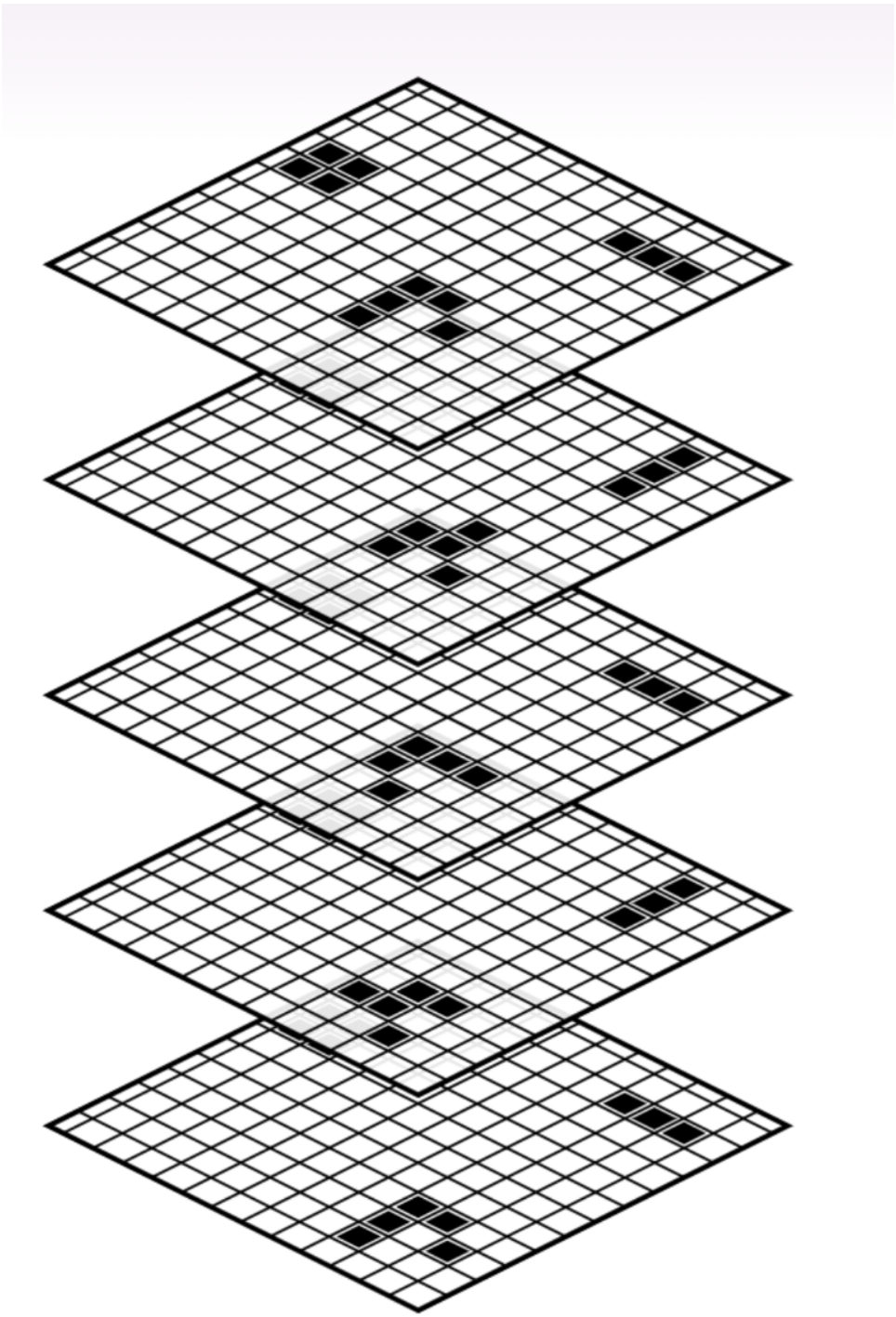
One analogy one could draw is that these histories are like the past light cones of a point in spacetime. When one point is before another point, then the backwards light cone of the earlier point is going to be a subset of the backwards light cone of the later point. This helps show why "before" can be like a subset relation.

We're also going to define orthogonality from history. We'll say that two partitions X and Y are **orthogonal**, written $X \perp^F Y$, if their histories are disjoint: $h^F(X) \cap h^F(Y) = \{\}$.

Now I'm going to go through an example.

2e. Game of Life

Let S be the set of all Game of Life computations starting from an $[-n, n] \times [-n, n]$ board.



Let $R = \{(r, c, t) \in Z^3 \mid 0 \le t \le n,\ |r| \le n - t,\ |c| \le n - t\}$ (i.e., cells computable from the initial $[-n, n] \times [-n, n]$ board). For $(r, c, t) \in R$, let $\ell(r, c, t) \subseteq S$ be the set of all computations such that the cell at row r and column c is alive at time t.

(Minor footnote: I've done some small tricks here in order to deal with the fact that the Game of Life is normally played on an infinite board. We want to deal with the finite case, and we don't want to worry about boundary conditions, so we're only going to look at the cells that are uniquely determined by the initial board. This means that the board will shrink over time, but this won't matter for our example.)

S is the set of all Game of Life computations, but since the Game of Life is deterministic, the set of all computations is in bijective correspondence with the set of all initial conditions. So $|S| = 2^{(2n+1)^2}$, the number of initial board states.

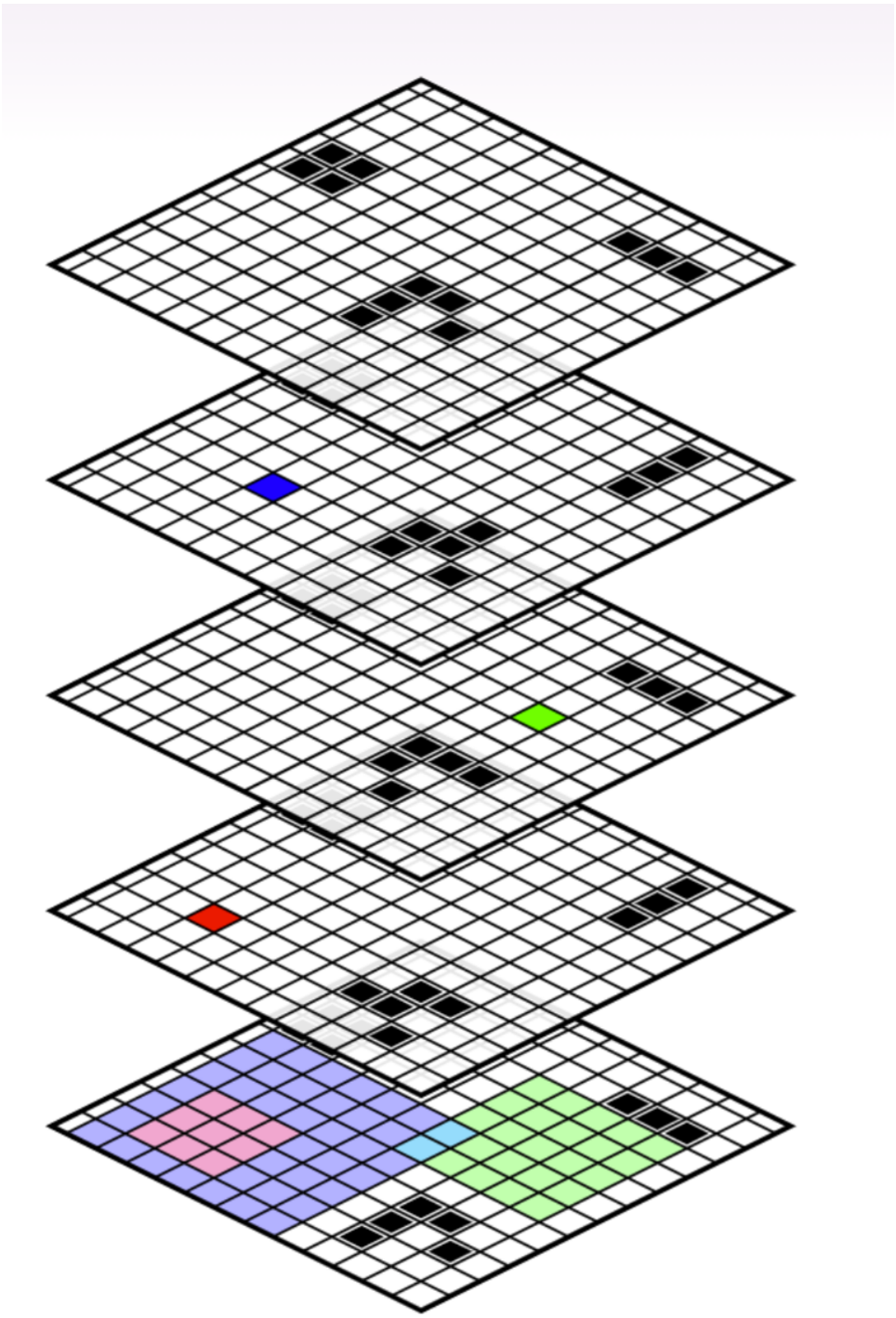
This also gives us a nice factorization on the set of all Game of Life computations. For each cell, there's a partition that separates out the Game of Life computations in which that cell is alive at time 0 from the ones where it's dead at time 0. Our factorization, then, will be a set of $(2n + 1)^2$ binary factors, one for each question of "Was this cell alive or dead at time 0?".

Formally: For $(r, c, t) \in R$, let $L_{(r,c,t)} = \{\ell(r, c, t), S \setminus \ell(r, c, t)\}$. Let $F = (S, B)$, where

$B = \{L_{(r,c,0)} \mid -n \leq r, c \leq n\}$.

There will also be other partitions on this set of all Game of Life computations that we can talk about. For example, you can take a cell and a time t and say, "Is this cell alive at time t?", and there will be a partition that separates out the computations where that cell is alive at time t from the computations where it's dead at time t.

Here's an example of that:



The lowest grid shows a section of the initial board state.

The blue, green, and red squares on the upper boards are (cell, time) pairs. Each square corresponds to a partition of the set of all Game of Life computations, "Is that cell alive or dead at the given time t ?"

The history of that partition is going to be all the cells in the initial board that go into computing whether the cell is alive or dead at time t. It's everything involved in figuring out that cell's state. E.g., knowing the state of the nine light-red cells in the initial board always tells you the state of the red cell in the second board.

In this example, the partition corresponding to the red cell's state is strictly before the partition corresponding to the blue cell. The question of whether the red cell is alive or dead is before the question of whether the blue cell is alive or dead.

Meanwhile, the question of whether the red cell is alive or dead is going to be *orthogonal* to the question of whether the green cell is alive or dead.

And the question of whether the blue cell is alive or dead is *not* going to be orthogonal to the question of whether the green cell is alive or dead, because they intersect on the cyan cells.

Generalizing the point, fix $X = L_{(r_X, c_X, t_X)}$, $Y = L_{(r_Y, c_Y, t_Y)}$, where $(r_X, c_X, t_X), (r_Y, c_Y, t_Y) \in R$. Then:

- $h^F(X) = \{L_{(r,c,0)} \in B \mid |r_X - r| \leq t_X, |c_X - c| \leq t_X\}$.
- $X <^F Y$ if and only if $t_X < t_Y$ and $|r_Y - r_X|, |c_Y - c_X| \leq t_Y - t_X$.
- $X \perp^F Y$ if and only if $|r_Y - r_X| > t_Y + t_X$ or $|c_Y - c_X| > t_Y + t_X$.

We can also see that the blue and green cells look *almost* orthogonal. If we condition on the values of the two cyan cells in the intersection of their histories, *then* the blue and green partitions become orthogonal. That's what we're going to discuss next.

> **David Spivak:** A priori, that would be a gigantic computation—to be able to tell me that you understand the factorization structure of that Game of Life. So what intuition are you using to be able to make that claim, that it has the kind of factorization structure you're implying there?
>
> **Scott:** So, I've defined the factorization structure.
>
> **David Spivak:** You gave us a certain factorization already. So somehow you have a very good intuition about *history*, I guess. Maybe that's what I'm asking about.
>
> **Scott:** Yeah. So, if I didn't give you the factorization, there's this obnoxious number of factorizations that you could put on the set here. And then for the history, the intuition I'm using is: "What do I need to know in order to compute this value?"
>
> I actually went through and I made little gadgets in Game of Life to make sure I was right here, that every single cell actually could in some situations affect the cells in question. But yeah, the intuition that I'm working from is mostly about the information in the computation. It's "Can I construct a situation where if only I knew this fact, I would be able to compute what this value is? And if I can't, then it can take two different values."
>
> **David Spivak:** Okay. I think deriving that intuition from the definition is something I'm missing, but I don't know if we have time to go through that.
>
> **Scott:** Yeah, I think I'm not going to here.

2b. Conditional Orthogonality

So, just to set your expectations: Every time I explain Pearlian causal inference to someone, they say that *d*-separation is the thing they can't remember. *d*-separation is a much more complicated concept than "directed paths between nodes" and "nodes without any common ancestors" in Pearl; and similarly, conditional orthogonality will be much more complicated than time and orthogonality in our paradigm. Though I do think that conditional orthogonality has a much simpler and nicer definition than *d*-separation.

We'll begin with the definition of conditional history. We again have a fixed finite set as our context. Let $F = (S, B)$ be a finite factored set, let $X, Y, Z \in \text{Part}(S)$, and let $E \subseteq S$.

The **conditional history** of X given E, written $h^F(X|E)$, is the smallest set of factors $H \subseteq B$ satisfying the following two conditions:

- For all $s, t \in E$, if $s \sim_b t$ for all $b \in H$, then $s \sim_X t$.
- For all $s, t \in E$ and $r \in S$, if $r \sim_{b_0} s$ for all $b_0 \in H$ and $r \sim_{b_1} t$ for all $b_1 \in B \setminus H$, then $r \in E$.

The first condition is much like the condition we had in our definition of history, except we're going to make the assumption that we're in E. So the first condition is: if all you know about an object is that it's in E, and you want to know which part it's in within X, it suffices for me to tell you which part it's in within each factor in the history H.

Our second condition is not actually going to mention X. It's going to be a relationship between E and H. And it says that if you want to figure out whether an element of S is in E, it's sufficient to parallelize and ask two questions:

- "If I only look at the values of the factors in H, is 'this point is in E' compatible with that information?"
- "If I only look at the values of the factors in $B \setminus H$, is 'this point is in E' compatible with that information?"

If both of these questions return "yes", then the point has to be in E.

I am not going to give an intuition about why this needs to be a part of the definition. I will say that without this second condition, conditional history would not even be well-defined, because it wouldn't be closed under intersection. And so I wouldn't be able to take the smallest set of factors in the subset ordering.

Instead of justifying this definition by explaining the intuitions behind it, I'm going to justify it by using it and appealing to its consequences.

We're going to use conditional history to define **conditional orthogonality**, just like we used history to define orthogonality. We say that X and Y are **orthogonal given** $E \subseteq S$, written $X \perp^F Y \mid E$, if the history of X given E is disjoint from the history of Y given E: $h^F(X|E) \cap h^F(Y|E) = \{\}$.

We say X and Y are **orthogonal given** $Z \in \text{Part}(S)$, written $X \perp^F Y \mid Z$, if $X \perp^F Y \mid z$ for all $z \in Z$. So what it means to be orthogonal given a partition is just to be orthogonal given each individual way that the partition might be, each individual part in that partition.

I've been working with this for a while and it feels pretty natural to me, but I don't have a good way to push the naturalness of this condition. So again, I instead want to appeal to the consequences.

2b. Compositional Semigraphoid Axioms

Conditional orthogonality satisfies the **compositional semigraphoid axioms**, which means finite factored sets are pretty well-behaved. Let $F = (S, B)$ be a finite factored set, and let $X, Y, Z, W \in \text{Part}(S)$ be partitions of S. Then:

- If $X \perp^F Y \mid Z$, then $Y \perp^F X \mid Z$. (*symmetry*)
- If $X \perp^F (Y \vee_S W) \mid Z$, then $X \perp^F Y \mid Z$ and $X \perp^F W \mid Z$. (*decomposition*)
- If $X \perp^F (Y \vee_S W) \mid Z$, then $X \perp^F Y \mid (Z \vee_S W)$. (*weak union*)
- If $X \perp^F Y \mid Z$ and $X \perp^F W \mid (Z \vee_S Y)$, then $X \perp^F (Y \vee_S W) \mid Z$. (*contraction*)
- If $X \perp^F Y \mid Z$ and If $X \perp^F W \mid Z$, then $X \perp^F (Y \vee_S W) \mid Z$. (*composition*)

The first four properties here make up the semigraphoid axioms, slightly modified because I'm working with partitions rather than sets of variables, so union is replaced with common refinement. There's another graphoid axiom which we're not going to satisfy; but I argue that we don't want to satisfy it, because it doesn't play well with determinism.

The fifth property here, composition, is maybe one of the most unintuitive, because it's not exactly satisfied by probabilistic independence.

Decomposition and composition act like converses of each other. Together, conditioning on Z throughout, they say that X is orthogonal to both Y and W if and only if X is orthogonal to the common refinement of Y and W.

2b. The Fundamental Theorem

In addition to being well-behaved, I also want to show that conditional orthogonality is pretty powerful. The way I want to do this is by showing that conditional orthogonality exactly corresponds to conditional independence in all probability distributions you can put on your finite factored set. Thus, much like *d*-separation in the Pearlian picture, conditional orthogonality can be thought of as a combinatorial version of probabilistic independence.

A **probability distribution on a finite factored set** $F = (S, B)$ is a probability distribution P on S that can be thought of as coming from a bunch of independent probability distributions on each of the factors in B. So $P(s) = \prod_{b \in B} P([s]_b)$ for all $s \in S$.

This effectively means that your probability distribution factors the same way your set factors: the probability of any given element is the product of the probabilities of each of the individual parts that it's in within each factor.

The **fundamental theorem of finite factored sets** says: Let $F = (S, B)$ be a finite factored set, and let $X, Y, Z \in \text{Part}(S)$ be partitions of S. Then $X \perp^F Y \mid Z$ if and only if for all probability distributions P

on F, and all $x \in X$, $y \in Y$, and $z \in Z$, we have $P(x \cap z) \cdot P(y \cap z) = P(x \cap y \cap z) \cdot P(z)$. I.e., X is orthogonal to Y given Z if and only conditional independence is satisfied across all probability distributions.

This theorem, for me, was a little nontrivial to prove. I had to go through defining certain polynomials associated with the subsets, and then dealing with unique factorization in the space of these polynomials; I think the proof was eight pages or something.

The fundamental theorem allows us to infer orthogonality data from probabilistic data. If I have some empirical distribution, or I have some Bayesian distribution, I can use that to infer some orthogonality data. (We could also imagine orthogonality data coming from other sources.) And then we can use this orthogonality data to get temporal data.

So next, we're going to talk about how to get temporal data from orthogonality data.

2b. Temporal Inference

We're going to start with a finite set Ω, which is our sample space.

One naive thing that you might think we would try to do is infer a factorization of Ω. We're not going to do that because that's going to be too restrictive. We want to allow for Ω to maybe hide some information from us, for there to be some latent structure and such.

There may be some situations that are distinct without being distinct in Ω. So instead, we're going to infer a factored set model of Ω: some other set S, and a factorization of S, and a function from S to Ω.

A **model** of Ω is a pair (F, f), where $F = (S, B)$ is a finite factored set and $f : S \to \Omega$. (f need not be injective or surjective.)

Then if I have a partition of Ω, I can send this partition backwards across f and get a unique partition of S. If $X \in \text{Parts}(\Omega)$, then $f^{-1}(X) \in \text{Parts}(S)$ is given by $s \sim_{f^{-1}(X)} t \Leftrightarrow f(s) \sim_X f(t)$.

Then what we're going to do is take a bunch of orthogonality facts about Ω, and we're going to try to find a model which captures the orthogonality facts.

We will take as given an **orthogonality database** on Ω, which is a pair $D = (O, N)$, where O (for "orthogonal") and N (for "not orthogonal") are each sets of triples (X, Y, Z) of partitions of Ω. We'll think of these as rules about orthogonality.

What it means for a model (F, f) to satisfy a database D is:

- $f^{-1}(X) \perp^F f^{-1}(Y) \mid f^{-1}(Z)$ whenever $(X, Y, Z) \in O$, and
- $\neg(f^{-1}(X) \perp^F f^{-1}(Y) \mid f^{-1}(Z))$ whenever $(X, Y, Z) \in N$.

So we have these orthogonality rules we want to satisfy, and we want to consider the space of all models that are consistent with these rules. And even though there will always be infinitely many models that are consistent with my database, if at least one is—you can always just add more information that you then delete with f—we would like to be able to sometimes infer that for all models that satisfy our database, $f^{-1}(X)$ is before $f^{-1}(Y)$.

And this is what we're going to mean by inferring time. If all of our models (F, f) that are consistent with the database D satisfy some claim about time $f^{-1}(X) <^F f^{-1}(Y)$, we'll say that $X <_D Y$.

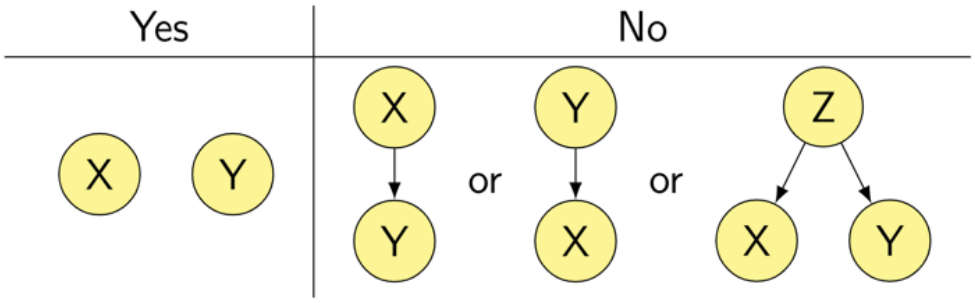
2e. Two Binary Variables (Pearl)

So we've set up this nice combinatorial notion of temporal inference. The obvious next questions are:

- Can we actually infer interesting facts using this method, or is it vacuous?
- And: How does this framework compare to Pearlian temporal inference?

Pearlian temporal inference is really quite powerful; given enough data, it can infer temporal sequence in a wide variety of situations. How powerful is the finite factored sets approach by comparison?

To address that question, we'll go to an example. Let X and Y be two binary variables. Pearl asks: "Are X and Y independent?" If yes, then there's no path between the two. If no, then there may be a path from X to Y, or from Y to X, or from a third variable to both X and Y.



In either case, we're not going to infer any temporal relationships.

To me, it feels like this is where the adage "correlation does not imply causation" comes from. Pearl really needs more variables in order to be able to infer temporal relationships from more rich combinatorial structures.

However, I claim that this Pearlian ontology in which you're handed this collection of variables has blinded us to the obvious next question, which is: is X independent of X XOR Y?

In the Pearlian world, X and Y were our variables, and X XOR Y is just some random operation on those variables. In our world, X XOR Y instead is a variable on the same footing as X and Y. The first

thing I do with my variables X and Y is that I take the product $X \times Y$ and then I forget the labels X and Y.

So there's this question, "Is X independent of X XOR Y?". And if X *is* independent of X XOR Y, we're actually going to be able to conclude that X is *before* Y!

So not only is the finite factored set paradigm non-vacuous, and not only is it going to be able to keep up with Pearl and infer things Pearl can't, but it's going to be able to infer a temporal relationship from only two variables.

So let's go through the proof of that.

2e. Two Binary Variables (Factored Sets)

Let $\Omega = \{00, 01, 10, 11\}$, and let X, Y, and Z be the partitions (/questions):

- $X = \{\{00, 01\}, \{10, 11\}\}$. (What is the first bit?)
- $Y = \{\{00, 10\}, \{01, 11\}\}$. (What is the second bit?)
- $Z = \{\{00, 11\}, \{01, 10\}\}$. (Do the bits match?)

Let $D = (O, N)$, where $O = \{(X, Z, \{\Omega\})\}$ and $N = \{(Z, Z, \{\Omega\})\}$. If we'd gotten this orthogonality database from a probability distribution, then we would have more than just two rules, since we would observe more orthogonality and non-orthogonality than that. But temporal inference is monotonic with respect to adding more rules, so we can just work with the smallest set of rules we'll need for the proof.

The first rule says that X is orthogonal to Z. The second rule says that Z is not orthogonal to itself, which is basically just saying that Z is non-deterministic; it's saying that both of the parts in Z are possible, that both are supported under the function f. The $\{\Omega\}$ indicates that we aren't making any conditions.

From this, we'll be able to prove that $X <_D Y$.

Proof. First, we'll show that that X is weakly before Y. Let (F, f) satisfy D. Let H_X be shorthand for $h^F(f^{-1}(X))$, and likewise let $H_Y = h^F(f^{-1}(Y))$ and $H_Z = h^F(f^{-1}(Z))$.

Since $(X, Z, \{\Omega\}) \in O$, we have that $H_X \cap H_Z = \{\}$; and since $(Z, Z, \{\Omega\}) \in N$, we have that $H_Z \neq \{\}$.

Since $X \leq_\Omega Y \vee_\Omega Z$—that is, since X can be computed from Y together with Z—$H_X \subseteq H_Y \cup H_Z$. (Because a partition's history is the smallest set of factors needed to compute that partition.)

And since $H_X \cap H_Z = \{\}$, this implies $H_X \subseteq H_Y$, so X is weakly before Y.

To show the strict inequality, we'll assume for the purpose of contradiction that $H_X = H_Y$.

Notice that Z can be computed from X together with Y—that is, $Z \leq_\Omega X \vee_\Omega Y$—and therefore $H_Z \subseteq H_X \cup H_Y$ (i.e., $H_Z \subseteq H_X$). It follows that $H_Z = (H_X \cup H_Y) \cap H_Z = H_X \cap H_Z$. But since H_Z is also disjoint from H_X, this means that $H_Z = \{\}$, a contradiction.

Thus $H_X \neq H_Y$, so $H_X \subset H_Y$, so $f^{-1}(X) <^F f^{-1}(Y)$, so $X <_D Y$. □

When I'm doing temporal inference using finite factored sets, I largely have proofs that look like this. We collect some facts about emptiness or non-emptiness of various Boolean combinations of histories of variables, and we use these to conclude more facts about histories of variables being subsets of each other.

I have a more complicated example that uses conditional orthogonality, not just orthogonality; I'm not going to go over it here.

One interesting point I want to make here is that we're doing temporal inference—we're inferring that X is before Y—but I claim that we're also doing conceptual inference.

Imagine that I had a bit, and it's either a 0 or a 1, and it's either blue or green. And these two facts are primitive and independently generated. And I also have this other concept that's like, "Is it grue or bleen?", which is the XOR of blue/green and 0/1.

There's a sense in which we're inferring X is before Y, and in that case, we can infer that blueness is before grueness. And that's pointing at the fact that blueness is more primitive, and grueness is a derived property.

In our proof, X and Z can be thought of as these primitive properties, and Y is a derived property that we're getting from them. So we're not just inferring time; we're inferring facts about what are good, natural concepts. And I think that there's some hope that this ontology can do for the statement "you can't really distinguish between blue and grue" what Pearl can do to the statement "correlation does not imply causation".

2b. Applications / Future Work / Speculation

The future work I'm most excited by with finite factored sets falls into three rough categories: inference (which involves more computational questions), infinity (more mathematical), and embedded agency (more philosophical).

Research topics related to inference:

- Decidability of Temporal Inference
- Efficient Temporal Inference
- Conceptual Inference
- Temporal Inference from Raw Data and Fewer Ontological Assumptions
- Temporal Inference with Deterministic Relationships
- Time without Orthogonality

- Conditioned Factored Sets

There are a lot of research directions suggested by questions like "How do we do efficient inference in this paradigm?". Some of the questions here come from the fact that we're making fewer assumptions than Pearl, and are in some sense more coming from the raw data.

Then I have the applications that are about extending factored sets to the infinite case:

- Extending Definitions to the Infinite Case
- The Fundamental Theorem of Finite-Dimensional Factored Sets
- Continuous Time
- New Lens on Physics

Everything I've presented in this talk was under the assumption of finiteness. In some cases this wasn't necessary—but in a lot of cases it actually was, and I didn't draw attention to this.

I suspect that the fundamental theorem can be extended to finite-dimensional factored sets (i.e., factored sets where |B| is finite), but it can not be extended to arbitrary-dimension factored sets.

And then, what I'm really excited about is applications to embedded agency:

- Embedded Observations
- Counterfactability
- Cartesian Frames Successor
- Unraveling Causal Loops
- Conditional Time
- Logical Causality from Logical Induction
- Orthogonality as Simplifying Assumptions for Decisions
- Conditional Orthogonality as Abstraction Desideratum

I focused on the temporal inference aspect of finite factored sets in this talk, because it's concrete and tangible to be able to say, "Ah, we can do Pearlian temporal inference, only we can sometimes infer more structure and we rely on fewer assumptions."

But really, a lot of the applications I'm excited about involve using factored sets to model situations, rather than inferring factored sets from data.

Anywhere that we currently model a situation using graphs with directed edges that represent information flow or causality, we might instead be able to use factored sets to model the situation; and this might allow our models to play more nicely with abstraction.

I want to build up the factored set ontology as an alternative to graphs when modeling agents interacting with things, or when modeling information flow. And I'm really excited about that direction.

Finite Factored Sets: Introduction and Factorizations

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is a longer, more mathematically dense introduction to "Finite Factored Sets." For a shorter introduction, see my Topos talk/transcript .

Abstract: We propose a new approach to temporal inference, inspired by the Pearlian causal inference paradigm—though quite different from Pearl's approach formally. Rather than using directed acyclic graphs, we make use of *factored sets*, which are sets expressed as Cartesian products. We show that finite factored sets are powerful tools for inferring temporal relations. We introduce an analog of *d*-separation for factored sets, *conditional orthogonality*, and we demonstrate that this notion is equivalent to conditional independence in all probability distributions on a finite factored set.

---

1. Introduction

1.1. Pearlian Causal Inference

Judea Pearl's theory of inferred causation (e.g., as presented in chapter 2 of *Causality: Models, Reasoning, and Inference*) was a deep advance in our understanding of the nature of time. The Pearlian paradigm allows us to infer causal relationships between variables using statistical data, and thereby infer temporal sequence—in defiance of the old adage that correlation does not imply causation.

In particular, given *a collection of variables* and *a joint probability distribution over those variables*, the Pearlian paradigm can often infer temporal relationships between the variables.

The joint probability distribution is usually what gets emphasized in discussions of Pearl's approach. Quite a bit of work is being done, however, by the assumption that we are handed "a collection of variables" to reason about. The Pearlian paradigm is not inferring temporal relationships from purely statistical data, but rather inferring temporal relationships from statistical data together with data about how to factorize the world into variables.[1]

A doctor who misdiagnoses their patient or misidentifies a symptom may base their subsequent reasoning on a wrong factorization of the situation into causally relevant variables. We would ideally like to build fewer assumptions like this into our model of inference, and instead allow the reasoner to figure such facts out, consider the merits of different factorizations into variables, etc.

Instead of beginning with a collection of variables and a joint probability distribution over those variables, one could imagine starting with just a finite sample space and a

probability distribution on that sample space. In this way, we might hope to do temporal inference purely using statistical data, without relying on *a priori* knowledge of a canonical way of factoring the situation into variables.

How might one do temporal inference without an existing factorization? One way might be to just consider all possible variables that can be defined on the sample space. This gives us one variable for each partition of the set.

However, when one tries to apply Pearl's methods to this collection of variables, one quickly runs into a problem: many of the variables definable on a fixed set are deterministic functions of each other. The Pearlian paradigm, as presented in the early chapters of *Causality*, lacks tools for performing temporal inference on variables that are highly deterministically related.[2]

We will introduce a new approach to temporal inference instead—one which is heavily inspired by the Pearlian paradigm, but approaches the problem with a very different formal apparatus, and does not make use of graphical models.

1.2. Overview

We'll begin by introducing the concept of a finite factored set, in Section 2. This will be our analogue of the directed acyclic graphs in Pearl's framework.

In Section 3, we will introduce the concepts of time and orthogonality, which can be read off of a finite factored set. In Pearl's framework, "time" corresponds to directed paths between nodes, and "orthogonality" corresponds to nodes that have no common ancestor.

In Section 4, we will introduce conditional orthogonality, which is our analogue of *d*-separation. We show that conditional orthogonality satisfies (a modified version of) the compositional semigraphoid axioms. We then (in Section 5) prove the fundamental theorem of finite factored sets, which states that conditional orthogonality is equivalent to conditional independence in all probability distributions on the finite factored set.

In Section 6, we discuss how to do temporal inference using finite factored sets, and give two examples. Finally, in Section 7 we discuss applications and future work, with an emphasis on temporal and conceptual inference, generalizing finite factored sets to the infinite case, and applications to embedded agency.

And here, we take our leave of Pearl. We've highlighted this approach's relationship to the Pearlian paradigm in order to motivate finite factored sets and explain how we'll be using them in this sequence. Formally, however, our approach is quite unlike Pearl's, and the rest of the sequence will stand alone.

2. Factorizations

Before giving a definition of finite factored sets, we will recall the definition of a partition, and give some basic notation related to partitions.

We do this for two reasons. First, we will use partitions in the definition of a factored set; and second, we want to draw attention to a duality between the notion of a partition, and the notion of a factorization.

2.1. Partitions

We begin with a definition of disjoint union.

Definition 1 (disjoint union). *Given a set* S *of sets, let* $\bigsqcup(S)$ *denote the set of all ordered pairs* (T, t)*, where* $T \in S$ *and* $t \in T$.[3]

Definition 2 (partition). *A partition of a set* S *is a set* $X \subseteq P(S)$ *of nonempty subsets of* S *such that the function* $\iota : \bigsqcup(X) \to S$ *given by* $\iota(x, s) = s$ *is a bijection.*[4]

Let $Part(S)$ *denote the set of all partitions of* S*. The elements of a partition are called parts.*

An equivalent definition of partition is often given: a partition is a set A of nonempty subsets of S that are pairwise disjoint and union to S. We choose the above definition because it will make the symmetry between partitions and factorizations more obvious.

Definition 3 (trivial partition). *A partition* X *of a set* S *is called trivial if* $|X| = 1$*.*

Definition 4. *Given a partition* X *of a set* S*, and an element* $s \in S$*, let* $[s]_X$ *denote the unique* $x \in X$ *such that* $s \in x$*.*

Definition 5. *Given a partition* X *of a set* S*, and elements* $s_0, s_1 \in S$*, we say* $s_0 \sim_X s_1$ *if* $[s_0]_X = [s_1]_X$*.*

Proposition 1. *Given a partition* X *of a set* S*,* $\sim_X$ *is an equivalence relation on* S*.*

Proof. Trivial. □

Definition 6 (finer and coarser). *We say that a partition* X *of* S *is finer than another partition* Y *of* S*, if for all* $s_0, s_1 \in S$*, if* $s_0 \sim_X s_1$*, then* $s_0 \sim_Y s_1$*.*

If X *is finer than* Y*, we also say* Y *is coarser than* X*, and we write* $X \geq_S Y$ *and* $Y \leq_S X$*.*

Definition 7 (discrete and indiscrete partitions). *Given a set* S*, let* $\mathrm{Dis}_S = \{\{s\} \mid s \in S\}$*.*

If S *is empty, let* $\mathrm{Ind}_S = \{\}$*, and if* S *is nonempty, let* $\mathrm{Ind}_S = \{S\}$*.*

Dis_S *is called the discrete partition, and* Ind_S *is called the indiscrete partition.*

Proposition 2. *For any set* S*,* $\geq_S$ *is a partial order on* $\mathrm{Part}(S)$*. Further, for all* $X \in \mathrm{Part}(S)$*,* $\mathrm{Dis}_S \geq_S X$ *and* $X \geq_S \mathrm{Ind}_S$*.*

Proof. Trivial. □

While both notations are sometimes used, it is more standard to draw the symbol in the opposite direction and have $X \leq Y$ when X is finer than Y. We choose to go against that standard because we want to think of partitions in part as the ability to distinguish between elements, and finer partitions correspond to greater ability to distinguish.[5]

Definition 8 (common refinement). *Given a set* C *of partitions of a fixed set* S*, let* $\bigvee_S(C)$ *denote the partition* $X \in \mathrm{Part}(S)$ *satisfying* $s_0 \sim_X s_1$ *if and only if* $s_0 \sim_c s_1$ *for all* $c \in C$*. Given* $X, Y \in \mathrm{Part}(S)$*, we let* $X \vee_S Y = \bigvee_S(\{X, Y\})$*.*

2.2. Factorizations

We start with a definition of Cartesian product.

Definition 9 (Cartesian product). *Given a set* S *of sets, let* $\prod(S)$ *denote the set of all functions* $f : S \to \bigsqcup(S)$ *such that for all* $T \in S$*,* $f(T)$ *is of the form* (T, t)*, for some* $t \in T$*.*

We can now give the definition of a factorization of a set.

Definition 10 (factorization). *A factorization of a set* S *is a set* $B \subseteq \text{Part}(S)$ *of nontrivial partitions of* S *such that the function* $\pi : S \to \prod(B)$*, given by* $\pi(s) = (b \mapsto (b, [s]_b))$*, is a bijection.*

Let $\text{Fact}(S)$ *denote the set of all factorizations of* S*. The elements of a factorization are called factors.*

In other words, a set of nontrivial partitions is a factorization of S if for each way of choosing one part from each factor, there exists a unique element of S in the intersection of those parts.

Notice the duality between the definitions of partition and factorization. We replace subsets with partitions, nonempty with nontrivial, and disjoint union with Cartesian product, and we reverse the direction of the function. We can think of a factorization of S as a way to view S as a product, in the same way that a partition was a way to view S as a disjoint union.

A factored set is just a set together with a factorization of that set.

Definition 11 (factored set). *A factored set* F *is an ordered pair* (S, B)*, such that* B *is a factorization of* S*.*

If $F = (S, B)$ *is a factored set, we let* $\text{set}(F) = S$*, and let* $\text{basis}(F) = B$*.*

An important fact about factored sets is that the factors are enough to distinguish distinct elements.

Proposition 3. *Given a factored set* $F = (S, B)$*, and elements* $s_0, s_1 \in S$*, if* $s_0 \sim_b s_1$ *for all* $b \in B$*, then* $s_0 = s_1$*.*

Proof. Let $F = (S, B)$ be a finite factored set, and let $s_0, s_1 \in S$ satisfy $s_0 \sim_b s_1$ for all $b \in B$.

Let $\pi : S \to \prod(B)$ be given by $\pi(s) = (b \mapsto (b, [s]_b))$, as in the definition of factorization.

Then $\pi(s_0) = (b \mapsto (b, [s_0]_b)) = (b \mapsto (b, [s_1]_b)) = \pi(s_1)$. Since π is bijective, this means

$s_0 = s_1$. □

2.3. Chimera Functions

The following theorem can be viewed as an alternate characterization of factorization. We will use this alternate characterization to define chimera functions, which will be useful tools for manipulating elements of factored sets.

Theorem 1. *Given a set* S*, a set* B *of nontrivial partitions of* S *is a factorization of* S *if and only if for every function* $g : B \to S$*, there exists a unique* $s \in S$ *such that for all* $b \in B$*,* $s \sim_b g(b)$*.*

Proof. First, we let B be a factorization of S, and let $g : B \to S$ be any function. We want to show that there exists a unique $s \in S$ such that for all $b \in B$, $s \sim_b g(b)$. Let $\pi : S \to \prod(B)$ be given by $\pi(s) = (b \mapsto (b, [s]_b))$, as in the definition of factorization. Note that π is bijective, and thus has an inverse.

Let $s = \pi^{-1}(b \mapsto (b, [g(b)]_b))$. Observe that this is well-defined, because $(b \mapsto (b, [g(b)]_b))$ is in fact in $\prod(B)$. We will show that $s \sim_b g(b)$ for all $b \in B$, and the uniqueness of this s will then follow directly from Proposition 3.

We have $\pi(s) = (b \mapsto [s]_b)$ by the definition of π. However, we also have $\pi(s) = (b \mapsto [g(b)]_b)$ by the definition of s. Thus, $b \mapsto [s]_b$ and $b \mapsto [g(b)]_b$ are the same function, so $[s]_b = [g(b)]_b$ for all $b \in B$, so $s \sim_b g(b)$ for all $b \in B$.

Conversely, let S be any set, and let B be any set of nontrivial partitions of S. Assume that for all $g : B \to S$, there exists a unique $s \in S$ satisfying $s \sim_b g(b)$ for $b \in B$. Again, let $\pi : S \to \prod(B)$ be given by $\pi(s) = (b \mapsto (b, [s]_b))$, as in the definition of factorization. We want to show that π is invertible.

First, we show that π is injective. Take an arbitrary $s_0 \in S$, and let $g : B \to S$ be the constant function satisfying $g(b) = s_0$ for all $b \in B$. Given another $s_1 \in S$, if

$\pi(s_0) = \pi(s_1)$, then $(b \mapsto [s_0]_b) = (b \mapsto [s_1]_b)$, so $[s_1]_b = [s_0]_b = [g(b)]_b$ for all $b \in B$, so $s_0 \sim_b s_1 \sim_b g(b)$ for all $b \in B$. Since there is a unique $s \in S$ satisfying $s \sim_b g(b)$ for all $b \in B$, this means $s_0 = s_1$. Thus π is injective.

To see that π is surjective, consider some arbitrary $h \in \prod(B)$. We want to show that there exists an $s \in S$ with $h = \pi(s)$.

For all $b \in B$, let $H_b \in b$ be given by $h(b) = (b, H_b)$, which is well-defined since $h \in \prod(B)$. Note that H_b is a nonempty subset of S, so there exists a function $g : B \to S$ with $g(b) \in H_b$ for all $b \in B$. Fix any such g, and let s satisfy $s \sim_b g(b)$ for all $b \in B$.

We thus have that for all $b \in B$, $h(b) = (b, H_b) = (b, [g(b)]_b) = (b, [s]_b) = \pi(s)(b)$, so $h = \pi(s)$. Thus π is surjective.

Since π is bijective, we have that B is a factorization of S. □

This also gives us that factors are disjoint from each other.

Corollary 1. *Given a factored set* $F = (S, B)$ *and distinct factors* $b_0, b_1 \in B$*,* $b_0 \cap b_1 = \{\}$*.*

Proof. Assume by way of contradiction that $T \in b_0 \cap b_1$. Since b_0 is nontrivial, there must be some other $T' \in b_0$ with $T \cap T' = \{\}$. Let $g : B \to S$ be any function such that $g(b_0) \in T'$ and $g(b_1) \in T$. Then there can be no s such that $s \sim_{b_0} g(b_0)$ and $s \sim_{b_1} g(b_1)$, since then s would be in both T and T'. This contradicts Theorem 1. □

We are now ready to define the chimera function of a factored set.

Definition 12 (chimera function). *Given a factored set* $F = (S, B)$*, the chimera function (of* F*) is the function* $\chi^F : (B \to S) \to S$ *defined by* $\chi^F(g) \sim_b g(b)$ *for all* $g : B \to S$ *and* $b \in B$*.*

The name "chimera function" comes from the fact that χ^F can be viewed as building an element of S by fusing together the properties of various different elements. Since we will often apply the chimera function to functions g that only take on two values, we will give notation for this special case.

Definition 13. *Given a factored set* $F = (S, B)$*, and a subset* $C \subseteq B$*, let* $\overset{F}{\chi_C} : S \times S \to S$ *be given by* $\overset{F}{\chi_C}(s, t) = \chi^F(g)$*, where* $g : B \to S$ *is given by* $g(b) = s$ *if* $b \in C$*, and* $g(b) = t$ *otherwise.*

For $T, R \subseteq S$*, we will write* $\overset{F}{\chi}_C(T, R)$ *for* $\{\overset{F}{\chi}_C(t, r) \mid t \in T, \quad r \in R\}$.

The following is a list of properties of $\overset{F}{\chi_C}$, which will be useful in later proofs. All of these properties follow directly from the definition of $\overset{F}{\chi_C}$.

Proposition 4. *Fix* $F = (S, B)$*, a factored set,* $C, D \subseteq B$*, and* $s, t, r \in S$*.*

1. $\overset{F}{\chi}_C(s, t) \sim_c s$ *for all* $c \in C$.
2. $\overset{F}{\chi}_C(s, t) \sim_b t$ *for all* $b \in B \setminus C$.
3. $\overset{F}{\chi_C}(s, s) = s$.
4. $\overset{F}{\chi_{B \setminus C}}(s, t) = \overset{F}{\chi_C}(t, s)$.
5. $\overset{F}{\chi_{C \cup D}}(s, t) = \overset{F}{\chi_C}(s, \overset{F}{\chi_D}(s, t))$.
6. $\overset{F}{\chi_{C \cap D}}(s, t) = \overset{F}{\chi_C}(\overset{F}{\chi_D}(s, t), t)$.
7. $\overset{F}{\chi_C}(\overset{F}{\chi_C}(s, t), r) = \overset{F}{\chi_C}(s, \overset{F}{\chi_C}(t, r)) = \overset{F}{\chi_C}(s, r)$.
8. $\overset{F}{\chi_C}(s, \overset{F}{\chi_D}(t, r)) = \overset{F}{\chi_D}(\overset{F}{\chi_C}(s, t), \overset{F}{\chi_C}(s, r))$.

9. $\chi^F_C(\chi^F_D(s,t),r) = \chi^F_D(\chi^F_C(s,r),\chi^F_C(t,r))$.
10. $\chi^F_B(s,t) = s$.
11. $\chi^F_{\{\}}(s,t) = t$.

Proof. Trivial. □

2.4. Trivial Factorizations

We now define a notion of a trivial factorization of a set, and show that every set has a unique trivial factorization.

Definition 14 (trivial factorization). *A factorization* B *of a set* S *is called trivial if* $|B| \leq 1$*. A factored set* (S, B) *is called trivial if* B *is trivial.*

Proposition 5. *For every set* S*, there exists a unique trivial factorization* B *of* S*. If* $|S| \neq 1$*, this trivial factorization is given by* $B = \{Dis_S\}$*, and if* $|S| = 1$*, it is given by* $B = \{\}$*.*

Proof. We start with the case where $|S| = 0$. The only partition of S is {}, so we only need to consider the sets of partitions {{}} and {} as potential factorizations. {{}} is vacuously a factorization of S by Theorem 1, since there are no functions from {{}} to S. {} is not a factorization by Theorem 1, since there is a function from {} to S, but there is no element of S. Thus, when $|S| = 0$, {{}} = $\{Dis_S\}$ is the unique trivial factorization of S.

Next, consider the case where $|S| = 1$. First, observe that the unique $s \in S$ vacuously satisfies $s \sim_b g(b)$ for all $g : \{\} \to S$ and $b \in \{\}$, since there is no $b \in \{\}$. Thus, by Theorem 1, {} is a factorization of S. Further, {} is the only factorization of S, since

there are no nontrivial partitions of S. Thus, when $|S| = 1$, $\{\}$ is the unique trivial factorization of S.

Next, we consider the case where $|S| \geq 2$. Observe that Dis_S is a nontrivial partition of S. Let $B = \{\text{Dis}_S\}$. We want to show that B is a factorization of S. By Theorem 1, it suffices to show that for all $g : B \to S$, there exists a unique $s \in S$ with $s \sim_{\text{Dis}_S} g(\text{Dis}_S)$. We can take $s = g(\text{Dis}_S)$, which clearly satisfies $s \sim_{\text{Dis}_S} g(\text{Dis}_S)$. This s is unique, since if $s' \sim_{\text{Dis}_S} g(\text{Dis}_S)$, then $s' \in [g(\text{Dis}_S)]_{\text{Dis}_S} = [s]_b = \{s\}$, so $s' = s$. Thus B is a factorization of S.

On the other hand, if $|S| \geq 2$, $\{\}$ is not a factorization of S, since if it were, Proposition 3 would imply that all elements of S are equal. Further, for any partition b of S, with $b \neq \{\text{Dis}_S\}$, there must exist $s_0, s_1 \in S$, with $s_0 \sim_b s_1$, but $s_0 \neq s_1$. Thus $\{b\}$ cannot be a factorization of S by Proposition 3. Thus when $|S| \geq 2$, Dis_S is the unique trivial factorization of S. □

2.5. Finite Factored Sets

This sequence will primarily be about finite factored sets.

Definition 15. *If* $F = (S, B)$ *is a factored set, the size of* F*, written* $\text{size}(F)$*, is the cardinality of* S*. The dimension of* F*, written* $\dim(F)$*, is the cardinality of* B*.* F *is called finite if its size is finite, and finite-dimensional if its dimension is finite.*

We suspect that the theory of infinite factored sets is both interesting and important. However, it is outside of the scope of this sequence, which will require finiteness for many of its key results.

Some of the definitions and results in this sequence will be given for finite factored sets, in spite of the fact that they could easily be extended to finite-dimensional or arbitrary factored sets. This is because they can often be extended in more than one way, and determining which extension is most natural requires further developing the theory of arbitrary factored sets.

Proposition 6. *Every finite factored set is also finite-dimensional.*

Proof. If $F = (S, B)$ is a factored set, B is a set of sets of subsets of S. Thus, $|B| \leq 2^{2^{|S|}}$.

□

This bound is horrible and will be improved in Proposition 9. First, however, we will take a look at the number of factorizations of a fixed finite set.

Proposition 7. *Let* $F = (S, B)$ *be a finite factored set. Then* $|S| = \prod_{b \in B} |b|$.

Proof. Trivial. □

Proposition 8. *If* $|S|$ *is equal to* 0, 1, *or a prime, the trivial factorization of* S *is the only factorization of* S.

Proof. If $|S| = 0$ or $|S| = 1$, then $|\text{Part}(S)| = 1$, so $B \subseteq \text{Part}(S)$ can have cardinality at most 1.

If $|S| = p$, a prime, then by Proposition 7, $|b|$ must divide p for all $b \in B$. Since factorizations cannot contain trivial partitions, this means $|b| = p$ for all $b \in B$. However, $\{\{s\} \mid s \in S\}$ is the only element of $\text{Part}(S)$ of cardinality p, so $|B| \leq 1$. □

On the other hand, in the case where $|S|$ is finite and composite, the number of factorizations of S grows very quickly, as seen in Table 1. Table 1 shows the number of factorizations of a set S with cardinality up to 25:

$\lvert S \rvert$	$\lvert \mathrm{Fact}(S) \rvert$	$\lvert S \rvert$	$\lvert \mathrm{Fact}(S) \rvert$
0	1	13	1
1	1	14	8648641
2	1	15	1816214401
3	1	16	181880899201
4	4	17	1
5	1	18	45951781075201
6	61	19	1
7	1	20	3379365788198401
8	1681	21	1689515283456001

9	5041		22	14079294028801
10	15121		23	1
11	1		24	4454857103544668620801
12	13638241		25	538583682060103680001

Given the naturalness of the notion of factorization, we were surprised to discover that this sequence did not exist on the On-Line Encyclopedia of Integer Sequences (OEIS). We added the sequence, A338681, on April 30, 2021.

To give one concrete example, the four factorizations of the set $\{0, 1, 2, 3\}$ are:

- $\{\{\{0\}, \{1\}, \{2\}, \{3\}\}\}$,
- $\{\{\{0, 1\}, \{2, 3\}\}, \{\{0, 2\}, \{1, 3\}\}\}$,
- $\{\{\{0, 1\}, \{2, 3\}\}, \{\{0, 3\}, \{1, 2\}\}\}$, and
- $\{\{\{0, 2\}, \{1, 3\}\}, \{\{0, 3\}, \{1, 2\}\}\}$.

Proposition 9. *Let* F *be a finite factored set.*

- *If* $\text{size}(F) = 0$, *then* $\dim(F) = 1$.
- *If* $\text{size}(F) = 1$, *then* $\dim(F) = 0$.
- *If* $\text{size}(F) = p$ *is prime, then* $\dim(F) = 1$.
- *If* $\text{size}(F) = p_0 \ldots p_{k-1}$ *is a product of* $k \geq 2$ *primes, then* $1 \leq \dim(F) \leq k$.

Proof. The first three parts follow directly from Proposition 5 and Proposition 8. For the fourth part, let $F = (S, B)$, and let $|S| = p_0 \ldots p_{k-1}$ be a product of $k \geq 2$ primes.

By Proposition 7, $|S| = \prod_{b \in B} |b|$. Consider an arbitrary $b \in B$. Since b is a nontrivial partition of a finite set S, $|b|$ is finite and $|b| \neq 1$. If $|b|$ were 0, then $|S|$ would be 0. Thus $|b|$ is a natural number greater than or equal to 2. B cannot be empty, since $|S| \neq 1$. If $|B|$ were greater than k, then we would be able to express $|S|$ as a product of more than k natural numbers greater than or equal to 2, which is clearly not possible since $|S|$ is a product of k primes. Thus $1 \leq \dim(F) \leq k$. □

In the next post, we will introduce the notions of the history of a partition, orthogonality between partitions, and time.

Acknowledgments: My thanks to Alex Appel, Ramana Kumar, Xiaoyu He, Tsvi Benson-Tilsen, Andrew Critch, Sam Eisenstat, Rob Bensinger, and Claire Wang for discussion and feedback on this sequence.

Footnotes

[1] Although I say "factorize" here, note that this will not be the kind of factorization that shows up in finite factored sets, because (as we will see) disjoint factors must be independent in a finite factored set. I appeal to the same concept in both contexts because factorization is just a very general and useful concept, rather than to indicate a direct connection.

[2] At least, it lacks such causal inference tools unless we assume access to interventional data.

[3] Note that this definition and Definition 9 could have been made more general by taking S to be a multiset.

[4] P(S) denotes the power set of S.

[5] In our view, "$Y \geq X$" is also a more natural way to visually represent a mapping between a three-part partition Y that is finer than a two-part partition X.

Finite Factored Sets: Orthogonality and Time

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

The main way we'll be using factored sets is as a foundation for talking about concepts like orthogonality and time. Finite factored sets will play a role that's analogous to that of directed acyclic graphs in Pearlian causal inference.

To utilize factored sets in this way, we will first want to introduce the concept of generating a partition with factors.

3.1. Generating a Partition with Factors

Definition 16 (generating a partition). *Given a finite factored set* $F = (S, B)$*, a partition* $X \in \text{Part}(S)$*, and a* $C \subseteq B$*, we say* C *generates* X *(in* F*), written* $C \vdash^F X$*, if* $\chi^F_C(x, S) = x$ *for all* $x \in X$*.*

The following proposition gives many equivalent definitions of $\vdash^F$.

Proposition 10. *Let* $F = (S, B)$ *be a finite factored set, let* $X \in \text{Part}(S)$ *be a partition of* S*, and let* C *be a subset of* B*. The following are equivalent:*

1. $C \vdash^F X$.
2. $\chi^F_C(x, S) = x$ *for all* $x \in X$.
3. $\chi^F_C(x, S) \subseteq x$ *for all* $x \in X$.
4. $\chi^F_C(x, y) \subseteq x$ *for all* $x, y \in X$.
5. $\chi^F_C(s, t) \in [s]_X$ *for all* $s, t \in S$.
6. $\chi^F_C(s, t) \sim_X s$ *for all* $s, t \in S$.

7. $X \leq_S \bigvee_S (C)$.

Proof. The equivalence of conditions 1 and 2 is by definition.

The equivalence of conditions 2 and 3 follows directly from the fact that $\overset{F}{\chi}_C(s,s) = s$ for all $s \in x$, so $\overset{F}{\chi}_C(x,S) \supseteq \overset{F}{\chi}_C(x,x) \supseteq x$.

To see that conditions 3 and 4 are equivalent, observe that since $S = \bigcup_{y \in X} y$, $\overset{F}{\chi}_C(x,S) = \bigcup_{y \in X} \overset{F}{\chi}_C(x,y)$. Thus, if $\overset{F}{\chi}_C(x,S) \subseteq x$, $\overset{F}{\chi}_C(x,y) \subseteq x$ for all $y \in X$, and conversely if $\overset{F}{\chi}_C(x,y) \subseteq x$ for all $y \in X$, then $\overset{F}{\chi}_C(x,S) \subseteq x$.

To see that condition 3 is equivalent to condition 5, observe that if condition 5 holds, then for all $x \in X$, we have $\overset{F}{\chi}_C(s,t) \in [s]_X = x$ for all $s \in x$ and $t \in S$. Thus $\overset{F}{\chi}_C(x,S) \subseteq x$. Conversely, if condition 3 holds, $\overset{F}{\chi}_C(s,t) \in \overset{F}{\chi}_C([s]_X, S) \subseteq [s]_X$ for all $s,t \in S$.

Condition 6 is clearly a trivial restatement of condition 5.

To see that conditions 6 and 7 are equivalent, observe that if condition 6 holds, and $s,t \in S$ satisfy $s \sim_{V_S(C)} t$, then $\overset{F}{\chi}_C(s,t) = t$, so $t = \overset{F}{\chi}_C(s,t) \sim_X s$. Thus $X \leq_S \bigvee_S(C)$.

Conversely, if condition 7 holds, then since $\overset{F}{\chi}_C(s,t) \sim_{V_S(C)} s$ for all $s,t \in S$, we have $\overset{F}{\chi}_C(s,t) \sim_X s$. □

Here are some basic properties of $\vdash^F$.

Proposition 11. *Let* $F = (S,B)$ *be a finite factored set, let* C *and* D *be subsets of* B*, and let* $X, Y \in \mathrm{Part}(S)$ *be partitions of* S*.*

1. *If* $X \leq_S Y$ *and* $C \vdash^F Y$*, then* $C \vdash^F X$*.*
2. *If* $C \vdash^F X$ *and* $C \vdash^F Y$*, then* $C \vdash^F X \vee_S Y$*.*

3. $B \vdash^F X$.

4. $\{\} \vdash^F X$ *if and only if* $X = \text{Ind}_S$.

5. *If* $C \subseteq D$ *and* $C \vdash^F X$, *then* $D \vdash^F X$.

6. *If* $C \vdash^F X$ *and* $D \vdash^F X$, *then* $C \cap D \vdash^F X$.

Proof. For the first 5 parts, we will use the equivalent definition from Proposition 10 that $C \vdash^F X$ if and only if $X \leq_S \bigvee_S(C)$.

Then 1 follows directly from the transitivity of $\leq_S$.

2 follows directly from the fact that any partition Z satisfies $X \vee_S Y \leq Z$ if and only if $X \leq Z$ and $Y \leq Z$.

3 follows directly from the fact that $\bigvee_S(B) = \text{Dis}_S$ by Proposition 3.

4 follows directly from the fact that $\bigvee_S(\{\}) = \text{Ind}_S$, together with the fact that $X \leq_S \text{Ind}_S$ if and only if $X = \text{Ind}_S$.

5 follows directly from the fact that if $C \subseteq D$, then $\bigvee_S(C) \leq \bigvee_S(D)$.

Finally, we need to prove part 6. For this, we will use the equivalent definition from Proposition 10 that $C \vdash^F X$ if and only if $\overset{F}{\chi}_C(s,t) \sim_X s$ for all $s,t \in S$. Assume that for all $s,t \in S$, $\overset{F}{\chi}_C(s,t) \sim_X s$ and $\overset{F}{\chi}_D(s,t) \sim_X s$. Thus, for all $s,t \in S$, $\overset{F}{\chi}_{C\cap D}(s,t) = \overset{F}{\chi}_C(\overset{F}{\chi}_D(s,t),t) \sim_X \overset{F}{\chi}_D(s,t) \sim_X s$. Thus $C \cap D \vdash^F X$. □

Our main use of $\vdash^F$ will be in the definition of the history of a partition.

3.2. History

Definition 17 (history of a partition). *Given a finite factored set* $F = (S,B)$ *and a partition* $X \in \text{Part}(S)$, *let* $h^F(X)$ *denote the smallest (according to the subset ordering)*

subset of B *such that* $h^F(X) \vdash^F X$.

The history of X, then, is the smallest set of factors $C \subseteq B$ such that if you're trying to figure out which part in X any given $s \in S$ is in, it suffices to know what part s is in within each of the factors in C. We can informally think of $h^F(X)$ as the smallest amount of information needed to compute X.

Proposition 12. *Given a finite factored set* $F = (S, B)$*, and a partition* $X \in \text{Part}(S)$*,* $h^F(X)$ *is well-defined.*

Proof. Fix a finite factored set $F = (S, B)$ and a partition $X \in \text{Part}(S)$, and let $h^F(X)$ be the intersection of all $C \subseteq B$ such that $C \vdash^F X$. It suffices to show that $h^F(X) \vdash^F X$; then $h^F(X)$ will clearly be the unique smallest (according to the subset ordering) subset of B such that $h^F(X) \vdash^F X$.

Note that $h^F(X)$ is a finite intersection, since there are only finitely many subsets of B, and that $h^F(X)$ is an intersection of a nonempty collection of sets since $B \vdash^F X$. Thus, we can express $h^F(X)$ as a composition of finitely many binary intersections. By part 6 of Proposition 11, the intersection of two subsets that generate X also generates X. Thus $h^F(X) \vdash^F X$. □

Here are some basic properties of history.

Proposition 13. *Let* $F = (S, B)$ *be a finite factored set, and let* $X, Y \in \text{Part}(S)$ *be partitions of* S.

1. *If* $X \leq_S Y$*, then* $h^F(X) \subseteq h^F(Y)$.
2. $h^F(X \vee_S Y) = h^F(X) \cup h^F(Y)$.
3. $h^F(X) = \{\}$ *if and only if* $X = \text{Ind}_S$.
4. *If* S *is nonempty, then* $h^F(b) = \{b\}$ *for all* $b \in B$.

Proof. The first 3 parts are trivial consequences of history's definition and Proposition 11.

For the fourth part, observe that $\{b\} \vdash^F b$ by condition 7 of Proposition 10. b is nontrivial, and since S is nonempty, b is nonempty. So we have $\neg(\{\} \vdash^F b)$ by part 4 of Proposition 11. Thus $\{b\}$ is the smallest subset of B that generates b. □

3.3. Orthogonality

We are now ready to define the notion of orthogonality between two partitions of S.

Definition 18 (orthogonality). *Given a finite factored set* $F = (S, B)$ *and partitions* $X, Y \in \text{Part}(S)$*, we say* X *is orthogonal to* Y *(in* F*), written* $X \perp^F Y$*, if* $h^F(X) \cap h^F(Y) = \{\}$*.*

If $\neg(X \perp^F Y)$*, we say* X *is entangled with* Y *(in* F*).*

We could also unpack this definition to not mention history or chimera functions.

Proposition 14. *Given a finite factored set* $F = (S, B)$*, and partitions* $X, Y \in \text{Part}(S)$*,* $X \perp^F Y$ *if and only if there exists a* $C \subseteq B$ *such that* $X \leq_S \bigvee_S(C)$ *and* $Y \leq_S \bigvee_S(B \setminus C)$*.*

Proof. If there exists a $C \subseteq B$ such that $X \leq_S \bigvee_S(C)$ and $Y \leq_S \bigvee_S(B \setminus C)$, then $C \vdash^F X$ and $B \setminus C \vdash^F Y$. Thus, $h^F(X) \subseteq C$ and $h^F(Y) \subseteq B \setminus C$, so $h^F(X) \cap h^F(Y) = \{\}$.

Conversely, if $h^F(X) \cap h^F(Y) = \{\}$, let $C = h^F(X)$. Then $C \vdash^F X$, so $X \leq_S \bigvee_S(C)$, and $B \setminus C \supseteq h^F(Y)$, so $B \setminus C \vdash^F Y$, so $Y \leq_S \bigvee_S(B \setminus C)$. □

Here are some basic properties of orthogonality.

Proposition 15. *Let* $F = (S, B)$ *be a finite factored set, and let* $X, Y, Z \in \text{Part}(S)$ *be partitions of* S*.*

1. *If* $X \perp^F Y$*, then* $Y \perp^F X$*.*
2. *If* $X \perp^F Z$ *and* $Y \leq_S X$*, then* $Y \perp^F Z$*.*

3. *If* $X \perp^F Z$ *and* $Y \perp^F Z$*, then* $(X \vee_S Y) \perp^F Z$*.*

4. $X \perp^F X$ *if and only if* $X = \text{Ind}_S$*.*

Proof. Part 1 is trivial from the symmetry in the definition.

Parts 2, 3, and 4 follow directly from Proposition 13. □

3.4. Time

Finally, we can define our notion of time in a factored set.

Definition 19 ((strictly) before). *Given a finite factored set* $F = (S, B)$*, and partitions* $X, Y \in \text{Part}(S)$*, we say* X *is before* Y *(in* F*), written* $X \leq^F Y$*, if* $h^F(X) \subseteq h^F(Y)$*.*

We say X *is strictly before* Y *(in* F*), written* $X <^F Y$*, if* $h^F(X) \subset h^F(Y)$*.*

Again, we could also unpack this definition to not mention history or chimera functions.

Proposition 16. *Given a finite factored set* $F = (S, B)$*, and partitions* $X, Y \in \text{Part}(S)$*,* $X \leq^F Y$ *if and only if every* $C \subseteq B$ *satisfying* $Y \leq_S \bigvee_S(C)$ *also satisfies* $X \leq_S \bigvee_S(C)$*.*

Proof. Note that by part 7 of Proposition 10, part 5 of Proposition 11, and the definition of history, C satisfies $Y \leq_S \bigvee_S(C)$ if and only if $C \supseteq h^F(Y)$, and similarly for X.

Clearly, if $h^F(Y) \supseteq h^F(X)$, every $C \supseteq h^F(Y)$ satisfies $C \supseteq h^F(X)$. Conversely, if $h^F(X)$ is not a subset of $h^F(Y)$, then we can take $C = h^F(Y)$, and observe that $C \supseteq h^F(Y)$ but not $C \supseteq h^F(X)$. □

Interestingly, we can also define time entirely as a closure property of orthogonality. We hold that the philosophical interpretation of time as a closure property on orthogonality is natural and transcends the ontology set up in this sequence.

Proposition 17. *Given a finite factored set* $F = (S, B)$*, and partitions* $X, Y \in \text{Part}(S)$*,* $X \leq^F Y$ *if and only if every* $Z \in \text{Part}(S)$ *satisfying* $Y \perp^F Z$ *also satisfies* $X \perp^F Z$*.*

Proof. Clearly if $h^F(X) \subseteq h^F(Y)$, then every Z satisfying $h^F(Y) \cap h^F(Z) = \{\}$ also satisfies $h^F(X) \cap h^F(Z) = \{\}$.

Conversely, if $h^F(X)$ is not a subset of $h^F(Y)$, let $b \in B$ be an element of $h^F(X)$ that is not in $h^F(Y)$. Assuming S is nonempty, b is nonempty, so we have $h^F(b) = \{b\}$, so $Y \perp^F b$, but not $X \perp^F b$. On the other hand, if S is empty, then $X = Y = \{\}$, so clearly $X \leq^F Y$. □

Here are some basic properties of time.

Proposition 18. *Let* $F = (S, B)$ *be a finite factored set, and let* $X, Y, Z \in \text{Part}(S)$ *be partitions of* S*.*

1. $X \leq^F X$.
2. *If* $X \leq^F Y$ *and* $Y \leq^F Z$*, then* $X \leq^F Z$.
3. *If* $X \leq_S Y$*, then* $X \leq^F Y$.
4. *If* $X \leq^F Z$ *and* $Y \leq^F Z$*, then* $(X \vee_S Y) \leq^F Z$.

Proof. Part 1 is trivial from the definition.

Part 2 is trivial by transitivity of the subset relation.

Part 3 follows directly from part 1 of Proposition 13.

Part 4 follows directly from part 2 of Proposition 13. □

Finally, note that we can (circularly) redefine history in terms of time, thus partially justifying the names.

Proposition 19. *Given a nonempty finite factored set* $F = (S, B)$ *and a partition* $X \in \text{Part}(S)$*,* $h^F(X) = \{b \in B \mid b \leq^F X\}$.

Proof. Since S is nonempty, part 4 of Proposition 13 says that $h^F(b) = \{b\}$ for all $b \in B$. Thus $\{b \in B \mid b \leq^F X\} = \{b \in B \mid \{b\} \subseteq h^F(X)\} = \{b \in B \mid b \in h^F(X)\} = h^F(X)$. □

In the next post, we'll build up to a definition of *conditional orthogonality* by introducing the notion of subpartitions.

Finite Factored Sets: Conditional Orthogonality

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

We now want to extend our notion of orthogonality to conditional orthogonality. This will take a bit of work. In particular, we will have to first extend our notions of partition *generation* and *history* to be defined on partitions of subsets of S.

4.1 Generating a Subpartition

Definition 20 (subpartition). *A subpartition of a set* S *is a partition of a subset of* S*. Let* $\text{SubPart}(S) = \bigcup_{E \subseteq S} \text{Part}(E)$ *denote the set of all subpartitions of* S*.*

Definition 21 (domain). *The domain of a subpartition* X *of* S*, written* $\text{dom}(X)$*, is the unique* $E \subseteq S$ *such that* $X \in \text{Part}(E)$*.*

Definition 22 (restricted partitions). *Given sets* S *and* E *and a partition* X *of* S*, let* $X|E$ *denote the partition of* $S \cap E$ *given by* $X|E = \{[e]_X \cap E \mid e \in E\}$*.*

Definition 23 (generating a subpartition). *Given a finite factored set* $F = (S, B)$*, and* $X \in \text{SubPart}(S)$*, and a* $C \subseteq B$*, we say* C *generates* X *(in* F*), written* $C \vdash^F X$*, if* $\chi_C^F(x, \text{dom}(X)) = x$ *for all* $x \in X$*.*

Note that this definition clearly coincides with Definition 16, when X has domain S. Despite the similarity of the definitions, the idea of generating a subpartition is a bit more complicated than the idea of generating a partition of S.

To see this, consider the following list of equivalent definitions. Notice that while the first five directly mirror their counterparts in Proposition 10, the last two (and especially the last one) require an extra condition.

Proposition 20. *Let* $F = (S, B)$ *be a finite factored set, let* $X \in \mathrm{SubPart}(S)$ *be a subpartition of* S*, let* $E = \mathrm{dom}(X)$ *be the domain of* X*, and let* C *be a subset of* B*. The following are equivalent.*

1. $C \vdash^F X$.
2. $\overset{F}{\chi}_C(x, E) = x$ for all $x \in X$.
3. $\overset{F}{\chi}_C(x, E) \subseteq x$ for all $x \in X$.
4. $\overset{F}{\chi}_C(x, y) \subseteq x$ for all $x, y \in X$.
5. $\overset{F}{\chi}_C(s, t) \in [s]_X$ for all $s, t \in E$.
6. $\overset{F}{\chi}_C(s, t) \in E$ and $\overset{F}{\chi}_C(s, t) \sim_X s$ for all $s, t \in E$.
7. $X \leq_E (\bigvee_S(C)|E)$ and $\overset{F}{\chi}_C(E, E) = E$.

Proof. The equivalence of conditions 1 and 2 is by definition.

The equivalence of conditions 2 and 3 follows directly from the fact that $\overset{F}{\chi}_C(s, s) = s$ for all $s \in x$, so $\overset{F}{\chi}_C(x, E) \supseteq \overset{F}{\chi}_C(x, x) \supseteq x$.

To see that conditions 3 and 4 are equivalent, observe that since $E = \bigcup_{y \in X} y$, $\overset{F}{\chi}_C(x, E) = \bigcup_{y \in X} \overset{F}{\chi}_C(x, y)$. Thus, if $\overset{F}{\chi}_C(x, E) \subseteq x$, $\overset{F}{\chi}_C(x, y) \subseteq x$ for all $y \in X$, and conversely if $\overset{F}{\chi}_C(x, y) \subseteq x$ for all $y \in X$, then $\overset{F}{\chi}_C(x, E) \subseteq x$.

To see that condition 3 is equivalent to condition 5, observe that if condition 5 holds, then for all $x \in X$, we have $\overset{F}{\chi}_C(s, t) \in [s]_X = x$ for all $s \in x$ and $t \in E$. Thus $\overset{F}{\chi}_C(x, E) \subseteq x$. Conversely, if condition 3 holds, $\overset{F}{\chi}_C(s, t) \in \overset{F}{\chi}_C([s]_X, E) \subseteq [s]_X$ for all $s, t \in E$.

Condition 6 is clearly a trivial restatement of condition 5.

To see that conditions 6 and 7 are equivalent, observe that if condition 6 holds, then $\chi_C^F(s,t) \in E$ for all $s,t \in E$, so $\chi_C^F(E,E) \subseteq E$, so $\chi_C^F(E,E) = E$. Further, if $s,t \in E$ satisfy $s \sim_{V_S(C)|E} t$, then $s \sim_c t$ for all $c \in C$, so $\chi_C^F(s,t) = t$, so $t = \chi_C^F(s,t) \sim_X s$. Thus $X \leq_E V_S(C)|E$.

Conversely, if condition 7 holds, then for all $s,t \in E$, we have $\chi_C^F(s,t) \sim_{V_S(C)} s$, so $\chi_C^F(s,t) \sim_{V_S(C)|E} s$, and thus $\chi_C^F(s,t) \sim_X s$. Further, clearly $\chi_C^F(E,E) = E$ implies $\chi_C^F(s,t) \in E$ for all $s,t \in E$. □

The first half of condition 7 in the above proposition can be thought of as saying that the values of factors in C are sufficient to distinguish between the parts of X.

The second half can be thought of as saying that no factors in C become entangled with any factors outside of C when conditioning on E. This second half is actually necessary (for example) to ensure that the set of all C that generate X is closed under intersection. As such, we will need this fact in order to extend our notion of history to arbitrary subpartitions.

Proposition 21. *Let* $F = (S, B)$ *be a finite factored set, let* C *and* D *be subsets of* B*, let* $X, Y, Z \in \text{SubPart}(S)$ *be subpartitions of S, and let* $\text{dom}(X) = \text{dom}(Y) = E$*.*

1. *If* $X \leq_E Y$ *and* $C \vdash^F Y$*, then* $C \vdash^F X$*.*
2. *If* $C \vdash^F X$ *and* $C \vdash^F Y$*, then* $C \vdash^F X \vee_E Y$*.*
3. $B \vdash^F X$*.*
4. $\{\} \vdash^F X$ *if and only if* $X = \text{Ind}_E$*.*
5. *If* $C \vdash^F X$ *and* $D \vdash^F X$*, then* $C \cap D \vdash^F X$ *and* $C \cup D \vdash^F X$*.*
6. *If* $X \subseteq Z$*, and* $C \vdash^F Z$*, then* $C \vdash^F X$*.*

Proof. The first 4 parts will use the equivalent definition from Proposition 20 that $C \vdash^F X$ if and only if $X \leq_S V_S(C)$. 1 and 2 are immediate from this definition.

3 follows directly from Definition 23.

4 follows directly from the fact that $\bigvee_S(\{\}) = \text{Ind}_S$, and $\text{Ind}_S|E = \text{Ind}_E$ so $X \leq_E \bigvee_S(C)|E$ if and only if $X = \text{Ind}_E$.

For part 5, we will use the equivalent definition from Proposition 20 that $C \vdash^F X$ if and only if $\overset{F}{\chi}_C(s,t) \in [s]_X$ for all $s,t \in E$. Assume that for all $s,t \in E$, $\overset{F}{\chi}_C(s,t) \in [s]_X$ and $\overset{F}{\chi}_D(s,t) \in [s]_X$. Thus, for all $s,t \in E$, $\overset{F}{\chi}_{C\cap D}(s,t) = \overset{F}{\chi}_C(\overset{F}{\chi}_D(s,t),t) \in [\overset{F}{\chi}_D(s,t)]_X = [s]_X$. Similarly, for all $s,t \in E$, $\overset{F}{\chi}_{C\cup D}(s,t) = \overset{F}{\chi}_C(s,\overset{F}{\chi}_D(s,t)) \in [s]_X$. Thus $C \cap D \vdash^F X$ and $C \cup D \vdash^F X$.

For part 6, we use the definition that $C \vdash^F X$ if and only if $\overset{F}{\chi}_C(x,y) \in x$ for all $x,y \in X$. Clearly if $X \subseteq Z$, and $\overset{F}{\chi}_C(x,y) \in x$ for all $x,y \in Z$, then $\overset{F}{\chi}_C(x,y) \in x$ for all $x,y \in X$.

Note that while the set of C that generate an $X \in \text{Part}(S)$ is closed under supersets, the set of C that generate an $X \in \text{SubPart}(S)$ is merely closed under union.

Further note that part 6 of Proposition 21 uses the subset relation on subpartitions, which is a slightly unnatural relation.

4.2 History of a Subpartition

Definition 24 (history of a subpartition). *Given a finite factored set* $F = (S,B)$ *and a subpartition* $X \in \text{SubPart}(S)$*, let* $h^F(X)$ *denote the smallest (according to the subset ordering) subset of* B *such that* $h^F(X) \vdash^F X$*.*

Proposition 22. *Given a finite factored set* $F = (S, B)$*,* $h^F : \text{SubPart}(S) \to P(B)$ *is well-defined, and if* X *is a partition of* S*, this definition coincides with Definition 17.*

Proof. Fix a finite factored set $F = (S, B)$ and a subpartition $X \in \text{SubPart}(S)$, and let $h^F(X)$ be the intersection of all $C \subseteq B$ such that $C \vdash^F X$. It suffices to show that $h^F(X) \vdash^F X$. Then $h^F(X)$ will clearly be the unique smallest (according to the subset ordering) subset of B such that $h^F(X) \vdash^F X$. The fact that this definition coincides with Definition 17 if $X \in \text{Part}(S)$ is clear.

Note that $h^F(X)$ is a finite intersection, since there are only finitely many subsets of B, and that $h^F(X)$ is a nonempty intersection since $B \vdash^F X$. Thus, we can express $h^F(X)$ as a (possibly empty) composition of finitely many binary intersections. By part 5 of Proposition 21, the intersection of two subsets that generate X also generates X. Thus $h^F(X) \vdash^F X$.

We will now give five basic properties of the history of subpartitions, followed by two more properties that are less basic.

Proposition 23. *Let* $F = (S, B)$ *be a finite factored set, let* $X, Y, Z \in \text{SubPart}(S)$ *be subpartitions of* S*, and let* $\text{dom}(X) = \text{dom}(Y) = E$*.*

1. *If* $X \leq_E Y$*, then* $h^F(X) \subseteq h^Y(Y)$*.*
2. $h^F(X \vee_E Y) = h^F(X) \cup h^F(Y)$*.*
3. *If* $X \subseteq Z$*, then* $h^F(X) \subseteq h^F(Z)$*.*
4. $h^F(X) = \{\}$ *if and only if* $X = \text{Ind}_E$*.*
5. *If* S *is nonempty, then* $h^F(b) = \{b\}$ *for all* $b \in B$*.*

Proof. Parts 1, 3, and 4 are trivial consequences of Proposition 21, and part 5 is just a restatement of part 4 of Proposition 13.

For part 2, first observe that $h^F(X \vee_E Y) \supseteq h^F(X) \cup h^F(Y)$, by part 1 of Proposition 21. Thus it suffices to show that $h^F(X) \cup h^F(Y) \supseteq h^F(X \vee_E Y)$, by showing that $h^F(X) \cup h^F(Y) \vdash^F X \vee_E Y$.

We will use condition 7 in Proposition 20. Clearly

$$X \leq_E (\bigvee_E (h^F(X)) \mid E)$$
$$\leq_E (\bigvee_S (h^F(X) \cup h^F(Y)) \mid E),$$

and similarly,

$$Y \leq_E (\bigvee_E (h^F(Y)) \mid E)$$
$$\leq_E (\bigvee_S e(h^F(X) \cup h^F(Y)) \mid E).$$

Thus, $X \vee_E Y \leq_E (\bigvee_S (h^F(X) \cup h^F(Y)) \mid E)$.

Next, we need to show that $\overset{F}{\chi}_{h^F(X) \cup h^F(Y)}(E, E) = E$. Clearly $E \subseteq \overset{F}{\chi}_{h^F(X) \cup h^F(Y)}(E, E)$.

Let s and t be elements of E, and observe that $\overset{F}{\chi}_{h^F(X) \cup h^F(Y)}(s, t) = \overset{F}{\chi}_{h^F(X)}(s, \overset{F}{\chi}_{h^F(Y)}(s, t))$. We have that $\overset{F}{\chi}_{h^F(Y)}(s, t) \in E$, since $\overset{F}{\chi}_{h^F(Y)}(E, E) = E$. Thus, we also have that $\overset{F}{\chi}_{h^F(X)}(s, \overset{F}{\chi}_{h^F(Y)}(s, t)) \in E$, since $\overset{F}{\chi}_{h^F(X)}(E, E) = E$. Thus, $\overset{F}{\chi}_{h^F(X) \cup h^F(Y)}(E, E) \subseteq E$.

Thus we have that $X \vee_E Y \leq_E (\bigvee_S (h^F(X) \cup h^F(Y)) \mid E)$ and $\overset{F}{\chi}_{h^F(X) \cup h^F(Y)}(E, E) = E$. Thus, by condition 7 in Proposition 20, $h^F(X) \cup h^F(Y) \vdash^F X \vee_E Y$, so $h^F(X \vee_E Y) = h^F(X) \cup h^F(Y)$. □

Lemma 1. *Let* $F = (S, B)$ *be a finite factored set, and let* $X, Y \in \text{Part}(E)$ *be subpartitions of* S *with the same domain. If* $h^F(X) \cap h^F(Y) = \{\}$, *then* $h^F(X) = h^F(X \mid y)$ *for all* $y \in Y$.

Proof. Let $F = (S, B)$ be a finite factored set, let $E \subseteq S$, and let $X, Y \in \text{Part}(E)$.

We start by showing that $(B \setminus h^F(X)) \vdash^F Y$ and $(B \setminus h^F(Y)) \vdash^F X$. Observe that $\chi_{B \setminus h^F(X)}(E, E) = \chi_{h^F(X)}(E, E) = E$. Further observe that $B \setminus h^F(X) \supseteq h^F(Y)$, so $\bigvee_S (B \setminus h^F(X)) \geq_S \bigvee_S (h^F(Y))$, so $(\bigvee_S (B \setminus h^F(X)) \mid E) \geq_E (\bigvee_S (h^F(Y)) \mid E) \geq_E Y$. Thus, $(B \setminus h^F(X)) \vdash^F Y$. Symmetrically, $(B \setminus h^F(Y)) \vdash^F X$.

Fix some $y \in Y$. We start by showing that $h^F(X) \supseteq h^F(X|y)$.

We have that $\chi^F_{B \setminus h^F(X)}(y, E) \subseteq y$, so $\chi^F_{h^F(X)}(E, y) \subseteq y$, so for all $x \in X$, we have $\chi^F_{h^F(X)}(x \cap y, y) \subseteq y$. We also have $\chi^F_{h^F(X)}(x \cap y, y) \subseteq \chi^F_{h^F(X)}(x, E) \subseteq x$. Thus $\chi^F_{h^F(X)}(x \cap y, y) \subseteq x \cap y$. Every element of $X|y$ is of the form $x \cap y$ for some $x \in X$, so we have $h^F(X) \vdash^F (X|y)$, so $h^F(X) \supseteq h^F(X|y)$.

Next, we need to show that $h^F(X) \subseteq h^F(X|y)$. For this, it suffices to show that $h^F(X|y) \vdash^F X$. Let s, t be arbitrary elements of E. It suffices to show that $\chi^F_{h^F(X|y)}(s, t) \in [s]_X$.

First, observe that since $(B \setminus h^F(Y)) \supseteq h^F(X) \supseteq h^F(X|y)$, we have that $\chi^F_{h^F(X|y)}(s, t) = \chi^F_{B \setminus h^F(Y)}(\chi^F_{h^F(X|y)}(s, t), t)$.

Let r be an arbitrary element of y. We thus have:

$$\begin{aligned}
\chi^F_{h^F(X|y)}(s, t) &= \chi^F_{B \setminus h^F(Y)}(\chi^F_{h^F(X|y)}(s, t), t) \\
&= \chi^F_{B \setminus h^F(Y)}(\chi^F_{h^F(Y)}(r, \chi^F_{h^F(X|y)}(s, t)), t) \\
&= \chi^F_{B \setminus h^F(Y)}(\chi^F_{h^F(X|y)}(\chi^F_{h^F(Y)}(r, s), \chi^F_{h^F(Y)}(r, t)), t).
\end{aligned}$$

Let $s' = \chi^F_{h^F(X|y)}(\chi^F_{h^F(Y)}(r, s), \chi^F_{h^F(Y)}(r, t))$. Note that $\chi^F_{h^F(Y)}(r, t)$ and $\chi^F_{h^F(Y)}(r, s)$ are both in y. Thus we have that $s' \in [\chi^F_{h^F(Y)}(r, s)]_{(X|y)}$. Since $(B \setminus h^F(Y)) \vdash^F X$, $\chi^F_{h^F(Y)}(r, s) = \chi^F_{B \setminus h^F(Y)}(s, r) \in [s]_X$. Thus $[\chi^F_{h^F(Y)}(r, s)]_{(X|y)} \subseteq [\chi^F_{h^F(Y)}(r, s)]_X = [s]_X$, so $s' \in [s]_X$.

We have that $\chi^F_{h^F(X|y)}(s,t) = \chi^F_{B \setminus h^F(Y)}(s',t)$. However, since $B \setminus h^F(Y) \vdash^F X$, we have $\chi^F_{B \setminus h^F(Y)}(s',t) \in [s']_X = [s]_X$. Thus, $h^F(X) \subseteq h^F(X|y)$, so $h^F(X) = h^F(X|y)$. □

Lemma 2. *Let* $F = (S, B)$ *be a finite factored set. Let* $E \subseteq S$ *and let* $X, Y \in \mathrm{Part}(E)$ *be subpartitions of* S *with the same domain. Then* $h^F(X \vee_E Y) = h^F(X) \cup \bigcup_{x \in X} h^F(Y \mid x)$.

Proof. Since $X \leq_E X \vee_E Y$, we have $h^F(X) \subseteq h^F(X \vee_E Y)$. Similarly, for all $x \in X$, since $Y|x \subseteq X \vee_E Y$, we have $h^F(Y|x) \subseteq h^F(X \vee_E Y)$. Thus, $h^F(X \vee_E Y) \supseteq h^F(X) \cup \bigcup_{x \in X} h^F(Y|x)$. We still need to show that $h^F(X \vee_E Y) \subseteq h^F(X) \cup \bigcup_{x \in X} h^F(Y|x)$.

We start with the special case where $|X| = 2$. Let $X = \{x_0, x_1\}$. In this case, we want to show that $h^F(X \vee_E Y) = h^F(X) \cup h^F(Y|x_0) \cup h^F(Y|x_0)$. Let $C = h^F(X)$, let $C_0 = h^F(Y|x_0)$, and let $C_1 = h^F(Y|x_1)$.

Consider arbitrary $s, t \in E$. Without loss of generality, assume that $s \in x_0$, and let $y = [s]_Y$. It suffices to show that $\chi^F_{C \cup C_0 \cup C_1}(s,t) \in x_0 \cap y$. Fix some $r \in x_1$.

$$\begin{aligned}
\chi^F_{C \cup C_0 \cup C_1}(s,t) &= \chi^F_{C_0}(s, \chi^F_C(s, \chi^F_{C_1}(s,t))) \\
&= \chi^F_{C_0}(s, \chi^F_C(s, \chi^F_C(r, \chi^F_{C_1}(s,t)))) \\
&= \chi^F_{C_0}(s, \chi^F_C(s, \chi^F_{C_1}(\chi^F_C(r,s), \chi^F_C(r,t)))).
\end{aligned}$$

Observe that $\chi^F_C(r,s)$ and $\chi^F_C(r,t)$ are both in x_1, so $\chi^F_{C_1}(\chi^F_C(r,s), \chi^F_C(r,t)) \in x_1$, and thus is in E. Combining this with the fact that $s \in x_0$ gives us that $\chi^F_C(s, \chi^F_{C_1}(\chi^F_C(r,s), \chi^F_C(r,t))) \in x_0$. Thus, since $s \in x_0 \cap y$,

$$\chi^F_{C \cup C_0 \cup C_1}(s,t) = \chi^F_{C_0}(s, \chi^F_C(s, \chi^F_{C_1}(\chi^F_C(r,s), \chi^F_C(r,t)))) \in x_0 \cap y.$$

Now, consider the case where $|X| \neq 2$. If $|X| = 0$, then $E = \{\}$, so all subpartitions involved are empty, and thus have the same (empty) history. If $|X| = 1$, let $X = \{E\}$. Then

$$\begin{aligned} h^F(X \vee_E Y) &= h^F(Y) \\ &= h^F(Y \mid E) \subseteq h^F(X) \cup h^F(Y \mid E) \\ &= h^F(X) \cup \bigcup_{x \in X} h^F(Y \mid x). \end{aligned}$$

Thus, we can restrict our attention to the case where $|X| \geq 3$.

Observe that $X \vee_E Y = \bigvee_E(\{(Y|x) \cup \{E \setminus x\} \mid x \in X\})$. Thus $h^F(X \vee_E Y) = \bigcup_{x \in X} h^F((Y|x) \cup \{E \setminus x\})$. However, from the case where $|X| = 2$, we have

$$\begin{aligned} h^F((Y \mid x) \cup \{E \setminus x\}) &= h^F(\{x, E \setminus x\} \vee_E ((Y \mid x) \cup \{E \setminus x\})) \\ &= h^F(\{x, E \setminus x\}) \cup h^F(\{E \setminus x\}) \cup h^F(Y \mid x). \end{aligned}$$

$h^F(\{E \setminus x\})$ is empty, so this gives us that $h^F(X \vee_E Y) = \bigcup_{x \in X}(h^F(Y|x) \cup h^F(\{x, E \setminus x\}))$. Since $\bigvee_E(\{\{x, E \setminus x\} \mid x \in X\}) = X$, $\bigcup_{x \in X} h^F(\{x, E \setminus x\}) = h^F(X)$, so we have $h^F(X \vee_E Y) = h^F(X) \cup \bigcup_{x \in X} h^F(Y|x)$. □

4.3 Conditional Orthogonality

We can also extend our notions of orthogonality and time to subpartitions.

Definition 25. *Let* $F = (S, B)$ *be a finite factored set. Let* $X, Y \in \mathrm{SubPart}(S)$ *be subpartitions of* S*. We write* $X \perp^F Y$ *if* $h^F(X) \cap h^F(Y) = \{\}$*, we write* $X \leq^F Y$ *if* $h^F(X) \subseteq h^F(Y)$*, and we write* $X <^F Y$ *if* $h^F(X) \subset h^F(Y)$*.*

We give this definition in general, but it is not clear whether orthogonality and time should be considered philosophically meaningful when the domains of the inputs differ from each other. Further, the temporal structure of subpartitions will mostly be outside the scope of this paper, and the orthogonality structure on subpartitions will mostly just be used for the following pair of definitions.

Definition 26 (conditional orthogonality given a subset). *Given a finite factored set* $F = (S, B)$*, partitions* $X, Y \in \text{Part}(S)$*, and* $E \subseteq S$*, we say* X *and* Y *are orthogonal given* E *(in* F*), written* $X \perp^F Y | E$*, if* $(X|E) \perp^F (Y|E)$*.*

Definition 27 (conditional orthogonality). *Given a finite factored set* $F = (S, B)$*, and partitions* $X, Y, Z \in \text{Part}(S)$*, if* $X \perp^F Y | z$ *for all* $z \in Z$*, then we say* X *and* Y *are orthogonal given* Z *(in* F*), written* $X \perp^F Y | Z$*.*

Unconditioned orthogonality can be thought of as a special case of conditional orthogonality, where you condition on the indiscrete partition.

Proposition 24. *Given a finite factored set* $F = (S, B)$ *and partitions* $X, Y \in \text{Part}(S)$*,* $X \perp^F Y$ *if and only if* $X \perp^F Y | \text{Ind}_S$*.*

Proof. If $S = \{\}$, then there is only one partition $X = \{\}$, and $X \perp^F X$ holds. Also, since Ind_S is empty, $X \perp^F X | \text{Ind}_S$ holds vacuously.

If $S \neq \{\}$, then $\text{Ind}_S = \{S\}$, so $X \perp^F Y | \text{Ind}_S$ if and only if $X \perp^F Y | S$ if and only if $X|S \perp^F Y|S$ if and only if $X \perp^F Y$. □

The primary combinatorial structure of finite factored sets that we will be interested in is the structure of orthogonality ($X \perp^F Y$), conditional orthogonality ($X \perp^F Y | Z$), and time ($X \leq^F Y$ and $X <^F Y$) on inputs that are partitions.

We now will show that conditional orthogonality satisfies (a slight modification of) the axioms for a compositional semigraphoid.

Theorem 2. *Let* $F = (S, B)$ *be a finite factored set, and let* $X, Y, Z, W \in \text{Part}(S)$ *be partitions of* S*.*

1. *If* $X \perp^F Y | Z$*, then* $Y \perp^F X | Z$*. (symmetry)*
2. *If* $X \perp^F (Y \vee_S W) | Z$*, then* $X \perp^F Y | Z$ *and* $X \perp^F W | Z$*. (decomposition)*
3. *If* $X \perp^F (Y \vee_S W) | Z$*, then* $X \perp^F Y | (Z \vee_S W)$*. (weak union)*

4. *If* $X \perp^F Y \mid Z$ *and* $X \perp^F W \mid (Z \vee_S Y)$, *then* $X \perp^F (Y \vee_S W) \mid Z$. *(contraction)*
5. *If* $X \perp^F Y \mid Z$ *and* $X \perp^F W \mid Z$, *then* $X \perp^F (Y \vee_S W) \mid Z$. *(composition)*

Proof. Symmetry is clear from the definition.

Decomposition and composition both follow directly from the fact that for all $z \in Z$, $h^F((Y \vee_S W)|z) = h^F((Y|z) \vee_z (W|z)) = h^F(Y|z) \cup h^F(W|z)$.

For weak union, assume that $X \perp^F (Y \vee_S W)|Z$. Thus, for all $z \in Z$, $h^F(X|z) \cap h^F((Y \vee_S W)|z) = \{\}$. In particular, this means that $h^F(X|z) \cap h^F(W|z) = \{\}$, so by Lemma 1, for all $w \in W$, $h^F(X|z) = h^F(X|w \cap z)$. Further, we have that for all $w \in W$, $h^F(Y|w \cap z) \subseteq h^F(Y \vee_S W|z)$. Thus, for all $w \in W$, $h^F(X|w \cap z) \cap h^F(Y|w \cap z) = \{\}$, which since every element of $W \vee_S Z$ is of the form $w \cap z$ for some $w \in W$ and $z \in Z$, means that $X \perp^F Y|(Z \vee_S W)$.

Finally, for contraction, assume that $X \perp^F Y|Z$ and $X \perp^F W|Z \vee_S Y$. Fix some $z \in Z$. We want to show that $h^F(X|z) \cap h^F((Y \vee_S W)|z) = \{\}$. We have that $h^F((Y \vee_S W)|z) = h^F((Y|z) \vee_z (W|z))$, and by Lemma 2, $h^F((Y|z) \vee_z (W|z)) = h^F(Y|z) \cup \bigcup_{y \in Y} h^F(W|(y \cap z))$. Thus, it suffices to show that $h^F(X|z) \cap h^F(Y|z) = \{\}$ and $h^F(X|z) \cap h^F(W|(y \cap z)) = \{\}$ for all $y \in Y$.

The fact that $h^F(X|z) \cap h^F(Y|z) = \{\}$ follows directly from $X \perp^F Y|Z$.

Fix a $y \in Y$. If $y \cap z = \{\}$, then $h^F(W|(y \cap z)) = \{\}$, so $h^F(X|z) \cap h^F(W|(y \cap z)) = \{\}$. Otherwise, we have $h^F(X|z) = h^F(X|(y \cap z))$ by Lemma 1, and we have that $h^F(X|(y \cap z)) \cap h^F(W|(y \cap z)) = \{\}$, since $X \perp^F W|Z \vee_S Y$, so we have $h^F(X|z) \cap h^F(W|(y \cap z)) = \{\}$.

Thus, $X \perp^F (Y \vee_S W)|Z$. □

The first four parts of Theorem 2 are essentially the semigraphoid axioms. The difference is that the semigraphoid axioms are normally defined as a ternary relation on

disjoint sets of variables. We use partitions instead of sets of variables, use common refinement instead of union, and have no need for the disjointness condition. The fifth part (composition) is a converse to the decomposition axiom that is sometimes added to define a compositional semigraphoid.

The results in this paper will not depend on the theory of compositional semigraphoids, so we will not need to make the analogy any more explicit, but it is nice to note the similarity to existing well-studied structures.

We also get a nice relationship between conditional orthogonality and the refinement order.

Proposition 25. *Let* $F = (S, B)$ *be a finite factored set, and let* $X, Y \in \text{Part}(S)$ *be partitions of* S. $X \perp^F X|Y$ *if and only if* $X \leq_S Y$.

Proof. If $X \perp^F X|Y$, then for all $y \in Y$, $h^F(X|y) = \{\}$, so $X|y = \text{ind}_y$, so for all $s, t \in y$, we have $s \sim_{X|y} t$, and thus $s \sim_X t$. Thus, for all $s, t \in S$, if $s \sim_Y t$, then $s \sim_X t$. Thus $X \leq_S Y$.

Conversely, if $X \leq_S Y$, observe that for all $y \in Y$, $X|y = \text{ind}_y$, so $h^F(X|y) = \{\}$. Thus, $X \perp^F X|Y$. □

In the next post, we will prove the fundamental theorem of finite factored sets, which says that conditional orthogonality exactly corresponds to conditional independence in all probability distributions that can be put on the relevant finite factored set.

Finite Factored Sets: Polynomials and Probability

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

In this post, given a finite factored set $F = (S, B)$, we will show how to associate each $E \subseteq S$ with a characteristic polynomial, Q^F_E. We will discuss how to factor these characteristic polynomials, and use these characteristic polynomials to build up to the fundamental theorem of finite factored sets, which associates conditional orthogonality with conditional independence in probability distributions.

5.1. Characteristic Polynomials

Definition 28. *Given a finite factored set* $F = (S, B)$*, let* Poly^F *denote the ring of polynomials with coefficients in* R *and variables in* $P(S)$*.*

Definition 29. *Given a finite factored set* $F = (S, B)$*, a* $p \in \text{Poly}^F$*, and an* $f : P(S) \to R$*, we write* $p(f) \in R$ *for the evaluation of* p *at* f*, computed by replacing each* $E \subseteq S$ *with* $f(E)$*.*

Definition 30. *Given a finite factored set* $F = (S, B)$*, and a polynomial* $p \in \text{Poly}^F$*,* $\text{supp}(p) \subseteq P(S)$ *denotes the set of all variables* $v \in P(S)$ *that appear in* p*.* $\text{supp}(p)$ *is called the support of* p*.*

Definition 31. *Given a finite factored set* $F = (S, B)$ *and an* $E \subseteq S$*, let* $Q^F_E \in \text{Poly}^F$ *be given by* $Q^F_E = \sum_{s \in E} \prod_{b \in B} [s]_b$*.* Q^F_E *is called the characteristic polynomial of* E *(in* F*).*

We will be building up to an understanding of how to factor Q^F_E into irreducibles. For that, we will first need to give some basic notation for manipulating polynomials in

Poly^F.

Definition 32. *Given a finite factored set* $F = (S, B)$, *an* $s \in S$, *and a* $C \subseteq B$, *let* $\text{mono}_C^F(s) \in \text{Poly}^F$ *be given by* $\text{mono}_C^F(s) = \prod_{b \in C} [s]_b$.

Definition 33. *Given a finite factored set* $F = (S, B)$, *an* $E \subseteq S$, *and a* $C \subseteq B$, *let* $\text{monos}_C^F(E) \in P(\text{Poly}^F)$ *be given by*

$\text{monos}_C^F(E) = \{\text{mono}_C^F(s) \mid s \in E\}$.

Definition 34. *Given a finite factored set* $F = (S, B)$, *an* $E \subseteq S$, *and a* $C \subseteq B$, *let* $\text{poly}_C^F(E) \in \text{Poly}^F$ *be given by* $\text{poly}_C^F(E) = \sum_{m \in \text{monos}_C^F(E)} m$.

Proposition 26. *Let* $F = (S, B)$ *be a finite factored set, and let* $E \subseteq S$. *Then* $Q_E^F = \text{poly}_B^F(E)$.

Proof. We start by showing that for all $s \neq t \in S$, $\text{mono}_B^F(s) \neq \text{mono}_B^F(t)$.

Let $s \neq t \in S$ be arbitrary. By Proposition 3, if $s \neq t$, there must be some $b \in B$ such that $[s]_b \neq [t]_b$. Then, note that $[s]_b \in \text{supp}(\text{mono}_B^F(s))$. If $[s]_b$ were also in $\text{supp}(\text{mono}_B^F(t))$, then t would be in both $[s]_b$ and $[t]_b$, contradicting the fact that these two sets are disjoint. Therefore $\text{mono}_B^F(s) \neq \text{mono}_B^F(t)$.

Thus $\text{monos}_B^F(E)$ has exactly one element for each element of E, so we have that $\sum_{m \in \text{monos}_B^F(E)} m = \sum_{s \in E} \text{mono}_B^F(s) = Q_E^F$. □

Proposition 27. *Let* $F = (S, B)$ *be a finite factored set, and let* $E_0, E_1 \subseteq S$ *be subsets of* S*. Let* $C_0, C_1 \subseteq B$ *be disjoint subsets of* B*. Let* $E_2 = \chi^F_{C_0}(E_0, E_1)$*, and let* $C_2 = C_0 \cup C_1$*. Then* $\text{poly}^F_{C_2}(E_2) = \text{poly}^F_{C_0}(E_0) \cdot \text{poly}^F_{C_1}(E_1)$.

Proof. For $i \in \{0, 1, 2\}$, let $M_i = \text{monos}^F_{C_i}(E_i)$. We will start by showing that $f : M_0 \times M_1 \to M_2$, given by $f(m_0, m_1) = m_0 m_1$, is a well-defined function and a bijection.

First, observe that it follows immediately from the definition that for all $s_0, s_1 \in S$, if $s_2 = \chi^F_{C_0}(s_0, s_1)$ we have that $\text{mono}^F_{C_0}(s_0) = \text{mono}^F_{C_0}(s_2)$, $\text{mono}^F_{C_1}(s_1) = \text{mono}^F_{C_1}(s_2)$, and $\text{mono}^F_{C_0}(s_2) \cdot \text{mono}^F_{C_1}(s_2) = \text{mono}^F_{C_2}(s_2)$. Combining these, we get that $\text{mono}^F_{C_0}(s_0) \cdot \text{mono}^F_{C_1}(s_1) = \text{mono}^F_{C_2}(\chi^F_{C_0}(s_0, s_1))$.

For all $(m_0, m_1) \in M_0 \times M_1$, there exists some $s_0 \in E_0$ such that $m_0 = \text{mono}^F_{C_0}(s_0)$, and some $s_1 \in E_1$ such that $m_1 = \text{mono}^F_{C_1}(s_1)$, and this gives us that $m_0 m_1 = \text{mono}^F_{C_0}(s_0)\text{mono}^F_{C_1}(s_1) = \text{mono}^F_{C_2}(\chi^F_C(s_0, s_1)) \in M_2$. Thus, f is well-defined.

To see that f is surjective, observe that for all $m_2 \in M_2$, there exists an $s_2 \in E_2$ such that $m_2 = \text{mono}^F_{C_2}(s_2)$, and there exist $s_0 \in E_0$ and $s_1 \in E_1$ such that $s_2 = \chi^F_C(s_0, s_1)$, and we have $f(\text{mono}^F_{C_0}(s_0), \text{mono}^F_{C_1}(s_1)) = m_2$.

To see that f is injective, observe that for $i \in \{0, 1\}$, for all $m_i \in M_i$, $\text{supp}(m_i) \subseteq \bigcup_{b \in C_i} b$. Further, $\bigcup_{b \in C_0} b$ and $\bigcup_{b \in C_1} b$ are disjoint. Thus, for all $m_0 \in M_0$ and $m_1 \in M_1$,

$\text{supp}(m_i) = \text{supp}(m_0 m_1) \cap \bigcup_{b \in C_i} b$.

This means that for all $m_0, m_0' \in M_0$ and $m_1, m_1' \in M_1$, if $m_0 m_1 = m_0' m_1'$, then $\text{supp}(m_0) = \text{supp}(m_0')$ and $\text{supp}(m_1) = \text{supp}(m_1')$. However, every monomial in M_0 or M_1 is just equal to the product of all variables in its support. Thus $m_0 = \prod_{v \in \text{supp}(m_0)} v = m_0'$ and $m_1 = \prod_{v \in \text{supp}(m_1)} v = m_1'$. Thus f is injective, and thus a bijection between $M_0 \times M_1$ and M_2.

Now, we have that

$$\begin{aligned}
\text{poly}_{C_0}^{F}(E_0) \cdot \text{poly}_{C_1}^{F}(E_1) &= \left(\sum_{m_0 \in M_0} m_0\right)\left(\sum_{m_1 \in M_1} m_1\right) \\
&= \sum_{m_0 \in M_0} \sum_{m_1 \in M_1} m_0 m_1 \\
&= \sum_{(m_0, m_1) \in M_0 \times M_1} m_0 m_1 \\
&= \sum_{(m_0, m_1) \in M_0 \times M_1} f(m_0, m_1) \\
&= \sum_{m_2 \in M_2} m_2 \\
&= \text{poly}_{C_2}^{F}(E_2).
\end{aligned}$$

□

Proposition 28. *Let* $F = (S, B)$ *be a finite factored set, and let* E *be a nonempty subset of* S*. If* p *divides* Q_E^F*, then* $p = r \cdot \text{poly}_C^F(E)$*, for some* $r \in R$ *and*

$C \subseteq B$.

Proof. Let $F = (S, B)$ be a finite factored set, and let E be a nonempty subset of S. Let $p, q \in \text{Poly}^F$ satisfy $pq = Q_E^F$. We thus must have $\text{supp}(p) \cup \text{supp}(q) = \text{supp}(Q_E^F)$.

If there were some $T \in \text{supp}(p) \cap \text{supp}(q)$, then the degree of T in Q_E^F would be at least 2, contradicting the definition of Q_E^F and Corollary 1. Thus, $\text{supp}(p) \cap \text{supp}(q) = \{\}$.

There can be no combining like terms, then, in the product pq. The monomial terms in Q_E^F are in bijective correspondence to the pairs of monomial terms in p and monomial terms in q.

In particular, this means that since all the coefficients in pq are equal to 1, all the coefficients in p must be equal to some $r \in R$, and all of the coefficients in q must be equal to $1/r$.

Further, for all $b \in B$, if $b \cap \text{supp}(p)$ is nonempty, $b \cap \text{supp}(q)$ must be empty, since otherwise Q_E^F would contain a term with two factors in b, which clearly never happens according to the definition of Q_E^F.

Since E is nonempty, for each $b \in B$ there must be some $T \in b \cap \text{supp}(Q_E^F)$. Thus at least one of $b \cap \text{supp}(p)$ and $b \cap \text{supp}(q)$ must be nonempty, so exactly one of $b \cap \text{supp}(p)$ and $b \cap \text{supp}(q)$ must be nonempty.

Let C be the set of all $b \in B$ such that $b \cap \text{supp}(p)$ is nonempty.

For every $b \in C$, every term of Q_E^F has exactly one factor in b. Thus, every term in p has exactly one factor in b. These cover all variables in the support of p, so each term

in p must have total degree |C|.

For each $m \in \text{monos}_C^F(E)$, m divides a term in Q_E^F. Since m has no common support with q, m must also divide a term in p. Thus $r \cdot m$ must be a term in p. Conversely, every term in p divides a term in Q_E^F, and thus must be in $\text{monos}_C^F(E)$. Thus every term in p is of the form $r \cdot m$ for some $m \in \text{monos}_C^F(E)$. Thus

$p = \sum_{m \in \text{monos}_C^F(E)} r \cdot m = r \cdot \text{poly}_C^F(E)$. □

5.2. Factoring Characteristic Polynomials

We will now show how to factor characteristic polynomials into irreducibles.

Definition 35. *Given a finite factored set* $F = (S, B)$*, and a nonempty subset* $E \subseteq S$*, let* $\text{Irr}^F(E) \subseteq P(B)$ *denote the set of all* $C \subseteq B$ *such that:*

1. C *is nonempty,*
2. $\chi_C^F(E, E) = E$*, and*
3. *there is no nonempty strict subset* $D \subset C$ *such that* $\chi_D^F(E, E) = E$*.*

Proposition 29. *Let* $F = (S, B)$ *be a finite factored set, and let* E *be a nonempty subset of* S*. Then* $\text{Irr}^F(E) \in \text{Part}(B)$*.*

Proof. Let $F = (S, B)$ be a finite factored set, and let E be a nonempty subset of S. It suffices to show that the sets in $\text{Irr}^F(E)$ are pairwise disjoint and cover B.

We start by showing that the set of all $C \subseteq B$ satisfying $\overset{F}{\chi_C}(E, E) = E$ is closed under intersection. Indeed, if $\overset{F}{\chi_{C_0}}(E, E) = E$ and $\overset{F}{\chi_{C_1}}(E, E) = E$, then

$$\overset{F}{\chi_{C_0 \cap C_1}}(E, E) = \overset{F}{\chi_{C_0}}(E, \overset{F}{\chi_{C_1}}(E, E)) = \overset{F}{\chi_{C_0}}(E, E) = E.$$

Next, observe that $\overset{F}{\chi_B}(E, E) = E$. Thus, for all $b \in B$, we can consider $C_b = \bigcap_{C \subseteq B, b \in C, \overset{F}{\chi_C}(E,E)=E} C$. Since C_b is an intersection of a finite nonempty collection of sets C satisfying $\overset{F}{\chi_C}(E, E) = E$, we have that $\overset{F}{\chi_{C_b}}(E, E) = E$. Further, $b \in C_b$, so C_b is nonempty.

Assume for the purpose of contradiction that there is some nonempty strict subset $D \subset C_b$ such that $\overset{F}{\chi_D}(E, E) = E$. If $b \in D$, then we have a contradiction by the definition of C_b. If $b \notin D$, then note that $\overset{F}{\chi_{B \setminus D}}(E, E) = E$, so $\overset{F}{\chi_{C_b \setminus D}}(E, E) = E$, and $C_b \setminus D$ is a nonempty strict subset of C_b that contains b, contradicting the definition of C_b.

Thus $C_b \in \mathrm{Irr}^F(E)$ for all $b \in B$, and since $b \in C_b$, this means that the sets in $\mathrm{Irr}^F(E)$ cover B.

Next, we need to show that the sets in $\mathrm{Irr}^F(E)$ are pairwise disjoint. Let $C_0, C_1 \in \mathrm{Irr}^F(E)$ be arbitrary distinct elements. We have that $\overset{F}{\chi_{C_0 \cap C_1}}(E, E) = E$, and $C_0 \cap C_1$ is a subset of C_0 and C_1, and thus a strict subset of at least one of them. Thus $C_0 \cap C_1$ is empty.

Thus $\mathrm{Irr}^F(E) \in \mathrm{Part}(B)$. □

The following two propositions constitute a factorization of $\overset{F}{Q_E}$ into irreducibles.

Proposition 30. *Let* $F = (S, B)$ *be a finite factored set, and let* E *be a nonempty subset of* S*. Then* $\overset{F}{Q_E} = \prod_{C \in \mathrm{Irr}^F(E)} \overset{F}{\mathrm{poly}_C}(E)$*.*

Proof. Let $F = (S, B)$ be a finite factored set, and let E be a nonempty subset of S. Let $n = |\mathrm{Irr}^F(E)|$, and let $\mathrm{Irr}^F(E) = \{C_0, \dots, C_{n-1}\}$. For $0 \leq k < n$, let $C_{\leq k} = \bigcup_{i=0}^{k} C_i$.

We will show by induction on k that $\prod_{i=0}^{k} \mathrm{poly}^F_{C_i}(E) = \mathrm{poly}^F_{C_{\leq k}}(E)$ for all $0 \leq k < n$.

If $k = 0$, the result is trivial, as $\prod_{i=0}^{0} \mathrm{poly}^F_{C_i}(E) = \mathrm{poly}^F_{C_0}(E) = \mathrm{poly}^F_{C_{\leq 0}}(E)$.

For $k > 0$, observe that C_k and $C_{\leq k-1}$ are disjoint, and that $E = \chi^F_{C_k}(E, E)$, thus by Proposition 27, we have $\mathrm{poly}^F_{C_k}(E) \cdot \mathrm{poly}^F_{C_{\leq k-1}}(E) = \mathrm{poly}^F_{C_{\leq k}}(E)$. Thus, by induction, we get $\prod_{i=0}^{k} \mathrm{poly}^F_{C_i}(E) = \mathrm{poly}^F_{C_{\leq k}}(E)$.

In the case where $k = n - 1$, this gives that $\prod_{C \in \mathrm{Irr}^F(E)} \mathrm{poly}^F_C(E) = \mathrm{poly}^F_B(E) = Q^F_E$. □

Proposition 31. *Let* $F = (S, B)$ *be a finite factored set, and let* E *be a nonempty subset of* S*. Then* $\mathrm{poly}^F_C(E)$ *is irreducible for all* $C \in \mathrm{Irr}^F(E)$*.*

Proof. Let $F = (S, B)$ be a finite factored set, let E be a nonempty subset of S, and let $C \in \mathrm{Irr}^F(E)$.

Assume for the purpose of contradiction that $p_0 \cdot p_1 = \mathrm{poly}^F_C(E)$, and that both p_0 and p_1 have nonempty support. By Proposition 28, we have that $p_i = r_i \cdot \mathrm{poly}^F_{C_i}(E)$, for some $r_0, r_1 \in R$, and $C_0, C_1 \subseteq B$.

We will first need to show that C_0 and C_1 are nonempty and disjoint. They must be nonempty, because p_0 and p_1 have nonempty support. Assume for the purpose of contradiction that $b \in C_0 \cap C_1$. Let s be an element of E, and note that for $i \in \{0, 1\}$,

we have $[s]_b \in \operatorname{supp} \operatorname{poly}^F_{C_i}(E)$. Thus $[s]_b$ must be degree at least 2 in $\operatorname{poly}^F_C(E)$, which contradicts the fact that every variable clearly has degree at most 1 in $\operatorname{poly}^F_C(E)$.

Next, we need to show that $C_0 \cup C_1 = C$. We already know that

$$\begin{aligned}\operatorname{supp}(\operatorname{poly}^F_C(E)) &= \operatorname{supp}(r_0 r_1 \operatorname{poly}^F_{C_0}(E) \operatorname{poly}^F_{C_1}(E)) \\ &= \operatorname{supp}(\operatorname{poly}^F_{C_0}(E)) \cup \operatorname{supp}(\operatorname{poly}^F_{C_1}(E)).\end{aligned}$$

Let s be an element of E. Given an arbitrary $b \in B$, we have that $b \in C$ if and only if $[s]_b \in \operatorname{supp}(\operatorname{poly}^F_C(E))$ if and only if $[s]_b \in \operatorname{supp}(\operatorname{poly}^F_{C_i}(E))$ for some $i \in \{0, 1\}$ if and only if $b \in C_0 \cup C_1$.

We now have that C_0 and C_1 are disjoint and that $C = C_0 \cup C_1$. Thus, by Proposition 27, we have that $\operatorname{poly}^F_{C_0}(E) \cdot \operatorname{poly}^F_{C_1}(E) = \operatorname{poly}^F_C(\chi^F_{C_0}(E, E))$. Thus $\operatorname{poly}^F_C(E) = r_0 r_1 \operatorname{poly}^F_C(\chi^F_{C_0}(E, E))$, so $\operatorname{monos}^F_C(E) = \operatorname{monos}^F_C(\chi^F_{C_0}(E, E))$.

Let $s_0, s_1 \in E$ be arbitrary, and let $s_2 = \chi^F_{C_0}(s_0, s_1)$. Note that $\operatorname{mono}^F_C(s_2) \in \operatorname{monos}^F_C(\chi^F_{C_0}(E, E)) = \operatorname{monos}^F_C(E)$, so there is some $s_3 \in E$ such that $\operatorname{mono}^F_C(s_2) = \operatorname{mono}^F_C(s_3)$. Thus $s_2 \sim_b s_3$ for all $b \in C$. However, we also have that $s_2 \sim_b s_1$ for all $b \in B \setminus C$, so $s_2 = \chi^F_C(s_3, s_1)$. Since $C \in \operatorname{Irr}^F(E)$, $\chi^F_C(E, E) = E$, so $s_2 = \chi^F_{C_0}(s_0, s_1) \in E$. Since s_0 and s_1 were arbitrary elements of E, we have that $\chi^F_{C_0}(E, E) = E$. Since C_0 is a nonempty strict subset of C, this contradicts the fact that $C \in \operatorname{Irr}^F(E)$.

Thus, $\text{poly}_C^F(E)$ is irreducible for all $C \in \text{Irr}^F(E)$. □

5.3. Characteristic Polynomials and Orthogonality

We can now give an alternate characterization of conditional orthogonality in terms of divisibility of characteristic polynomials.

Lemma 3. *Let* $F = (S, B)$ *be a finite factored set, and let* $X, Y, Z \in \text{Part}(S)$ *be partitions of* S*. The following are equivalent.*

1. $X \perp^F Y \mid Z$.
2. Q_z^F *divides* $Q_{x \cap z}^F \cdot Q_{y \cap z}^F$ *for all* $x \in X$, $y \in Y$, *and* $z \in Z$.
3. $Q_z^F \cdot Q_{x \cap y \cap z}^F = Q_{x \cap z}^F \cdot Q_{y \cap z}^F$ *for all* $x \in X$, $y \in Y$, *and* $z \in Z$.

Proof. Clearly condition 3 implies condition 2. We will first show that condition 1 implies condition 3, and then show that condition 2 implies condition 1.

Let $F = (S, B)$, and let $X, Y, Z \in \text{Part}(S)$ satisfy $X \perp^F Y \mid Z$. Consider an arbitrary $x \in X$, $y \in Y$, and $z \in Z$. We want to show that $Q_z^F \cdot Q_{x \cap y \cap z}^F = Q_{x \cap z}^F \cdot Q_{y \cap z}^F$.

Let $C = h^F(X|z)$. Clearly $C \vdash^F X|z$. We thus have that $\chi_C^F(z, z) = z$, so $\chi_{B \setminus C}^F(z, z) = z$. We also have that $h^F(Y|z) \subseteq B \setminus C$, so $Y|z \leq_z (\bigvee_S (B \setminus C))|z$. These two together give that $B \setminus C \vdash^F Y|z$.

Since $C \vdash^F X|z$, we have that $\chi_C^F(x \cap z, z) = x \cap z$. Thus, by Proposition 27, we have that $\text{poly}_C^F(x \cap z) \cdot \text{poly}_{B \setminus C}^F(z) = Q_{x \cap z}^F$. Similarly, since $B \setminus C \vdash^F Y|z$, we have that $\text{poly}_C^F(z) \cdot \text{poly}_{B \setminus C}^F(y \cap z) = Q_{y \cap z}^F$.

Since $\chi_C^F(x \cap z, y \cap z) \subseteq \chi_C^F(x \cap z, z) = x \cap z$, and $\chi_C^F(x \cap z, y \cap z) \subseteq \chi_C^F(z, y \cap z) = y \cap z$, we have $\chi_C^F(x \cap z, y \cap z) \subseteq x \cap y \cap z$. We also have that

$$\begin{aligned} \chi_C^F(x \cap z, y \cap z) &\supseteq \chi_C^F(x \cap y \cap z, x \cap y \cap z) \\ &\supseteq x \cap y \cap z. \end{aligned}$$

Thus $\chi_C^F(x \cap z, y \cap z) = x \cap y \cap z$.

By Proposition 27, this gives that $\text{poly}_C^F(x \cap z) \cdot \text{poly}_{B \setminus C}^F(y \cap z) = Q_{x \cap y \cap z}^F$.

Finally, since $\chi_C^F(z, z) = z$, we have that $\text{poly}_C^F(z) \cdot \text{poly}_{B \setminus C}^F(z) = Q_z^F$.

Thus, $Q_z^F \cdot Q_{x \cap y \cap z}^F$ and $Q_{x \cap z}^F \cdot Q_{y \cap z}^F$ are both equal to

$\text{poly}_C^F(x \cap z) \cdot \text{poly}_{B \setminus C}^F(y \cap z) \cdot \text{poly}_C^F(z) \cdot \text{poly}_{B \setminus C}^F(z)$.

Thus, condition 1 implies condition 3. It remains to show that condition 2 implies condition 1.

Fix $F = (S, B)$, and $X, Y, Z \in \text{Part}(S)$, and let Q_z^F divide $Q_{x \cap z}^F \cdot Q_{y \cap z}^F$ for all $x \in X$, $y \in Y$, and $z \in Z$. Assume for the purpose of contradiction that it is not the case that $X \perp^F Y | Z$. Thus, there exists some $z \in Z$ such that $h^F(X|z) \cap h^F(Y|z) \neq \{\}$. Let $z \in Z$ and $b \in B$ satisfy $b \in h^F(X|z) \cap h^F(Y|z) \neq \{\}$.

Let $C \subseteq B$ be such that $b \in C$ and $C \in \text{Irr}^F(z)$, and let $p = \text{poly}_C^F(z)$. Thus, p is an irreducible factor of Q_z^F.

Either p divides $Q^F_{x \cap z}$ for all $x \in X$ or p divides $Q^F_{y \cap z}$ for all $y \in Y$, since otherwise there would exist an $x \in X$ and a $y \in Y$ such that p divides neither $Q^F_{x \cap z}$ nor $Q^F_{x \cap z}$, but does divide their product, contradicting the fact that p is irreducible, and thus prime.

Assume without loss of generality that p divides $Q^F_{x \cap z}$ for all $x \in X$. Fix an $x \in X$. Let us first restrict attention to the case where $x \cap z$ is nonempty.

Let $Q^F_{x \cap z} = p \cdot q$. By Proposition 28, $p = r_0 \cdot \text{poly}^F_{C_0}(x \cap z)$ and $q = r_1 \cdot \text{poly}^F_{C_1}(x \cap z)$ for some $r_0, r_1 \in R$ and $C_0, C_1 \subseteq B$. We will show that $C_0 = C, C_1 = B \setminus C$, and $r_0 = r_1 = 1$.

Let s be an element of $x \cap z$. Then for all $b \in B$, $b \in C$ if and only if $[s]_b \in \text{supp}(p)$ if and only if $[s]_b \in \text{supp}(\text{poly}^F_{C_0}(x \cap z))$ if and only if $b \in C_0$. Thus $C_0 = C$.

For all $b \in B \setminus C$, we have $[s]_b \in \text{supp}(Q^F_{x \cap z})$ and $[s]_b \notin \text{supp}(p)$, so $[s]_b \in \text{supp}(q)$, so $b \in C_1$. Similarly, for all $b \in C_1$, $[s]_b \in \text{supp}(q)$, so $[s]_b \notin \text{supp}(p)$, so $b \in B \setminus C$. Thus $C_1 = B \setminus C$.

Since p and $\text{poly}^F_{C_0}(x \cap z)$ both have all coefficients equal to 1, we have $r_0 = 1$. Thus, $p = \text{poly}^F_C(x \cap z)$.

Similarly, since all the coefficients of p are 1 and all the coefficients of $Q^F_{x \cap z}$ are 1, all the coefficients of q are 1, so $r_1 = 1$. Thus, $q = \text{poly}^F_{B \setminus C}(x \cap z)$.

We thus have that $Q^F_{x \cap z} = \text{poly}^F_C(z) \cdot \text{poly}^F_{B \setminus C}(x \cap z)$.

In the case where $x \cap z$ is empty, we also have $Q^F_{x \cap z} = \text{poly}^F_C(z) \cdot \text{poly}^F_{B \setminus C}(x \cap z)$, since both sides are 0.

By Proposition 27, $Q^F_{x \cap z} = \text{poly}^F_B(\chi^F_C(z, x \cap z))$. Thus, $\text{monos}^F_B(x \cap z) = \text{monos}^F_B(\chi^F_C(z, x \cap z))$, so $x \cap z = \chi^F_C(z, x \cap z) = \chi^F_{B \setminus C}(x \cap z, z)$.

Since $x \cap z = \chi^F_{B \setminus C}(x \cap z, z)$ for all $x \in X$, we have that $B \setminus C \vdash^F X|z$. However, this contradicts the fact that $b \notin B \setminus C$, and $b \in h^F(X|z)$.

Thus, condition 2 implies condition 1. □

5.4. Probability Distributions on Finite Factored Sets

The primary purpose of all this discussion of characteristic polynomials has been to build up to thinking about the relationship between orthogonality and probabilistic independence. We will now discuss probability distributions on finite factored sets.

Recall the definition of a probability distribution.

Definition 36. *Given a finite set* S*, a probability distribution on* S *is a function* $P : P(S) \to R$ *such that*

1. $P(E) \geq 0$ *for all* $E \subseteq S$*,*
2. $P(\{\}) = 0$*,*
3. $P(S) = 1$*, and*
4. $P(E_0 \cup E_1) = P(E_0) + P(E_1)$ *whenever* $E_0, E_1 \subseteq S$ *satisfy* $E_0 \cap E_1 = \{\}$*.*

A probability distribution on a finite factored set F is a probability distribution on its underlying set that also satisfies another condition, which represents the probability distribution coming from a product of distributions on the underlying factors.

Definition 37. *Given a finite factored set* $F = (S, B)$*, a probability distribution on* F *is a probability distribution* P *on* S *such that for all* $s \in S$*, we have* $P(\{s\}) = \prod_{b \in B} P([s]_b)$.

Proposition 32. *Given a finite factored set* $F = (S, B)$*, a probability distribution on* S *is a probability distribution* P *on* F *if and only if* $P(E) = Q^F_E(P)$ *for all* $E \subseteq S$.

Proof. If $P(E) = Q^F_E(P)$ for all $E \subseteq S$, in particular this means that $P(\{s\}) = Q^F_{\{s\}}(P) = (\prod_{b \in B}[s]_b)(P) = \prod_{b \in B} P([s]_b)$ for all $s \in S$.

Conversely, if $P(\{s\}) = \prod_{b \in B} P([s]_b)$ for all $s \in S$, then for all $E \subseteq S, P(E) = \sum_{s \in E} \prod_{b \in B} P([s]_b) = (\sum_{s \in E} \prod_{b \in B}[s]_b)(P) = Q^F_E(P)$. □

5.5. The Fundamental Theorem of Finite Factored Sets

We are now ready to state and prove the fundatmental theorem of finite factored sets.

Theorem 3. *Let* $F = (S, B)$ *be a finite factored set, and let* $X, Y, Z \in \text{Part}(S)$ *be partitions of* S*. Then* $X \perp^F Y \mid Z$ *if and only if for all probability distributions* P *on* F *and all* $x \in X$, $y \in Y$*, and* $z \in Z$*, we have* $P(x \cap z) \cdot P(y \cap z) = P(x \cap y \cap z) \cdot P(z)$.

Proof. We already have by Lemma 3 that if $X \perp^F Y|Z$, then for all $x \in X$, $y \in Y$, and $z \in Z$, $Q^F_z \cdot Q^F_{x \cap y \cap z} = Q^F_{x \cap z} \cdot Q^F_{y \cap z}$. Thus for any probability distribution P on F, we have

$$\begin{aligned} P(z) \cdot P(x \cap y \cap z) &= Q^F_z(P) \cdot Q^F_{x \cap y \cap z}(P) \\ &= Q^F_{x \cap z}(P) \cdot Q^F_{y \cap z}(P) \\ &= P(x \cap z) \cdot P(y \cap z). \end{aligned}$$

Conversely, assume that for all probability distributions P on F, and all $x \in X$, $y \in Y$, and $z \in Z$, we have $P(x \cap z) \cdot P(y \cap z) = P(x \cap y \cap z) \cdot P(z)$.

If S is empty, then $\{\}$ is the unique partition of S, and we have $\{\} \perp^F \{\}|\{\}$. Thus, we can restrict our attention to the case where S is nonempty.

Fix an arbitrary $x \in X$, $y \in Y$, and $z \in Z$. Let $q = Q^F_{x \cap z} \cdot Q^F_{y \cap z} - Q^F_{x \cap y \cap z} \cdot Q^F_z$. We will first show that $q(f) = 0$ for all $f : P(S) \to R^{>0}$.

Given an arbitrary $f : P(S) \to R^{>0}$, we can define $P_f : P(S) \to R$ by $P_f(E) = Q^F_E(f)/Q^F_S(f)$, and we will show that P_f is a distribution on F.

P_f is well-defined because $Q^F_S(f)$ is a nonempty sum of products of positive real numbers, and thus positive. Further, since $Q^F_E(f)$ is a sum of products of positive real numbers, $P_f(E) \geq 0$ for all $E \subseteq S$. Since $Q^F_{\{\}} = 0$, we also have $P_f(\{\}) = 0$. Clearly $P_f(S) = 1$. Finally, for all $E_0, E_1 \subseteq S$ with $E_0 \cap E_1 = \{\}$, we have

$$\begin{aligned} P_f(E_0 \cup E_1) &= Q^F_{E_0 \cup E_1}(f) / Q^F_S(f) \\ &= (Q^F_{E_0}(f) + Q^F_{E_1}(f)) / Q^F_S(f) \\ &= P_f(E_0) + P_f(E_1). \end{aligned}$$

Therefore P_f is a distribution on S. We still need to show that P_f is a distribution on F.

Observe that for all $s \in S$ and $b \in B$, since $\overset{F}{\chi}_{\{b\}}([s]_b, S) = [s]_b$, we have that $\overset{F}{Q}_{[s]_b}(f) = \overset{F}{\text{poly}}_{\{b\}}([s]_b) \cdot \overset{F}{\text{poly}}_{B \setminus \{b\}}(S)$, and since $\overset{F}{\chi}_{\{b\}}(S, S) = S$, we have that $\overset{F}{Q}_S(f) = \overset{F}{\text{poly}}_{\{b\}}(S) \cdot \overset{F}{\text{poly}}_{B \setminus \{b\}}(S)$. Thus, we have that

$$\begin{aligned} P_f([s]_b) &= \overset{F}{\text{poly}}_{\{b\}}([s]_b)(f) / \overset{F}{\text{poly}}_{\{b\}}(S)(f) \\ &= f([s]_b) / \overset{F}{\text{poly}}_{\{b\}}(S)(f). \end{aligned}$$

Thus, for all $s \in S$,

$$\begin{aligned} \prod_{b \in B} P_f([s]_b) &= \left(\prod_{b \in B} f([s]_b)\right) / \left(\prod_{b \in B} \overset{F}{\text{poly}}_{\{b\}}(S)(f)\right) \\ &= \overset{F}{Q}_{\{s\}}(f) / \overset{F}{Q}_S(f) \\ &= P_f(\{s\}). \end{aligned}$$

Thus P_f is a distribution on F.

It follows that $P_f(x \cap z) \cdot P_f(y \cap z) = P_f(x \cap y \cap z) \cdot P_f(z)$. We therefore have that

$$\begin{aligned} q(f) &= \overset{F}{Q}_{x \cap z}(f) \cdot \overset{F}{Q}_{y \cap z}(f) - \overset{F}{Q}_{x \cap y \cap z}(f) \cdot \overset{F}{Q}_z(f) \\ &= \left(P_f(x \cap z) \cdot P_f(y \cap z) - P_f(x \cap y \cap z) \cdot P_f(z)\right) \cdot \overset{F}{Q}_S(f)^2 \\ &= 0 \cdot \overset{F}{Q}_S(f)^2 \\ &= 0. \end{aligned}$$

Thus, q is a polynomial that is zero on an open subset of inputs, so q is the zero polynomial. Thus $\overset{F}{Q}_{x \cap z} \cdot \overset{F}{Q}_{y \cap z} - \overset{F}{Q}_z \cdot \overset{F}{Q}_{x \cap y \cap z} = 0$, so $\overset{F}{Q}_z \cdot \overset{F}{Q}_{x \cap y \cap z} = \overset{F}{Q}_{x \cap z} \cdot \overset{F}{Q}_{y \cap z}$. Since $x \in X$, $y \in Y$, and $z \in Z$ were arbitrary, by Lemma 3, we have $X \perp^F Y | Z$. $\square$

In the next two posts, we will introduce temporal inference using finite factored sets, and discuss future potential research directions.

Finite Factored Sets: Inferring Time

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

The fundamental theorem of finite factored sets tells us that (conditional) orthogonality data can be inferred from probabilistic data. Thus, if we can infer temporal data from orthogonality data, we will be able to combine these to infer temporal data purely from probabilistic data. In this section, we will discuss the problem of inferring temporal data from orthogonality data, mostly by going through a couple of examples.

6.1. Factored Set Models

We'll begin with a sample space, Ω.

Naively, one might except that temporal inference in this paradigm involves inferring a factorization of Ω. What we'll actually be doing, however, is inferring a factored set *model* of Ω. This will allow for the possibility that some situations are distinct without being distinct in Ω—that there can be latent structure not represented in Ω.

Definition 38 (model). *Given a set Ω, a model of Ω is a pair $M = (F, f)$, where F is a finite factored set and $f : \text{set}(F) \to \Omega$ is a function from the set of F to Ω.*

Definition 39. *Let S and Ω be sets, and let $f : S \to \Omega$ be a function from S to Ω.*

Given a $\omega \in \Omega$, we let $f^{-1}(\omega) = \{s \in S \mid f(s) = \omega\}$.

Given an $E \subseteq \Omega$, we let $f^{-1}(E) = \{s \in S \mid f(s) \in E\}$.

Given an $X \in \text{Part}(\Omega)$, we let $f^{-1}(X) \in \text{Part}(S)$ be given by

$f^{-1}(X) = \{f^{-1}(x) \mid x \in X, f^{-1}(x) \neq \{\}\}$.

Definition 40 (orthogonality database). *Given a set Ω, an orthogonality database on Ω is a pair $D = (O, N)$, where O and N are both subsets of $\text{Part}(\Omega) \times \text{Part}(\Omega) \times \text{Part}(\Omega)$.*

Definition 41. *Given an orthogonality database* $D = (O, N)$ *on a set* Ω*, and partitions* $X, Y, Z \in \mathrm{Part}(\Omega)$*, we write* $X \perp_D Y \mid Z$ *if* $(X, Y, Z) \in O$*, and we write* $X \rightleftharpoons_D Y \mid Z$ *if* $(X, Y, Z) \in N$*.*

Definition 42. *Given a set* Ω*, a model* $M = (F, f)$ *of* Ω*, and an orthogonality database* $D = (O, N)$ *on* Ω*, we say* M *models* D *if for all* $X, Y, Z \in \mathrm{Part}(\Omega)$*,*

1. *if* $X \perp_D Y \mid Z$ *then* $f^{-1}(X) \perp^F f^{-1}(Y) \mid f^{-1}(Z)$*, and*
2. *if* $X \rightleftharpoons_D Y \mid Z$ *then* $\neg(f^{-1}(X) \perp^F f^{-1}(Y) \mid f^{-1}(Z))$*.*

Definition 43. *An orthogonality database* D *on a set* Ω *is called consistent if there exists a model* M *of* Ω *such that* M *models* D*.*

Definition 44. *An orthogonality database* D *on a set* Ω *is called complete if for all* $X, Y, Z \in \mathrm{Part}(\Omega)$*, either* $X \perp_D Y \mid Z$ *or* $X \rightleftharpoons_D Y \mid Z$*.*

Definition 45. *Given a set* Ω*, an orthogonality database* D *on* Ω*, and* $X, Y \in \mathrm{Part}(\Omega)$*, we say* $X <_D Y$ *if for all models* (F, f) *of* Ω *that model* D*, we have* $f^{-1}(X) <^F f^{-1}(Y)$*.*

6.2. Examples

Example 1. *Let* $\Omega = \{00, 01, 10, 11\}$ *be the set of all bit strings of length* 2*. For* $i \in \{0, 1\}$*, let* $x_i = \{i0, i1\}$ *be the event that the first bit is* i*, and let* $y_i = \{0i, 1i\}$ *be the event that the second bit is i. Let* $X = \{x_0, x_1\}$ *and let* $Y = \{y_0, y_1\}$*.*

Let $v_0 = \{00, 11\}$ *be the event that the two bits are equal, let* $v_1 = \{01, 10\}$ *be the event that the two bits are unequal, and let* $V = \{v_0, v_1\}$*.*

Let $D = (O, N)$*, where* $O = \{(X, V, \{\Omega\})\}$ *and* $N = \{(V, V, \{\Omega\})\}$.

Proposition 33. *In Example 1,* D *is consistent.*

Proof. First observe that $F = (\Omega, \{X, V\})$ is a factored set, and so $M = (F, f)$ is a model of Ω, where f is the identity on Ω. It suffices to show that M models D.

Indeed $h^F(X) = \{X\}$, and $h^F(V) = \{V\}$, so $X \perp^F V$, so $f^{-1}(X) \perp^F f^{-1}(V) | f^{-1}(\{\Omega\})$.

Further, it is not the case that $V \perp^F V$, since $V \neq \text{Ind}_\Omega$. Thus it is not the case that $f^{-1}(V) \perp^F f^{-1}(V) | f^{-1}(\{\Omega\})$.

Thus M satisfies all of the conditions to model D, so D is consistent. □

Proposition 34. *In Example 1,* $X <_D Y$.

Proof. Let (F, f) be any model of Ω that models D. Let $F = (S, B)$. For any $A \in \text{Part}(\Omega)$, let $H_A = h^F(f^{-1}(A))$. Our goal is to show that H_X is a strict subset of H_Y.

First observe that $X \leq_\Omega Y \vee_\Omega V$, so for any $s, t \in S$, if $s \sim_{f^{-1}(Y)} t$ and $s \sim_{f^{-1}(V)} t$, then $f(s) \sim_Y f(t)$ and $f(s) \sim_V f(t)$, so $f(s) \sim_X f(t)$, so $s \sim_{f^{-1}(X)} t$. Thus $f^{-1}(X) \leq_S f^{-1}(Y) \vee_S f^{-1}(V)$.

It follows that $H_X \subseteq h^F(f^{-1}(Y) \vee_S f^{-1}(V)) = H_Y \cap H_V$. However, since $X \perp_D V | \{\Omega\}$, we have that $H_X \cap H_V = \{\}$, so $H_X \subseteq H_Y$.

By swapping X and V in the argument above, we also get that $H_V \subseteq H_Y$. Since $V \rightleftharpoons_D V | \{\Omega\}$, we have that $H_V \neq \{\}$. Thus H_V contains some element b. Observe that $b \notin H_X$, but $b \in H_Y$. Thus H_X is a strict subset of H_Y, so $f^{-1}(X) <^F f^{-1}(Y)$.

Since (F, f) was an arbitrary model of Ω that models D, this implies that $X <_D Y$. □

Example 2. *Let* $\Omega = \{ 000 , 001 , 010 , 011 , 100 , 101 , 110 , 111 \}$ *be the set of all bit strings of length* 3*. For* $i \in \{ 0 , 1 \}$*, let* $x_i = \{ i00 , i01 , i10 , i11 \}$ *be the event that the first bit is* i*, let* $y_i = \{ 0i0 , 0i1 , 1i0 , 1i1 \}$ *be the event that the second bit is* i*, and let* $z_i = \{ 00i , 01i , 10i , 11i \}$ *be the event that the third bit is* i*. Let* $X = \{ x_0 , x_1 \}$*, let* $Y = \{ y_0 , y_1 \}$*, and let* $Z = \{ z_0 , z_1 \}$.

Let $v_0 = \{ 000 , 001 , 110 , 111 \}$ *be the event that the first two bits are equal, let* $v_1 = \{ 010 , 011 , 100 , 101 \}$ *be the event that the first two bits are unequal, and let* $V = \{ v_0 , v_1 \}$.

Let $D = (O , N)$*, where* $O = \{ (X , V , \{ \Omega \}) , (X , Z , Y) , (V , Z , Y) \}$ *and* $N = \{ (X , Z , \{ \Omega \}) , (V , Z , \{ \Omega \}) , (Z , Z , Y) \}$.

Proposition 35. *In Example 2,* D *is consistent.*

Proof. Let $S = \Omega \cup \{00, 01, 10, 11\}$ be the set of all bit strings of length either 2 or 3.

For $i \in \{0, 1\}$, let $x_i' = \{i00, i01, i10, i11, i0, i1\}$ be the event that the first bit is i, and let $X' = \{x_0', x_1'\}$.

For $i \in \{0, 1\}$, let $y_i' = \{0i0, 0i1, 1i0, 1i1, 0i, 1i\}$ be the event that the second bit is i, and let $Y' = \{y_0', y_1'\}$.

Let $v_0' = \{000, 001, 110, 111, 00, 11\}$ be the event that the first two bits are equal, let $v_1' = \{010, 011, 100, 101, 01, 10\}$ be the event that the first two bits are unequal, and let $V' = \{v_0', v_1'\}$.

For $i \in \{0,1\}$, let $z_i' = \{00i, 01i, 10i, 11i\}$ be the event that the third bit exists and is i, let $z_2' = \{00, 01, 10, 11\}$ be the event that there are only two bits, and let $Z' = \{z_0', z_1', z_2'\}$.

Let $B = \{X', V', Z'\}$. Clearly, (S, B) is a finite factored set.

Let $f : S \to \Omega$ be given by $f(s) = s$ if $s \in \Omega$, $f(00) = 000$, $f(01) = 011$, $f(10) = 100$, and $f(11) = 111$, so f copies the last bit on inputs of length 2, and otherwise leaves the bit string alone. We will show that (F, f) models D.

First, observe that $f^{-1}(X) = X'$, $f^{-1}(Y) = Y'$, $f^{-1}(V) = V'$, and $f^{-1}(Z) = \{\{000, 010, 100, 110, 00, 10\}, \{001, 011, 101, 111, 01, 11\}\}$.

It is easy to verify that $h^F(X') = \{X'\}, h^F(V') = \{V'\}, h^F(Y') = \{X', V'\}$, and $h^F(f^{-1}(Z)) = B$. From this, we get that $X' \perp^F V'$ holds, but $X' \perp^F f^{-1}(Z)$ and $V' \perp^F f^{-1}(Z)$ do not hold.

Next, observe that for $i \in \{0,1\}$, $X'|_{y_i} = V'|_{y_i} = \{\{0i0, 0i1, 0i\}, \{1i0, 1i1, 1i\}\}$. It is easy to verify that $h^F(X'|_{y_i}) = h^F(V'|_{y_i}) = \{X', V'\}$.

Also, observe that $f^{-1}(Z)|_{y_0} = \{\{000, 100, 00, 10\}, \{001, 101\}\}$, and observe that $f^{-1}(Z)|_{y_1} = \{\{010, 110\}, \{011, 111, 01, 11\}\}$. It is easy to verify that $h^F(f^{-1}(Z)|_{y_0}) = h^F(f^{-1}(Z)|_{y_1}) = \{Z'\}$.

From this, we get that $X' \perp^F f^{-1}(Z)|Y'$ and $V' \perp^F f^{-1}(Z)|Y'$ hold, and $f^{-1}(Z) \perp^F f^{-1}(Z)|Y'$ does not hold.

Thus, (F, f) models D, so D is consistent. □

Proposition 36. *In Example 2,* $X <_D Y <_D Z$.

Proof. Let (F, f) be any model of Ω that models D. Let $F = (S, B)$. For any $A \in \text{Part}(\Omega)$, let $H_A = h^F(f^{-1}(A))$. Our goal is to show that H_X is a strict subset of H_Y and that H_Y is a strict subset of H_Z.

First observe that $X \leq_\Omega Y \vee_\Omega V$, so $f^{-1}(X) \leq_S f^{-1}(Y) \vee f^{-1}(V)$, so $H_X \subseteq H_Y \cup H_V$. Since $X \perp_D Y | \{\Omega\}$, $H_X \cap H_V = \{\}$, so $H_X \subseteq H_Y$. Symmetrically, $H_V \subseteq H_Y$, so $H_X \cup H_V \subseteq H_Y$.

Similarly, $Y \leq_\Omega X \vee_\Omega V$, so $H_Y \subseteq H_X \cup H_V$. Thus $H_Y = H_X \cup H_V$.

We also know that H_X and H_V are nonempty, because $X \rightleftharpoons_D Z | \{\Omega\}$ and $Y \rightleftharpoons_D Z | \{\Omega\}$.

Thus H_X is a strict subset of H_Y, so $X <_D Y$.

Let $C \subseteq B$ be arbitrary such that $H_X \cap C$ and $H_V \cap (B \setminus C)$ are both nonempty. Fix some $b_X \in H_X \cap C$ and $b_V \in H_V \cap (B \setminus C)$.

Since $b_X \in H_X$, there must exist $s_0, s_1 \in S$ such that $s_0 \sim_b s_1$ for all $b \in B \setminus \{b_X\}$, but not $s_0 \sim_{f^{-1}(X)} s_1$. Thus it is not the case that $f(s_0) \sim_X f(s_1)$. Without loss of generality, assume that $f(s_0) \in x_0$ and $f(s_1) \in x_1$.

Similarly, since $b_V \in H_V$, there must exist $t_0, t_1 \in S$ such that $t_0 \sim_b t_1$ for all $b \in B \setminus \{b_V\}$, but not $t_0 \sim_{f^{-1}(V)} t_1$. Again, without loss of generality, assume that $f(t_0) \in v_0$ and $f(t_1) \in v_1$.

For $i, j \in \{0, 1\}$, let $r_{ij} = \overset{F}{\chi}_{H_X}(s_i, t_j)$.

Next, observe that $r_{ij} \sim_{f^{-1}(X)} s_i$, so $f(r_{ij}) \sim_X f(s_i) \in x_i$, so $f(r_{ij}) \in x_i$. Similarly, $f(r_{ij}) \in v_j$, so $f(r_{ij}) \in x_i \cap v_j$. Thus, if $i = j, f(r_{ij}) \in y_0$, and if $i \neq j$, $f(r_{ij}) \in y_1$.

Further, observe that $\overset{F}{\chi_C}(r_{00}, r_{11}) = r_{01}$, since r_{00} and r_{11} agree on all factors other than b_X and b_V. In particular, this means that $\overset{F}{\chi_C}(f^{-1}(y_0), f^{-1}(y_0)) \neq f^{-1}(y_0)$. Similarly, since $\overset{F}{\chi_C}(r_{01}, r_{10}) = r_{00}$, we have that $\overset{F}{\chi_C}(f^{-1}(y_1), f^{-1}(y_1)) \neq f^{-1}(y_1)$.

We will use this to show that for any $y \in f^{-1}(Y)$ and $A \in \text{Part}(y)$, either $h^F(A) \cap H_Y = \{\}$, or $H_Y \subseteq h^F(A)$. This is because $h^F(A) \vdash^F A$, so $\overset{F}{\chi_{h^F(A)}}(y, y) = y$, so by the above argument, if $h^F(A) \cap H_X$ is nonempty, then $H_V \subseteq h^F(A)$, which since H_V is nonempty means $h^F(A) \cap H_V$ is nonempty, so $H_X \subseteq h^F(A)$, so $H_Y \subseteq h^F(A)$. Symmetrically, we also have that if $h^F(A) \cap H_V$ is nonempty, then $H_Y \subseteq h^F(A)$. Thus, if $h^F(A) \cap H_Y$ is nonempty, then either $h^F(A) \cap H_X$ or $h^F(A) \cap H_V$ is nonempty, so $H_Y \subseteq h^F(A)$.

Note that for any $y \in f^{-1}(Y)$, two of the elements among the four r_{ij} defined above are in y, and those two elements are in different parts in $f^{-1}(X)$, so $f^{-1}(X)|y$ has at least two parts, so $h^F(f^{-1}(X)|y)$ is nonempty. However, $h^F(f^{-1}(X)|y) \subseteq h^F(f^{-1}(X) \vee_S f^{-1}(Y)) = H_Y$. Thus, $h^F(f^{-1}(X)|y) \cap H_Y \neq \{\}$, so $H_Y \subseteq h^F(f^{-1}(X)|y)$, so $h^F(f^{-1}(X)|y) = H_Y$. Symmetrically, $h^F(f^{-1}(V)|y) = H_Y$.

In particular, this means that $h^F(f^{-1}(Z)|y) \cap H_Y = \{\}$, since $X \perp_D Z|Y$.

Since $X \rightleftharpoons_D Z|\{\Omega\}$, there exists some $b_Z \in H_X \cap H_Z$. Since $b_Z \in H_Z$, there exist $u_0, u_1 \in S$ such that $u_0 \sim_b u_1$ for all $b \in B \setminus \{b_Z\}$, but it is not the case that $u_0 \sim_{f^{-1}(Z)} u_1$. Without loss of generality, assume that $f(u_0) \in z_0$ and $f(u_1) \in z_1$. Let $y = [u_0]_{f^{-1}(Y)}$.

Let b_y be an arbitrary element of H_Y. Since $b_Y \in H_Y$, there exist $q_0, q_1 \in S$ such that $q_0 \sim_b q_1$ for all $b \in B \setminus \{b_Y\}$, but it is not the case that $q_0 \sim_{f^{-1}(Y)} q_1$. Without loss of generality, assume that $q_0 \in y$ and $q_1 \notin y$.

Consider $p_0 = \overset{F}{\chi}_{H_Y}(q_0, u_0) = \overset{F}{\chi}_{H_Y}(q_0, u_1)$. Since $q_0 \in y$, $p_0 \in y$. Since u_0 is also in y, $\overset{F}{\chi}_{h^F(f^{-1}(Z)|y)}(p_0, u_0) \sim_{f^{-1}(Z)} p_0$. However, since $h^F(f^{-1}(Z)|y) \cap H_Y = \{\}$, we have $\overset{F}{\chi}_{h^F(f^{-1}(Z)|y)}(p_0, u_0) = u_0$, so $u_0 \sim_{f^{-1}(Z)} p_0$.

If u_1 were in y, we would similarly have $u_1 \sim_{f^{-1}(Z)} p_0$, which would contradict the fact that it is not the case that $u_0 \sim_{f^{-1}(Z)} u_1$. Thus $u_1 \notin y$.

Next, consider $p_1 = \overset{F}{\chi}_{H_Y}(q_1, u_0) = \overset{F}{\chi}_{H_Y}(q_1, u_1)$. Since $q_1 \notin y$, $p_1 \notin y$. Since u_1 is also not in y, $\overset{F}{\chi}_{h^F(f^{-1}(Z)|(S \setminus y))}(p_1, u_1) \sim_{f^{-1}(Z)} p_1$. However, since $h^F(f^{-1}(Z)|(S \setminus y)) \cap H_Y = \{\}$, we have $\overset{F}{\chi}_{h^F(f^{-1}(Z)|(S \setminus y))}(p_1, u_1) = u_1$, so $u_1 \sim_{f^{-1}(Z)} p_1$.

Thus, it is not the case that $p_0 \sim_{f^{-1}(Z)} p_1$. However, we constructed p_0 and p_1 such that $p_0 \sim_b p_1$ for all $b \neq b_Y$. Thus $b_Y \in H_Z$. Since b_Y was arbitrary in H_Y, we have that $H_Y \subseteq H_Z$. Finally, we need to show that this subset relation is strict.

Since $Z \rightleftharpoons_D Z|Y$, there is some y such that $h^F(f^{-1}(Z)|y) \neq \{\}$. Let b be any element of $h^F(f^{-1}(Z)|y)$. Since $h^F(f^{-1}(Z)|y) \cap H_Y = \{\}$, $b \notin H_Y$. However, $b \in h^F(f^{-1}(Z)|y) \subseteq h^F(f^{-1}(Z) \vee_S f^{-1}(Y)) = h_Z \cup H_Y$. Therefore $b \in H_Z$. Thus H_Y is a strict subset of H_Z, so $Y <_D Z$. □

In the next post, we'll discuss applications and future research directions.

Finite Factored Sets: Applications

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

We will now discuss several different applications and directions for future work on finite factored sets. We will divide these research directions into three categories: 'Inference,' 'Infinity,' and 'Embedded Agency.'

This post will be much more speculative than the previous posts. It is very likely that some of these avenues for research will turn out to be dead ends, and some of the claims made here may not hold up to further investigation.

7.1. Inference

Decidability of Temporal Inference

In Finite Factored Sets: Inferring Time, we described a combinatorial problem of inferring temporal relations from an orthogonality database. However, it is not clear whether the question "Does a given temporal relation follow from a given orthogonality database?" is decidable.

One way we could hope to decide whether a temporal relation follows from some orthogonality database D over Ω would be to simply check all factored set models of Ω that model D up to a given size, and see whether the temporal relation always holds. For this to work, we would need an upper bound on the size of factored sets that we need to consider, as a function of the size of Ω. (Note that the existence of such a bound would not mean that there are no models larger than this upper bound. Rather, it would mean that every model larger than this will have all of the same temporal relations as some smaller model.)

Efficient Temporal Inference

Assuming temporal inference is computable, we would further like to be able to infer temporal relations from an orthogonality database *quickly*.

The naive way to get negative results in temporal inference (i.e., to show that certain temporal relations need not hold) would be to search over the space of models. Without the upper bound discussed above, however, this method would only ever yield negative results.

The naive way to get positive results would be to formalize the kind of reasoning used to prove Propositions 34 and 36, and search over proofs of this form. It is unclear whether this method can be made efficient.

Alternatively, we could hope to develop some new results and refine our understanding of temporal inference to the point where an alternative method can be made efficient.

Temporal Inference from Raw Data and Fewer Ontological Assumptions

In the Pearlian causal inference paradigm, we can infer temporal relationships from joint distributions on a collection of variables.

In Pearl's paradigm, however, this data is already factored into a collection of variables at the outset. Further, the Pearlian paradigm does not make explicit the assumptions that go into this factorization.

Our paradigm instead starts from a distribution on some set of observably distinct worlds. This approach allows us to make fewer ontological assumptions; we don't need to take for granted a particular way the world should be factored into variables. Thus, one might hope that the factored sets paradigm could be used to infer time or causality more directly from raw probabilistic data.

Causality, Determinism, and Abstraction

Another issue with the Pearlian causal inference paradigm is that it does not work well in cases where some of the variables are (partially) deterministic functions of each other. Our paradigm has determinism and abstraction built in, so it can be used to infer time in situations where the Pearlian paradigm might not apply.

Conceptual Inference

In Example 1, we can infer that $X <_D Y$. We can think of this fact as being about time. However, we can also think of it as being about which concepts are more natural or fundamental. In that example, X and V were more primitive variables, while Y was a more derived variable that was computed from X and V.

Suppose we had a symbol that was either 0 or 1, chosen according to some probability, and was also colored either blue or green, chosen independently according to some other probability. We can reason about this symbol using concepts like color or number. Alternatively, we could define a new concept *bleen* meaning "the symbol is either blue and 0, or green and 1," and *grue*, meaning "the symbol is either green and 0, or blue and 1," and use these two concepts instead.

We want to say that color and number are in some sense better or more useful concepts, while bleen and grue are less useful. Finite factored sets help give formal content to the idea that color and number are more primitive, while bleen and grue

are more derived; and this primitiveness seems to point at part of what it means to be a good concept for the purpose of thinking about the world.

Inferring Time without Orthogonality

In this sequence, we have focused on inferring time from an orthogonality database. Such a database may have been inferred in turn from observed independence and dependence facts drawn from a probability distribution.

We could instead consider inferring time directly from a probability distribution. Cutting out the orthogonality database in this way could even allow us to infer time from a probability distribution that has no nontrivial conditional independencies at all.

To see why it might be possible to infer time without any orthogonality, consider a set Ω, and a model of Ω, (F, f), where $F = (S, B)$ has n binary factors, and $|\Omega| > n + 1$.

There are n degrees of freedom in an arbitrary probability distribution on F, and thus at most n degrees of freedom in a probability distribution P on Ω that comes from a probability distribution on F. However, there are $|\Omega - 1|$ degrees of freedom in an arbitrary distribution on Ω.

As such, the probability distribution on Ω will lie on some surface without full dimension in the space of probability distributions on Ω, which could be used to infer some of the properties of F.

However, if $|\Omega|$ is much smaller than $|S|$, and f is chosen at random, it is unlikely that there will be any conditional orthogonality relations on partitions of Ω at all (other than the trivial conditional orthogonality relations that come from one partition being finer than another).

Inferring Conditioned Finite Factored Sets

If we modify the temporal inference definition to instead allow for f to be a partial function from S to Ω, we get a new, weaker model of temporal inference. This can be thought of as allowing for the possibility that our distribution on Ω passes through some filter that only shows us some of the observably distinct worlds.

7.2. Infinity

The Fundamental Theorem of Finitely Generated Factored Sets

Throughout this sequence, we have assumed finiteness fairly gratuitously. It is likely that many of the results can be extended to arbitrary finite sets. However, this generalization will not be immediate. Indeed, even history is not well-defined on arbitrary factored sets.

One intermediate possibility is to consider finite-dimensional factored sets. In this case, history would be well-defined, but our proof of the fundamental theorem would not directly generalize. However, we conjecture that the finite-dimensional analogue of the fundamental theorem would in fact hold.

Conjecture 1. *Theorem 3 can be generalized to finite-dimensional factored sets.*

On the other hand, we do not expect the fundamental theorem to generalize to arbitrary factored sets. To see why, consider the following example.

Example 3. *Let* $F = (S, B)$*, where* $S = P(N)$*,*

$b_n = \{\{s \in S \mid n \in s\}, \{s \in S \mid n \notin s\}\}$*, and* $B = \{b_n \mid n \in N\}$*.*

Let $X = \{\{\{\}\}, S \setminus \{\{\}\}\}$*, and let* $Y = \{\{N\}, S \setminus \{N\}\}$*.*

In this example, it seems that in the correct generalization of orthogonality to arbitrary factored sets, we likely want to say that X is not orthogonal to Y. However, it also seems like we want to say that in every distribution on F, at least one of {{}} and {N} has probability zero, so this should give a counterexample to the fundamental theorem.

Even without the fundamental theorem, we believe that orthogonality and time in arbitrary-dimensional factored sets will be important and interesting.

Orthogonality and Time in Arbitrary Factored Sets

In the infinite-dimensional case, it is not even clear how we should define orthogonality, time, and conditional orthogonality. There are three main contenders.

First, we could say that (sub)partitions X and Y are orthogonal if there exist disjoint $C_X, C_Y \subseteq B$ such that $C_X \vdash^F X$ and $C_Y \vdash^F Y$. We could then define time as a closure property on orthogonality.

Second, we could just define the history of a (sub)partition X to be the intersection of all $C \subseteq B$ such that $C \vdash^F X$, and leave the definitions of orthogonality and time alone. This second option has some unintuitive behavior. Consider the following example.

Example 4. *Let* $F = (S, B)$*, where* $S = P(N)$*,*

$b_n = \{\{s \in S \mid n \in s\}, \{s \in S \mid n \notin s\}\}$*, and* $B = \{b_n \mid n \in N\}$*.*

Let $Z = \{\{s \in S \mid |s| < \infty\}, \{s \in S \mid |s| = \infty\}\}$*.*

In this example, Z is orthogonal to itself according to the second option, in spite of having more than one part. However, it is possible that this is a feature, rather than a bug, since it seems to interact nicely with Kolmogorov's zero–one law.

Third, we could define a way to flatten factored sets by merging some of the factors into their common refinement, and we could say X and Y are orthogonal given Z in F if X and Y are orthogonal given Z in some finite-dimensional flattening of F.

The main difference between the first and third options comes from the case where Z has infinitely many parts. In the third option, we must fix a single finite-dimensional flattening such that $X|z$ and $Y|z$ have disjoint histories for all $z \in Z$.

We are most optimistic about the third option, because we conjecture that it can satisfy the compositional semigraphoid axioms, while the other two options cannot. It is also possible that other options give the compositional semigraphoid axioms for partitions with finitely many parts, but not general partitions.

Continuity and Physics

A major reason why we are interested in exploring arbitrary-dimensional factored sets is because it could allow us to talk about continuous time.

The Pearlian paradigm takes advantage of the parenthood relationship between nodes to make inferences. E.g., the nodes are thought of as probabilistic functions of their parents, and the existence of edges between nodes is a central part of temporal inference.

In the factored set paradigm, there is no mention of parenthood; instead, $\leq^F$ is both reflexive and transitive, and so can be thought of as an ancestry relation. Further, by working with arbitrary partitions rather than a fixed collection of variables, we allow for "zooming in" on our variables.

These two properties together suggest that the factored set paradigm is much closer to being able to talk about continuous time, if the theory can be extended naturally to

infinite dimensions.

As pointed out by Eliezer Yudkowsky in his blog post "Causal Universes,'' physics looks an awful lot like a continuous analogue of Pearlian causal diagrams. We are thus hopeful that when extended to arbitrary dimensions, factored sets could provide a useful new way of looking at physics.

7.3. Embedded Agency

Embedded Observations

We can use finite factored sets to build a new way of thinking about observations.

Definition 46 (observes an event). *Let* $F = (S, B)$ *be a finite factored set. Let* A *and* W *be partitions of* S*, and let* E *be a subset of* S*. Let* X_E *be the partition of* S *given by* $X_E = \{S\}$ *if* $E = \{\}$ *or* $E = S$*, and* $X_E = \{E, S \setminus E\}$ *otherwise. We say* A *observes* E *with respect to* W *(in* F*) if the following two conditions hold.*

1. $A \perp^F X_E$.
2. $A \perp^F W \mid S \setminus E$.

A can be thought of as an agent, with the different parts in A representing options available to A. E represents some fact about the world. W can be thought of as some high-level world model. We will especially think of W as a model that captures all of the information about the world that the agent cares about.

When we say that A observes E, this does not necessarily mean that E holds. Rather, we are saying that A can safely assume that E holds. A can safely make this assumption if it is the case that A's choice can't affect whether E holds, and if, when E does not hold, A's choice can have no effect on any part of the world that A cares about. This is exactly what is represented by the two conditions in Definition 46.

In Drescher's transparent Newcomb thought experiment, the agent cannot be said to observe the contents of the box, because the first condition in Definition 46 is violated. In Nesov's counterfactual mugging thought experiment, the agent cannot be said to observe the result of the coin flip, because the second condition is violated.

We can extend this definition to give a notion of an agent observing a partition rather than an event.

Definition 47 (observes a partition). *Let* $F = (S, B)$ *be a finite factored set. Let* A*,* W*, and* X *be partitions of* S*. Let* $X = \{x_0, \dots, x_{n-1}\}$*. We say* A *observes* X *with respect to* W *(in* F*) if* $A \perp^F X$ *and there exist partitions of* S*,* A_i *for* $i \in \{0, \dots, n-1\}$ *such that*

1. $A = \bigvee_S(\{A_i \mid i \in \{0, \dots, n-1\}\})$.
2. $A_i \perp^F W \mid S \setminus x_i$.

Saying that A observes X is roughly saying that A can be divided into subagents, where each subagent observes a different part in X.

Counterfactability

The factored set paradigm also has some interesting things to say about counterfactuals. The chimera functions can be thought of representing a way of taking counterfactuals.

Given a finite factored set $F = (S, B)$, $C \subseteq B$, and $s, t \in S$, let $X_C = \bigvee_S(C)$.

We can think of $\chi_C^F(s, t)$ as the result of starting with t, then performing a counterfactual surgery that changes the value of X_C to match its value in s.

Unfortunately, while we can tell this story for X_C, we cannot tell the same story for an arbitrary partition of S.

Definition 48 (counterfactability). *Given a finite factored set* $F = (S, B)$*, a partition* $X \in \text{Part}(S)$ *is called counterfactable (in* F*) if* $X = \bigvee_S(h^F(X))$*.*

When a partition X is counterfactable, the chimera function gives a well-defined way to start with an element of S, and change it by changing what part in X it is in.

Being counterfactable is rather strong, but we have a weaker notion of relative counterfactability.

Definition 49 (relative counterfactability). *Given a finite factored set* $F = (S, B)$*, a partition* $X \in \text{Part}(S)$ *is called counterfactable relative to another partition* $W \in \text{Part}(S)$ *(in* F*) if* $\bigvee_S(h^F(X)) \perp^F W|X$*.*

X is counterfactable relative to W if X screens off the history of X from W. This means that if we want to counterfact on the value of X, we can safely counterfact on the finer partition $\bigvee_S(h^F(X))$. As long as we only care about what part in W the result is in, choices about which subpart in $\bigvee_S(h^F(X))$ to counterfact will not matter, so we can think of counterfacting on the value of X as well-defined up to the partition W.

This notion of counterfactability explains why counterfactuals sometimes seem clear, and other times they do not seem well-defined. In the factored set ontology, sometimes partitions are not counterfactable because they are not fine enough to fully specify all the effects of the counterfactual.

Cartesian Frames

The factored set paradigm can be seen as capturing many of the benefits of the Cartesian frame paradigm. We have already seen this in part in our discussion of embedded observations. We feel that the factored set paradigm successfully captures a meaningful notion of time, while the Cartesian frame paradigm mostly fails at this goal.

The connection between factored sets and Cartesian frames is rather strong. For example, a 2-dimensional factored set model of a set W is in effect a Cartesian frame over W. The only difference is that the factored set model forgets which factor is the agent, and which factor is the environment. When one Cartesian frame over W is a multiplicative subagent of another, we can construct a 3-dimensional factored set model of W, with the subagent represented by one of the factors, and the superagent represented by a pair of the factors.

Unraveling Causal Loops

Whenever an agent makes a decision, there is a temptation to think of the effects of the decision as causally "before" the decision being made. This is because the agent uses its model of the effects as an input when making the decision. This causes a problem, because the effects of the decision can of course also be seen as causally after the decision being made.

On our view, part of what is going on is that there is a distinction between the agent's model of the effects, and the effects themselves. The problem is that the agent's model of the effects is highly entangled with the actual effects, which is why we feel tempted to combine them in the first place.

One way to model this situation is by thinking of the agent's model of the effects as being a coarser version of the actual world state after the decision. It is thus possible for the model of the effects to be before the decision, which is before the effects themselves.

By allowing for some variables to be coarsenings or refinements of other variables, the factored set paradigm possibly gives us the tools to be able to straighten out these causal loops.

Conditional Time

We can define conditional time similarly to how we define conditional orthogonality.

Definition 50 (conditional time). *Given a finite factored set* $F = (S, B)$*, partitions* $X, Y \in \text{Part}(S)$*, and* $E \subseteq S$*, we say that* X *is before* Y *given* E *(in* F*), written* $X \leq^F Y \mid E$*, if* $h^F(X|E) \subseteq h^F(Y|E)$*.*

It is not clear if this notion has any important philosophical meaning, but it seems plausible that it does. In particular, this notion could be useful for reasoning about situations where time appears to flow in multiple directions at different levels of description, or under different assumptions. Incorporating conditional time could then be used to flatten some causal loops.

Logical Causality

Upon discovering logical induction, one of the first things we considered was the possibility of inferring logical causality using our probabilities on logical sentences. We considered doing this using the Pearlian paradigm, but it now seems like that approach was doomed to fail, because we had many deterministic relationships between our variables.

The factored set paradigm seems much closer to allowing us to correctly infer logical causality from logical probabilities, but it is still far from ready.

One major obstacle is that the factored set paradigm does not have a reasonable way to think about the uniform distribution on a four-element set. The independence structure of the uniform distribution on a four-element set is not a compositional semigraphoid, because if we take X, Y, and Z to be the three partitions that partition

the four-element set into two parts of size two, then X is independent of Y and of Z, but not independent of the common refinement of Y and Z.

Since the uniform distribution on a four-element set will likely (approximately) show up many times in logical induction, it is not clear how to do the causal inference.

Orthogonality as Simplifying Assumptions for Decisions

While we largely have been thinking of orthogonality as a property of the world, one could also think of orthogonality as something that an agent assumes to make decisions.

For example, when an agent is looking at a coin that came up heads, the agent might make the assumption that its decision has no effect on the worlds in which the coin came up tails. This assumption might only be approximately true, but part of being an embedded agent is working with approximations. Orthogonality seems like a useful language for some of the simplifying assumptions agents might make.

Conditional Orthogonality and Abstractions

Given some complicated structure X, one might want to know when a simpler structure Y is a good abstraction for X. One desirable property of an abstraction is that Y screens off X from all of the properties of the world that an agent cares about, W. In this way, by thinking in terms of Y, the agent does not risk missing any important information.

We could also consider weaker notions than this, by taking W to just be that which the agent cares about within a certain context in which the agent is using the abstraction.

This is all very vague and rough, but the point is that conditional orthogonality seems related to what makes a good abstraction, so being able to talk about conditional orthogonality and abstractions together seems like it could prove useful.

Saving Time

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

For the last few years, a large part of my research motivation has been directed at trying to save the concept of time—save it, for example, from all the weird causal loops created by decision theory problems. This post will hopefully explain why I care so much about time, and what I think needs to be fixed.

Why Time?

My best attempt at a short description of time is that **time is causality**. For example, in a Pearlian Bayes net, you draw edges from earlier nodes to later nodes. To the extent that we want to think about causality, then, we will need to understand time.

Importantly, **time is the substrate in which learning and commitments take place**. When agents learn, they learn over time. The passage of time is like a ritual in which opportunities are destroyed and knowledge is created. And I think that many models of learning are subtly confused, because they are based on confused notions of time.

Time is also crucial for thinking about agency. My best short-phrase definition of agency is that **agency is time travel**. An agent is a mechanism through which the future is able to affect the past. An agent models the future consequences of its actions, and chooses actions on the basis of those consequences. In that sense, the consequence *causes* the action, in spite of the fact that the action comes earlier in the standard physical sense.

Problem: Time is Loopy

The main thing going wrong with time is that it is “loopy.”

The primary confusing thing about Newcomb's problem is that we want to think of our decision as coming “before” the filling of the boxes, in spite of the fact that it physically comes after. This is hinting that maybe we want to understand some other "logical" time in addition to the time of physics.

However, when we attempt to do this, we run into two problems: Firstly, we don't understand where this logical time might come from, or how to learn it, and secondly, we run into some apparent temporal loops.

I am going to set aside the first problem and focus on the second.

The easiest way to see why we run into temporal loops is to notice that it seems like physical time is at least a little bit entangled with logical time.

Imagine the point of view of someone running a physics simulation of Newcomb's problem, and tracking all of the details of all of the atoms. From that point of view, it seems like there is a useful sense in which the filling of the boxes comes before an agent's decision to one-box or two-box. At the same time, however, those atoms compose an agent that shouldn't make decisions as though it were helpless to change anything.

Maybe the solution here is to think of there being many different types of "before" and "after," "cause" and "effect," etc. For example, we could say that X is before Y from an agent-first perspective, but Y is before X from a physics-first perspective.

I think this is right, and we want to think of there as being many different systems of time (hopefully predictably interconnected). But I don't think this resolves the whole problem.

Consider a pair of FairBot agents that successfully execute a Löbian handshake to cooperate in an open-source prisoner's dilemma. I want to say that each agent's cooperation causes the other agent's cooperation in some sense. I could say that relative to each agent the causal/temporal ordering goes a different way, but I think the loop is an important part of the structure in this case. (I also am not even sure which direction of time I would want to associate with which agent.)

We also are tempted to put loops in our time/causality for other reasons. For example, when modeling a feedback loop in a system that persists over time, we might draw structures that look a lot like a Bayes net, but are not acyclic (e.g., a POMDP). We could think of this as a projection of another system that has an extra dimension of time, but it is a useful projection nonetheless.

Solution: Abstraction

My main hope for recovering a coherent notion of time and unraveling these temporal loops is via abstraction.

In the example where the agent chooses actions based on their consequences, I think that there is an abstract model of the consequences that comes causally before the choice of action, which comes before the actual physical consequences.

In Newcomb's problem, I want to say that there is an abstract model of the action that comes causally before the filling of the boxes.

In the open source prisoners' dilemma, I want to say that there is an abstract proof of cooperation that comes causally before the actual program traces of the agents.

All of this is pointing in the same direction: We need to have coarse abstract versions of structures come at a different time than more refined versions of the same structure. Maybe when we correctly allow for different levels of description having different links in the causal chain, we can unravel all of the time loops.

But How?

Unfortunately, our best understanding of time is Pearlian causality, and Pearlian causality does not do great with abstraction.

Pearl has Bayes nets with a bunch of variables, but when some of those variables are coarse abstract versions of other variables, then we have to allow for determinism, since some of our variables will be deterministic functions of each other; and the best parts of Pearl do not do well with determinism.

But the problem runs deeper than that. If we draw an arrow in the direction of the deterministic function, we will be drawing an arrow of time from the more refined version of the structure to the coarser version of that structure, which is in the opposite direction of all of our examples.

Maybe we could avoid drawing this arrow from the more refined node to the coarser node, and instead have a path from the coarser node to the refined node. But then we could just make another copy of the coarser node that is deterministically downstream of the more refined node, adding no new degrees of freedom. What is then stopping us from swapping the two copies of the coarser node?

Overall, it seems to me that Pearl is not ready for some of the nodes to be abstract versions of other nodes, which I think needs to be fixed in order to save time.

[AN #163]: Using finite factored sets for causal and temporal inference

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Alignment Newsletter is a weekly publication with recent content relevant to AI alignment around the world. Find all Alignment Newsletter **resources here**. In particular, you can look through **this spreadsheet** of all summaries that have ever been in the newsletter.

Audio version **here** (may not be up yet).

Please note that while I work at DeepMind, this newsletter represents my personal views and not those of my employer. This newsletter is a combined summary + opinion for the **Finite Factored Sets sequence** by Scott Garrabrant. I (Rohin) have taken a lot more liberty than I usually do with the interpretation of the results; Scott may or may not agree with these interpretations.

Motivation

One view on the importance of deep learning is that it allows you to automatically *learn* the features that are relevant for some task of interest. Instead of having to handcraft features using domain knowledge, we simply point a neural net at an appropriate dataset and it figures out the right features. Arguably this is the *majority* of what makes up intelligent cognition; in humans it seems very analogous to **System 1**, which we use for most decisions and actions. We are also able to infer causal relations between the resulting features.

Unfortunately, **existing models** of causal inference don't model these learned features -- they instead assume that the features are already given to you. Finite Factored Sets (FFS) provide a theory which can talk directly about different possible ways to featurize the space of outcomes and still allows you to perform causal inference. This sequence develops this underlying theory and demonstrates a few examples of using finite factored sets to perform causal inference given only observational data.

Another application is to **embedded agency** (**AN #31**): we would like to think of "agency" as a way to featurize the world into an "agent" feature and an "environment" feature, that together interact to determine the world. In **Cartesian Frames** (**AN #127**), we worked with a function $A \times E \to W$, where pairs of (agent, environment) together determined the world. In the finite factored set regime, we'll think of A and E as features, the space $S = A \times E$ as the set of possible feature vectors, and $S \to W$ as the mapping from feature vectors to actual world states.

What is a finite factored set

Generalizing this idea to apply more broadly, we will assume that there is a set of possible worlds Ω, a set S of arbitrary elements (which we will eventually interpret as

feature vectors), and a function f : S → Ω that maps feature vectors to world states. Our goal is to have some notion of "features" of elements of S. Normally, when working with sets, we identify a feature value with the set of elements that have that value. For example, we can identify "red" as the set of all red objects, and in **some versions of mathematics**, we define "2" to be the class of all sets that have exactly two elements. So, we define a feature to be a *partition* of S into subsets, where each subset corresponds to one of the possible feature values. We can also interpret a feature as a *question* about items in S, and the values as possible *answers* to that question; I'll be using that terminology going forward.

A finite factored set is then given by (S, B), where B is a set of **factors** (questions), such that if you choose a particular answer to every question, that uniquely determines an element in S (and vice versa). We'll put aside the set of possible worlds Ω; for now we're just going to focus on the theory of these (S, B) pairs.

Let's look at a contrived example. Consider S = {chai, caesar salad, lasagna, lava cake, sprite, strawberry sorbet}. Here are some possible questions for this S:

- **FoodType**: Possible answers are Drink = {chai, sprite}, Dessert = {lava cake, strawberry sorbet}, Savory = {caesar salad, lasagna}
- **Temperature**: Possible answers are Hot = {chai, lava cake, lasagna} and Cold = {sprite, strawberry sorbet, caesar salad}.
- **StartingLetter**: Possible answers are "C" = {chai, caesar salad}, "L" = {lasagna, lava cake}, and "S" = {sprite, strawberry sorbet}.
- **NumberOfWords**: Possible answers are "1" = {chai, lasagna, sprite} and "2" = {caesar salad, lava cake, strawberry sorbet}.

Given these questions, we could factor S into {FoodType, Temperature}, or {StartingLetter, NumberOfWords}. We *cannot* factor it into, say, {StartingLetter, Temperature}, because if we set StartingLetter = L and Temperature = Hot, that does not uniquely determine an element in S (it could be either lava cake or lasagna).

Which of the two factorizations should we use? We're not going to delve too deeply into this question, but you could imagine that if you were interested in questions like "does this need to be put in a glass" you might be more interested in the {FoodType, Temperature} factorization.

Just to appreciate the castle of abstractions we've built, here's the finite factored set F with the factorization {FoodType, Temperature}:

F = ({chai, caesar salad, lasagna, lava cake, sprite, strawberry sorbet}, {{{chai, sprite}, {lava cake, strawberry sorbet}, {caesar salad, lasagna}}, {{chai, lava cake, lasagna}, {sprite, strawberry sorbet, caesar salad}}})

To keep it all straight, just remember: a **factorization** B is a set of **questions** (factors, partitions) each of which is a set of **possible answers** (parts), each of which is a set of elements in S.

A brief interlude

Some objections you might have about stuff we've talked about so far:

Q. Why do we bother with the set S -- couldn't we just have the set of questions B, and then talk about answer vectors of the form (a1, a2, … aN)?

A. You could in theory do this, as there is a bijection between S and the Cartesian product of the sets in B. However, the problem with this framing is that it is hard to talk about other derived features. For example, the question “what is the value of B1+B2” has no easy description in this framing. When we instead directly work with S, the B1+B2 question is just another partition of S, just like B1 or B2 individually.

Q. Why does f map S to Ω? Doesn’t this mean that a feature vector uniquely determines a world state, whereas it’s usually the opposite in machine learning?

A. This is true, but here the idea is that the set of features together captures *all* the information within the setting we are considering. You could think of feature vectors in deep learning as only capturing an important subset of all of the features (which we’d have to do in practice since we only have bounded computation), and those features are not enough to determine world states.

Orthogonality in Finite Factored Sets

We’re eventually going to use finite factored sets similarly to Pearlian causal models: to infer which questions (random variables) are conditionally independent of each other. However, our analysis will apply to arbitrary questions, unlike Pearlian models, which can only talk about independence between the predefined variables from which the causal model is built.

Just like Pearl, we will talk about *conditioning on evidence*: given evidence e, a subset of S, we can “observe” that we are within e. In the formal setup, this looks like erasing all elements that are not in e from all questions, answers, factors, etc.

You might think that "factors" are not analogous to nodes or random variables in a Pearlian model. However, this isn't right, since we’re going to assume that all of our factors are *independent* from each other, which is usually not the case in a Pearlian model. For example, you might have a Pearlian model with two binary variables, e.g. “Variable Rain causes Variable Wet Sidewalk”; these are obviously not independent. The corresponding finite factored set would have *three* factors: “did it rain?”, “if it rained did the sidewalk get wet?” and “if it didn’t rain did the sidewalk get wet?” This way all three factors can be independent of each other. We will still be able to ask whether Wet Sidewalk is independent of Rain, since Wet Sidewalk is just another question about the set S -- it just isn’t one of the underlying factors anymore.

The point of this independence is to allow us to reason about *counterfactuals*: it should be possible to say “imagine the element s, except with underlying factor b2 changed to have value v”. As a result, our definitions will include clauses that say “and make sure we can still take counterfactuals”. For example, let’s talk about the “history” of a question X, which for now you can think of as the “factors relevant to X”. The *history* of X given e is the smallest set of factors such that:

1) if you know the answers to these factors, then you can infer the answer to X, and

2) any factors that are *not* in the history are independent of X. As suggested above, we can think of this as being about counterfactuals -- we’re saying that for any such factor, we can counterfactually change its answer and this will remain consistent with the evidence e.

(A technicality on the second point: we'll never be able to counterfactually change a factor to a value that is never found in the evidence; this is fine and doesn't prevent things from being independent.)

Time for an example! Consider the set S = {000, 001, 010, 011, 100, 101, 110, 111} and the factorization {X, Y, Z}, where X is the question "what is the first bit", Y is the question "what is the second bit", and Z is the question "what is the third bit". Consider the question Q = "when interpreted as a binary number, is the number >= 2?" In this case, the history of Q given no evidence is {X, Y} because you can determine the answer to Q with the combination of X and Y. (You can still counterfact on anything, since there is no evidence to be inconsistent with.)

Let's consider an example with evidence. Suppose we observe that all the bits are equal, that is, e = {000, 111}. Now, what is the history of X? If there wasn't any evidence, the history would just be {X}; you only need to know X in order to determine the value of X. However, suppose we learned that X = 0, implying that our element is 000. We can't counterfact on Y or Z, since that would produce 010 or 001, both of which are inconsistent with the evidence. So given this evidence, the history of X is actually {X, Y, Z}, i.e. the entire set of factors! If we'd only observed that the first two bits were equal, so e = {000, 001, 110, 111}, then we *could* counterfact on Z and the history of X would be {X, Y}.

(Should you want more examples, here are two **relevant** **posts**.)

Given this notion of "history", it is easy to define orthogonality: X is orthogonal to Y given evidence e if the history of X given e has no overlap with the history of Y given e. Intuitively, this means that the factors relevant to X are completely separate from those relevant to Y, and so there cannot be any entanglement between X and Y. For a *question* Z, we say that X is orthogonal to Y given Z if X is orthogonal to Y given z, for every possible answer z in Z.

Now that we have defined orthogonality, we can state the *Fundamental Theorem of Finite Factored Sets*. Given some questions X, Y, and Z about a finite factored set F, X is orthogonal to Y given Z if and only if in every probability distribution on F, X is conditionally independent of Y given Z, that is, P(X, Y | Z) = P(X | Z) * P(Y | Z).

(I haven't told you how you put a probability distribution on F. It's exactly what you would think -- you assign a probability to every possible answer in every factor, and then the probability of an individual element is defined to be the product of the probabilities of its answers across all the factors.)

(I also haven't given you any intuition about why this theorem holds. Unfortunately I don't have great intuition for this; the proof has multiple non-trivial steps, each of which I locally understand and have intuition for... but globally it's just a sequence of non-trivial steps to me. Here's an attempt, which isn't very good: we specifically defined orthogonality to capture *all* the relevant information for a question, in particular by having that second condition requiring that we be able to counterfact on other factors, and so it intuitively makes sense that if the relevant information doesn't overlap, then there can't be a way for the probability distribution to have interactions between the variables.)

The fundamental theorem is in some sense a *justification* for calling the property "orthogonality" -- if we determine just by studying the structure of the finite factored set that X is orthogonal to Y given Z, then we know that this implies conditional independence in the "true" probability distribution, whatever it ends up being.

Pearlian models have a similar theorem, where the graphical property of d-separation implies conditional independence.

Foundations of causality and time

You might be wondering why we have been calling the minimal set of relevant factors "history". The core philosophical idea is that, if you have the right factorization, then "time" or "causality" can be thought of as flowing in the direction of larger histories. Specifically, we say that X is "before" Y if the history of X is a subset of the history of Y. (We then call it "history" because every factor in the history of X will be "before" X by this definition.)

One intuition pump for this is that in physics, if an event A causes an event B, then the past light cone of A is a subset of the past light cone of B, and A happens before B in every possible reference frame.

But perhaps the best argument for thinking of this as causality is that we can actually use this notion of "time" or "causality" to perform causal inference. Before I talk about that, let's see what this looks like in Pearlian models.

Strictly speaking, in Pearlian models, the edges do not *have* to correspond to causality: formally they only represent conditional independence assumptions on a probability distribution. However, consider the following Cool Fact: for some Pearlian models, if you have observational data that is generated from that model, you can recover the exact graphical structure of the generating model just by looking at the observational data. In this case, you really are inferring cause-and-effect relationships from observational data! (In the general case where the data is generated by an arbitrary model, you can recover a lot of the structure of the model but be uncertain about the direction of some of the edges, so you are still doing *some* causal inference from observational data.)

We will do something similar: we'll use our notion of "before" to perform causal inference given observational data.

Temporal inference: the three dependent bits

You are given statistical (i.e. observational) data for three bits: X, Y and Z. You quickly notice that it is always the case that Z = X xor Y (which implies that X = Y xor Z, and Y = Z xor X). Clearly, there are only two independent bits here and the other bit is derived as the xor of the two independent bits. From the raw statistical data, can you tell which bits are the independent ones, and which one is the derived one, thus inferring which one was *caused* by the other two? It turns out that you can!

Specifically, you want to look for which two bits are *orthogonal* to each other, that is, you want to check whether we approximately have P(X, Y) = P(X) P(Y) (and similarly for other possible pairings). In the world where two of the bits were generated by a biased coin, you will find exactly one pair that is orthogonal in this way. (The case where the bits are generated by a fair coin is special; the argument won't work there, but it's in some sense "accidental" and happens because the probability of 0.5 is very special.)

Let’s suppose that the orthogonal pair was (X, Z). In this case, we can *prove* that in *every* finite factored set that models this situation, X and Z come “before” Y, i.e. their histories are strict subsets of Y’s history. Thus, we’ve inferred causality using only observational data! (And unlike with Pearlian models, we did this in a case where one “variable” was a deterministic function of two other “variables”, which is a type of situation that Pearlian models struggle to handle.)

Future work

Remember that motivation section, a couple thousand words ago? We talked about how we can do causal inference with learned featurizations and apply it to embedded agency. Well, we actually haven’t done that yet, beyond a few examples of causal inference (as in the example above). There is a lot of future work to be done in applying it to the case that motivated it in the first place. The author wrote up potential future work **here**, which has categories for both causal inference and embedded agency, and also adds a third one: generalizing the theory to infinite sets. If you are interested in this framework, there are many avenues for pushing it forward.

FEEDBACK

I'm always happy to hear feedback; you can send it to me, **Rohin Shah**, by **replying to this email**.

PODCAST

An audio podcast version of the **Alignment Newsletter** is available. This podcast is an audio version of the newsletter, recorded by **Robert Miles**.

AXRP Episode 9 - Finite Factored Sets with Scott Garrabrant

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Google Podcasts link

This podcast is called AXRP, pronounced axe-urp and short for the AI X-risk Research Podcast. Here, I (Daniel Filan) have conversations with researchers about their papers. We discuss the paper and hopefully get a sense of why it's been written and how it might reduce the risk of artificial intelligence causing an existential catastrophe: that is, permanently and drastically curtailing humanity's future potential.

Being an agent can get loopy quickly. For instance, imagine that we're playing chess and I'm trying to decide what move to make. Your next move influences the outcome of the game, and my guess of that influences my move, which influences your next move, which influences the outcome of the game. How can we model these dependencies in a general way, without baking in primitive notions of 'belief' or 'agency'? Today, I talk with Scott Garrabrant about his recent work on finite factored sets that aims to answer this question.

Topics we discuss:

Daniel Filan: Before we begin, a note about this episode. More than other episodes, it assumes knowledge of the subject matter during the conversation. So, although we'll repeat some basic definitions, before listening there's a good chance that you'll want to watch or read something explaining the mathematics of finite factored sets. The description of this episode will contain links to resources that I think do this well. Now, on to the interview.

Daniel Filan: Hello everybody. Today, I'll be speaking with Scott Garrabrant. Scott is a researcher at the Machine Intelligence Research Institute or MIRI for short. And prior to that, he earned his PhD in Mathematics at UCLA studying combinatorics. Today, we'll be talking about his work on finite factored sets. For links to items that we'll be discussing, you can check the description of this episode, and you can read a transcript at axrp.net.

Relation to Pearlian causality and abstraction

Daniel Filan: Scott, welcome to AXRP. I guess we're going to start off talking about the finite factored sets work, but to start off that starting off, you've kind of compared this to… Or I think it's sort of meant to be somehow in the same vein of Judea Pearl's work on causality, where you have this directed acyclic graph of nodes and arrows. And the nodes are things that might happen and the arrows are one thing causing another kind of. So I'm wondering, what's good about Pearlian causality? Why does it deserve to be developed on. Let's just start with that. What's good about Pearlian causality?

Scott Garrabrant: What's good about Pearlian causality. So specifically I want to draw attention to the fact that I'm talking about kind of earlier Pearlian stuff, like Pearl has a bunch of stuff. I'm talking about the stuff that you'll find in chapter two of the book Causality specifically. I mean, so basically it's a framework that allows you to take statistical data and from it infer temporal structure on the variables that you have. Which is just really useful for a lot of concepts, there are a lot of purposes. It's a framework that allows you to kind of just go from pure probabilities to having an actual structure as to what's going on and causality, which then lets you answer some questions about interventions possibly and things like that.

Daniel Filan: So if it's so great, why do we need to do any more work? Why can't we just all read this book and go home?

Scott Garrabrant: So my main issue is kind of a failure to work well with abstraction. And so we have these situations, possibly coming from decision theory, where we want to model agents that are making some choice and they have some effect on the world. And it makes sense to model these kinds of things with causality, which is not directly using Pearlian causal inference, but using just kind of the general framework of causality, where we kind of draw these directed acyclic graphs with arrows, that kind of represent effects that are happening.

Scott Garrabrant: And you run into these problems where if you have an agent and, for example, it's being simulated by another agent, then there's this desire to put multiple different copies of the same structure in multiple different places in your causal story. And to me it feels like this is really pointing towards needing the ability to have some of your nodes, some of your variables in your causal diagrams be abstract copies of other nodes and variables. And there's an issue, the Pearlian paradigm doesn't really work well with being able to have some of your variables be abstract copies of others.

Daniel Filan: So what's an example of a place where you'd want to have like multiple copies of the same structure in different places. If you could spell that out.

Scott Garrabrant: Yeah, I can give more specific direct examples that kind of are made up examples, but really I want to claim that you can see a little bit of this any time you have any agent that's making a decision. Agents will make decisions based on their model of the world. And then they'll make that decision - based on their model of the consequences of their actions, and they'll make a decision and they'll take an action. And then those consequences of their action will actually take place in the real world. And so you can kind of see that there's the agent's model of what will happen that is kind of causing the agent's choice, which kind of causes what actually happens. And there's this weird relationship between the agent's model of what will happen and what actually happens that could be well-described as the agent's model is kind of an abstract version of the actual future.

Daniel Filan: Okay. And why can't we have abstractions in Pearlian causality?

Scott Garrabrant: So the problem, I think, lies with what happens when you have some variables that are kind of deterministic functions of others. And so if you have an abstract refined version - if you have a refined version of a variable and you have another coarse abstract version of the same variable, you can kind of view the coarse version that has less detail as a deterministic function of the more refined variable. And then Pearl has some stuff that allows for determinism in the structure. But the part that I really like doesn't really have a space for having some of your variables be deterministically or partially deterministically related. And in the parts of Pearl where some of your variables can be deterministically related, the ability to do inference is much worse. The ability to infer causality from the statistical data.

Daniel Filan: Yeah. I guess to what degree are we dealing with strict determinism here? Because guesses can sometimes be wrong, right? If I'm thinking about the necessity of abstraction here as "I have models about things", it's not really the case that my model is a deterministic function of reality, right?

Scott Garrabrant: Yeah. This is right. I mean, there's a story where you kind of can have some real deterministic functions where you have multiple copies of the same algorithm in different places in space time or something but... Yeah, I feel I'm kind of dancing around my true crux with the Pearlian paradigm. And I don't know, I'm not very good at actually pointing at this thing. But my true crux feels something like variable non-realism, where in the Pearlian world, we have all these different variables and you have a structure that kind of looks like this variable listens to this variable. And then this variable listens to this other variable to kind of determine what happens. And in my world, I'm kind of saying that...

Scott Garrabrant: I'm kind of in an ontology in which there's nothing real about the variables beyond their information content. And so if you had two variables, one of which is a copy of the other, it wouldn't really make sense to talk about like... Yeah, if X and X' were copies of each other, it wouldn't really make sense to ask is Y looking at X or looking at X' if they're actually the same information content. And so kind of philosophically, I think that the biggest difference between my framework and Pearl's framework is something about denying the realism of the variables. Yeah, that didn't really answer your question about, do we really get determinism? I mean, I think that systems that have a lot of determinism are useful for models. We have systems that don't have real determinism, but we also don't actually analyze our systems in full detail.

Scott Garrabrant: And so I can have a high level model of the situation. And while this calculator is not actually deterministically related to this other calculator, relative to my high level model it kind of is. And so even if we don't get real determinism in the real world, in high-level models it feels still useful to be able to work okay with determinism. But I don't know, I think that focusing... I don't know, in some sense, I want to say determinism is kind of the real crux, but in another sense I want to say that it's distracting from the real crux. I'm not really sure.

Daniel Filan: It also seems like one issue - let's say I'm playing Go. And I'm thinking about what's going to happen when I make some move. And then I play the move and then that thing happens. Well, if we say that my model of what's happening is an abstraction of what actually happens, it's a function of what actually happens, then it's the case that there's sort of an arrow from what actually happens to what I think is going happen. And then there's an arrow from that to what I do, because that's what

caused me to make the decision and then there's an arrow from what I do to what actually happens because that's how normal things work. But then you have a loop which you're sort of not allowed to have in Pearl's framework. So that seems like kind of a problem.

Scott Garrabrant: Yeah. I think that largely my general, or a large part of my research motivation for the last, I'm not sure how long, I think at least three years has been a lot towards trying to fix this problem where you have a loop. Thinking about decision theory in ways that people were talking about decision theory in, I don't know, around 2016 or something. There was stuff that involved, well, what happens when you take DAGs [directed acyclic graphs], but you have loops and stuff like this. And I had this glimmer of hope around, well, maybe we can not have loops, when I realized that in a lot of stories like this, the loop is kind of caused by wanting to have… by conflating a variable with an abstract copy of the same variable. And that's not what you did. What you did was you drew an arrow from the thing that actually happens to your model.

Scott Garrabrant: And I think that… So in my framework there aren't going to be arrows, but to the extent that there are arrows like this, it makes more sense to draw the arrow from the coarse model to what actually happens. And to the extent that the coarse model is like a noisy approximation of what actually happens, you kind of won't actually get an arrow there or something. But in my world, the coarser descriptions of what's going on will kind of necessarily be no later than a more refined picture of the same thing. And so I think that I more want to say, you don't want to draw that same kind of arrow between the real world and the model. And kind of, if I really wanted to do all this with graphs, I would say you should at least draw an undirected arrow.

Scott Garrabrant: That's kind of representing their logical entanglement. That's not really causal or something like that. That's not the approach that I take. The approach that I take is to kind of throw all the graphs away. But yeah, so I largely gained some hope that these weird situations that felt like they were happening all over the place and setting up decision theory problems and agents that trust themselves and think about themselves and do all sorts of reasoning about themselves. I gained some hope that all of the weird loopy stuff that's going on there might be able to be made less loopy via somehow combining temporal reasoning with abstraction. And that's largely a good description of a lot of what I've been working on for many years.

Daniel Filan: All right. I guess if I think concretely about this coarse versus abstract thing. Imagine I'm like… So basically think about the Newcomb scenario. So Newcomb's problem is, there's this really smart agent called Omega. And Omega is simulating me, Daniel Filan. And Omega gives me the choice to just take one box that has an unknown amount of money to me in it, unknown to me. And two boxes where I take the unknown amount of money and also a box that contains $1,000. And I can just see that it definitely has $1,000. And Omega is a weird type of agent that says, "I figured out what you would do. And if you would've taken one box, I put a lot of money in the one box, but if you would have taken both then I put almost none in it."

Daniel Filan: And then I take one or the other. So what should I think of… It seems like in your story, there's an abstract variable, maybe in a non-realistic kind of way, because we're variable non-realists, but there's some kind of abstract variable that's Omega's prediction of what I do, and then there's what I actually do. And if it's the case that Omega just correctly guesses, always correctly predicts whether I'm going to one box or two box. What should I think of… In what way is this abstraction lossy? What extra information is there when I actually take the one or the two?

Scott Garrabrant: So I guess if Omega is just completely predicting you correctly. Then I don't want to say that… I kind of want to say, well, there's a variable-ish thing that is what you do. And it goes into Omega filling the boxes. And it also goes into you choosing the boxes. And it's kind of at one point in logical time or something like that. I think that the need for actual abstraction can be seen more in a situation where you, Daniel, can also partially simulate Omega and learn some facts about what Omega is going to predict about you. So maybe… Yeah, Omega is doing some stuff to predict you and you're also simulating Omega. And in your simulation of Omega, you can see predictions about stuff that you're currently doing or about to do and stuff like that.

Scott Garrabrant: And then there's this weird thing where now you have these weird loops between your action and your action. So in the situation where Omega was kind of opaque to you, the intuition goes against what we normally think of as time, but we didn't necessarily get loops, but in the intuition where you're able to kind of see Omega's prediction of you, things are kind of necessarily lossy because you could kind of diagonalize against predictions of you and kind of because it doesn't fit inside you. So, let's see. If you're looking at a prediction of yourself and you have some program trace, which is your computation, and you're working with an object that is a prediction of yourself and maybe it's a very good prediction of yourself. You're not actually going to be able to fully simulate every little part of the program trace because it's kind of contained inside your program trace.

Scott Garrabrant: And so there's a sense in which I want to say, there's some abstract facts that may be our predictions or proofs about what Daniel will do, and that can kind of live inside Daniel's computation. And then Daniel's actual program trace, is this more refined picture of the same thing. And so, I think the need for actually having different levels of abstraction that are at different times is coming more from situations that are kind of actually loopy as opposed to in the Newcomb problem you described. The only reason that it feels loopy is because we have this logical time, then we also have physical time and they seem to go in different directions.

Partitions and factors

Daniel Filan: So now we've gotten a bit into the motivation. What is a finite factored set?

Scott Garrabrant: Okay. So there's this thing - I don't know, I guess I first want to recall the definition of a partition of a set. So a partition of a set is a set of subsets of that set. So we'll start with an original set S and a partition of S is a set of subsets of S, such that each of the sets is non-empty and the sets are pairwise disjoint, so they don't have any common intersection. And when you union all the sets together, you get your original set S. So it's kind of a way to take your set and view it as a disjoint union.

Daniel Filan: Yeah. I kind of think of it as like dividing up a set into parts and that way of dividing it up is a partition.

Scott Garrabrant: Yeah. And so I introduced this concept called a factorization, which can be thought of as a multiplicative version of a partition where you kind of… In the partition story, you put the sets next to each other and you union them together to get the whole thing. And in a factorization, I instead want to kind of multiply your different sets together. And so the way I define a factorization of a set S

is it's a set of non-trivial partitions of S such that for each way of choosing a single part from each of these partitions, there will be a unique element of S that's in the intersection of those parts. And so the same way that you can view partition as a disjoint union, you can view a factorization as a… or sorry, a partition is a way to view S as a disjoint union, a factorization of S is a way to view S as a product.

Daniel Filan: Okay. And so to make that concrete, an example that I like to have in my head is suppose we have points on a 2D plane, and we imagine the points have an X coordinate and a Y coordinate. So one partition of the plane is I can divide the plane up into sets where the X coordinate is zero, sets where the X coordinate is one, sets where the X coordinate is two. And those look like lines that are perpendicular to the X axis. And none of those lines intersect. And every point has some X coordinate.

Daniel Filan: So it's this set of lines that together cover the plane, that's one partitioning, the X partitioning. And there's another one for values of Y, right? Which look like horizontal lines that are various amounts up or down. And so once you have the X partitioning and the Y partitioning, any point on the plane can be uniquely identified by which part of the X partitioning it is and which part of the Y partitioning it is because it just tells you how much to the right of the origin are you, how much above the origin are you, that just picks out a single point. I'm wondering, do you think that's a good kind of intuition to have.

Scott Garrabrant: Yeah, I think that's a great example. To say a little bit more there. So your original set S in that example you just gave is going to be the entire Cartesian plane, the set of all ordered pairs, like (X coordinate, Y coordinate). And then your factorization is going to be a set B, which is going to have just two elements. And the two elements are the partition according to what is the Y coordinate, and the partition according to what is the X coordinate. You can kind of view partitions as questions. And so in general, if I have a set like the Cartesian plane, and I want to specify a partition, one quick way to do that is I can just ask a question. I can say, what is the X coordinate? And that question kind of corresponds to the partition that breaks things up according to their X coordinate.

Daniel Filan: Okay. And the one slightly misleading thing about that example is that there are an infinite number of points in the X, Y plane. But of course, we're talking about finite factored sets. So S only has a finite number of points.

Scott Garrabrant: Yeah. So we're talking about finite factored sets. So in general, I'll want to work with a pair (S, B). Where S is a set, a finite set, and B is a factorization of S.

Daniel Filan: Why choose the letter B, for a factorization?

Scott Garrabrant: It's for basis.

Daniel Filan: Mmm.

Scott Garrabrant: Yeah. It's for basis. Largely, while I'm thinking of the elements of B as partitions of S, I'm also kind of just thinking of them as elements, just out on their own that are kind of representing the different basis elements. Yeah, so it actually looks a lot like a basis, because for any point in S you can kind of uniquely specify it by specifying its value on each of the basis partitions.

Daniel Filan: Okay. So yeah, this gets into a question I have which is, how should I think about the factors here? Like these finite factored sets, I guess they're supposed

to represent what's going on in the world. How should I think about factors in general or partitions, I can think about as questions. These factors in B, how should I think about those?

Scott Garrabrant: Yeah, I think I didn't explicitly say, but the elements of B, we'll call factors. We use the word factor for the elements of B. So I almost want to say the factors are kind of a preferred basis for… Yeah, the factors kind of form a preferred basis for your set of possibilities. And so if I consider the set of all bit strings of length five, right. There are 32 elements there. And if I wanted to specify an element, I can do so in lots of different ways. But there's something intuitive about this choice of breaking my 32 elements into what's the first bit, what's the second bit, what's the third bit, what's the fourth bit, what's the fifth bit. So it's kind of… It's a set of questions that I can ask about my element, that uniquely specify what element it is.

Scott Garrabrant: And also for any way of answering the questions, there's going to be some element that works with those answers. And so factorization is kind of just like, it's a combinatorial thing that could be used for many different things. But one way to think about it is kind of, you're making a choice that is a preferred basis, a preferred set of variables to break things up into and you're thinking of those as kind of primitive and then other things as built up from that. So properties like do the first two bits match is then thought of as built up from what is the first bit and what is the second bit. And so it's kind of a choice of what comes first, a choice of what are the primitive variables in your structure.

Daniel Filan: Okay. So if I'm trying to think about maybe some kind of decision problem where I'm going to do something, and then you're going to do something and then another thing's going to happen. And I want to model that whole situation with a finite factored set. Instead of thinking about modeling which thing is going to happen to me today, if I want to model an evolving situation, how should I think about what the set S is and what the factors should be?

Scott Garrabrant: Yeah. So factorization is very general and actually, I use the word factorization in multiple different contexts when talking about this kind of thing. But in the question that you're asking, to say some more background stuff. So I'm going to introduce this theory of time that has a background structure that looks like a factored set rather than a background structure that looks like a DAG, looks like a directed acyclic graph, like in the Pearlian case. And so specifically if I'm using a factored set and I am using that to interpret… I'm using that to kind of describe some causal situation, I am not going to have nodes. I'm not going to have factors that kind of correspond to the nodes that you would have in the Pearlian world. Instead, I am mostly going to be thinking of the factors as independent sources of randomness.

Scott Garrabrant: I'm hesitant there, because a lot of my favorite parts of the framework aren't really about probability. But if I'm thinking about it in a temporal inference setting where I'm like getting the statistical distribution of… Where I'm getting a statistical distribution, then I'm thinking of the factors as basically independent sources of randomness. And so if we have some variable X, and then we have some later variable Y that can take on some values and partially is going to be a function of X, then we won't have a factor for X and a factor for Y, we'll have a factor for maybe that which kind of goes into what X is. And then we'll have another factor that's like the extra randomness that went into the computation of Y once you already knew X. And when we think of factored sets as related to probability, we're actually going to always want our factors to be independent. And so the factors can't really be put on things like X and Y when there's a causal relationship between X and Y.

Daniel Filan: So it almost sounds like the factors are supposed to be somehow initial data, like the problem set-up or something, the specification of what's going on, but kind of an initial specification?

Scott Garrabrant: Yeah, indeed, if you take my theory of time. So I have way of taking a factored set and taking an arbitrary partition on your set S. And then I'm going to be able to specify time, specify when some partitions are before or after other partitions. The factors will be partitions with the property that you can't have anything else that's strictly before it besides deterministic things. And so there's a sense in which the factors are initial.

Scott Garrabrant: Yeah, the factors are basically the initial things in the notion of time that I want to create out of factored sets. But also intuitively, it feels like they're initial. It feels like they're the primitive things that came first and then everything else was built out of them.

Daniel Filan: Yeah, I think primitive is a better word than initial.

Scott Garrabrant: Yeah, they're - In the poset of divisibility, right? I wanted to say that one is initial, not the primes. But the primes have the property that you can't really have anything else before them. Yeah, I don't know, primitive is like prime is... Yeah.

Daniel Filan: Yeah. So in the sense they're like when you're dividing numbers, whole numbers, you get to primes and then you get nowhere else, and they're primitive in that sense.

Scott Garrabrant: Right.

Orthogonality and time

Daniel Filan: Yeah. So you talk about the history of variables as all the initial factors that you need to specify what a variable is. And then you have this definition of orthogonality. And people can read the paper. I don't know if they'll be able to read the paper by the time this goes live, but they'll be able to read something to learn about what orthogonality is.

Scott Garrabrant: Yeah, the talk and the transcript of the talk that, for example, appears on the MIRI blog or on LessWrong, it has all of the statements of everything that I find important. And so you shouldn't have to wait for a paper. I think that, modulo the fact that you'd have to prove things yourself, that has all the important stuff.

Daniel Filan: So, people, I think for me at least, it's easy to read the definition of orthogonality and still not really know how to think about it, right? So how should people think about what's orthogonal?

Scott Garrabrant: So in the temporal inference thing where we're going to be connecting up these combinatorial structures with probability distributions, orthogonality, it's going to be equivalent to independence. And so, one way to think of orthogonality is... I kind of took out a combinatorial fragment of independence, where I'm not actually working with probabilities but I am working with a thing that is representing independence. Another way to think of it is like when two partitions are orthogonal, you should expect that if you come in and tweak one of them, it will have

no effect on the other one, they're separated. And specifically in the factored set framework, orthogonality means they do not have a common factor. Where these factors can be thought of as sources of randomness or sources of something. If they do not share a common factor, then they're in some sense separated.

Daniel Filan: Yeah. So I'll just say one example that helped me understand it is this XY-plane example where the factors were, you're partitioning up according to the X coordinate. And you can also partition up according to what the Y coordinate is. So you can have one different partition, that's are you on the left-hand side or the right-hand side of the Y-axis. That's a partition of the plane. It's like a variable. It's like are you on the left, or are you on the right. And there's a second variable, that's are you above the X-axis, or are you below the X axis. Right? That also partitions the plane up. And my understanding is that in this factored set, those two partitions, or you could think of them as variables, are orthogonal, and that hopefully gives people a sense. It's also kind of nice because they're literally - if you think about the dividing lines, they're literally orthogonal in that case, and that's maybe not a coincidence.

Scott Garrabrant: Yeah. So historically when I was developing factored set stuff, I was actually working with things that looked like I have two partitions and there's something nice when it's the case that they're both kind of mutually… Or sorry, for any way of choosing a value of this X partition and also choosing the value of the Y partition, those two parts will intersect. So in your example, right, the point is that all four quadrants exist, and this is like there's a sense in which this is saying that the two partitions aren't really stepping on each other's toes.

Scott Garrabrant: If we specify the value of this X partition, it doesn't stop us from doing whatever we want in terms of specifying the value of the Y partition. And largely orthogonality is a step more than that, where a consequence of being orthogonal, this is they're not going to step on each other's toes. But I have this extra thing, which is really just the structure of the factorization, but there's this extra thing beyond just not stepping on each other's toes, which is coming from some sort of theory of intervention or something. You can view the factored set as a theory of intervention because the factored set basically allows you to take your set and view its elements like tuples. And when your elements are like tuples, you can imagine going in and changing one value and not the others. And so, orthogonality looks like it's not just X and Y are compatible, like compatible in all ways and assign values to each of them, but it's also, when you mess with one, it doesn't really change the other.

Daniel Filan: Cool. So another question about orthogonality. So in the talk that listeners can watch or read a transcript of, you sort of say, "Okay, we have this definition of orthogonality and this definition of conditional orthogonality, which is a little bit more complicated but kind of similar." You then talk about inference in the real world. So we sort of imagine that you're observing things that are roughly like… You don't observe the underlying set, but you observe things that are roughly like these factors. And somehow you get evidence about what things are orthogonal to each other. And from that, you go on and infer a whole bunch of stuff, or you can sometimes. And you give an example of how that might work. How would I go about gathering this orthogonality data of what things in the world are orthogonal to what other things?

Scott Garrabrant: So the default is definitely passing through this thing that I call the fundamental theorem. So the fundamental theorem says that conditional orthogonality is equivalent to, for all probability distributions that you can put on your factored set which respect the factorization, your variables are conditionally

independent. I don't know, I phrased that as for all probability distributions, but you can quickly jump that for a probability distribution in general position.

Scott Garrabrant: And so the basic thing to do is if you have access to a distribution over the things, over the elements of your set or over something that's a function of the elements of your set, if you have access to a distribution, it's kind of a reasonable assumption to say that if you have a lot of data and it looks like these two variables are independent, then you assume that they are orthogonal in whatever underlying structure produced that distribution. And if they're not independent, then you see they're not orthogonal.

Scott Garrabrant: And it's just - the default way to get these things is via taking some distribution, which could be coming from a bunch of samples, or it could be a Bayesian distribution. But I think that orthogonality is something that you can basically observe through its connection to independence versus time is not as much.

Daniel Filan: I guess this works when my probability distribution has this form where the original factors in your set B, like the primitive variables, have to be independent in your probability distribution. And if they're independent in that distribution, then conditional independence is conditional orthogonality. How would I know if my distribution had that nice structure?

Scott Garrabrant: You can just interpret all of the independence as orthogonality and see whether it contradicts itself, right? It's like you might have a distribution that can't really be well-described using this thing. And one thing you might do is you might develop an orthogonality database where you keep track of all of the orthogonalities that you observed. And then you notice that there's some orthogonalities that you observed that are incompatible with coming from something like this. Yeah, I'm not sure I fully understand the question.

Using finite factored sets

Daniel Filan: I guess what I'm asking is imagine I'm in a situation, right? And I don't already have the whole finite factored set structure of the world. I'm wondering how I go about getting it. Especially if the world is supposed to be my life or something. So I don't get tons of reruns. Maybe this isn't what it's supposed to do.

Scott Garrabrant: Basically the answer you'd give here is similar to the answer you'd give to the same question about Pearlian Causality. And I think that largely the temporal inference story makes the most sense in a context that's very like you have a repeated trial that you can repeat an obnoxious number of times, and then you can get a bunch of data. And you're trying to tell a story about this trial that you repeated. And yeah, so the story that I tell makes the most sense in a situation that's like that.

Scott Garrabrant: I'm excited for a lot of things about factored sets that are not about just doing temporal inference from a probability distribution. And those feel like they play a lot more nicely with… Sorry, the applications that feel like they're about embedded agents to me are different from the applications that feel like they're about temporal inference. Because it feels like you need something like this frequentist, lots of repetition to be able to get a distribution in order to be able to do a lot of stuff with temporal inference. Or at least doing it the naive way, maybe you can build up more stuff.

Daniel Filan: So embedded agency is your term for something being an agent in a world where your thinking processes are just part of the physical world and can be modified and modeled and such. Is that a fair summary?

Scott Garrabrant: Yeah.

Daniel Filan: Okay, so the applications of the finite factored set framework to embedded agency, are you thinking of that more as a way to model things?

Scott Garrabrant: Yeah, I'm thinking of it mostly like basically, there's a lot of ways in which people model agents using graphs where the edges in the graph represent information flow or causal flow or something, that all feel they're entangled with this Pearlian causality story. And often I think that pictures like this fail to be able to handle abstraction correctly, right? I have a node that represents my agent. And I don't really have room for another node that represents a coarser version of my agent because if I did, which one gets the arrow out of it and such. And so largely, my hope in embedded agency is just all the places where we want to draw graphs, instead, maybe we can draw a factored set and this will allow things to play more nicely with abstraction and playing nicely with abstraction feels like a major bottleneck for embedded agency.

Why not infinite factored sets?

Daniel Filan: Okay. Yeah, I'll ask more about that a bit later. Yeah, so I guess I want to ask more questions about the finite factored set concept itself. Why is it important that it's finite?

Scott Garrabrant: So I can give an example where you should not expect the fundamental theorem to hold in the cases where it's infinite. So the thing where independence exactly corresponds to orthogonality, in the infinite case, one shouldn't expect that to hold. And it might be that you can save it by saying, "Well, now we can't take arbitrary partitions. We can only take petitions that have a certain shape."

Daniel Filan: Sort of like measurability criteria?

Scott Garrabrant: Sort of like measurability criteria, but measurability is actually not going to suffice for this. I can give an example. To give an example, if you imagine the infinite factored set that is countable bit strings.

Daniel Filan: So this is just like infinite sequences of ones and zeros, the set of all of them?

Scott Garrabrant: Yeah. So you have the set of all infinite sequences of ones and zeros, and you have the obvious factorization on this set, which is you have one factor for each bit. And then there's a partition that is asking, "Is the infinite bit string, the all zero string?" And there's the partition that's asking, "Is the infinite bit string, the all one string?" These are two partitions, let's call them X and Y.

Daniel Filan: Okay.

Scott Garrabrant: And it turns out that for any probability distribution you can put on this factored set that respects the factorization. At least one of this partitions is going to have to be - either it's going to be the case that the all zero string is probability zero or the all one string is probability zero.

Daniel Filan: Yep.

Scott Garrabrant: Because all of the bits have to be independent. And then you'd be able to conclude that the question, "Is it the all zero string?" And "Is it the all one string?" Those two questions will have to be independent in all distributions on the structure, but it really doesn't make sense to call them orthogonal.

Daniel Filan: Why doesn't it?

Scott Garrabrant: Why doesn't it make sense to call them orthogonal? It doesn't make sense to call them orthogonal because all of our factors go into the computation of that fact. And so, if you're thinking of orthogonality as, they can be computed using disjoint collections of factors, you can't really compute whether something's the all zero string or whether something's all the one string without seeing all the factors. I mean, you can't compute it because the first time you see a one, you can say, "All right, I'll stop looking." But in my framework, that doesn't count. You have to specify upfront all the bits you're going to use. And so, I don't know, I think there's some hope to being able to save all of this. And I haven't done that yet. And there's another obstacle to infinity, which is even the notion of thinking about the history of a partition, that's not even going to be well-defined in the infinite case.

Daniel Filan: Okay, and so the history of a partition in the finite case, it was just the smallest set of factors that determined - if I have a partition with elements like X1 through Xk or something, the history of the partition is just all of the basic factors, all the things in my set B where if you think of the things in the set B, its variables, if I know the value for all the variables, it's the smallest set of variables, such that if I know that the value for all of them, I can tell what partition element I'm in for the thing that I'm looking for the history of. So it's the smallest set of, the smallest amount of initial information or something, that's specifying the thing I'm interested in. Is that what a history is?

Scott Garrabrant: Yeah, that's right.

Daniel Filan: And it's a bit worrisome because usually there's not a smallest set of sets of… Right? That's not obviously well-defined.

Scott Garrabrant: Yeah. So, specifically smallest in the subset ordering. And so the history of a partition is a set of factors. And if you take any set of factors that would be sufficient to determine the value of that partition, the answer to that question, what part of it's in, if you have any set of factors that were sufficient to compute the value of X, then necessarily that set of factors must be a superset of the history. And so it's not smallest by cardinality, it's smallest by the subset ordering. And showing that this is well-defined involves basically just showing that sets of factors that are sufficient to determine the value of X are closed under intersection. So if I have two different sets of factors and it's possible to compute the value of X with either of those sets of factors, then it's possible to compute the value of X using their intersection.

Scott Garrabrant: But actually this is only true for finite intersections. And so if I have a partition and you're able to compute the value from either of these two sets of factors, then you're able to compute the value from their intersection. But if I have an infinite class of things then I can't necessarily compute the value from their intersection. To see an example of this, if you, again, look at infinite bit strings with the obvious factorization and you look at the partition, are there finitely many ones? If you take any infinite tail of your bit string that's sufficient to determine, are there

finitely many ones? But if you take the intersection of all of these infinite tails, you get the empty set.

Daniel Filan: Yeah, so one way I'm now thinking of this is the problem with infinite sets is that they have these things that are analogous to tail events or something in normal probability theory where you depend on all of these infinite number of things, the limit of it. But the limit is actually zero, but you can't exactly-

Scott Garrabrant: Yeah, there's this - actually, what I think is a coincidence.

Daniel Filan: Okay.

Scott Garrabrant: But you could do this naive thing where you say, "Well, just take the intersection and call that the history." And then the history of the question, "Are there finitely many ones?", would then be the empty set and then a lot of properties would break. But an interesting coincidence here or something, I don't know, I don't think it's a coincidence, but I don't like this definition. I don't like the definition of generalizing to the infinite case by just defining the history to be the intersection. But if you were to do that, then you would get that the question, "Are there finitely many ones?", is orthogonal to itself because it has empty history. And what kind of partitions are orthogonal to themselves? They're deterministic ones where one of the parts has probability one and all the others have probability zero.

Scott Garrabrant: And the Kolmogorov zero-one law says that properties like the question, "Are there finitely many ones?", if all of the individual things are independent, are necessarily probability zero or one. And so there's this thing where if you define history in this way, if you naively extend history to the infinite case by just taking the intersection, even though it's not closed under intersection, you actually get something that feels like it gives the right answer because of the Kolmogorov zero-one law.

Daniel Filan: Yeah, but it's going through some weird steps to get the right answer.

Scott Garrabrant: Yeah.

Daniel Filan: By the way, I don't know, there are a variety of these things called zero-one laws in probability theory, and if listeners want to think about knowledge and changing over time, I don't know. Some of these zero-one laws are fun to mull over and think about how they apply to your life or something. Oh, do you have comments on that claim?

Scott Garrabrant: No, I think Kolmogorov's zero-one law is really interesting and I would recommend it to people who like interesting things.

Limits of, and follow-up work on, finite factored sets

Daniel Filan: All right. So back on the finite factored sets, they're sort of a way of modeling some types of worlds. Or some sorts of ways the world can be. Are there any worlds that can't be modeled by finite factored sets?

Scott Garrabrant: Yeah.

Daniel Filan: I guess infinite ones, but ignoring that for this second.

Scott Garrabrant: So there's this issue that's similar to an issue in Pearl where we kind of - when you look at distributions that are coming from a finite factored set or from a DAG, we're looking at probabilities in general position. And so it doesn't really make sense to have a probability one half or one fourth. Although-

Daniel Filan: What do you mean by probability in general position?

Scott Garrabrant: So in both my world and Pearl's world, we want to specify a structure.

Daniel Filan: Yep.

Scott Garrabrant: And then we have all of the probability distributions that are compatible with that structure. And these give you a manifold or something of different probability distributions. And then some measure zero subset of them have special coincidences. And when I say in general position, I mean, you don't have any of those special coincidences. And so an example of a special coincidence is any time you have a probability of one half or one fourth or something like that, that's a special coincidence. But to a Bayesian, that might happen because of principle of indifference or something. Or to a limited Bayesian that kind of doesn't really know what… It feels like principle of indifference advises having probabilities that are rational numbers, but probabilities that are rational numbers lead to coincidences in independence that don't arise from orthogonality. And so there's a sense in which my framework and the Pearlian framework don't believe in rational probabilities as something that just happens, or something.

Daniel Filan: Yeah. I mean, it's even more concerning because if I think that I'm a computer and I think I assign probabilities to things, well, the probabilities I assign will be numbers that are the output of some computation. And there are a countably infinite number of computations, but there are an uncountably infinite number of real values. So somehow, if I'm only looking at things in general position, I'm ruling out all of the things that I actually could ever output.

Scott Garrabrant: Yeah, so I have a little bit of a fragment of where I want to go with trying to figure out how to deal with the fact that my system and Pearl's system don't believe in rational probabilities, which is to define something, and this is going to be informal. Well, it'll be formal but wrong. To define something that's like, you take a structure that is a factored set together with a group of symmetries on the factored set that allows you to maybe swap two of the parts within a partition or swap two of the partitions with each other, or swap two of the factors with each other, right? So you can have some symmetry rules. So if you, for example, considered bit strings of length five again, you could imagine a factored set that it separates into the five bit locations, is this bit zero or one, but then it also has the symmetry of you can swap any of the digits with each other that can also do swapping zero with one. But for now, I'll just think about swapping any of the digits with each other.

Scott Garrabrant: And then it's subject to the structure that this is this factored set and the swapping rule for this group of symmetries, then the set of distributions that are compatible with that structure will not be just things in which the bits are independent from each other, it'll be any distribution in which the bits are IID, so independent and also identical. And so you could do that. So you could say, "Well, now my model of what might happen, the thing that I'm going to try to infer from my probability distribution is a factored set together with a group of symmetries on that

factored set." And I mean, you're not going to at least naively get the fundamental theorem the same way. And so I'm not sure what happens if you try to do inference on something like this, but if you want to allow for things like rational probabilities, then maybe something like that would be helpful.

Daniel Filan: All right, so I'd like to ask a question about the form of the framework. So when I think of models of causality or of time, the two prior ones that I think of are, we talked about Pearl's work on directed acyclic graphs and do-operators and such, where you can draw a graph, and also Einstein's work on special and general relativity, where you have this space time thing that's very geometric, very curved and you have this time direction, which is special and kind of different from the spatial directions. Those were all really geometric and included some nice pictures. Finite factored sets does not have many pictures. Why not? I really liked pictures.

Scott Garrabrant: Yeah. I mean, I think it has something to do with the variable non-realism, where it feels like the points or nodes in your pictures or something - if I take a Pearlian DAG and there are 10 nodes in it and even if I assume that they're all just binary facts, then now I have, well, you take 1024 different ways that the world can be. And then you take Bell number of 1024 different possible variables that I can define on this, which is obnoxiously huge and then there's not as much of a useful interpretation of the arrows that connect them up or something, it's something to do with variable non-realism, where there's a sense in which Pearl's kind of starting from a collection of variables, which is a way to factor the world into some small object. And because I'm not starting with that, my world is kind of a lot larger.

Scott Garrabrant: I don't know. Another thing that I'd call a theory of time is people talk about time with entropy. That's another example that doesn't feel as visual.

Daniel Filan: Yeah, that's true.

Scott Garrabrant: And I think that that's a lot more variable free and that's maybe part of why.

Daniel Filan: Yeah. It's also the case that once you have variables, they have these relations in terms of their histories and such, you could draw them in a DAG or something.

Scott Garrabrant: Yeah. It's like the structure of an underlying finite factored set is very trivial or something. It's - Pearl has a DAG, and if you wanted to draw a finite factored set as a DAG, it would just be a bunch of nodes that are not connected at all that each have their own independent sources of randomness. And then if you wanted to, you could maybe draw an arrow from these nodes to all the different things that you can compute using them. Or if you wanted to say, oh, let's just have the basic variables then it's just these nodes.

Daniel Filan: Yeah. But I guess somehow if you want to talk about the structure that lets you talk about variables, but you don't want to talk about variables. I guess that's less amenable to pictures perhaps.

Scott Garrabrant: Yeah. I mean, I don't feel like physics and Pearl have pictures for necessarily the exact same reason or something, and I'm kind of just, graphs got lucky in the way that they're easy to visualize or something like that.

Daniel Filan: Yeah, that might be right. It's also true that graphs are not - simple graphs are easy to visualize, but there are a lot of non-planar graphs that are kind of a

pain to draw. Yeah. So, a related question complaint I have is a lot of this work seems like it could be category theory.

Scott Garrabrant: Yeah. It could be category theory.

Daniel Filan: Yes. So partitions are basically functions from a set to a different set and the parts are just all the things that have the same value of the function?

Scott Garrabrant: Yeah. Partitions is kind of the information content of a function out that you get by kind of ignoring the target and only looking at the source.

Daniel Filan: Yeah. And it seems like there are probably nice definitions of factors and such and… You know, there's lots of duality and category theory has pictures. It's also a little bit nice in that I have to admit, looking at sets of sets of sets of sets can be a little bit confusing after a while in the way that categories can have some nice language for that. So why isn't everything category theory, even though category theory is objectively great?

Scott Garrabrant: Yeah. I mean, it goes further than that. I actually know most of the category theory story and I've worked out a lot of it and kind of went with the combinatorialist aesthetic with the presentation anyway. So what are my reasons? One reason is because I kind of trust my category theory taste less and I kept on changing things or something in a way that I was not getting stuff done in terms of actually getting the product out or something by working in category theory. And so it was kind of oh yeah, I'll punt that to the future.

Scott Garrabrant: Another reason is because it feels the system just really doesn't have prerequisites and by phrasing everything in terms of category theory, you're kind of adding artificial prerequisites that maybe make the thing prettier, but you actually, you know what a set is, you can kind of go through all the proofs or something. That's not entirely true, but it's because I'm working in a system that has very few prerequisites, the extra cost of prerequisites is higher. The marginal cost of adding prerequisites is higher.

Scott Garrabrant: Another reason was I was just really shocked by the sequence that counts the number of factorizations not showing up on OEIS. So yeah, if you take an n-element set and you count how many factorizations there are on the n-element set, you get a sequence and there's this Online Encyclopedia of Integer Sequences that has 300,000 sequences and it does not have this sequence in spite of having a bunch of lower quality sequences.

Scott Garrabrant: And I was very surprised by this fact, and it feels like a very objective test. I'm not a particularly scholarly person. It's hard for me to figure out what people have already done. And I was just pretty blown away by the fact that this thing didn't show up on OEIS. And so I kind of stuck with the combinatorialist thing because it had that objective thing for the purpose of being able to do an initial sell or something.

Daniel Filan: Okay.

Scott Garrabrant: Yeah. Those are most of my reasons. I think that - I haven't worked out all the category theory, but I think it will end up being pretty nice. In fact, I think that just even the definition of conditional orthogonality, I think can be made to look relatively nice categorically, and it's via a path that is pretty unclear from the definition that I give in the talk or the post but there's an alternative definition that

kind of looks like if you want to do orthogonality and you want to condition on some fact about the world, the first thing you do is you take your original factored set and you kind of take the minimal flattening of it. So that the thing you want to condition on is kind of rectangular in your factored set and then you - where by flattening I mean, you merge some of the factors together.

Daniel Filan: Okay.

Scott Garrabrant: And if you take the minimal factoring and then you ask whether your partitions are orthogonal in the minimal factoring, that corresponds to conditional orthogonality. And so I think that categorically there's a nice definition here. But I definitely agree about the category theory aesthetic and I think that it actually is a good direction to go here that I may or may not try to do myself, but if somebody was super interested in trying to convert everything to category theory, I could talk to them about it.

Daniel Filan: So speaking of that, I'm wondering, what follow-up work are you excited about being done here? And do you think that this kind of development is going to look more like showing nice things within this framework - making it categorical or showing the decidability of inference in finite factored sets? Or do you think it's going to look more like iterating on some of the definitions and tweaking the framework a little bit until it's the right framework?

Scott Garrabrant: Yeah. I mean the category theory thing a little bit does fall into tweaking the framework until it's the right framework. Although it's a little different, I have applications I'm excited about in both spaces.

Scott Garrabrant: Yeah, if I were to list applications that I expect that I'm not personally going to do, that seem like projects that would be interesting for people to pick up, one of them would be converting everything to category theory. One of them would be figuring out all the infinite case stuff and looking at applications to physics. I think that there's a non-trivial chance of some pretty good applications in physics that would come out of figuring out all the infinite case stuff. Because I think that factored sets are actually a lot closer to being able to give you something like continuous time then the Pearlian stuff. Yeah.

Scott Garrabrant: So one would be the physics thing. One would be basically trying to do computational stuff in terms of, I kind of just have a couple proofs of concept of how to do temporal inference. And I think you said, showing decidability of temporal inference is a thing, where really it's… I think that somebody should be able to actually search over the space of a certain flavor of proof and be able to actually come up with examples of temporal inference that come from this where you take in some orthogonality data and are able to infer time from them. And I think that there's a computational question here that I think, I might be wrong, might be able to at least be able to produce some good examples, even if it's not actually doing temporal inference in practice, and so I'd be excited about something that.

Scott Garrabrant: I would be excited about somebody trying to extend to symmetric finite factored sets, which is the thing I was talking about earlier, about dealing with rational probabilities. I think that of these that I listed, the one that I'm most likely to want to try to work on myself is the symmetric factored sets thing. Because I think that could actually have applications to the kind of embedded agency type stuff I'd want to work with. But for the most part I'm expecting to myself think in terms of applications, as opposed to think in terms of extending the theory, and all the things

that I kind of said, were all forms of extending the theory either by tweaking stuff or by kind of putting stuff on top of it. I think it's mostly just putting stuff on top of it.

Scott Garrabrant: I think I don't say that much. Sorry, I think that there aren't that many knobs to twiddle with the basic thing. You could have some new orientation on it, but I think it will be basically the exact same thing. I think that the parts… The way that I defined it, there was only one factorization on a zero element set. Maybe it would be nicer if there were infinitely many factorizations on zero. The definitions might be slightly different, but I think it's basically the same core thing.

Scott Garrabrant: I think that I'm mostly thinking that the baseline I have is kind of correct enough for the kind of thing that I want to do with it, that I don't expect it to be a whole new thing. I expect it to be built on top, and there's different levels of built on top.

Relevance to embedded agency and x-risk

Daniel Filan: So I guess I'd now like to pivot into a bit of a more general discussion about your research and your research taste. How do you see the work on finite factored sets as contributing to reducing existential risk from artificial intelligence? If you see it as doing that?

Scott Garrabrant: I think that a lot of it factors through trying to become less confused about agency and embedded agency, which I don't know, I have opinions in both directions about the usefulness of this. Sometimes I'm feeling like, yeah this isn't going to be useful and I should do something else. And then sometimes I'm just interacting with questions that are a lot more direct and noticing how a lot of the kind of questions that I'm trying to figure out for embedded agency actually feel like bottlenecks to be able to say smart things about things that feel more direct.

Daniel Filan: Can you give an example of that?

Scott Garrabrant: I don't know. Evan says some things about myopia.

Daniel Filan: That's Evan Hubinger.

Scott Garrabrant: Right. And that feels a lot more direct, trying to get a system that's kind of optimizing locally and not looking far ahead and stuff like that.

Scott Garrabrant: And I feel like, in wanting to think about what this even means, I notice myself wanting to have a better notion of time and better notion of things about the boundaries between agent and environment and all of this stuff. And so I don't know… That's an example of something that feels kind of more direct, myopia feels like something that could be very useful if it could be implemented correctly and could be understood correctly.

Scott Garrabrant: And when I try to think about things that are more direct than embedded agency, I feel like I hit the same kind of cruxes and that working with embedded agency feels like it's more directed at the cruxes, even though it's less directed at the actual application of the thing in a way that I expect to be useful.

Daniel Filan: In the myopia example, I think the first-pass solution would be look, there's just physical time [that] basically exists. And we're just going to say, okay, I want an AI system. I want it to care about what's going to happen in the next 10

seconds of physical time and not things that don't happen within the next 10 seconds of physical time. Do you think that's unsatisfactory?

Scott Garrabrant: Yeah. So, I mean, I think that you can't really look at a system and try to figure out whether it's optimizing for the next 10 seconds or not. And I think that - the answer that I actually gave with the myopia thing was a little off because I was actually remembering a thought about myopia, but it wasn't about time. It was about - just time was the thing that I said in that thing.

Scott Garrabrant: It was more about counterfactuals and more about the boundary between where the agent is or something like that. But I don't know. I still think the example works. I mean, I think that it comes down to you want to be able to look at a system and try to figure out what it's optimizing for. And if you have the ability to do that, you can check whether it's optimizing for the next 10 seconds, but in general, you don't really have the ability to do that.

Scott Garrabrant: Figure out what it's trying to do or something like that. And I think that… Yeah, how do I get at the applications? Okay. So one thing I think is that in trying to figure out how the system works, it is useful to try to understand what concepts it's using and stuff like this. And I think that the strongest case I can kind of make for factored sets is that I think that there's a sense in which factored sets is also the theory of conceptual inference. And I think that this could be helpful for looking at systems or trying to do oversight of systems that you want to be able to look at a thing and figure out what it's optimizing for.

Daniel Filan: In what ways would you say it's a theory of conceptual inference?

Scott Garrabrant: Well, one way to look at the diff between factored sets and Pearl is that we're kind of not starting from a world factored into variables instead we're inferring the variables ourselves. And so there's a sense in which if you try to do Pearlian style analysis on a collection of variables, but you messed it up, and instead of having a variable for what number this… I have a number and it's either zero or one and it's also either blue or green. And I can also invent this concept called grue, which is a green zero or a blue one. And instead of thinking in terms of what's the number and is it blue, you can think of what's the number and is it grue, and maybe if you're working in the latter framework, you're kind of using the wrong concepts and you will not be able to pull out all the useful stuff you'd be able to if you were using the right concepts.

Scott Garrabrant: And factored sets kind of has a proof of concept towards being able to distinguish between blue and grue here, where the point is, in this situation, if the number is kind of independent of the color and you're working with the concept of number and the concept of grue-ness, you have this weird thing where it looks there's a connection between number and grue-ness, but it also is the case that if I invent the concept of number xor grue-ness, I kind of invent color, and color lets me factor the situation more and see that maybe you should think of it as the number and the color are primitive properties like we were saying before, and grue-ness is a derived property.

Scott Garrabrant: And so there's a sense in which earlier things are more primitive, and it's not just earlier things, I think there was more than just that. But there's a sense in which because I'm not taking my variables or my concepts as given, I am also doing some inferring which concepts are good.

Daniel Filan: So somehow it strikes me that inferring which concepts are good, is a related, but different problem to inferring which concepts a system is using.

Scott Garrabrant: I don't know, there's stuff that you like to think about that involve kind of having separate neurons as part of it. And I think there's a sense in which it might be that we're confused when we're looking at a neural net because we're thinking of the neurons as more independent things, when really they could be a transform similar to the blue/grue thing from some other thing that is actually happening and being able to have objective notions of what's going on there - being able to have a computation and having there be a preferred basis that causes things to be able to factor more or something feels… Yeah, so I guess I'm concretely pointing at the picture of factorization into neurons in the result of a learned system might be similar to grue.

Daniel Filan: Yeah, it's interesting in that people have definitely thought about this problem, but all the work on it seems kind of hacky to me. So for instance, so I know Chris Olah and collaborators now or formerly at OpenAI, have done a lot of stuff on using non-negative matrix factorization to kind of get out the linear combinations of neurons that they think are important. And the reason they use non-negative matrix factorization, as far as, I might be getting this wrong, but as far as I can tell it's because it kind of gets good results sort of, rather than a theory of non-negativity or something.

Daniel Filan: Or a similar thing is there's some work about exactly trying to figure out whether the concepts in neural networks are on the neurons or whether they're these linear combinations of neurons, but the way they do it, which again, I'm going to sound critical here. It's a good first pass, but a lot of this work is, okay, we're going to make a list of all of the concepts. And now we're going to test if a neuron has one of the concepts which I've decided really exists, and we're going to check random combinations of neurons and see if they have concepts, which I've decided exist and which does better.

Daniel Filan: Yeah. There's definitely something unsatisfying about this. Maybe I'm not aware of more satisfying work. Yeah. It does seem there's some problem there.

Scott Garrabrant: And again, I think that you're not going to be directly applying the kind of math that I'm doing, but it feels I kind of have a proof of concept for how one might be able to think of blueness as a statistical property, blueness versus grue-ness as a statistical property. It's something that you can kind of get from raw data.

Scott Garrabrant: And I don't know, I feel there's a lot of hope in something that. But that's also not my main motivation. That was a side effect of trying to do the embedded agency stuff. But it's kind of not a side effect because I think that the fact that I'm trying to do a bunch of embedded agency stuff and then I… I was trying to figure out stuff related to time and related to decision theory and agents modeling themselves and each other. And I feel like I stumbled into something that might be useful for identifying good concepts, like blue.

Scott Garrabrant: And I think that that stumbling is part of the motivation. I don't know, that stumbling is part of the reason why I'm thinking so abstractly. That's not a motivation for thinking about embedded agency. That's a motivation for thinking as abstractly as I am, because you might get far reaching consequences out of the abstraction.

How Scott researches

Daniel Filan: All right, so I guess a few other questions to kind of get at this. What do you do? What is a day of Scott researching look like?

Scott Garrabrant: I mean, recently it's been thinking about presentation of factored set stuff. Often it involves thinking in Overleaf or something where I'm just writing some stuff up and then I have thoughts as a consequence of the writing. Often it looks like talking to people about different formalisms and different weird philosophy. Yeah, I don't know.

Daniel Filan: So you're thinking about presentation of this work? What are you trying to get right or not get wrong? What are the problems that you're trying to solve in the presentation?

Scott Garrabrant: I think that a large part of the presentation thing is I want to wrap everything up so that it feels like something that can just be used without thinking about it too much or something that.

Scott Garrabrant: Part of the presentation is some hope that maybe it can have large consequences to the way that people think about structure learning. But mostly it's kind of having it be a basic tool that I can then kind of… I've kind of locked in some of the formalism such that I don't have to think about these details as much and I can think about the things that are built on top of them.

Scott Garrabrant: I don't know, I think that in thinking about this presentation or something, it's not where the interesting work is done. I think that the interesting… The part that had a lot of interesting meat in terms of actually how research is done was a lot of the stuff that I did late last year, which was kind of, okay I finally wrapped up Cartesian frames. What is it missing? And it largely was - I had this orientation that was Cartesian frames kind of feel like they're doing the wrong thing similar to… Or sorry, like… All right. So here's a story that I can kind of tell, which is I was looking at Cartesian frames, which is some earlier work from last year. And part of the thing was you viewed this world as a binary function from an agent's choice and an environment's choice, or an agent's way of being, or an agent's action to the environment's way of being… Sorry, cross the environment's way of being to the full world state. And a large part of the motivation was around taking some stuff that was kind of treated as primitive and making it more derived. In particular I was trying to make time more derived and some other things, but I was trying to make time feel more derived so that I can kind of do some reductionism or something.

Scott Garrabrant: And at the end of Cartesian frames, I was unsatisfied because it felt like the binary function… The function from A cross E to W was itself derived, but not treated that way.

Scott Garrabrant: When I look at a function from A cross E to W, I don't want to think of it as a function. I want to think of it as like, well, there's this object A cross E, and there's this object W, and there's a relation between them. And then there's that relation kind of satisfies the axioms of a function, which is for each way of choosing an A cross E there exists a W. But then I also wanted to say, well, it's not just a function from A cross E to W, where A cross E is a single object.

Scott Garrabrant: There's this other thing, which is I have this space, A cross E, and I'm specifically viewing it as a product of A and E. And what was going on there was, it

felt like in my function from A cross E to W, I did not just have it's a function not a relation, I also have this system of kind of interventions, where I could imagine tweaking the A bit, and tweaking the E bit independently. And the product A cross E as an object, A cross E, it has the structure of a product. And I was trying to figure out what was going on there in a way that - the same way that you can view the function as just a relation that satisfies some extra conditions.

Scott Garrabrant: I wanted to view the product as some extra conditions. And those extra conditions were basically what kind of grew into me being really interested in understanding the combinatorial notion of orthogonality. And so, I was dissatisfied with something being not quite philosophically right or not quite derived enough or something, and I double clicked on that a bunch.

Daniel Filan: Okay, so another question that I want to ask is, so you work at MIRI, the Machine Intelligence Research Institute, and, I think, among people who are trying to reduce existential risks from AI, as a shorthand, people often talk about the MIRI way of viewing things and the thoughts that MIRI has, or something. I also work at CHAI, so CHAI is the Center for Human Compatible AI and sometimes people talk about the CHAI way of doing things and that always makes me mad because I'm an individual, damn it. How do you think, if people are modeling you as just one of the MIRI people, basically identical to Abram Demski, but with a different hairstyle, what do you think people will get wrong?

Scott Garrabrant: Yeah, so I've been doing a lot of individual work recently, so it's not like I'm working very tightly with a bunch of people, but there is still something to be said for, well, even if people aren't working tightly together, they have similar ways of looking at things.

Daniel Filan: Maybe in a world where they really understand Abram and the rest of the people.

Scott Garrabrant: Yeah, I can point at concrete disagreements or differences in methodology or something. I think that Abram and I have some disagreements about time. I think that there's a thing where Abram, I think Abram more than anybody else, is taking logical induction seriously and kind of doing a bunch of work in the field that is generated by and exemplified by logical induction. And I look at logical induction and I'm like, "You're just putting all this stuff on top of time and I don't know what time is yet. I need to go back and reinvent everything, because it's built on top of something and I don't like what it's built on top of."

Scott Garrabrant: And so, I end up being a lot more disconnected and a lot more pushing towards… I don't know, I think that Abram's work will tend to be more on the surface feel directed towards the thing that he's trying to do or something. And I will kind of just keep going backwards into the abstract or something. I also think that there are a lot of similarities between the way of thinking that people in MIRI share, and also some large subset of people in AI safety in general, that they're just a bunch of people that I can kind of predictably expect, if I come up with a new insight and I want to communicate it to them, it'll go kind of well and quickly, not just because they're smart, because they're on the same background, there's less inferential gap. But yeah, people are individuals.

Daniel Filan: That's true. Yeah, so speaking of this desire to make things derived, like where does time come from and such? What do you think you're happy to see as just primitive?

Scott Garrabrant: I don't think it's a what. I think it's something like taking something that you're working with that you think is important and doing reductionism on it, is a useful tool when you have something that is both critical, like you need to understand this thing and there's actual mutual information between this thing that I'm holding and stuff that I care about. And also, it feels like this thing that I'm holding has all these mistakes in it, or has all these inconsistencies, right?

Scott Garrabrant: It's like, why be interested in something like decision theory? Well, decisions are important and also, if you zoom in at them and look at the edge cases, you can kind of see they're built on top of something that feels kind of hacky. And then a thing that you can do is you can say, "Well, what are they built out of?" Yeah, you can try to do some sort of reductionism. And so, it's more a move for when things aren't clicking together nicely. I don't think of reductionism as get down to the atoms, I think of reductionism as the pieces don't fit together correctly, go down one more step and see what's going on or something.

Daniel Filan: So, a related question, it might be too direct, but suppose a listener wants to develop an inner Scott, they want to be able to, "What would Scott say or think about such and such topic?" Just restricted to the topic of reducing existential risk from AI. What do you think the most important opinions and patterns of thought are to get right, that you haven't already explicitly said?

Scott Garrabrant: So, it depends on whether they want an inner Scott for predicting Scott or whether they want an inner Scott for just generally giving them useful ideas or something, in the space of being a thing to bounce things off of and say, "I want to understand X more." One question is, "What would Scott say about X?" And it's not actually important that it matches and it's more important, does it generate useful thoughts? Which is generally what I do with my models of people. I have inner people and then sometimes I find out that they don't exactly match the outer people. And I don't care that much, because their main purpose is to give me thoughts, so I want to make it better, but it's not for prediction, it's for ideas.

Scott Garrabrant: So, I have an inner Scot and my inner Scott is kind of a little bit being rewritten now, because a large part of my inner Scott was kind of identifying with logical induction. And I actually do this for a lot of people. I think about their thought patterns as in relation to things that they've developed. And so, if I were to tell that story, I would say things like part of the thing in logical induction is that you don't make the sub-agents have full stories. A large part of what's going on in logical induction is, it's a way of ensembling different opinions where you don't require that each individual opinion can answer all the questions. You allow them to specialize and you allow them to fail to be able to model things in some domains. And you want them to be able to track what they fail to be able to model. And so, I have a lot of that going on, where I just have kind of boxed fake frameworks in my head where I'm just very comfortable drawing analogies to all sorts of stuff.

Scott Garrabrant: And I don't know. I, for example, wrote a blog post on what does the Magic: The Gathering color wheel say on AI safety or something. I do that kind of thing, where I'm just like, "Here's a model, it's useful for me to be able to think with" or something. And I'm not trusting it in these ways, but I'm kind of trusting it as being generative in certain ways. And I keep on working with it as long as it's generative. Yeah, what am I saying? I'm saying that I tend to think that if something is fruitful and creating good ideas, but also being wrong in lots of ways, I wouldn't say don't mess with it, but if messing with it breaks it, undo that messing with it and let it be wrong

and still be fruitful or something. And so, I tend to work with obviously wrong thoughts or something like that.

Daniel Filan: Okay. Another question about intellectual production is, there's this idea of complements to something, or complements to some production process or things that are not exactly, maybe I partially mean inputs, or inputs to the process or things separate from the process, that make the process better. So, in the case of Scott Garrabrant doing research, what are the best complements to it?

Scott Garrabrant: I think that isolation has been pretty good actually. Does that count? Is that a complement? Is that the type of complement?

Daniel Filan: I just realized that I've been kind of confused about what complement is, but it's at least an input. What's been good about isolation?

Scott Garrabrant: I don't know. I think I've just largely been thinking by myself for the last year, as opposed to thinking with other people and it felt like it was good for me for this year or something. I might want to go back to something else. And why? I mean, I think there is a thing where I have in the past made mistakes of the form, trying to average myself with the people around me in terms of what to think about things, and this is dampening, just because of the law of large numbers, I guess.

Daniel Filan: It's the heat equation, right? Everyone averages everyone else, and things become uniform.

Scott Garrabrant: Right. There's a sense in which working with other people is grounding in that it keeps on giving feedback on things, but I don't know, grounding has good things and has bad things associated with it. And one of the ways in which it has bad things associated with it is that, I don't know, it's like things can flow less or something.

Daniel Filan: Yeah, it's funny, I don't exactly know what grounding is in the social sense. I recently read a good blog post about it, but I totally forgot. I mean, there is one sense in which, so if I think about literal physical grounding in electronics, the point of that is to equalize your electrical potential with the electrical potential of the ground, so that you don't build up this big potential difference and then have someone else touch it and touch the ground and have some crazy thing happen. But as long as you have the same potential as the ground, it means that there's not a net force for charges to move from the ground to you or vice versa, but it does mean that things can sort of move in both ways.

Daniel Filan: And, I don't know, I think maybe there's an analogy here of something about being averaged with a bunch of people, I guess, it sort of forces you to develop a communication protocol or common language or something that somehow facilitates flow of ideas or whatever. And just directly, because other people have ideas and you average them into you as part of like some kind of average, maybe literal averaging is not right. I'm kind of blabbering on. Does any of that resonate?

Scott Garrabrant: I'm not sure.

Daniel Filan: Okay, so we can move on. Is there anything else that I should have asked?

Scott Garrabrant: Yeah. I mean, I have a large space of thoughts, which I don't even know what exactly I'd say next or something, about how I plan to use factored sets, I

think, because I think it actually does differ quite a bit from the use case in the paper slash video slash whatever blog post. That's one thing that comes to mind. Let me keep thinking for more. Yeah, I guess that's the main thing that comes to mind.

Daniel Filan: Okay, how do you plan to use it?

Scott Garrabrant: I mean, so one piece of my plan is that, I talk a bunch about probability and I don't really plan on working with probability very much in the future. A large part of the thing is I'm kind of pulling out a combinatorial fragment of probability, or combinatorial fragment of independence, so that I can avoid thinking of things with probability or something like that. I don't really talk about probability in Cartesian frames and a lot of the stuff in Cartesian frames I think can, slash, I hope to, port over to factored sets. It's largely, there are lots of places where I'd want to draw a DAG, but I'd want to never mention probability. Or maybe I could mention probability, maybe I could think of things as sometimes being grounded in probability sometimes not, but I want to draw DAGs all over the place and I have this suspicion that, well, maybe places where I'm drawing DAGs, I could instead think in terms of factored sets.

Scott Garrabrant: Although, I have to admit, DAGs are still useful, I have this example where I can infer some time in factored sets that's not the one that I give in the talk. And when I think about it, I have a graph in my head, that's like the Pearlian picture, which maybe has something to do with the fact that you're saying that Pearl can be visualized, but it definitely feels like I haven't fully ported my head over to thinking in factored set world, which seems like a bad sign, but it also doesn't seem like that strong a bad sign, because it's new.

Daniel Filan: Yeah. I mean, graphs do have this nice, somehow if you want to understand dependence, it's just so easy to say this thing depends on this thing, which depends on these three things and it's very nice to draw that as a graph.

Scott Garrabrant: Yeah. I mean, I'm more thinking in terms of screening off. Screening off is a nice picture where you imagine getting in on a path and you kind of block the path and you kind of condition on something on the path and then information can't flow across anymore.

Daniel Filan: Yeah, so a variable screens something off. Yeah, can you just say what screening off means?

Scott Garrabrant: I mean, the way I'm using it here, I'm mostly just saying X screens off Y from Z, if Y and Z are orthogonal, given X, where orthogonal could mean many different things. It could mean graphs, it could mean independence, it could mean the thing in factoring sets.

Daniel Filan: Yeah. Yeah, I guess the idea of paths and graphs, gives you this kind of nice way of thinking about screening off.

Scott Garrabrant: Yeah. So, I do feel like I can't really picture conditional orthogonality as well as I can picture D-separation, even though I can give a definition that's shorter than D-separation that captures a lot of the things.

Relation to Cartesian frames

Daniel Filan: So yeah, speaking of Cartesian frames, so Cartesian frames is I guess, a framework you worked on, as you said last year, and one thing that existed in

Cartesian frames was it had this notion of sub-agency where it was, if you had an agent in an environment, you'd kind of talk about what it meant to view it as like somehow a collection or a composition of sub-agents.

Daniel Filan: So yeah, this question, we're just going to assume that listeners basically get the definition of that and you can skip ahead to the last question if you want to look that up or don't want to bother with that. But I'm wondering, so these finite factored sets, it's kind of easy to see the world being a product of the agent and the environment, that's kind of like this factorization thing. I'm wondering how you think the notion of the sub-agents thing goes into it? Because that was the thing I was kind of… It was kind of the most interesting part of Cartesian frames.

Scott Garrabrant: Yeah, so I think that I actually do have to say something about the definition of sub-agent to answer this, which is in Cartesian frames, I gave multiple definitions of sub-agent and one of them was kind of opaque, but very short. And I kind of mutually justified things as like, "Ah, this is pointing to something you care about, because it agrees with this other definition." But it's really carving it at the joints, because it's so simple.

Scott Garrabrant: So, it shouldn't be clear why this is a sub-agent, but in Cartesian frames, I had a thing that was like C is a sub-agent of D if every morphism from C to $\perp$ factors through D. And when you think about $\perp$, there's a sense in which you can kind of think of $\perp$ as like the world, because $\perp$ is the thing where the agent kind of just gets to choose a world and the environment doesn't do anything. And so, you can view this thing as saying every morphism from C to $\perp$ factors through D and I think this translates pretty nicely. It's not symmetric in the Cartesian frame thing, but I think it translates pretty nicely to, D screens off C from the world. C is a sub-agent of D, means that D screens off C from the world. In the Cartesian frame thing, it's not a symmetric notion. When I convert to factored sets, it becomes a symmetric notion and maybe there's something lossy there.

Daniel Filan: And by screening off, you mean just conditional orthogonality?

Scott Garrabrant: I mean conditional orthogonality. I'm saying "factoring through", saying a function factors through an object is similar to a screening off notion. And the way that I define sub-agency in factored sets looks like this. So, I can say more about what I say about the world. The world is maybe some high level world model that we care about, so we have some partition of our finite factored set W, which is kind of representing stuff that we care about. And we have some partition that's maybe like, you have some partition D which corresponds to the super-agent and the choices made by the super-agent. And we also have some partition C, which corresponds to the choices made by our sub-agent. And so, you can think of maybe D as a large team and C is like one sub part of that team.

Scott Garrabrant: And if you imagine that the large team has this channel through which it interacts with the world, and that D kind of represents the output of the large team to the world, but then internally, it has some internal discussions, but those internal discussions never kind of leave the team's internal discussion platform or whatever. Then there's a sense in which, if I want to, if I know, if the team is a very tight team and C doesn't really have any interaction with the world besides through the official channels that are D, then if I want to know about the world, once I know about the output of D, learning more about C doesn't really help, which is saying that C is orthogonal to the world, given D.

Daniel Filan: How is that symmetric, because normally, if X is orthogonal to Y given Z, it's not also the case that Y is orthogonal to Z given X, right?

Scott Garrabrant: Sorry, it's symmetric with respect to… You're not replacing the given.

Daniel Filan: Oh, okay.

Scott Garrabrant: By symmetric, it's symmetric with respect to swapping… Yeah, when I said symmetric, obviously I should've meant, it's symmetric with respect to swapping C with W and D is in the middle, which, what does it mean to swap C with W?

Daniel Filan: That's kind of strange.

Scott Garrabrant: It's capturing something about being a sub-agent means that the interface of the super-agent is kind of screening off all of your stuff. And one way to see this is if we weren't working with, I was thinking of this in terms of like W is everything we care about. But if we weren't thinking about W is everything we care about, if we just took any partition X and any other partition Y and we let C be equal to X, and we let D be equal to the common refinement of X and Y, then D screes off C from the world.

Scott Garrabrant: So if you take any two choices, any two partitions, and you just put them together, you get a super-agent under this definition. And you can kind of combine any partitions that are kind of representing some choices or something maybe, and you can combine them and you get super-agents. But that's super-agent with respect to the whole world. And then as you take a more restricted world, now you can have sub-agents that are not just one piece of many pieces, but instead, maybe the sub-agent can have some internal thoughts that don't actually affect the world and are not captured in the super-agent, which is maybe only capturing some more external stuff.

Daniel Filan: Yeah. I guess one thing that comes to mind is that… Yeah, so we have this weird thing where the definition of a sub-agent you could swap out the sub-agent with the rest of the world, because we were thinking of an agent as like a partition, probably not all, just the choice of an agent, maybe, as a partition, but probably not all partitions, not all variables, should get to count as being an agent, right? And I'm wondering if there's some restrictions you could place on what counts as an agent at all, that would break that symmetry?

Scott Garrabrant: Yeah, you could. I don't actually have a good reason to want to here, I think. I think that part of what I'm trying to build up is to not have to make that choice of what counts as an agent and what doesn't. I don't know, I can define this sub-agent thing and I can define things like this partition observes this other partition, so embedded observation, which I haven't explained. And I feel like it's useful that I can give these definitions and they extend to these other partitions that we don't want to think of as agentic either. I actually feel a little more confused in the factored set world about how to define agents than I did before the factored set world, because if I'm trying to define agency my kind of go to thing is agency is time travel. It's this mechanism through which the future affects the past through the agent's modeling and optimization.

Scott Garrabrant: And now I'm like, well, part of the point of factored sets is I was trying to actually understand the real time, such that time travel doesn't make sense

as much anymore. And one hope that I have for saving this definition and thinking about what is an agent in the factored set world, is the factored set world leads to multiple different ways of defining time. And so, just like we want to say there's some sort of internal to the agent notion of time, where it feels like the fact that I'm going to eat some food causes me to drive to the store or something. So like in internal to the agent there's some time and then there's also the time of physics. And so, one way you could think of agency is where there's kind of different notions of time that disagree.

Scott Garrabrant: And there's a hope for having a good system of different notions of time in factored sets that comes from the fact that we can just define conditional time, the same way we define conditional orthogonality. We can just imagine taking a factored set, taking some condition, and now we have a new structure of time in the conditioned object and it might disagree. And so, you might be able to say something like agents will tend to have different versions of their time disagree with each other, and this might be able to be made formal. I don't know, this is a vague hope.

How to follow Scott's work

Daniel Filan: All right, cool. Maybe that gives people ideas for how to extend this or for work to do. So yeah, I guess the final question I would like to ask is, if people have listened to this and they're interested in following you and your work, how should they do so?

Scott Garrabrant: Yeah, so specifically for finite factored sets, everything that I've put out so far is on LessWrong. And so, you could Google some combination of my name, Scott Garrabrant and LessWrong and finite factored sets. And I intend for that to be true in the future. I intend to keep posting stuff on LessWrong related to this and probably related to future stuff that I do. Yeah, I tend to have big chunks of output rarely. Yeah, I haven't posted much on LessWrong since posting Cartesian frames and I'm currently planning on posting a bunch more in the near future related to factored sets and that'll all be on LessWrong.

Daniel Filan: All right, Scott, thanks for being on the podcast and to the listeners. I hope you join us again.

Scott Garrabrant: Thank you.

Daniel Filan: This episode is edited by Finan Adamson. The financial costs of making this episode are covered by a grant from the Long Term Future Fund. To read a transcript of this episode, or to learn how to support the podcast, you can visit axrp.net. Finally, if you have any feedback about this podcast, you can email me at feedback@axrp.net.

Finite Factored Sets in Pictures

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

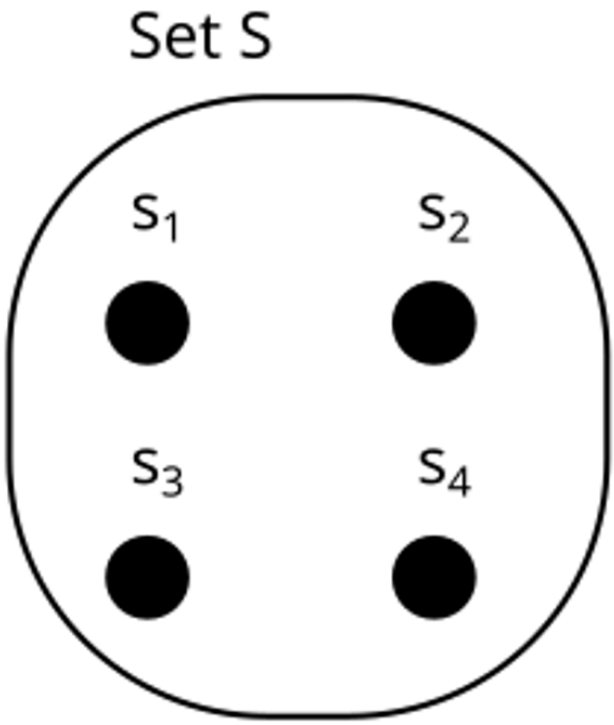
Finite factored sets are a new paradigm for talking about causality. You can use them to do some cool things you can't do with Pearl's causal graphs, for example inferring a causal arrow between two binary variables.

Also, finite factored sets are a really neat mathematical structure: they are a way of taking a **set** and expressing it as a *product of some factors.* Set factorizations are analogous to integer factorizations, in the same way that set partitions are analogous to integer partitions.

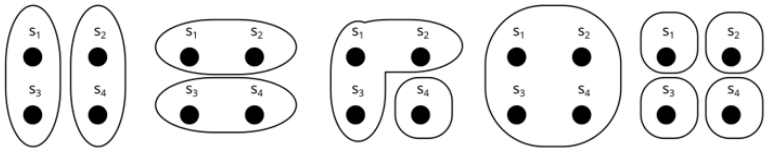
So, here is my current understanding of finite factored sets, in pictures.

1. What are Set Factorizations?

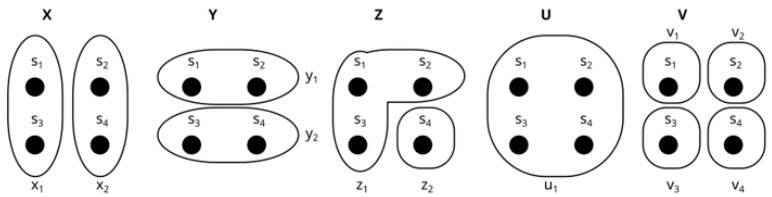
What do these "factored sets" look like? Let's start with a set S and factor it.



The first concept we need is a partition of a set S. A partition is a way of chopping up S into subsets (called *parts*). Here are a few examples of partitions:

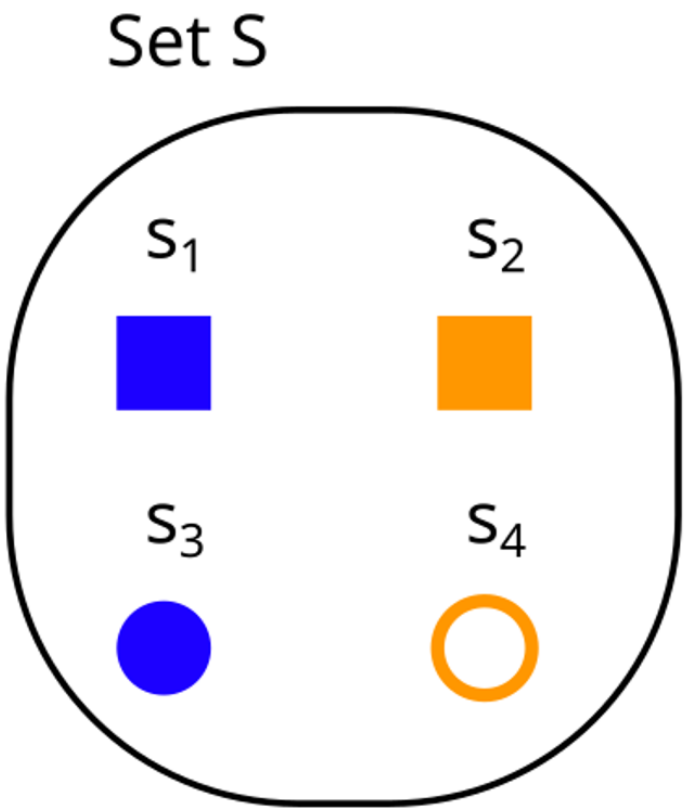


We usually call the partitions X, Y, Z, U, V, or W, and their parts $x_i, y_i, \ldots$ like this:

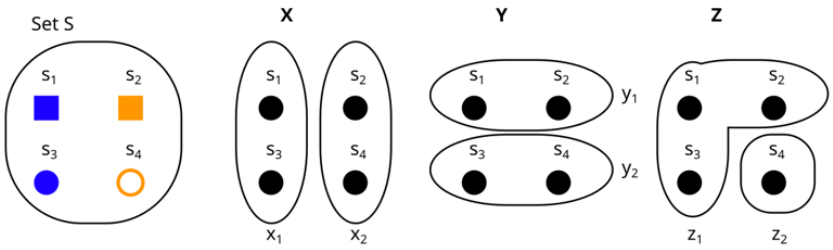


U is called the *trivial partition*. It only has one part.

We can think of **partitions as properties, or variables** over our set. For example, consider a set like this:



and compare it to the partitions X, Y and Z from above:

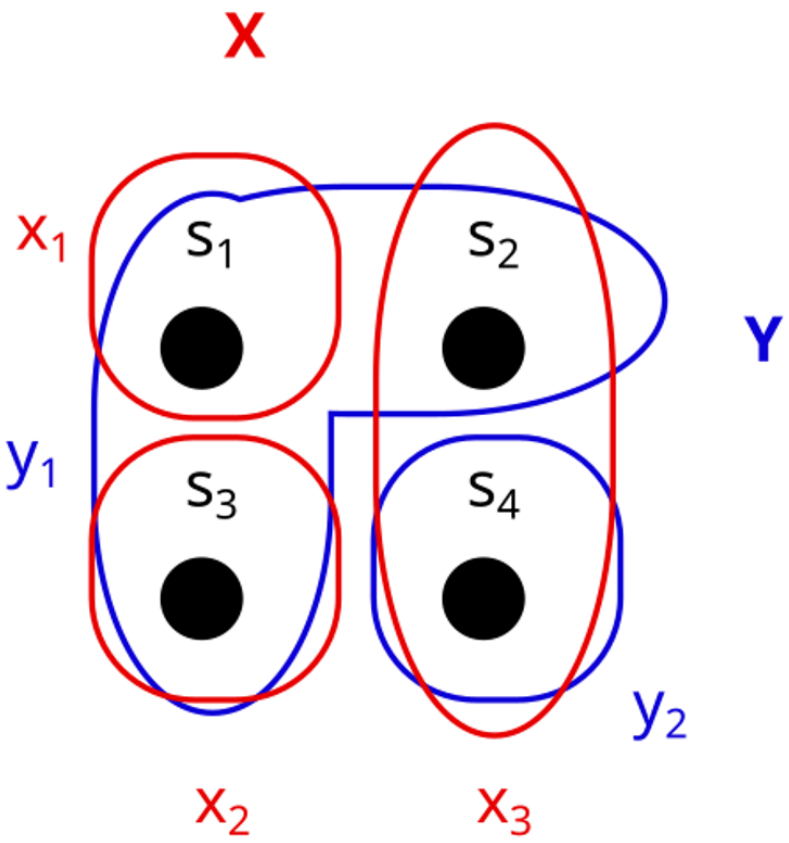


Then

- The partition X is the property “color”, with x_1 = blue and x_2 = orange.
- The partition Y is the property “form” with y_1 = square, and y_2 = circle.
- The partition Z is the property “filled” with z_1 = yes, and z_2 = no.

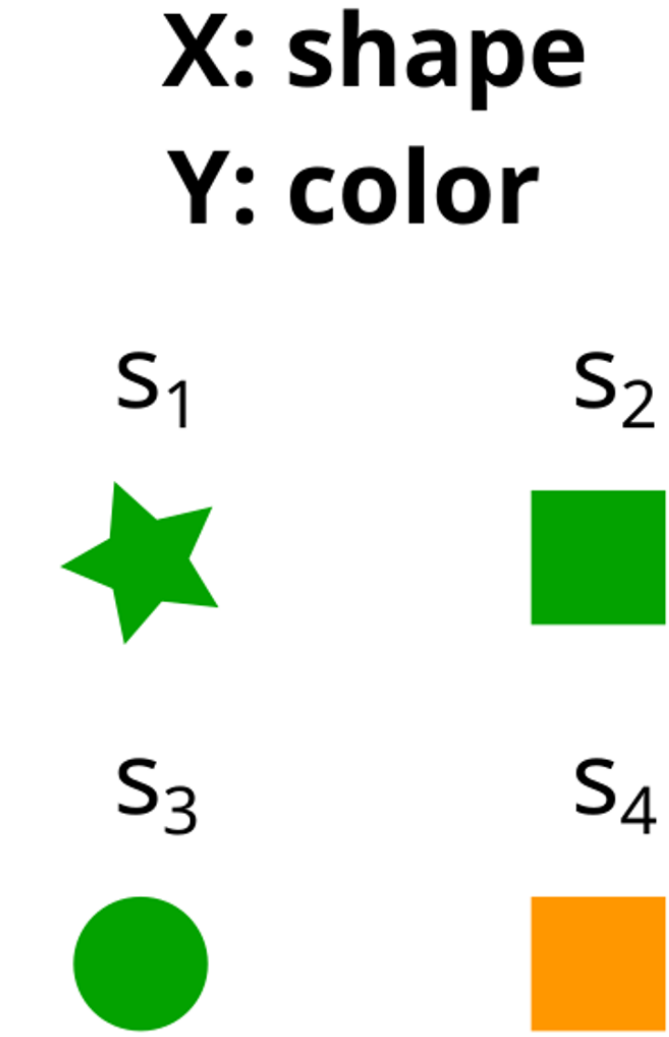
Exercise

Consider these two partitions X and Y on the set S. What would it look like to represent them as properties (e.g. X = shape, Y = color) instead?



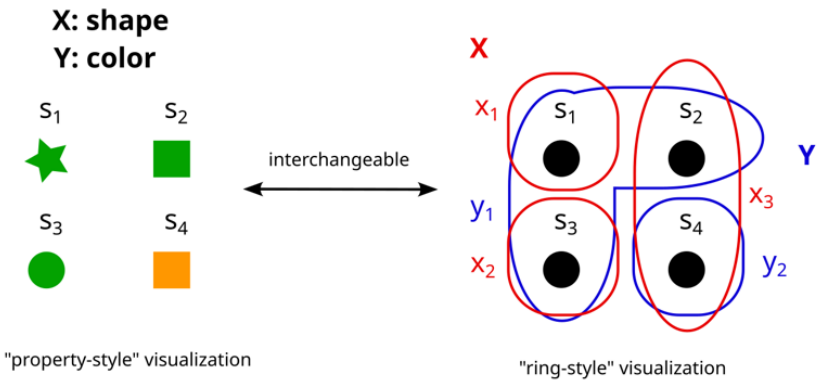
Spoiler space

It could look something like this:



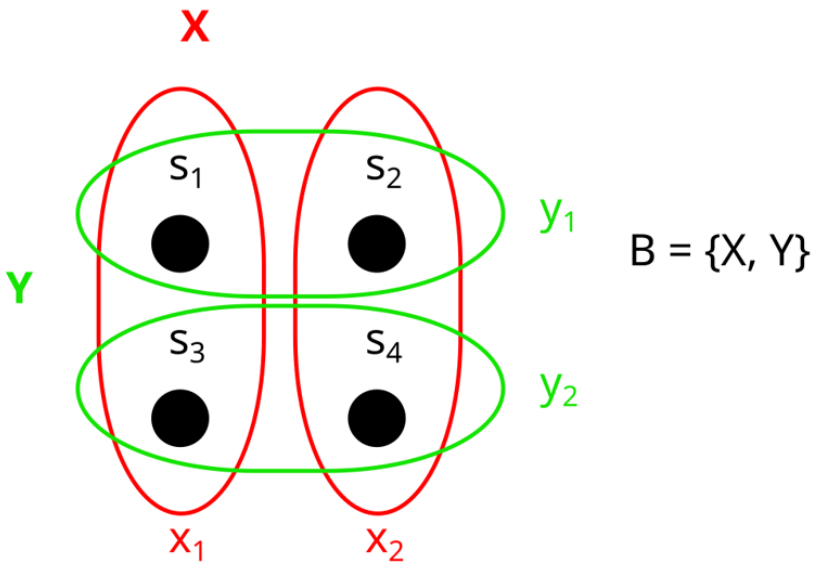
Here X = shape = $\{x_1, x_2, x_3\}$ = {star, circle, square}, and Y = color = $\{y_1, y_2\}$ = {green, orange}.

I hope you can see how partitions and properties are basically the same thing. In the rest of this post, I will use “partitions” and “properties” interchangeably. Sometimes I will use the ring-style visualization of partitions, and sometimes the property style, depending on what I find more intuitive in any given example.



Now we can define **set factorizations**:

A **factorization** B of our set S looks like this:



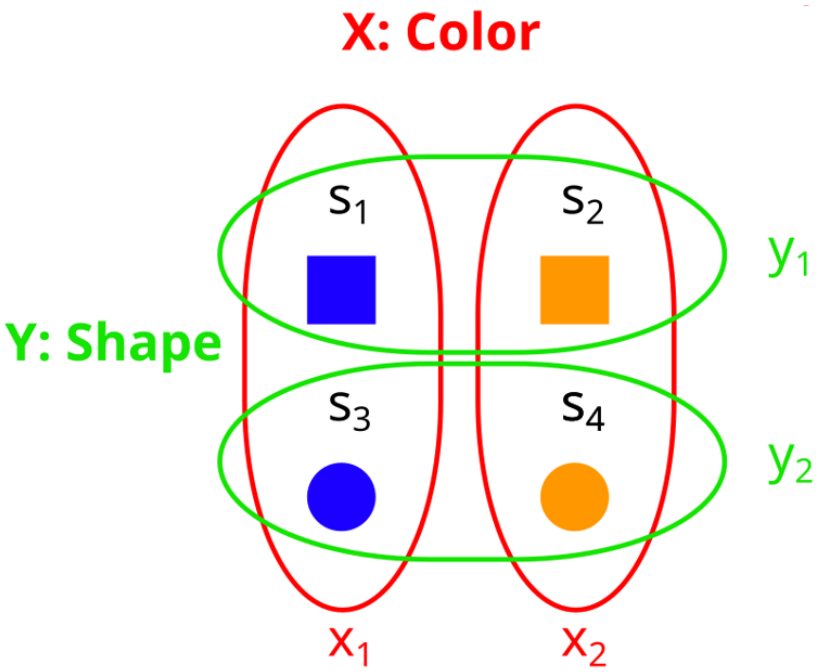
A factorization is a set $B = \{b_1, b_2, \dots, b_n\}$ of partitions (called *factors*). In this case $B = \{b_1, b_2\} = \{X, Y\} = \{\{x_1, x_2\}, \{y_1, y_2\}\}$.

But it can't just be *any* set of partitions. In the following sections, I will explain the two conditions that B needs to fulfill in order to count as a factorization:

1. There is a unique element for all combinations of properties
2. No factor is trivial

1. There is a unique element for all combinations of properties

Let's look at our partitions in terms of properties again:



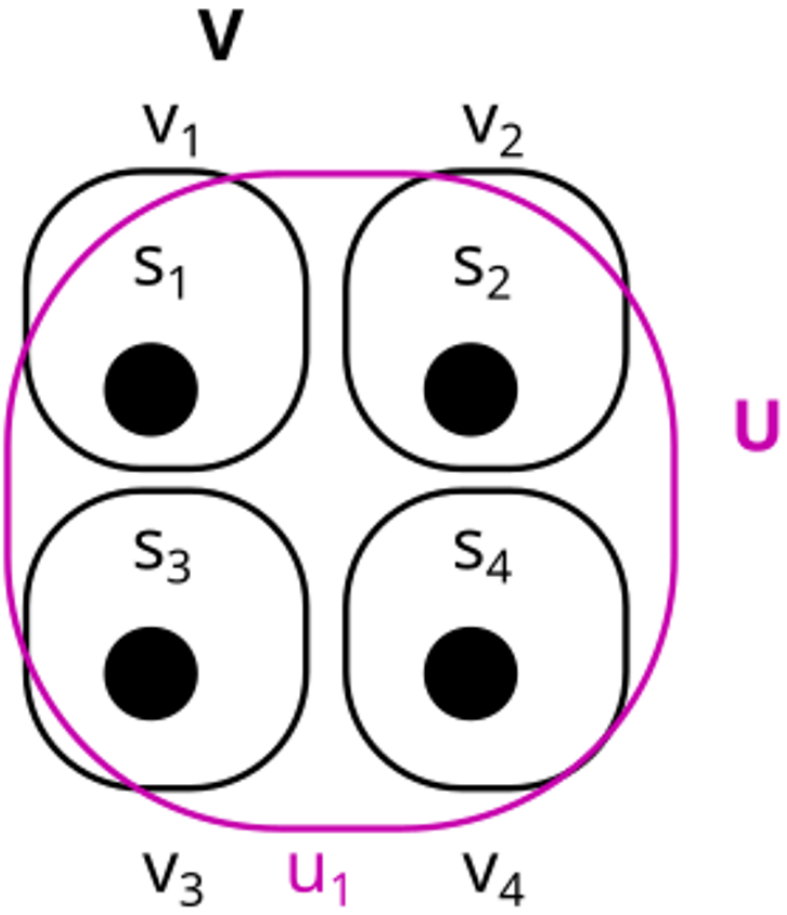
What we need in order for B to be a factorization, is that for all combinations(x_i, y_j) of properties (for example (x_1, y_2), which is (blue, circle)), there is a unique element with these properties.

We can see that this is the case here: We have exactly one blue square, exactly one orange square, exactly one blue circle, and exactly one orange circle.

To express it more mathematically: For $B = \{X_1, X_2, \dots, X_n\}$ to be a factorization, we need that for all $x_1 \in X_1, x_2 \in X_2, \dots x_n \in X_n$, it holds that the intersection of $\{x_1, x_2, \dots, x_n\}$ contains exactly one element. This means the cartesian product of our factors is bijective to the set S, which justifies that we say we can "express S as the *product* of our factors".

2. No factor is trivial

Here is an example of a non-factorization:

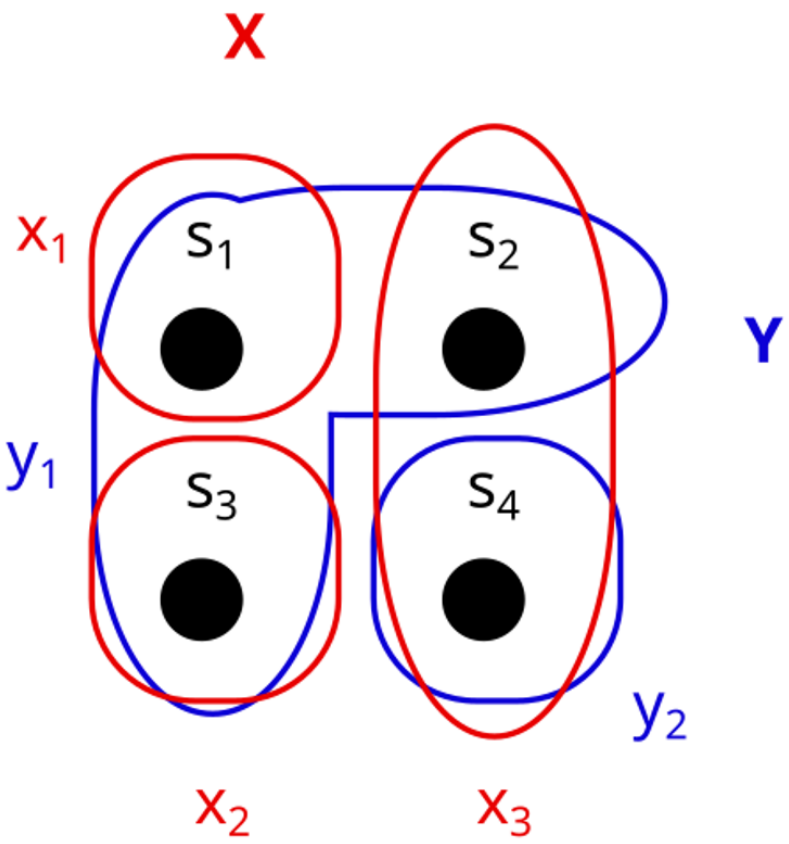


B = {U, V} is not a factorization here, because U is trivial and factors aren't allowed to be trivial.

This is analogous to integer factorization, where we don't count 1 as a factor. For example, for the integer 6 we say the factorizations are {6} and {2,3}, and don't mention {6,1} and {6,1,1} and {6,1,1,1} and so on.

Exercise

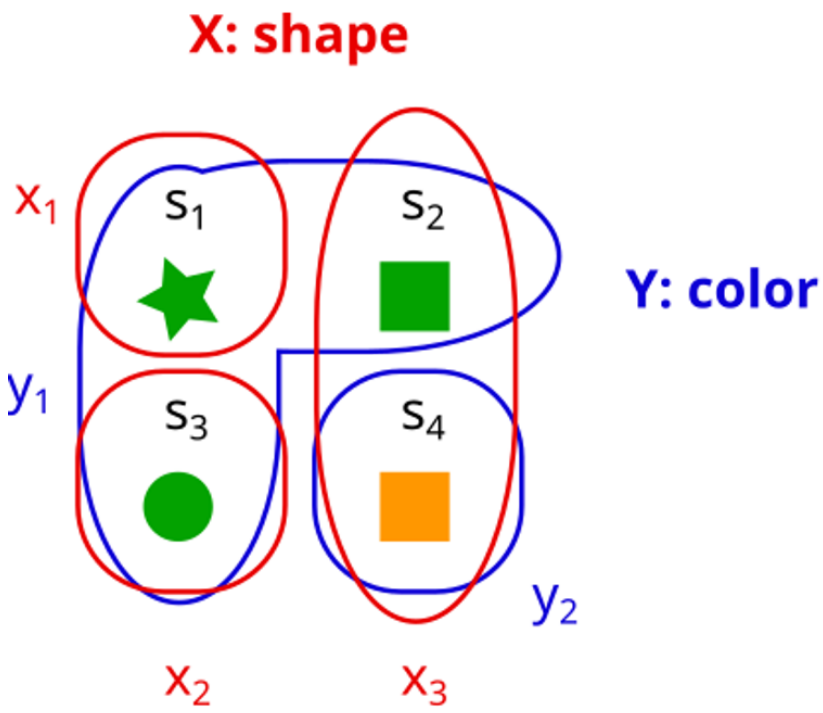
What about this? Is B = {X, Y} a factorization here? (take a moment to think for yourself)



more spoiler space

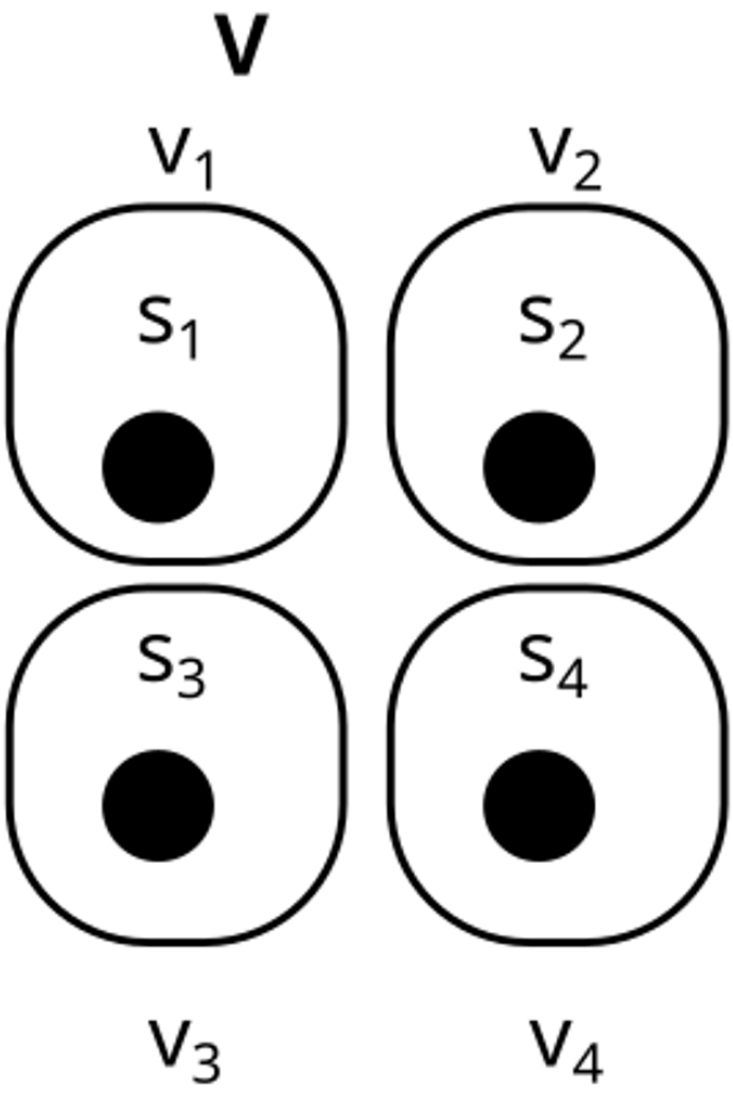
... No. B is not a factorization.

Why not? Let’s look at it in terms of properties again:



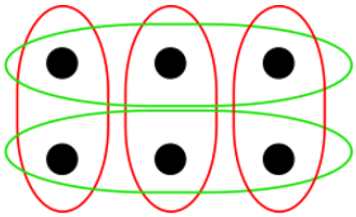
We can see that there is not a unique element for every combination of properties. For example, there is no orange star, and also no orange circle (i.e. $x_1 \cap y_2$ and $x_2 \cap y_2$ are empty).

What about this one? Is B = {V} a factorization?

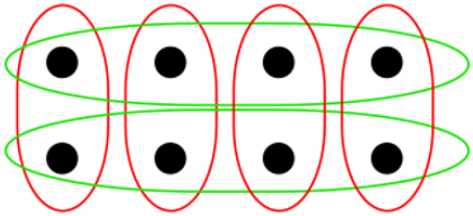


… yes it is. This one is called *trivial factorization*. Each set can be factorized as B = $\{b_1\}$ with b_1 being the maximally separating partition.

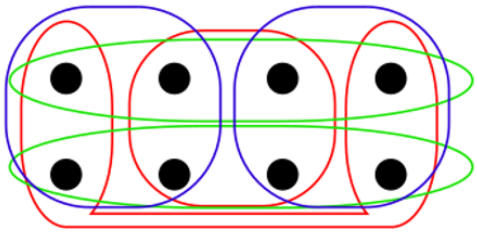
If you play around a bit with sets of different sizes, you will see that the possible set factorizations correspond to the integer factorizations of the set's size [1]:



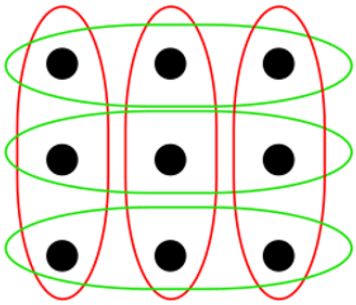
$6 = 2*3$



$8 = 2*4$



$8 = 2*2*2$



$9 = 3*3$

In particular, sets with a prime number of elements only have the trivial factorization.

This concludes the examples for factorizations. Hopefully you now have some grasp on how factoring a set works.

If we have a tuple F = (S, B) of a set S and a factorization B of S, then we call F a **factored set**. In this post I will assume that all sets are finite, and use “factored set” synonymously with “finite factored set”

2. What does this have to do with Causality? - The Building Blocks

In this section, I will introduce three building blocks: three structures on factored sets, that will help us make the connection to causality later.

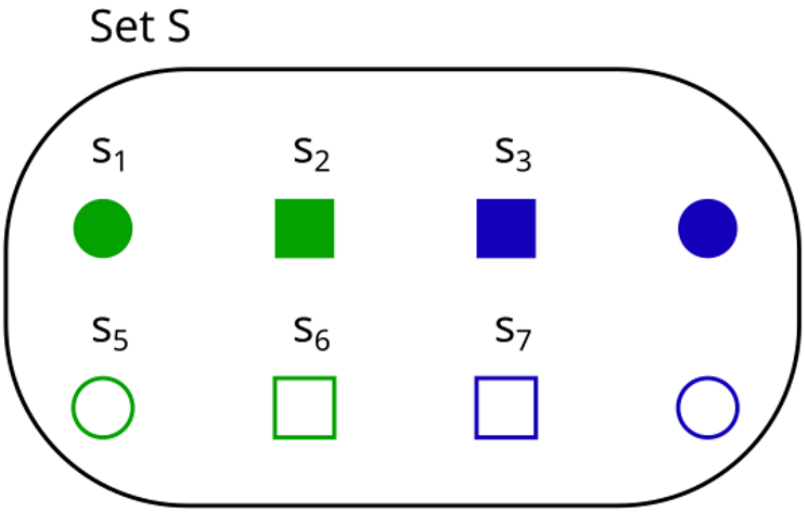
In section 3, I will then use these building blocks to model the causal structure behind a probability distribution with factored sets, similar to causal graphs.

The three building blocks are:

1. The **history** of a partition X is related to a random variable’s set of **ancestors** in a causal graph
2. **Orthogonality** of two partitions X, Y means they have no shared history. In the Pearl-paradigm we know that two variables X and Y have no common ancestors if and only if they are independent. Analogously, Scott Garrabrant proved that in the factored set paradigm, two variables are **orthogonal if and only if they are independent**. [2]
    - **Conditional orthogonality** of two partitions: In the Pearl-paradigm we know that two variables X and Y are d-separated if and only if they are conditionally independent (proof here and here). Analogously, Scott proved that in the factored set paradigm, two variables are conditionally orthogonal if and only if they are conditionally independent. [2]
3. **"Time"**: Saying a partition A is *before* B is related to a **causal path** going from A to B in a causal graph.

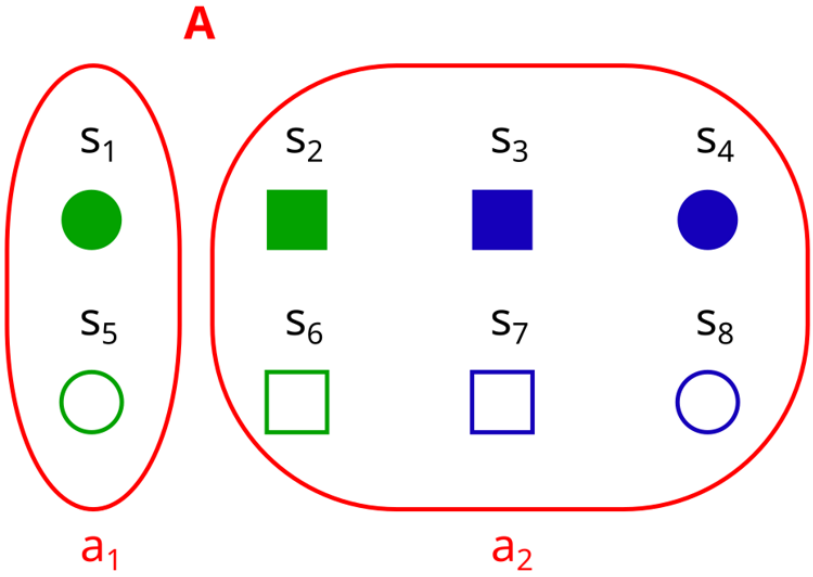
History

Consider an 8-element set S, which is factorized into the factors “color”, “shape” and “fill”, like this:



Here, the factored set F is F = (S, B) with the factorization B = {color, shape, fill}.

Now, let’s consider a partition/property A - which does not need to be a factor! (i.e. A does not have to be color, shape or fill here):



Now assume we know some properties of an element s, and want to figure out if it is in a_1 or in a_2. The fill doesn’t matter for this, so the minimum required properties for finding out are {color, shape}.

If we know that the color is blue, then the color would be enough to determine that we are in a_2, but in order to *reliably* find out if we are in a_2, we need both color and shape.

This is what we will call the **history** of A: We say that in a factored set F, the history $h^F(A)$ of a property A is the set of properties we need in order to figure out A.

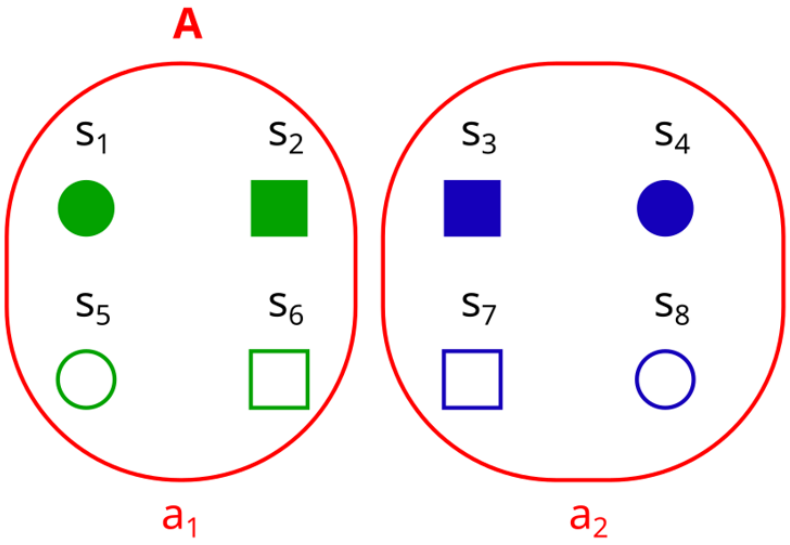
Once we build up factored sets as a model for a causal structure, the history of a partition A will correspond to the set of ancestors of a random variable in a causal graph.

Note that I represent A as these red rings, and {color, shape, fill} as properties. I could just as well represent A as a property too (e.g. different sizes), but I prefer this representation because it distinguishes the factors in our factorization {color, shape, fill}

from the variable A whose history we want to find.

Exercise

The history $h^F(A)$ of a property A is the set of properties we need in order to figure out A. So, what is the $h^F(A)$ here?

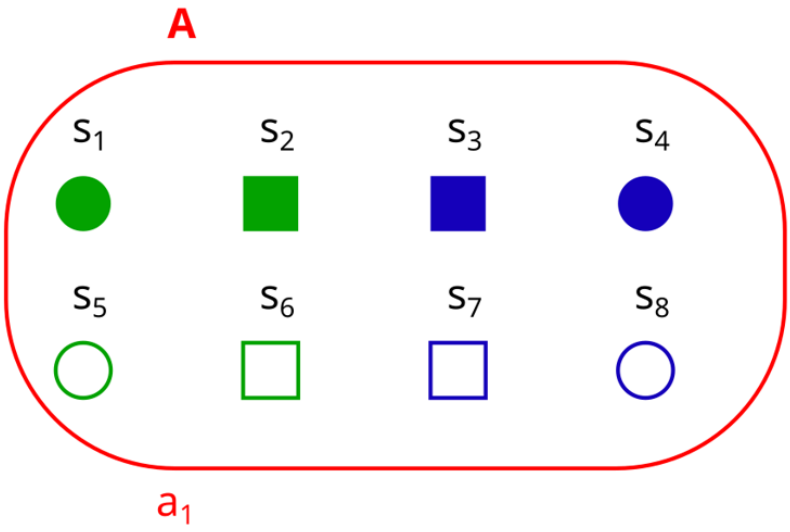


$h^F(A)$ = {color}

You only need to know the color in order to tell if an element is in a_1 or in a_2.

Notice that in this case, A = color. So $h^F(A)$= {color} is the same as h^F(color) = {color}. Which is basically just saying that you just need to know the color in order to find out the color. In general, for every factor b in our factorization, it holds that $h^F(b)$ = {b}.

Another exercise: What if A is the trivial partition? What is $h^F(A)$ here?



Here, the history is empty! ($h^F(A) = \emptyset$) We don't need to know any properties, because we *already know* that every s is in a_1.

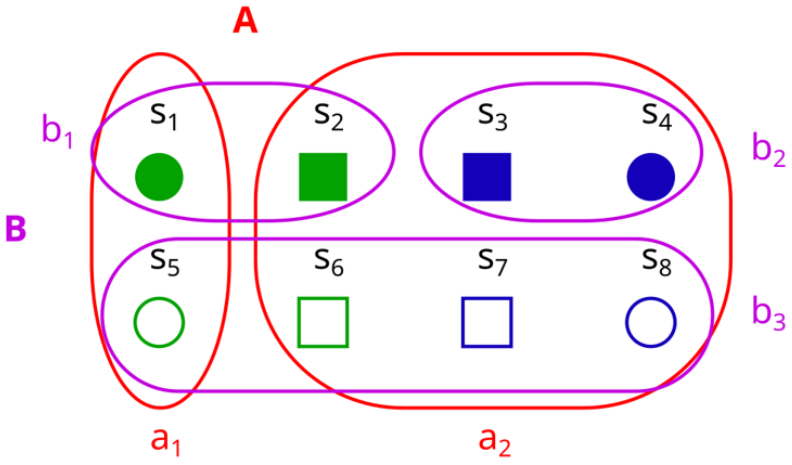
Orthogonality

Now we have defined history, we can define **orthogonality**, which is closely related to independence of random variables!

We say that two partitions A, B are *orthogonal*, if their **histories don't overlap**.

Exercise

Are A and B orthogonal here?



No. The history of A is $h^F(A)$ = {color, shape}. The history of B is $h^F(B)$= {color, fill}, so the histories overlap.

In the Pearl-paradigm we know that two variables X and Y have no common ancestors if and only if they are independent. Analogously, Scott Garrabrant proved that when we use a factored set to model causal structure (I'll explain how to do that in section 3), then two variables are **orthogonal if and only if they are independent**. [2]

Note that in any factorization, the factors are all orthogonal to each other, because $h^F(b) = \{b\}$ for any factor b (we only need color to infer color, remember) so $h^F(b_1) \cap h^F(b_2) = \emptyset$ if $b_1 \neq b_2$.

Scott also defines **conditional orthogonality** as an analog of d-separation. I won't define conditional orthogonality here in order to keep things simple, but you can find the definition here.

In the Pearl-paradigm there is a theorem called **soundness and completeness of d-separation**: Two variables X and Y are d-separated with regard to a set of variables Z if and only if X and Y are conditionally independent given Z (proof here and here)

Scott's central result is the analog of this theorem in the factored set paradigm:

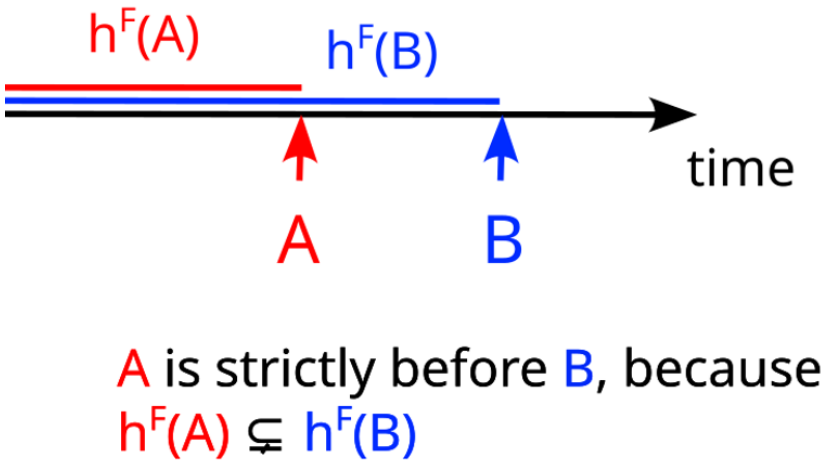
Two variables X, Y are conditionally orthogonal with regard to a set of variables Z if and only if X and Y are conditionally independent given Z. (Technically it's slightly more complicated, but this is the gist [2])

"Time"

- We say that a partition A is **weakly before** B if A's history of A is a *subset or equal* to B's history (i.e. $h^F(A) \subseteq h^F(B)$).
- We say that A is **strictly before** B if A's history is a *strict subset* of B's history (i.e. $h^F(A) \subsetneq h^F(B)$).

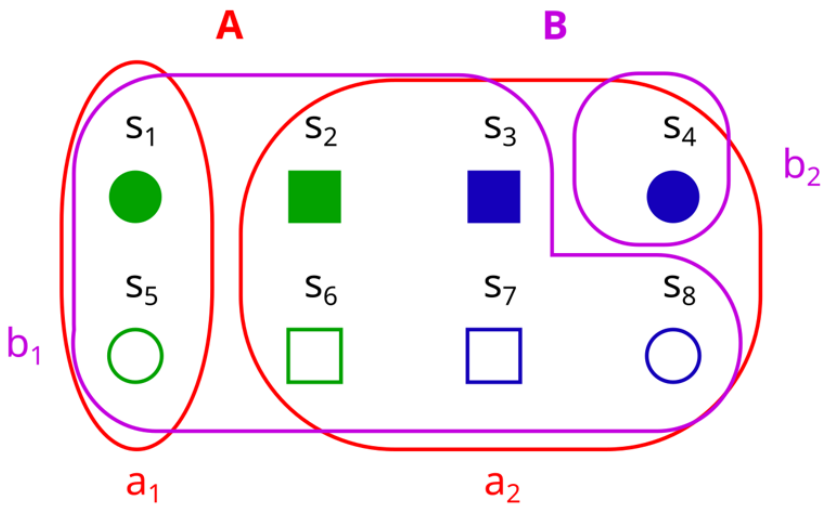
This notion of "time" is closely related to the concept of a causal arrow going from A to B in a causal graph.

You can imagine A's history like "everything that comes before A in time", so if everything that's in A's history is also in B's history then A is before B:



Exercise

Is A or B weakly or strictly before the other here? (i.e. is one of the histories of A or B a subset of the other? Reminder: history = set of properties needed to infer our partition)



A is strictly before B! $h^F(A) = \{color, shape\}$ and $h^F(B) = \{color, shape, fill\}$, so $h^F(A) \subsetneq h^F(B)$.

Now we have the building blocks to use factored sets for causal inference!

3. Causal Inference using Factored Sets

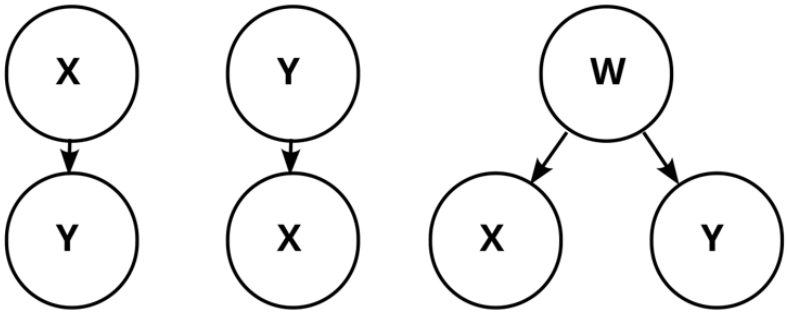
In this section I will walk through an example of inferring causality from data using factored sets.

Consider an experiment in which we collect 2 bits. The distribution P looks like this:

P(00) = 1%
P(01) = 9%
P(10) = 81%
P(11) = 9%

Let's say X is the first bit, and Y is the second bit.

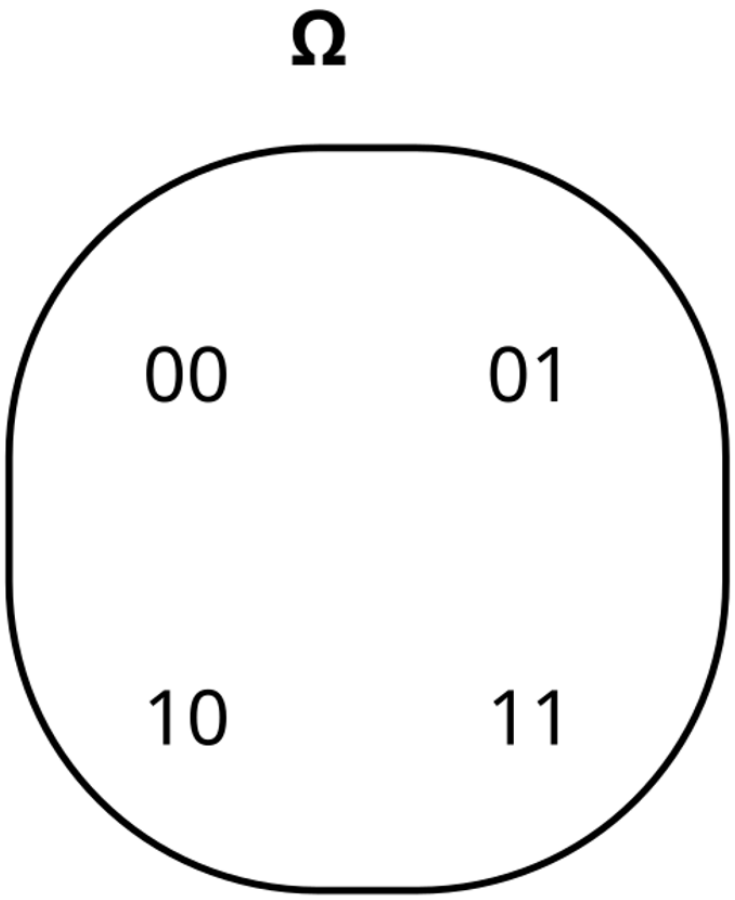
In the causal graph paradigm, we would observe that X and Y are dependent ($P(X = 0) = 10\% \neq P(X = 0|Y = 0) = 8\%$). Thus we are not able to distinguish between these three causal graphs:



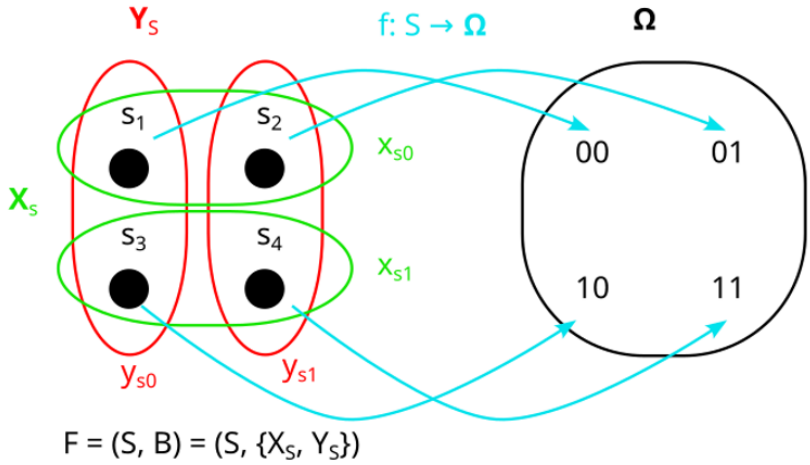
We don’t know whether X causes Y, Y causes X, or there is some common factor W that causes both.

How do we look at this in the Factored Set Paradigm?

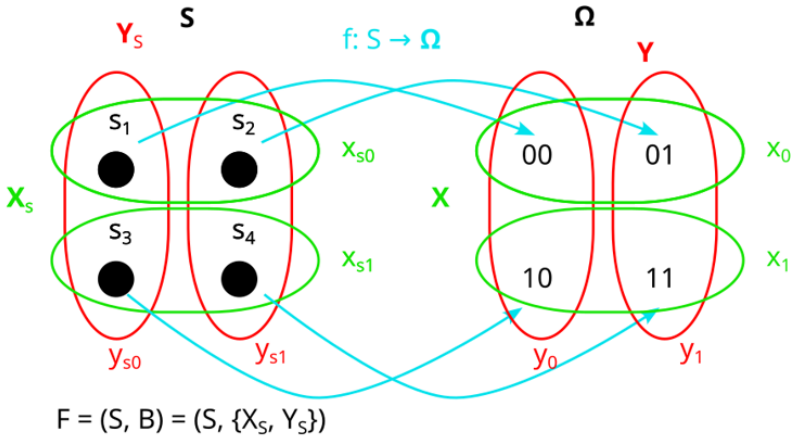
In the factored set paradigm, start with the sample space Ω of our distribution P:



We then say a **model** of the distribution is a factored set F = (S, B) and a function f : S → Ω:



X_S and Y_S are the "preimages" of X and Y under f. By that I mean that all their parts x_{Si} and y_{Sj} are the preimages of x_i and y_j respectively, under f. That looks as follows:



(Note that in this case f is bijective, but in general f does not have to be bijective. If we allow f to be non-bijective, then the framework works in more generality because we can describe some processes that we couldn't otherwise describe. [3])

However, we can't just use *any* factorization B. In order for (F, f) to count as a **model** of our distribution P, it needs to be such that the *dependencies and independencies of our distribution are represented in the factorization*.

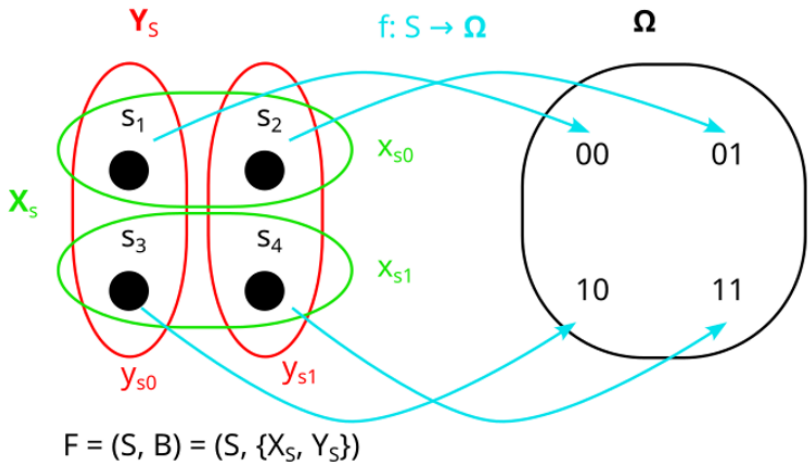
That means:

- Two variables X and Y are independent in P if and only if X_S and Y_S are **orthogonal** in F (i.e. their histories don't overlap)
- Two variables X and Y are **dependent** in P if and only if X_S and Y_S are **not orthogonal** in F

Remember, our probability distribution P was

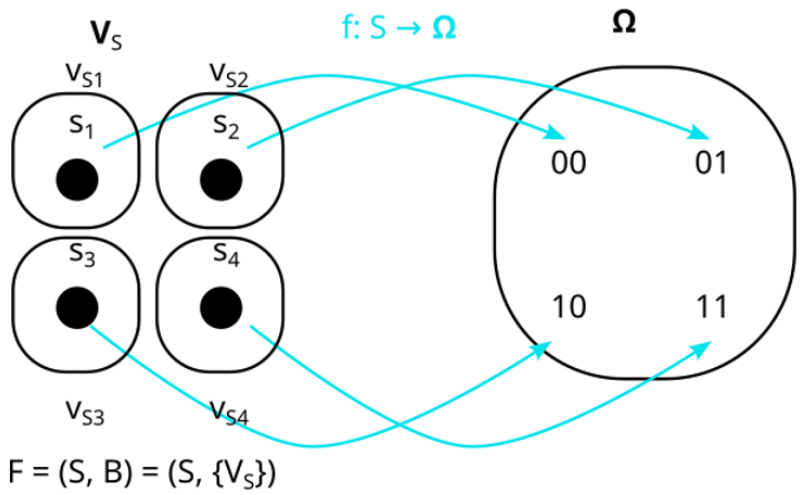
P(00) = 1%
P(01) = 9%
P(10) = 81%
P(11) = 9%

Is the following actually a model of P?



No, it is not: X, and Y are **dependent** in P, but X_S and Y_S are **orthogonal** in F! (Note that orthogonality depends on what model we are in/what factorization we use.)

Here is a really tricky one: Is this a model of P?



Also no, but this is hard to see, so let's walk through it.

Consider the variable Z = X XOR Y

We don't *have to* express our distribution in terms of X and Y, we can just as well define it in terms of X and Z. And then it looks like this:

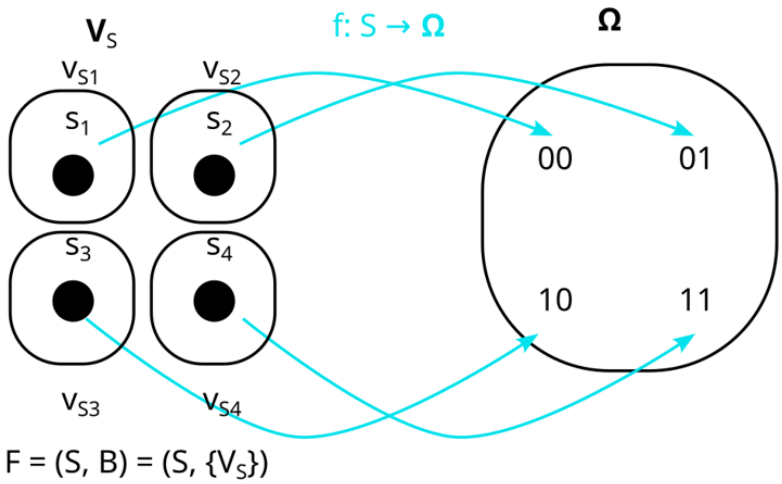
P(00) = P(X=0, Z=0) = 1%
P(01) = P(X=0, Z=1) = 9%
P(10) = P(X=1, Z=1) = 81%
P(11) = P(X=1, Z=0) = 9%

We can see that **Z and X are independent**! (because$P(Z = 0) = 1\% + 9\% = 10\%$, and also$P(Z = 0|X = 0) = \frac{9\%}{9\%+1\%} = 10\%$, and

$P(Z = 0|X = 1) = \frac{81\%}{81\%+9\%} = 10\%$)

What does this mean for our model? It means that Z_S and X_S need to be orthogonal.

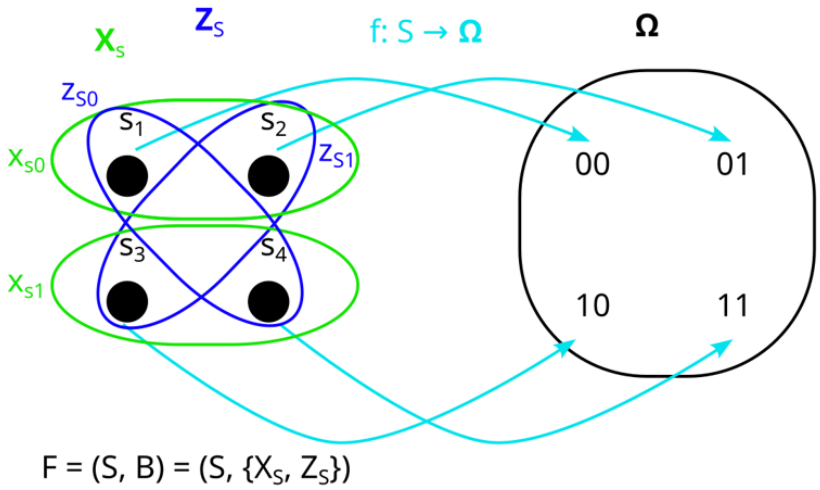
Are Z_S and X_S orthogonal in our model? Here it is again:



No, because the history for both X_S and Z_S is {V}, so X_S and Z_S have an overlapping history.

So F = (S, {V_S}) is also not a model of P.

Does our distribution P have a model at all?
Yes - here is a model that actually works for P:

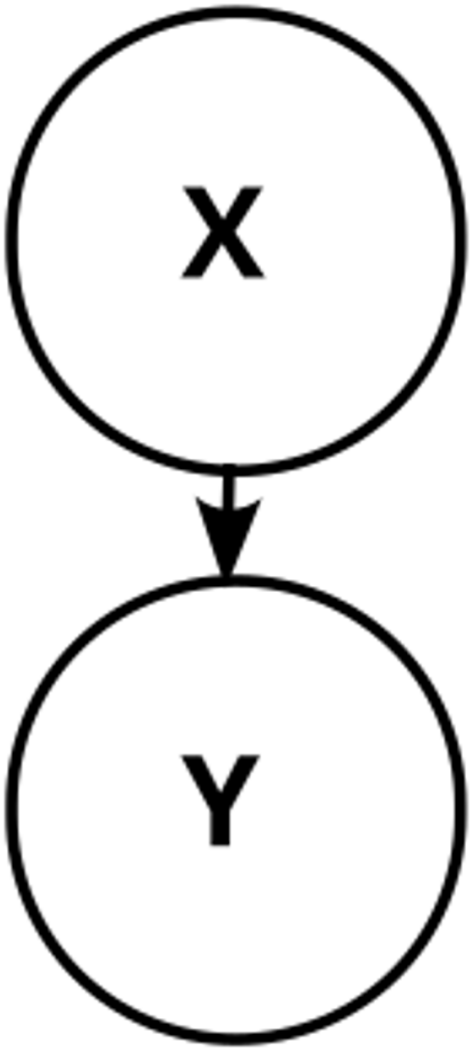


Here $h^F(X_S) = \{X_S\}$ and $h^F(Z_S) = \{Z_S\}$, and $h^F(Y_S) = \{X_S, Z_S\}$.

This means X_S and Z_S are *orthogonal*, which matches our observation that X and Z are *independent* in P. Also X_S and Y_S are *not orthogonal*, which matches our observation that X and Y are *dependent* in P.

It also means that $h^F(X_S) \subsetneq h^F(Y_S)$, so X_S is strictly before Y_S.

If X_S is strictly before Y_S, then the causal arrow goes from X to Y, so we have found a causal direction! It's this one:



From what we know so far, this causal direction $X \to Y$ only holds for this particular model (F, f), but in fact, you can also prove that $X \to Y$ holds for *any* model of this distribution P.

So, we just inferred causality from observational (as opposed to interventional) data, in a way that Pearl's causal models wouldn't have inferred!

Sanity-checking the result in Pearl's paradigm

I have encountered a lot of skepticism that we can infer the causality $X \to Y$ here. So I'm going to switch back to the Pearl paradigm, and explain why X causes Y if our distribution is P, and we can actually infer that from only observational data without needing interventional data.

This section will assume you know how to determine (in-)dependence in probability distributions and in causal graphs. (If you don't, you can either just believe me, or learn about it here and here).

Again, say X is the first bit, Y is the second bit, and Z = X XOR Y. Here is our distribution again, in table-form:

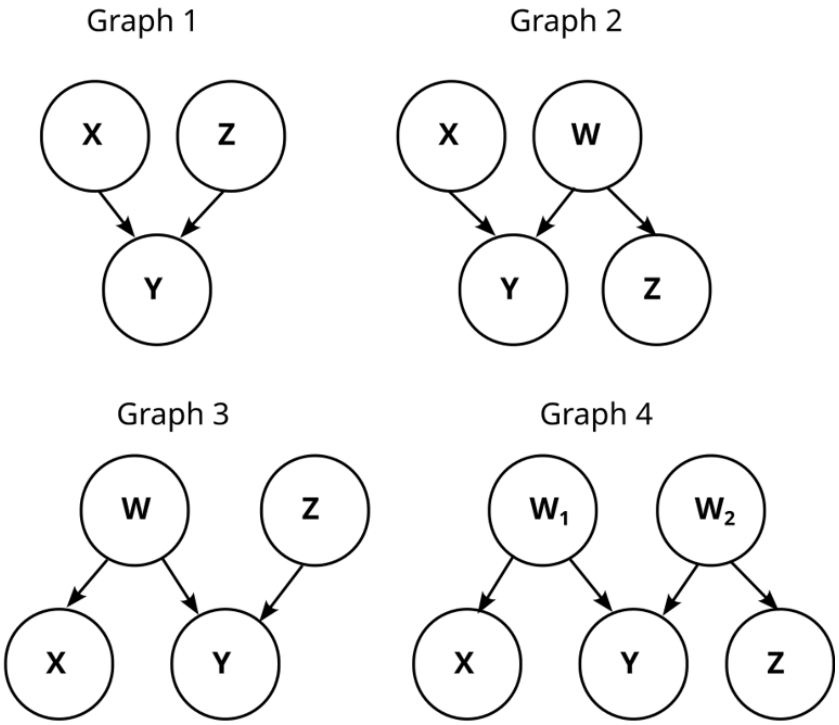
X	Y	Z	P(X, Y, Z)
0	0	0	1%

0	1	1	9%
1	0	1	81%
1	1	0	9%
Any other combination of X, Y, and Z			0%

The (in-)dependencies we can read from this are:

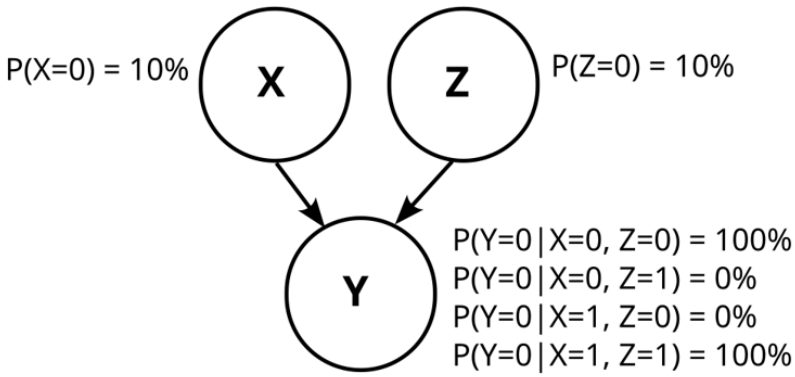
- X and Y are **dependent**
- X and Z are **independent**
- Y and Z are **dependent**

The possible causal graphs which fulfill these dependencies and independencies are these four:

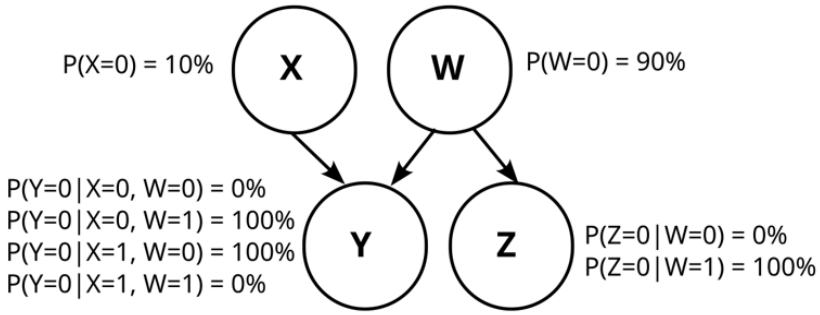


So, we know that Y does not cause X! This matches our finite factored set finding that X causes Y, but it's weaker. Can we also infer that X causes Y?

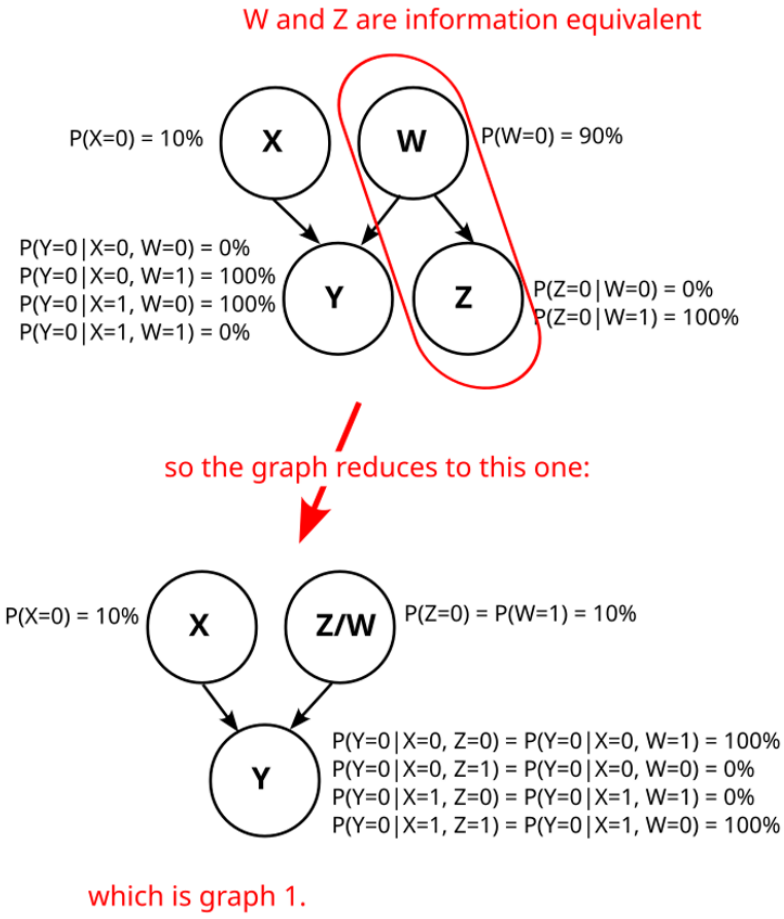
Let's concretize the above graphs by adding the conditional probabilities. Graph 1 then looks like this:



Graph 2 is somewhat trickier, because W is not uniquely determined. But one possibility is like this:



Note that W is just the negation of Z here (W = ¬Z). Thus, W and Z are information equivalent, and that means graph 2 is actually just graph 1.



Can we find a different variable W such that graph 2 does *not* reduce to graph 1? I.e. can we find a variable W such that Z is not deterministic given W?

No, we can't. To see that, consider the distribution

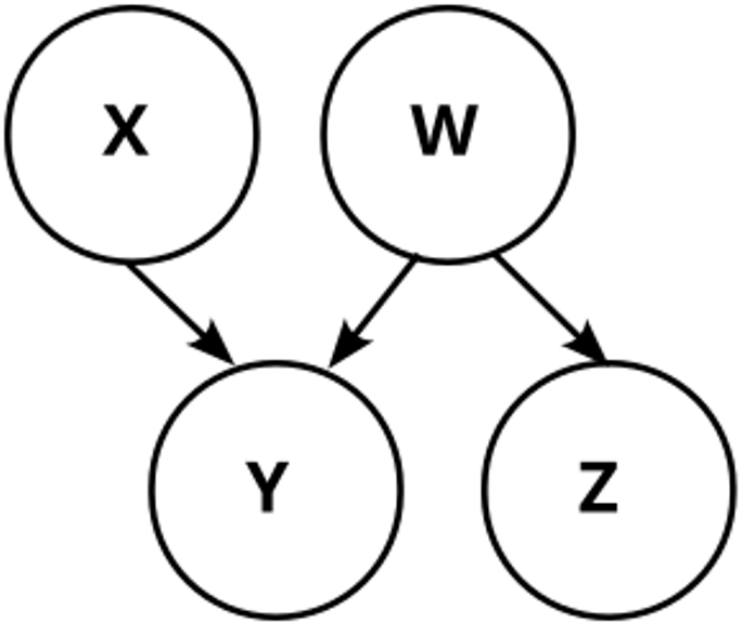
$$P(Z|X,$$

. By definition of Z, we know that

$$P(Z|X,Y,W) = \begin{cases} 1 & \text{if } Z = X \text{ XOR } Y, \\ 0 & \text{otherwise.} \end{cases}.$$

In other words, $P(Z|X,Y,W)$ is deterministic.

We also know that W d-separates Z from X,Y in graph 2:



This d-separation implies that Z is independent from X, Z given W:

$$P(Z \mid W) = P(Z \mid W, X, Y)$$

As

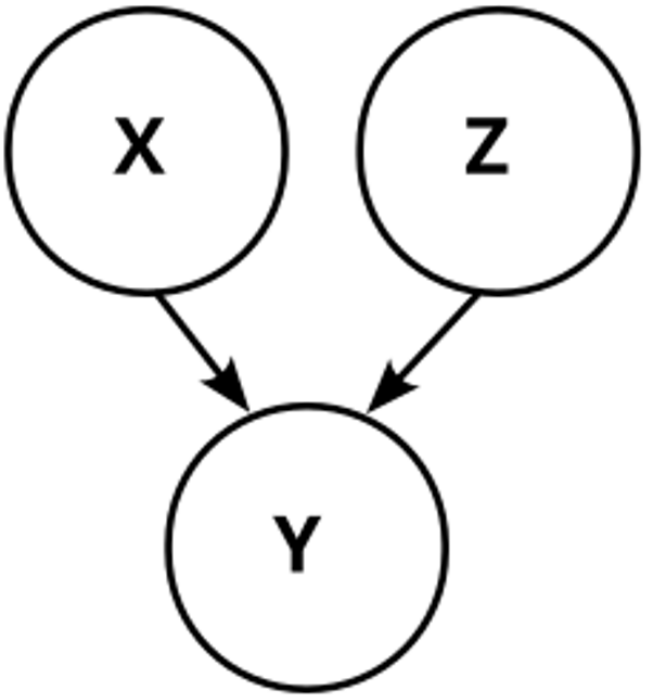
$$P(Z \mid X,$$

is deterministic,

$$P(Z \mid W)$$

also has to be deterministic.

So, graph 2 always reduces to graph 1, no matter how we choose W. Analogously, graph 3 and graph 4 also reduce to graph 1, and we know that our causal structure is graph 1:



Which means we also know that X causes Y.

The reason why we usually wouldn't have found this causal direction using causal graphs is that we *wouldn't even have considered* Z as potentially interesting. This is what factored sets give us: They make us consider every possible way of defining variables, so we **don't miss out on any information** that may be hidden if we just look at a predetermined set of variables.

Summary

Set factorizations are a way of expressing sets as a **product of some factors**, similar to how integer factorization is about expressing integers as a product of some factors.

We can define a **history** on them, that tells us which properties came "before" other properties. We say that two variables are **orthogonal** if they have no shared history. Using these notions of history and orthogonality, we can define a mathematical structure called **model** of a probability distribution. With this model, we can do causal inference (inferring causal structure from data).

Factored sets let us infer causal relations that we usually wouldn't have found using causal graphs. For example, if we have two binary variables X and Y, and X is independent from X XOR Y, then we can infer the causal direction $X \to Y$.

Further Reading

I hope you got a bit of a grasp on finite factored sets, and see why they are really neat. If you want to read more, the best entry point is probably this edited transcript from a talk by Scott Garrabrant.

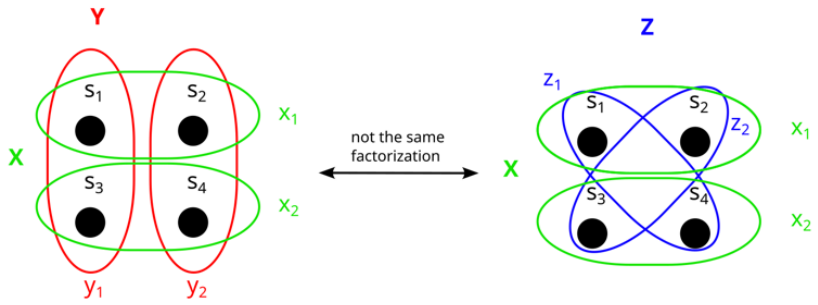
For a non-mathematical intuition how Scott relates the concepts of time, causality, abstraction, and agency, see his Saving Time post.

I haven't looked closely into AI alignment specific applications of factored sets, but it looks like they can be used to better talk about embedded agency, decision theory, and ELK.

This post is a result of a distillation workshop led by John Wentworth at SERI MATS. I'd like to thank Leon Lang, Scott Garrabrant, Matt MacDermott, Jesse Hoogland, and Marius Hobbhahn for feedback and discussions on this post.

1. ^

   Note that the number of set factorizations of an n-element set is not the same as the number of integer factorizations of n, because elements are distinguishable, so for example these two factorizations do not count as the same factorization:

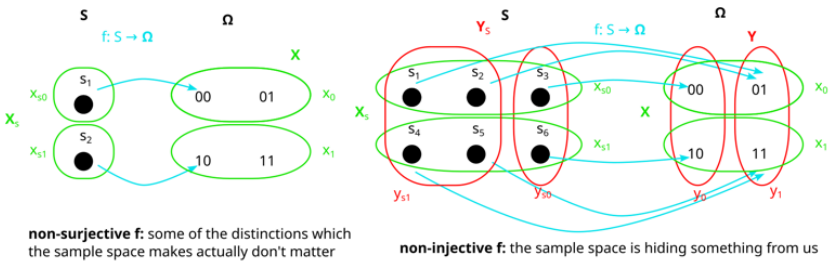


2. ^

   Actually, it's somewhat more complicated than "X and Y are (conditionally) orthogonal if and only if they are independent". The full version is more like "If X_S and Y_S are partitions on a set S, which has a mapping $f : S \to \Omega$ to the sample space, then X_S and Y_S are (conditionally) orthogonal if and only if the images X and Y of X_S and Y_S are (conditionally) independent". But for the sake of this explanation, if you just remember that "orthogonality ⇔ independence", that's enough.

3. ^

   Even if f is not bijective, the "preimages" X_S and Y_S of X and Y are always well-defined partitions. Here are two examples in which f is not bijective:



non-surjective f: some of the distinctions which the sample space makes actually don't matter

non-injective f: the sample space is hiding something from us

Exploring Finite Factored Sets with some toy examples

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://tm.kehrenberg.net/a/finite-factored-set-example/

Seeing as there is little secondary literature for the Finite Factored Set formalism, I thought I'd write up my experience of exploring it through some toy examples that are classic examples in the Pearlian paradigm. My goal was to see how these models that I understood very well in the Pearlian paradigm would work in Finite Factored Sets.

As a warning, this doesn't make any use of the more exciting properties of Finite Factored Sets. It's just an exploration of how this formalism handles the mundane stuff. This also means that I'm using the factored set *directly*, without the abstractions of the orthogonality database. Which I think is fine here, because these are tiny toy examples whose structure is fully known. (However, it's possible that I've missed the entire point of Finite Factored Sets.)

The first example is the 3-variable collider that is very central to Pearl's formalism. It is given by the following Causal Diagram:

A, B, and C are all binary variables (0=false, 1=true).

The intended meaning of a Causal Diagram (or rather, *function causal model*) is that the value of a node x_i is given by a deterministic function that takes as input the parents, $pa(x_i)$, (indicated by the arrows) and an "error", or "noise", variable u_i that is governed by a probability distribution that is independent from the other error/noise variables: $x_i = f_i(pa(x_i), u_i)$. Thus, the value of C is given by $c = f_c(a, b, u_C)$ where u_C is noise or uncertainty that is not explicitly modeled, which we can visualize like this:

We could also split up A and B into a deterministic and a random part, but as they are root nodes, there is little point. It would just be $a = f_a(u_A)$.

The Pearlian formalism runs on *graphs*, but Finite Factored Sets run on the *set* S of all possible outcomes – the sample space. So, the goal is now to construct a sample space that is consistent to the above graph. After that, we'll find a *factorization* of that sample space.

I think it should be clear that to cover the whole sample space S, it is sufficient to consider all possible combinations of the outcomes of A, B, and U_C (but not C),

because if we know the value of these three, then we also know the value of C, via f_C.

So, we simply define S as the Cartesian product of the sets of possible values of A and B: $A = \{0, 1\}$, $B = \{0, 1\}$, and the possible values of U_C: U_C, which I'll leave undefined for the moment (except to note that it must be a finite set):

$$S = A \times B \times U_C, \quad (a, b, u_C) \in S$$

(We use the lowercase letters to represent elements of sets that make up the Cartesian product: $a \in A$, $b \in B$, and $u_C \in U_C$.)

Then – as is custom in the formalism of Finite Factored Sets – the variables A, B, U_C are defined as *partitions* on S. For example, A is a partition consisting of two parts: 1) the set of elements of S where $a = 0$ and 2) those where $a = 1$:

$$A = \{\{(a, b, u_C) \in S | a = 0\}, \{(a, b, u_C) \in S | a = 1\}\}$$

This captures exactly what the variable A is supposed to represent: the question of whether the first element of the Cartesian product is 0 or 1. B is defined analogously:

$$B = \{\{(a, b, u_C) \in S | b = 0\}, \{(a, b, u_C) \in S | b = 1\}\}$$

And U_C as well:

$$U_C = \{\{(a, b, u_C) \in S | u_C = 0\}, \ldots\}$$

with an as of yet undefined number of parts in the partition.

Now, given the way we constructed S, the set of partitions $G = \{A, B, U_C\}$ is a factorization of S: $F = (S, G)$, because any element of S can be uniquely identified by knowing in which part of A, B, and U_C it is, because S is just the Cartesian product of A , B, and U_C.

Now that we have the set, let's look at the probability distribution over S. Let $P(A = a')$ be shorthand for $P(\{(a, b, u_C) \in S | a = a'\})$, i.e., the probability that the outcome lands

in the subset $\{(a, b, u_C) \in S | a = a'\}$ of the sample space S, where the first element of the Cartesian product is equal to a'. We define $P(B = b)$ and $P(U_C = u_c)$ analogously. Finally, let $P(a, b, u_C)$ refer to the probability of the individual outcome $(a, b, u_C) \in S$.

The fact that we want our finite factored set model to be consistent with the Causal Diagram above, implies some things about P. In particular, the diagram implies that A, B, and U_C should be independent, which means that the joint probability should factorize like this:

$$P(a, b, u_C) = P(A = a)P(B = b)P(U_C = u_C)$$

But this is exactly how our (finite) set S factorizes! Thus, P factorizes the same way that S does.

This concludes the translation of the graphical model into a semantically-equivalent finite factored set. To recap what we have accomplished: we turned the original model with the variables A, B, and C into one with three independent variables: A, B, and U_C. And defined C as a deterministic function of these. We then constructed a finite factored set with the factors A, B, and U_C.

Now, let's define C on S as well. We again define it as a partition of S:

$$C = \{\{(a, b, u_C) \in S | f_C(a, b, u_C) = 0\},$$

$$\{(a, b, u_C) \in S | f_C(a, b, u_C) = 1\}\}$$

We simply partition S depending on the output of the function f_C. This also allows us to define $P(C = c)$ as $P(\{(a, b, u_C) \in S | f_C(a, b, u_C) = c\})$.

In the Pearlian formalism, we can read off the fact that A and B are independent from the graph structure. With Finite Factored sets, we have to look at *histories*. The history of a partition (aka a variable) is roughly speaking the minimal set of factors in the factorization G that is sufficient to fully specify the partition (aka variable). A and B are factors in their own right, so their history consists just of themselves: $h(A) = \{A\}$ and $h(B) = \{B\}$, because surely knowing A is enough to know A. As $h(A) \cap h(B) = \emptyset$, we can

conclude that A and B are orthogonal, but this is just because we *defined* the factorization that way, so this is no new information. Still, it's a good consistency check.

The history of C is more interesting, but still pretty trivial to determine. As long as f_C makes use of all its arguments and is generally non-pathological, all the factors are needed to determine C. So: $h(C) = \{A, B, U_C\}$. This implies $h(A) \subset h(C)$ and $h(B) \subset h(C)$ which implies A and B are both *strictly before* C in the time defined by our factorization. This is a non-trivial result which matches what we would have expected from the graphical model that I drew in the beginning. So, that's good.

But what if we condition on C? Colliders are special because they have $A \perp B$ but $A \not\perp B|C$. Can we recover this results from our finite factored set model?

To compute conditional orthogonality, we have to compute *conditional histories*. In order to condition on C, we need the histories conditioned on the subsets of S where $C = 0$ and where $C = 1$. We'll call these subsets C_0 and C_1:

$$C_0 := \{(a, b, u_C) \in S | f_C(a, b, u_C) = 0\}$$

$$C_1 := \{(a, b, u_C) \in S | f_C(a, b, u_C) = 1\}$$

Let's start with the latter: let's determine $h(A|C_1)$ – the conditional history of A given C_1. However, the definition of conditional history is not as straightforward as the definition of history. I'm reproducing it here:

> Let $F = (S, G)$ be a finite factored set, let $X, Y, Z \in \text{Part}(S)$, and let $E \subseteq S$. We use $s \sim_X t$ to say that two elements s and t are in the same part in X. The conditional history of X given E, written $h^F(X|E)$, is the smallest set of factors $H \subseteq G$ satisfying the following two conditions:
>
> 1. For all $s, t \in E$, if $s \sim_b t$ for all $b \in H$, then $s \sim_X t$.
> 2. For all $s, t \in E$ and $r \in S$, if $r \sim_{b_0} s$ for all $b_0 \in H$ and $r \sim_{b_1} t$ for all $b_1 \in G \setminus H$, then $r \in E$.

The first condition is easy. It just says that the factors in $h(A|C_1)$ should be sufficient to pin down A in C_1. We can safely assume that A itself will be enough. So, is $H := h(A|C_1)$ just {A} again? Well there is that second condition. To show its effect, let's assume that indeed $H = \{A\}$. The condition then specifies something that has to be true for all $s, t \in C_1$ and $r \in S$. However, given the assumption we just made, I can construct a counterexample to the stated condition. (Thus showing by contradiction that $h(A|C_1)$ is not just {A}.)

To make things concrete, I'll give a concrete definition of how C is computed. First, let's say that U_C is the result of a 6-sided die; so: $U_C = \{1, 2, 3, 4, 5, 6\}$. We'll then say that C is XOR of A and B, except when U_C rolls a 6, in which case C is just 0:

$$f_C(a, b, u_C) = \begin{cases} a \oplus b & \text{if } u_C \neq 6 \\ 0 & \text{else.} \end{cases}$$

The counterexample is then:

$$s = (a_s, b_s, u_{Cs}) = (1, 0, 1)$$

$$t = (a_t, b_t, u_{Ct}) = (0, 1, 2)$$

$$r = (a_r, b_r, u_{Cr}) = (1, 1, 2)$$

Let's first confirm that s and t are indeed in C_1. To be in C_1, we need $f_C(a, b, u_C) = 1$. This is true for both, because $1 \oplus 0 = 1$ and $0 \oplus 1 = 1$ and in neither case is $u_C = 6$.

Now, in the if-statement in the second condition, we first ask whether r and s can be distinguished by the factors in H. We're assuming for now that H consists only of A, so the question becomes whether r and s can be distinguished by looking at A. The answer is *no*, because in both cases, the first entry is 1 (so, $a = 1$). Thus, as far as the factors in H are concerned r and s are the same.

Next we must investigate r and t in light of the the factors that are *not* in H. In our case, that's B and U_C. As we can see, r and t are indeed indistinguishable under B and

U_C because r and t only differ in the *first* entry of the tuple (which corresponds to A). The if-statement is thus true, and so according to the condition, we should have $r \in C_1$. However, this is not the case, because $1 \oplus 1 = 0$, and so r is in C_0. In other words, we have a contradiction and $h(A|C_1)$ can't consist of only A.

So, what does the second condition in the definition of *conditional history* look out for? It seems to want to prevent that both H and its complement are *incomplete* when it comes to being able to answer the question of whether an element is in C_1 or not. That is, if both the factors within H and the factors without seem to indicate that an element r is in C_1, then – the definition says – it should really be in C_1. The factors should *not* be split in such a way that neither H nor its complement ($G \setminus H$ where G is the set of all factors) can reconstruct C; otherwise, the border itself leaks information about the likely values of factors.

The problem can be easily fixed by adding B and U_C to $h(A|C_1)$ as well, so that H has all the factors needed to fully pin down the border of C_1: $h(A|C_1) = \{A, B, U_C\}$. Via symmetry, we also get $h(A|C_0) = \{A, B, U_C\}$ and also the same for $h(B|C_{0,1})$. We thus definitely don't have $h(A|C_{0/1}) \cap h(B|C_{0/1}) = \emptyset$ anymore. And so, A and B are *not* orthogonal given C.

This means we have recovered the two most important independence statements about the collider: $A \perp B$ and $A \not\perp B|C$, as well as $C \not\perp A$ and $C \not\perp B$ (just from the fact that A and B are strictly before C in time). What remains to be confirmed are $C \not\perp A|B$ and $C \not\perp B|A$. I'll leave that as an exercise for the reader.

As our next example, we'll have a look at the 3-variable chain:

Separating out the noise/uncertainty:

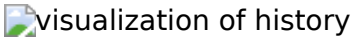
I won't repeat all the steps now and just skip to constructing the sample space S as the Cartesian product of the possible values of A, U_B, and U_C. The elements of S are then tuples like this one:

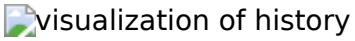
$$(a, u_B, u_C) \in S$$

We define the partitions A, U_B, and U_C on this in the obvious way, and so they are our factorization of S. The variables B and C are defined as partitions of S according to some deterministic function of (a, u_B) and (a, u_B, u_C) respectively. For the histories, it follows then that $h(A) = \{A\}$, $h(B) = \{A, U_B\}$, and $h(C) = \{A, U_B, U_C\}$, which implies $A < B < C$, as one would expect.

(It might seem remarkable here that we can reconstruct the exact order of A, B, and C , but that is only because we *defined* S that way. Nothing interesting has happened yet. This is just a self-consistency check.)

I tried visualizing these histories but had limited success:


visualization of history


visualization of history

The interesting independence statement in this model is $A \perp C|B$. So, to investigate this, let's look at $h(C|B_1)$ – the conditional history of C on the subset of S where B = 1. The first step is to clarify that C is computed via B:

$$f_C(a, u_B, u_C) = g_C(f_B(a, u_B), u_C)$$

That is, a and u_B are only used via f_B. But if we already know the output of f_B (because we're in the subset $B_1 \subset S$ where $f_B = 1$), then we only need u_C to compute the value of C. Thus, it would be consistent with *condition 1* of the conditional histories definition if we had $h(C|B_1) = \{U_C\}$. But is it also consistent with *condition 2*?

In condition 2, we have first this part: "if $r \sim_{U_C} s$ ..." However, U_C has no bearing on whether r is in B_1 or not, because f_B doesn't take U_C as an input. So, we can ignore that part. Without that part we are left with: if $r \sim_X t$ for all $X \in \{A, U_B\}$, then $r \in B_1$. Is this true for all $t \in B_1$ and $r \in S$? Yes, because what it is saying is, if r seems to be in

B_1 according to A and U_B, then it really is in B_1. And that is true, because A and U_B already fully determine B, so if they say you are in B_1, then you are.

So, we avoided the trap we had above, where neither H nor its complement could reconstruct the boundary of the set we were conditioning on. Here, A and U_B (which are both not in H) are able to reconstruct the boundary of B_1 perfectly, so condition 2 is fulfilled. Thus, $h(C|B_1) = \{U_C\}$ (and analogously $h(C|B_0) = \{U_C\}$), which implies that $h(A|B_{0/1}) \cap h(C|B_{0/1}) = \emptyset$. (I didn't check that $h(A|B_{0/1})$ doesn't contain U_C, but a quick argument shows that it won't: U_C is neither needed to pin down A nor to pin down the border of $B_{0/1}$.) So, $A \perp C|B$ as expected.

There is one interesting 3-variable model left – the common cause:

The reader may want to try this out for themself before reading on.
Splitting off the randomness, we get:

As before, we can construct a sample space that is factorized by $G = \{U_B, A, U_C\}$, giving us the finite factored set $F = (S, G)$. The histories for A, B, and C are then again obvious: $h(A) = \{A\}$, $h(B) = \{U_B, A\}$, and $h(C) = \{A, U_C\}$. We can see that B and C are not independent, because their histories both include A. But they also don't have a definite temporal order; we can neither say $B <^F C$ nor $C <^F B$.

From Pearl's theory, we expect that B and C become independent when conditioned on A. So let's look at those conditional histories. As we know all the tricks by now, it will be a very high-level analysis.

We recall the first rule of conditional histories: the history should contain those factors that are needed to pin down the variable *given that we already know the value of the conditioned variable*. If we know the value of A, then it suffices to know U_B in order to know B. So, the conditional history, H, of B given that we know the value of A, contains U_B at the least.

The second rule of conditional histories demands that either H or its complement, $G \backslash H$, (or both) is on its own sufficient to determine the value of the conditioned variable (A in our case). Assuming $H = \{U_B\}$, the complement of H, $\{A, U_C\}$, contains A itself, and so is definitely sufficient to determine A. Thus, when conditioning on values for A, $H = \{U_B\}$ is a permitted history by the second rule.

By symmetry, we get $\{U_C\}$ for the conditional history for C. This all then implies $B \perp C|A$ as expected.

Conclusions

I hope this was useful to someone. And I hope I didn't completely mess up the intended use of this formalism.

One thing I appreciate about this formalism is that I find it easy to drop to the base level (the sample space with the factorization) to explicitly check my higher-level thoughts when I get confused. It's nice to have that solid ground level available whenever I need it.

The rule about conditional histories is not exactly *easy*, but it feels closer to a fundamental law than the *d-separation* rules of the Pearlian paradigm, which always felt a bit arbitrary.

Finally, I still kind of think that a DAG is a really nice way of visualizing dependencies and independencies of random variables. I wonder if there is a visualization that feels more native to Finite Factored Sets while still looking nice.

A simple example of conditional orthogonality in finite factored sets

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Recently, MIRI researcher Scott Garrabrant has publicized his work on finite factored sets. It allegedly offers a way to understand agency and causality in a set-up like the causal graphs championed by Judea Pearl. Unfortunately, the definition of conditional orthogonality is very confusing. I'm not aware of any public examples of people demonstrating that they understand it, but I didn't really understand it until an hour ago, and I've heard others say that it went above their heads. So, I'd like to give an example of it here.

In a finite factored set, you have your base set S, and a set B of 'factors' of your set. In my case, the base set S will be four-dimensional space - I'm sorry, I know that's one more dimension than the number that well-adjusted people can visualize, but it really would be a much worse example if I were restricted to three dimensions. We'll think of the points in this space as tuples (x_1, x_2, x_3, x_4) where each x_i is a real number between, say, -2 and 2 [footnote 1]. We'll say that X_1 is the 'factor', aka partition, that groups points together based on what their value of x_1 is, and similarly for X_2, X_3, and X_4, and set $B = \{X_1, X_2, X_3, X_4\}$. I leave it as an exercise for the reader to check whether this is in fact a finite factored set. Also, I'll talk about the 'value' of partitions and factors - technically, I suppose you could say that the 'value' of some partition at a point is the set in the partition that contains the point, but I'll use it to mean that, for example, the 'value' of X_1 at point (x_1, x_2, x_3, x_4) is x_1. If you think of partitions as questions where different points in S give different answers, the 'value' of a partition at a point is the answer to the question.

[EDIT: for the rest of the post, you might want to imagine S as points in space-time, where x_4 represents the time, and (x_1, x_2, x_3) represent spatial coordinates - for example, inside a room, where you're measuring from the north-east corner of the floor. In this analogy, we'll imagine that there's a flat piece of sheet metal leaning on the floor against two walls, over that corner. We'll try conditioning on that - so, looking only at points in space-time that are spatially located on that sheet - and see that distance left is no longer orthogonal to distance up, but that both are still orthogonal to time.]

Now, we'll want to condition on the set $E = \{(x_1, x_2, x_3, x_4) | x_1 + x_2 + x_3 = 1\}$. The thing with E is that once you know you're in E, x_1 is no longer independent of x_2, like it was

before, since they're linked together by the condition that $x_1 + x_2 + x_3 = 1$. However, x_4 has nothing to do with that condition. So, what's going to happen is that conditioned on being in E, X_1 is orthogonal to X_4 but not to X_2.

In order to show this, we'll check the definition of conditional orthogonality, which actually refers to this thing called conditional history. I'll write out the definition of conditional history formally, and then try to explain it informally: the conditional history of X given E, which we'll write as h(X|E), is the smallest set of factors $H \subseteq B$ satisfying the following two conditions:

1. For all $s, t \in E$, if $s \sim_b t$ for all $b \in H$, then $s \sim_X t$.
2. For all $s, t \in E$ and $r \in S$, if $r \sim_b s$ for all $b \in H$ and $r \sim_{b'} t$ for all $b' \in B \setminus H$, then $r \in E$.

Condition 1 means that, if you think of the partitions as carving up the set S, then the partition X doesn't carve E up more finely than if you carved according to everything in h(X|E). Another way to say that is that if you know you're in E, knowing everything in the conditional history of X in E tells you what the 'value' of X is, which hopefully makes sense.

Condition 2 says that if you want to know if a point is in E, you can separately consider the 'values' of the partitions in the conditional history, as well as the other partitions that are in B but not in the conditional history. So it's saying that there's no 'entanglement' between the partitions in and out of the conditional history regarding E. This is still probably confusing, but it will make more sense with examples.

Now, what's conditional orthogonality? That's pretty simple once you get conditional histories: X and Y are conditionally orthogonal given E if the conditional history of X given E doesn't intersect the conditional history of Y given E. So it's saying that once you're in E, the things determining X are different to the things determining Y, in the finite factored sets way of looking at things.

Let's look at some conditional histories in our concrete example: what's the history of X_1 given E? Well, it's got to contain X_1, because otherwise that would violate condition 1: you can't know the value of X_1 without being told the value of X_1, even once you know you're in E. But that can't be the whole thing. Consider the point

s = $(0.5, 0.4, 0.4, 0.7)$. If you just knew the value of X_1 at s, that would be compatible with s actually being $(0.5, 0.25, 0.25, 1)$, which is in E. And if you just knew the values of X_2, X_3, and X_4, you could imagine that s was actually equal to $(0.2, 0.4, 0.4, 0.7)$, which is also in E. So, if you considered the factors in $\{X_1\}$ separately to the other factors, you'd conclude that s could be in E - but it's actually not! This is exactly the thing that condition 2 is telling us can't happen. In fact, the conditional history of X_1 given E is $\{X_1, X_2, X_3\}$, which I'll leave for you to check. I'll also let you check that the conditional history of X_2 given E is $\{X_1, X_2, X_3\}$.

Now, what's the conditional history of X_4 given E? It has to include X_4, because if someone doesn't tell you X_4 you can't figure it out. In fact, it's exactly $\{X_4\}$. Let's check condition 2: it says that if all the factors outside the conditional history are compatible with some point being in E, and all the factors inside the conditional history are compatible with some point being in E, then it must be in E. That checks out here: you need to know the values of all three of X_1, X_2, and X_3 at once to know if something's in E, but you get those together if you jointly consider those factors outside your conditional history, which is $\{X_1, X_2, X_3\}$. So looking at $(0.5, 0.4, 0.4, 0.7)$, if you only look at the values that aren't told to you by the conditional history, which is to say the first three numbers, you can tell it's not in E and aren't tricked. And if you look at $(0.5, 0.25, 0.25, 0.7)$, you look at the factors in $\{X_4\}$ (namely X_4), and it checks out, you look at the factors outside $\{X_4\}$ and that also checks out, and the point is really in E.

Hopefully this gives you some insight into condition 2 of the definition of conditional history. It's saying that when we divide factors up to get a history, we can't put factors that are entangled by the set we're conditioning on on 'different sides' - all the entangled factors have to be in the history, or they all have to be out of the history.

In summary: $h(X_1|E) = h(X_2|E) = \{X_1, X_2, X_3\}$, and $h(X_4|E) = \{X_4\}$. So, is X_1 orthogonal to X_2 given E? No, their conditional histories overlap - in fact, they're identical! Is X_1 orthogonal to X_4 given E? Yes, they have disjoint conditional histories.

Some notes:

- In this case, X_1 was already orthogonal to X_4 before conditioning. It would be nice to come up with an example where two things that weren't already orthogonal become so after conditioning. [EDIT: see my next post]
- We didn't really need the underlying set to be finite for this example to work, suggesting that factored sets don't really need to be finite for all the machinery Scott discusses.
- We did need the range of each variable to be bounded for this to work nicely. Because all the numbers need to be between -2 and 2, once you're in E, if $x_1 = 2$ then x_2 can't be bigger than 1, otherwise x_3 can't go negative enough to get the numbers to add up to 1. But if they could all be arbitrary real numbers, then even once you were in E, knowing x_1 wouldn't tell you anything about x_2, but we'd still have that X_1 wasn't orthogonal to X_2 given E, which would be weird.

[^1] I know what you're saying - "That's not a finite set! Finite factored sets have to be finite!" Well, if you insist, you can think of them as only the numbers between -2 and 2 with two decimal places. That makes the set finite and doesn't really change anything. (Which suggests that a more expansive concept could be used instead of finite factored sets.)

A second example of conditional orthogonality in finite factored sets

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Yesterday, I wrote a post that gave an example of conditional non-orthogonality in finite factored sets. I encourage you to read that post first. However, I'm kind of dissatisfied with it because it doesn't show any interesting cases of conditional orthogonality (despite the title seeming to promise that). So I'd like to show you one today.

First, let's imagine that Alice is a person who has some height. Bob and Charlie both measure her height, and take note of the measurements. However, their measuring instruments have independent sources of error, such that neither gets exactly the right answer. In this world, Bob's measurement is not independent of Charlie's measurement, because they'll both be pretty close - the error isn't that big. However, once you know Alice's height, they will be independent, because given that knowledge, learning Bob's measurement doesn't tell you anything about Charlie's measurement. Below, we'll see how we can formalize that in the language of finite factored sets.

Our finite factored set will be the set of tuples $(a, \epsilon_b, \epsilon_c)$, and the basis factors will be A

, which partitions the tuples by their value of a, E_B, which partitions them by their

value of ϵ_b, and E_C, which partitions them by their value of ϵ_c. These represent Alice's height, and the error that Bob's and Charlie's machines respectively introduce. Note that you might have imagined we'd have the basic factors as Alice's height, Bob's measurement, and Charlie's measurement, but then these wouldn't be probabilistically or logically independent, and so would violate the assumptions that go into modelling things as finite factored sets. [footnote 1]

Next, we'll define the function $b(a, \epsilon_b, \epsilon_c) = a + \epsilon_b$, which gives the height that Bob

measures, and the partition B which groups tuples with the same value of b together.

Similarly, we'll define $c(a, \epsilon_b, \epsilon_c) = a + \epsilon_c$, which tells us the height that Charlie

measures, and the partition C that groups tuples together by their value of c.

What's the history of B? Well, it's the smallest set of factors such that if we know the

'value' of the factors, then we know the 'value' of B, and that's $\{A, E_B\}$. Similarly, the

history of C is $\{A, E_C\}$. So B's and C's histories have A in common, and therefore aren't orthogonal.

Now, let's consider the set $A_2 = \{(a, \epsilon_b, \epsilon_c) \mid a = 2\}$, which represents the worlds where Alice is 2 metres tall, and check out the conditional histories. The conditional history of B in A_2 is the smallest set of factors such that once you're in A_2, knowing the 'values' of those factors tells you the 'value' of B, and that includes all the factors that are 'entangled' with those factors by the set A_2 - for more detail, check out the previous post. In this case, the conditional history of B is just $\{E_B\}$: Once you're in A_2, knowing ϵ_b is enough to tell you the value of b. Furthermore, the only thing you need to know to figure out whether something's in A_2 is a, so $\{E_B\}$ also satisfies the second condition: if the 'value' of E_B at some tuple is compatible with being in A_2 (which is always true), and the 'values' of A and E_C are jointly compatible with being in A_2, then you must be in A_2. Similarly, the conditional history of C given A_2 is $\{E_C\}$. So, the conditional histories don't intersect, and B is orthogonal to C given A_2.

Hopefully this post was useful both in giving you a better sense of conditional orthogonality, and in illustrating how to model things with finite factored sets.

[footnote 1] Note that we could 'change coordinates' and have the underlying set be tuples (a, b, c) - Alice's height, Bob's measurement, and Charlie's measurement - and the factors being:

- A, the partition of points according to their value of a
- E_B, the partition of points according to their value of $b - a$
- E_C, the partition of points according to their value of $c - a$

This would give exactly the same results as those in the main post!

Counterfactability

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post will assume an understanding of the finite factored set ontology. It will be more speculative that the main FFS sequence, and I will leave out some proofs. It seems likely that I will later regret some of the definitions laid out here. I will be particularly sloppy in defining evidential and causal counterfactuals, because this post is not advocating for them, only demonstrating that it is possible to represent them within our ontology. This post was (approximately) written a year ago. I am currently in the process of moving a bunch of my ideas from the last year to LessWrong.

Nontechnical Summary

The main thing in this post is introduce a new concept of counterfactability. A counterfactable event E is one that screens off its own history (i.e. everything usptream of E) from everything that you care about.

Decision theory is easy when considering choices that represent counterfactable events. When trying to take counterfactuals on non-counterfactable events, things are under-defined. This is because non-counterfactable events have artificially low resolution. It is like asking what would happen if I either took action X or took action Y. The question does not carve reality at the joints. Different worlds were artificially merged. It is as though details about the event were forgotten.

We can reframe questions in decision theory as follows: When given a non-counterfactable event E, how do we add some more details to E to form a counterfactable E', while only adding details that feel like they were artificially forgotten? In this new frame, we are settled on the question of how to take counterfactuals on counterfactable things, and are only asking "What is the counterfactable thing we were meant to be counteracting on?"

Both CDT and EDT can be reframed as providing an answer to this question. They are giving the wrong answers, but this fact shows that the reframe is sufficient to capture both CDT and EDT. Further, we can see that when choosing between counterfactable choices, CDT and EDT give the same answer. I believe the correct way to find the natural events to counteract on requires introspection on the gears of the agent's cognition, and will not be either of the CDT and EDT extremes.

Finally, I show that this concept of counterfactability is sufficiently natural that it applies outside the context of decision theory, and use it as a lens on the eliciting latent knowledge problem.

Counterfactability

Let $F = (S, B)$ be a finite factored set, and let E be a nonempty proper subset of S, and let W be a partition of S.

We say that E is counterfactable relative to W if for all $X \in \text{Part}(S)$, if $X \leq^F \{E, S \setminus E\}$, then $X \perp W \mid E$. (E screens off its own history from W.)

We will generally think of W as a high level description of the world that contains all of the features we care about.

Whenever E is counterfactable relative to W, we can define a counterfactual function $\text{do}_E^W : S \to W$ given by $\text{do}_E^W(s) = [\chi^F_{h^F(\{E,S\setminus E\})}(e, s)]_W$, where $e \in E$. In order for do_E^W to be well defined, we need this to be independent of the choice of $e \in E$.

Claim: When E is counterfactable relative to W, do_E^W is well defined.

Proof: Let $H = h^F(\{E, S \setminus E\})$, and let $X = \bigvee_S(H)$. Observe that $H = h^F(X)$. Consider $e_0, e_1 \in E$, and let $s_0 = \chi^F_H(e_0, s)$ and $s_1 = \chi^F_H(e_1, s)$. Assume for the purpose of contradiction that $[s_0]_W \neq [s_1]_W$. Since $X \leq^F \{E, S \setminus E\}$, we have $X \perp W \mid E$. Thus, $h^F(X|E)$ and $h^F(W|E)$ are disjoint.

Since $[s_0]_W \neq [s_1]_W$, and $s_0, s_1 \in E$, there must be some $b \in h^F(W|E)$ such that $s_0 \nsim_b s_1$, and observe that further, we must have $b \in H$ and $b \notin h^F(X|E)$.

Since $b \in H$, we have $b \leq_S X$, so $b|E \leq_E X|E$, so $h^F(b|E) \subseteq h^F(X|E)$, so $b \notin h^F(b|E)$. Note that $B \setminus h^F(b|E) \vdash^F (b|E)$, since $\chi^F_{B\setminus h^F(b|E)}(E, E) = E$, and $b \in B \setminus h^F(b|E)$, so $b|E \leq_E \bigvee_S(B \setminus h^F(b|E))|E$. Thus $\{\} = h^F(b|E) \cap (B \setminus h^F(b|E)) \vdash^F (b|E)$, so $b|E$ is a singleton. This contradicts the fact that s_0 and s_1 are in different parts in $b|E$. □

W here can be thought of as a high level world model up to which we want our counterfactuals to be well defined. If we take $W = \{\{s\} | s \in S\}$, then we will have that

E is counterfactable relative to W if and only if there exists an $s \in S$ and a $C \subseteq B$ such that $e \in E$ if and only if $e \sim_b s$ for all $b \in C$.

So, in the FFS framework, we have this simple notion of counterfactability (relative to W), together with a way of counterfacting (up to W) on any counterfactable event.

Extending Beyond the Counterfactable

The question then becomes "How do you counterfact on non-counterfactable events?" I will start by presenting two (in my opinion) bad strategies for extending counterfactuals to (some) non-counterfactable events. These two strategies will give us evidential counterfactuals and causal counterfactuals. Both will require some extra structure beyond the finite factored set structure in order to be defined.

We will see that counterfactibility is actually a very strong notion, and for any counterfactable E, if you sample an $s \in S$, and then counterfact on E, you will get the same distribution on W, regardless of whether you use do_E^W, evidential counterfactuals, or causal counterfactuals. If we take W to be the level sets of a utility function, this leads the result that (updateless) CDT is the same as (updateless) EDT whenever the agent is choosing between counterfactable events.

However, the interesting part is that both evidential and causal counterfactuals (while agreeing with the standard intuitions) will be defined in terms of do^W. In both cases, we will be given a not necessarily counterfactable E, and then (possibly randomly) find a counterfactable $E' \subseteq E$, and counterfact on E' instead. Thus, the whole of evidential and causal counterfactuals can be summarized as "Given an event E, first find the correct counterfactable subset E', and then counterfact on E'."

Note that the above is not saying much, since we could just take E' to be a singleton. The interesting part is that both evidential and causal counterfactuals can be viewed as $do_{E'}^W$ for E' no later than E ($\{E', S \setminus E'\} \leq^F \{E, S \setminus E\}$). Thus, we are not just forcing evidential and causal counterfactuals to fit our notion of counterfactuals by counterfacting on an overly specific description of the whole world. We are counterfacting on a local event, no later than E itself.

This gives a new orientation on the problem of counterfactuals. We are given an event E that we want to counterfact on, and unfortunately it is not counterfactable, because it is not specific enough, so we instead need to counterfact on some subset

E' that is both counterfactable and no later than E. Evidential and causal counterfactuals are just (bad) ways of choosing that subset. Now instead of trying to figure out how to define counterfactuals, we can instead think of ways to choose a counterfactable subset of the event we want to counterfact on.

The emphasis on E' that are no later than E is important, because it is capturing that non-counterfactable events are somewhat artificially not specific enough to counteract on. It is as though they were constructed by unioning together counterfactable events. The "were constructed by" is important here. Events can be expressed as unions of counterfactable events in many ways, but it feels more like the resolution was artificially removed, when we express the event as as union of events that came weakly earlier. Sometimes, I know how to counterfact on "I do X," and I know how to counterfact on "I do Y," but I get confused when I try to counterfact on "I do X or I do Y," because the resolution was artificially lowered. This is not to say we can just not lower the resolution. This is the curse of embedded agency.

To make it more pithy, "Counterfactuals are sometimes under-defined for events that are not at their native resolution." I actually like the analogy with native resolution on a monitor here. When the image has less resolution than the monitor, there isn't a well defined best way to display it. If the image has resolution that is an integer fraction the monitor (in each dimension), there is a well defined best way to display it, but that is because dividing by an integer corresponds in this analogy to taking out a factor, and thus having a smaller history.

Defining Decision Theories

Evidential Counterfactuals

Defining evidential counterfactuals will require more than just a finite factored set $F = (S, B)$. We will also need a probability distribution P on F that is nowhere zero. Recall that a probability distribution on a finite factored set is the product distribution of a separate probability distribution on each of the factors.

Given a nonempty proper subset $E \subseteq S$, we will sample a subset of E as follows. Let $X_E = \bigvee_S(h^F(\{E, S \setminus E\}))|E$. Note that X_E is a partition of E. For each $E' \in X_E$, sample x with probability $P(E'|E)$. Note that the sum of these probabilities will be 1.

Further, note that the E' sampled by the above procedure will always be a subset of E, and will always satisfy $\{E', S \setminus E'\} \preceq^F \{E, S \setminus E\}$.

Further, E' will be counterfactable relative to W for all $W \in \text{Part}(S)$. This is because if we take $s \in E'$ and $C = h^F(\{E, S \setminus E\})$, we have that $e \in E'$ if and only if $e \sim_b$ for all $b \in C$.

Thus, we have successfully specified a (randomized) procedure, which given an $E \subseteq S$, produces an $E' \subseteq E$ that is counterfactable and no later than E.

That gives our our evidential counterfactuals: $EC_E^{(W,P)} : S \to \Delta W$, which are given by setting $EC_E^{(W,P)}(s)(w)$ to the probability that $do_{E'}^{W}(s) = w$, where E' is defined as above from E and P.

Note that even the concept of evidential counterfactuals is going against the native ontology of evidential decision theory. Evidential decision theory doesn't really talk about interventions, and the type signatures above are about taking an intervention on an initial $s \in S$.

However, note that if you sample an $s \in S$ according to P, and then sample a $w \in W$ according to $EC_E^{(W,P)}(s)$, you will end up sampling each $w \in W$ with probability $P(w|E)$, so you get the same end result as if you just conditioned on E.

However, the evidential counterfactuals we define here also have the nice property that counterfacting on E will always leave unchanged all variables that are orthogonal to $\{E, S \setminus E\}$, so our counterfactuals are local in a sense.

Causal Counterfactuals

We will now define causal counterfactuals similarly to the above evidential counterfactuals. We will again need some extra structure. This time, we will imagine that our finite factored set $F = (S, B)$ was constructed from some Pearlian causal DAG, D.

For this, we will first need to describe a procedure for constructing a FFS from a Pearlian DAG. We will take one factor for each node in our DAG. The factor corresponding to the node v will have one part for each function from assignments of

states to the parents of v to assignments of a state to v. Note that when we construct a factored set in this way, for each $s \in S$, we have a well defined state for each node, which can be recursively defined using the functions you get by projecting onto each factor.

Note that this also means that for each node v, we get a function f_v which takes in an element of S, and outputs the state that element assigns to v.

We will not be able to describe causal counterfactuals for general subsets of S. For any set of nodes V, and any assignment of states t, where t(v) is a state of v for each $v \in V$, we can take $E_{(V,t)} = \{s \in S | \forall v \in V, f_v(s) = t(v)\}$. We will only be able to define a causal counterfactual for subsets of this form. These can be thought of events that can be described by assigning states to some set of nodes.

If we have $E = E_{(V,t)}$ of this form, we can define E' to to be the set of all elements such that for all $v \in V$, the factor corresponding to v has the value corresponding to the constant function const t(v).

Observe that $E' \subseteq E$, that $\{E', S \setminus E'\} \leq^F \{E, S \setminus E\}$, and that E' is counterfactable relative to W for all $W \in \text{Part}(S)$.

If $E = E_{(V,t)}$ as above, let $CC_E^{(W,D)}(s) = do_{E'}^{W}(s)$.

This is basically saying that we are take an E which corresponds to a collection of nodes having a specific assignment of states. We can't directly counterfact on that event, because there are many different assignments of the parents of the nodes that could result in those states, so instead we counterfact on the nodes being constantly equal those states, independent of the states of their parents.

CDT=EDT (for Counterfactable Events)

We have defined evidential and causal counterfactuals, and we can use them to now define EDT and CDT. Let $F = (S, B)$ be a finite factored set. Let W be a partition of S, and let $U : W \to [0, 1]$ be a utility function.

(We are assuming here that W has enough resolution to capture the agent's utility. For example, we could start with a utility function on S, and define W to be exactly the partition into level sets of that utility function.)

Let A be a partition of S, representing the agent's action. Let P be a distribution on F, representing the agent's beliefs. (We are ignoring any observations here, so A is more like a space of updateless policies.)

We can then define the EDT choice, which is the element $E \in A$ that maximizes the expectation of $U(w)$, where s is sampled according to P, and w is sampled according to $EC_E^{(W,P)}(s)$.

Similarly, if we have a Pearlian DAG D, and F was generated from D, and for all $E \in A$, E is of the form $E_{(V_E,t_E)}$, then we can define the CDT choice to be the $E \in A$ which maximizes the expectation of $U(EC_E^{(W,D)}(s))$, where s is sampled according to P.

Finally, if E is counterfactable for all $E \in A$, we can define a third decision theory, where we choose the $E \in A$, which maximizes the expectation of $U(do_E^W(s))$, where s is sampled according to P.

Note that if E is counterfactable for all $E \in A$, then this third decision theory will give the same result as EDT. Further if CDT is also well defined, all three decision theories will give the same result. CDT, EDT, and the third decision theory need not counterfact on the same events (i.e. the E' might be different, but the decisions will end up the same).

This is mostly saying that counterfactability is very strong: once you have counterfactable events, decision theory is over-determined.

In all of the above, we are doing updateless versions of the decision theories. Our agent is not making any observations.

Other Counterfactuals

I described evidential and causal counterfactuals above, not because I think they are the right way to take counterfactuals, but because I wanted to demonstrate that they

both fit into the framework where when counterfacting on E, you apply $do^{w}_{E'}$ for some counterfactable $E' \subset E$, no later than E itself. The fact that you have to pass to an E' comes from the fact that E was not actually at the native resolution of the action being counterfacted on.

There are other ways we could select an E', that look more at the gears of the process by which the decision is being made. When I am in a prisoner's dilemma with someone with similar psychology, part of my decision making process is happening on the part of me that is shared with my opponent, while some of my decision making process uses methods that are unique to me. Thus, my opponent's action is partially downstream from the calculation I am currently doing, and partially independent of it. Determining how much of the decision is in each part is difficult, and will not just be one extreme (EDT) or the other (CDT).

Counterfactability and ELK

An event is counterfactable if it screens off its own history from everything you care about. Dealing with non-counterfactable events is confusing, so instead of dealing with a non-counterfactable event E, we would rather deal with a counterfactable $E' \subseteq E$, and luckily, there is always a counterfactable E' that uses no more information that E. This story is sufficiently natural that it also applies outside of decision theory.

Say you have some opaque machine learning system that gives some output. Let $\{E_1, \dots, E_n\}$ partition the possible worlds according to the output of the system. The history of the output is all the information/thinking/computation/knowledge that goes into computing the output. (This is not the history according to FFS taken literally, but there is an analogy here that is deep, and part of the motivation for defining the FFS toy model.)

There are a bunch of details in the history of the output that do not make it into the output. This is fine. However, it is scary when there are details in the history of the output that are about things we care about, that do not make it into the output (or into our understanding of the output). Thus, we would like it if (our understanding of) the output of our ML system successfully screened off its own history from that which we care about.

Unfortunately, we have an opaque ML system that does not have this property, and thus will not have to tell us if it is trying to deceive us. What can we do?

We would like to add some additional notes to the output of the system, without necessarily changing the original system. These notes refine the partition of possible worlds according to output, and thus the worlds corresponding to a given output-notes

pair will be a subset of the worlds corresponding to just the output. We are finding a subset E_i' of our original E_i by adding more information through the notes.

We would like E_i' to be counterfactable, meaning that it contains all the information in its own history that is relevant to what we care about. Luckily, there is this intuition that this shouldn't be that hard. All the information is already there. We just need to pull it out, we shouldn't need to add any new information to do this.

All this is to say that we would like our system to instead of outputting E, output a counterfactable $E' \subseteq E$, and luckily there should always be a counterfactable E' no later than E. (I am being sloppy here with conflating the output and our understanding of the output, but still there is a rhyme in the structure that is hard to deny.)

I think that the thing that is going on here is that FFS gives us a nice definition of a good summary: Y is a good summary of X if Y screens off X from everything you care about. The goal of informed oversight is to have systems that output good summaries of themselves. Without this, the overseer cannot evaluate the consequences of the output in an unbiased way. Similarly, when considering non-counterfactable actions, an agent cannot judge the consequences of those actions in an unbiased way.

Countably Factored Spaces

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

A followup to Scott's "Finite Factored Sets", specifically the Applications part at the end where he talked about the infinite case.

As it turns out, there's a natural-seeming (at least to my mathematical tastes) way to generalize all the finite factored set stuff to the infinite case.

To be perfectly clear about what is being claimed:

1: It is possible to deal with arbitrary compact metric spaces, instead of just finite products of finite sets, with a minimum of fuss. In particular this means that we can have countably many axes/basis factors/coordinates now!

2: Orthogonality, time, history, and all that remain perfectly well-defined in the infinite-coordinate case, no meaningful differences crop up in comparision to how the finite case works.

4: Of the three ways to implement the extension to the infinite case which were suggested in Scott's post, it is closest in spirit to the second approach, with having the history of X be the intersection of all the sets $C \subseteq B$ s.t. $C \vdash X$. The proffered counterexample from Scott's post is forbidden by restricting the partitions to "sufficiently nice" ones, which makes everything Just Work.

5: Yes, the semigraphoid axioms, and everything not specifically mentioned, work out, with absolutely no special tricks needed besides the single starting restriction on only using "sufficiently nice" partitions and sets.

6: I was only able to get one direction of the fundamental theorem to work out (the conditional orthogonality to conditional probabilistic independence direction), since the nice attributes of ordinary vanilla polynomials don't last when you start mucking about with uncountable sums of countable products of variables. It's fairly plausible that the missing direction of the fundamental theorem does work out and I'm just not a good enough mathematician to show it, feel free to try it yourself. I suspect it'll require a different proof path than the original, though.

7: Daniel Filan has, completely independently of me, generalized all results in "Finite Factored Sets" to countable sets (note that in that case, there can still only be finitely many factors, while this version can have countably many factors. Daniel's version of things is unconstrained by the compactness restriction that I have, though.) Apparently everything works out perfectly, including the fundamental theorem, although the direction of it which I failed to show was also tricky for them, and TurnTrout contributed several essential pieces of the proof for that.

Let's get started.

Why Nice Partitions? What is "Nice" Anyways?

So, first off: where someone working in combinatorics sees a partition of a set, a topologist sees a quotient space.

If you've got a space S, then a partition X induces a function $\sim_X: S \to X$, which maps a point to the unique equivalence class it's in. Then, we just need to stick a topology on the set X, which ends up being "put as many open sets in X as you can while keeping the function $\sim_X$ continuous".

Basically, taking a quotient is the process of taking a space and using an equivalence relation to go "let's make a new space where all things in the same equivalence class is treated as the same point". Quotient spaces are generally written as something like $S/\sim$.

So, if you're trying to generalize finite factored sets to the infinite case, and are working with various sorts of partitions, looking at the quotient spaces of the partitions is a very natural thing to do.

The teensy little problem is that taking quotients is just not a very well-behaved operation topologically. Operations like disjoint sum and products are very good at preserving toplogical properties. You take two "nice" spaces, for many definitions of "nice", take a product, and it'll probably be "nice" as well, for many definitions of "nice". Quotients... not so much. If you're dealing with arbitrary quotients, you can whip up some pretty hideous-looking spaces. The natural next question is whether there's some sort of topological property that is preserved by "sufficiently nice" quotients.

To cash out what "sufficiently nice" means, Hausdorffness is a very basic property on topological spaces, that's held by most spaces encountered in typical mathematical practice (type theory excepted), and topology gets a lot stranger if it isn't present. It's "given two distinct points, there are disjoint open neighborhoods around the two points". Basically,

for any two distinct points x and y, it should be possible to come up with *some* notion of "close to x" and "close to y" that are mutually exclusive. So, we could demand that our quotient spaces at least be this nice.

Much nicer than that are compact metrizable spaces. Approximately, any space with a notion of distance, and for all ϵ, you can cover the space with finitely many patches of size ϵ. Well, actually, this isn't true, compactness is stronger than that, but I think it gets the spirit across. Examples of such spaces are finite batches of points, the space of all finite and infinite bitstrings, the space of probability distributions over a 256-dimensional hypersphere, the Mandelbrot set, and the space of all closed subsets of the interval [0,1]. They're extraordinarily nicely-behaved topologically while still managing to cover a healthy chunk of spaces that might be encountered in practice.

And, as it turns out, there is a *lovely* theorem that, for compact metrizable spaces S, if the quotient space $S/\sim$ happens to be Hausdorff (ie, not terrible), the quotient space will also be compact metrizable (about as nice as possible). Or at least, that's what Math StackExchange says.

The rough idea behind the proof is that quotients of compact spaces are always compact, so that leaves metrizability. Using compactness of S and Hausdorffness of $S/\sim$, it's possible to show the intermediate result that, for every set $K \subseteq S$ that's a union of equivalence classes and closed, you can take an open neighborhood of K and it will contain a smaller open neighborhood which *remains* open when shoved through $\sim$. Using this ability to craft open neighborhoods that remain open even after applying $\sim$, you can push enough open sets forward into the quotient space $S/\sim$ to show that it's second-countable and regular (the regularity argument was skipped in the linked post but it's not hard to fill in). And now, you can invoke the Urysohn Metrization Theorem to show the quotient space is metrizable.

A nifty related theorem that we won't use anywhere is that every compact metric space arises as a quotient of the space of infinite bitstrings. $\{0,1\}^\omega$ is all you need.

So... what happens if we try to make all the stuff from "Finite Factored Sets" work with compact metrizable spaces, and we just restrict the sorts of partitions we're dealing with to the nice ones? (those where the resulting quotient space $S/\sim_X$ is Hausdorff.) Well then, everything works out exactly as you'd expect and you have to change no important definitions, except one direction of the fundamental theorem gets really dang hard and I gave up on it. The nice behavior of the quotient spaces works miracles to tame the infinite case.

Let's start running through the Finite Factored Sets posts, flagging everywhere which requires special care. For all results not specifically mentioned, assume it works out perfectly with very little effort involved in cleaning things up for the infinite case. As usual, feel free to gloss over proofs if you want, but at least read the discussion and theorem statements.

Countably Factored Spaces: Introduction and Factorization

Not much really changes here, except now we aren't dealing with just a batch of sets, we've got some topological structure on them too. And the permissible quotients have to respect the topology appropriately.

The first differences start showing up around definition 8. Since we aren't permitting just any old partition anymore, we need to put in some work to check that the intersection of nice partitions is a nice partition, especially since there can be uncountably many partitions!

Proposition 1: *Given a compact metrizable space* S*, and a set of partitions* C *s.t. for all* $c \in C$, $S/\sim_c$ *is Hausdorff, the partition* $\bigvee C$*, defined by* $s_0 \sim_{\bigvee C} s_1 \leftrightarrow \forall c \in C : s_0 \sim_c s_1$*, has the property that* $S/\sim_{\bigvee C}$ *is Hausdorff.*

Proof: Fix two distinct points in $S/\sim_{\bigvee C}$. These correspond to equivalence classes in S, so we'll write them as $[s_0]_{\bigvee C}$ and $[s_1]_{\bigvee C}$ for some distinguished s_0 and s_1. From this point on, a superscript of p on an equivalence class like this means we'll be treating it as a point in a quotient space, a superscript of e means we'll be treating it as an subset of S .

Our task is to find open neighborhoods for those two points in $S/\sim_{\bigvee C}$ which don't overlap. Since they're distinct points in the quotient, the corresponding equivalence classes in S are distinct, s_0 and s_1 are in different equivalence classes. There has to be *some* $c^* \in C$ s.t. $[s_0]^e_{c^*} \neq [s_1]^e_{c^*}$ (because otherwise, s_0 and s_1 would be in the same equivalence class according to $\sim_{\bigvee C}$, which is impossible).

What now? Well, now that we've got a c^* that thinks that s_0 and s_1 are noticeably different, make a function $\phi : S/\sim_{\bigvee C} \to S/\sim_{c^*}$ defined as $\phi([s]^p_{\bigvee C}) := [s]^p_{c^*}$. Basically, you can think of a point in $S/\sim_{\bigvee C}$ aka an equivalence class under $\bigvee C$, as being associated with one equivalence class for each $c \in C$, by how the intersection equivalence class is defined. And that tells you what point to map it to in $S/\sim_{c^*}$.

Next up, we'll show that $\phi \circ \sim_{\bigvee C} = \sim_{c^*}$. This is easy, we just take some $s \in S$, and go:

$$\phi(\sim_{\bigvee C}(s)) = \phi([s]^p_{\bigvee C}) = [s]^p_{c^*} = \sim_{c^*}(s)$$

Done. Alright, now that we've got enough setup out of the way, we can start building our disjoint open neighborhoods. The two points $[s_0]^p_{c^*}$ and $[s_1]^p_{c^*}$ are distinct in $S/\sim_{c^*}$, because they correspond to the equivalence classes $[s_0]^e_{c^*}$ and $[s_1]^e_{c^*}$ in S which are distinct. So, they've got some disjoint open neighborhoods O_0 and O_1, in $S/\sim_{c^*}$, since it's Hausdorff.

We will now let our disjoint open neighborhoods of $[s_0]^p_{\bigvee C}$ and $[s_1]^p_{\bigvee C}$ be $\phi^{-1}(O_0)$ and $\phi^{-1}(O_1)$. They're disjoint because the preimage of any two disjoint sets is disjoint. They contain the requisite points because $\phi([s_0]^p_{\bigvee C}) = [s_0]^p_{c^*} \in O_0$, and the same for the other one. And they're open because their preimage under the quotient function $\sim_{\bigvee C}$ is open. To demonstrate this, we have

$$\sim^{-1}_{\bigvee C}(\phi^{-1}(O_0)) = (\phi \circ \sim_{\bigvee C})^{-1}(O_0) = \sim^{-1}_{c^*}(O_0)$$

Because we proved that the functions compose like that, and also, since $\sim_{c^*}$ is a continuous function $S \to S/\sim_{c^*}$, the preimage of an open is an open. And we're done! We got disjoint open neighborhoods for any two distinct points. □

So, this is very nice. You can intersect nicely-behaved partitions however you wish, and it'll still stay as a nicely-behaved partition. There's very strong closure properties.

The next time we hit something nontrivial is around definition 10. A batch of partitions was defined to factorize a set iff the function mapping a point to the tuple of equivalence classes it's in is a bijection. But wait, we have topological structure now! As it turns out, amazingly enough, if that function $\pi : S \to \prod_{b \in B} S/\sim_b$ defined as $\pi(s) = \lambda b.\sim_b (s)$ is a bijection, it's actually going to be a homeomorphism! That's the topology version of an isomorphism, it's a continuous function with a continuous inverse, like the mapping back and forth between a donut and a coffee cup. The spaces have identical topological structure and we can consider them as basically the same.

This is quite nifty. Leaving S with whatever topological structure it had originally, and equipping it with the product topology you get from the product of the various quotient spaces you made, are the *same thing*. So, our restriction on partitions implies the *only* factorizations available are those of the form "look at your space S and realize it *already* had the topological structure of a cartesian product from the start, and the quotient functions are just projecting down on the various coordinates."

Since the definition of a finite factored set was essentially a pair of a set, and the batch of quotients/partitions which let you view the set as a product of other sets, this lets us analogously define a countably factored space as a pair of a nice-enough topological space, and the batch of quotients/partitions which let you view the space as a product of other topological spaces.

Definition 1: Countably Factored Space

A countably factored space is a (S, B) *pair, where* S *is a compact metrizable space, and* B *is a set of partitions* $\sim_b$ *s.t.* $S/\sim_b$ *is a compact metrizable space for all* b*, and the function* $S \to \prod_{b \in B} S/\sim_b$ *given by* $\lambda s . (\lambda b . \sim_b (s))$ *is a bijection (and actually, a homeomorphism, the topological structure is preserved, though this is nonobvious)*

Proposition 2: *If the function* $\pi : \lambda s . (\lambda b . \sim_b (s))$ *is a bijection between* S *(a compact metrizable space) and* $\prod_{b \in B} S/\sim_b$*, with all the* $S/\sim_b$ *being compact metrizable, then it's a homeomorphism.*

We'll do this by showing that π is continuous, and then using the fact that π is a continuous bijection to show that π^{-1} is continuous. We'll use x for a point in the product space, and x_b for the b'th coordinate.

First up, let's get a form for π^{-1}. The form of the inverse is $\pi^{-1}(x) = \bigcap_{b \in B} \sim_b^{-1} (x_b)$. Map a batch of equivalence classes to their intersection. If there were multiple points in that intersection, then applying π would map them to the same point (they're in the same list of equivalence classes), contradicting injectivity of π. And if there were no points in that intersection, then there'd be no point in S with that particular combination of equivalence classes, contradicting surjectivity of π.

To show that π is continuous, take some base open set O for $\prod_{b \in B} S/\sim_b$. By how the product topology works, there are finitely many b_n and open sets $O_n \subseteq S/\sim_{b_n}$, s.t. $O = O_1 \times O_2 .. \times O_n \times \prod_{b \in B/\{b_1 ... b_n\}} S/\sim_b$.

By how π^{-1} is defined, and the fact that the set O factorizes into the product of a bunch of other sets O_b for each of the coordinates, we have

$$\pi^{-1}(O) = \bigcap_{b \in B} \sim_b^{-1} (O_b)$$

And then, unpack what the form of the various O_b are.

$$= \bigcap_{b \in \{b_1 ... b_n\}} \sim_b^{-1} (O_n) \cap \bigcap_{b \in B/\{b_1 ... b_n\}} \sim_b^{-1} (S/\sim_b)$$

$$= \bigcap_{b \in \{b_1 ... b_n\}} \sim_b^{-1} (O_n) \cap \bigcap_{b \in B/\{b_1 ... b_n\}} S = \bigcap_{b \in \{b_1 ... b_n\}} \sim_b^{-1} (O_n)$$

And bam, we've written the preimage of our base open, through the function π, as a finite intersection of preimages of open sets through continuous functions $\sim_b$. So it's a finite intersection of opens, and so is open. The preimage of all base opens is open.

This shows that π is continuous, because you can write any open in the product space as a union of base opens, and the preimage of a union is the union of the preimages, so it'd be the union of a bunch of open sets, ie, open.

Now that we know π is continuous, we'll show that π^{-1} is. To do this, we'll show that the preimage of any closed set through π^{-1} is closed (as that's equivalent to the preimage of any open set being open, ie, continuity.) The inverse of

π^{-1} is π, so we need to show that for any closed set in S, applying π produces a closed set in the product space, and we'll have a continuous inverse.

This holds because, for any closed set $C \subseteq S$, since S is a compact metrizable space, C is compact. Also, $\pi(C)$ is compact, because applying a continuous function to a compact set makes a compact set. And, arbitrary products of Hausdorff spaces (our quotient spaces) are Hausdorff, and in any Hausdorff space, compact sets are closed, so $\pi(C)$ is closed.

Bam, π^{-1} is continuous, and since we showed that π is, it's a homeomorphism. □

There's another result that I should mention now, because it's implicitly used in a lot of upcoming arguments, and it justifies the use of the term "countably factored space".

Proposition 3: *Any compact metrizable space* S *can only have countably many nontrivial factors.*

Assume this is false, and there are uncountably many nontrivial factors. We can identify our compact metrizable space S with $\prod_{b \in B} S/\sim_b$ by Proposition 2, there's the same topology. All factors are nontrivial, so each space $S/\sim_b$ has two distinct points in them. And all the $S/\sim_b$ are Hausdorff. Since it's an uncountable product, we can call upon the power of MathOverflow to conclude that $\prod_{b \in B} S/\sim_b$ isn't a first-countable space. But all compact metrizable spaces are first-countable, contradiction. □

Past this point, everything works out precisely as it did in the original post.

Countably Factored Spaces: Conditional Orthogonality

We're going to skip proving that history exists the first time around, and all that ordinary orthogonality stuff, because Scott's posts had to run through the proofs a second time in greater generality for conditional everything, so we might as well just show that conditional history exists (and the other associated results), and go "conditioning on the entire space S recovers the ordinary case" so we only have to prove things once.

If a statement is not mentioned here or proved, assume it works out perfectly straightforwardly with no issues and requiring no special tricks. I'll only be focusing on the propositions which take a bit of work to generalize to the infinite case.

There is something important to note in this section. We can't condition on an arbitrary subset of S, it has to be closed! This is the same sort of topological restriction as the one we imposed on the quotients. Any closed subset of a compact metrizable space is compact metrizable too, so it's motivated by the same reasoning as what motivated the restriction on the quotients. It's far far easier to prove this fact though, since the argument for metrizability is just "if you've got a metric on the full space, the subspace inherits that metric". As usual, we'll restrict subpartitions to those which make the quotient space Hausdorff.

Our first order of business is going to be showing that if you've got a permissible partition X, then $\sim_X$ restricted to the closed set E is a permissible partition of E. We'll actually show something considerably stronger, that the spaces $E/\sim_{X|E}$ and $\sim_X(E)$ (as a subspace of $S/\sim_X$) are homeomorphic. In other words, it doesn't matter whether you slice E out first and then take a quotient of it, or take the quotient first and slice the image of E out of the quotient space, you'll get the same space. So, in particular, you can go "$S/\sim_X$ is compact-metrizable, and $\sim_X(E)$ happens to be closed if E is, so these two isomorphic spaces are also compact-metrizable". And this means that taking permissible quotients of a subspace works perfectly well and causes no issues.

Proposition 4: *For a compact metrizable space* S*, closed subset* E*, and partition* X *s.t.* S/ $\sim_X$ *is Hausdorff, then* E/ $\sim_{X|E}$ *and* $\sim_X$ (E) *are homeomorphic.*

Our attempted homeomorphism $\phi : E/\sim_{X|E} \to \sim_X (E)$ will be mapping the point $[s]^p_{X|E}$ (where $s \in E$) to $[s]^p_X$.

First off, we've gotta show injectivity. Let's say $[s_0]^p_{X|E}$ and $[s_1]^p_{X|E}$ are distinct in E/ $\sim_{X|E}$. Then $\phi([s_0]^p_{X|E}) = [s_0]^p_X$, and similar for s_1. If these two points were identical, $[s_0]^p_X = [s_1]^p_X$, then the sets $[s_0]^e_X$ and $[s_1]^e_X$ would be identical in S, and so $s_0 \sim_X s_1$, but $s_0, s_1 \in E$, so $s_0 \sim_{X|E} s_1$, and so $[s_0]^p_{X|E} = [s_1]^p_{X|E}$, but they're distinct points, contradiction.

Now for surjectivity. Fix some $[s]^p_X \in \sim_X (E)$. Then there's some $s' \in E$ where $\sim_X (s') = [s]^p_X$. So, in particular, $[s']^p_X = [s]^p_X$. This point s', due to being in E, maps to the point $[s']^p_{X|E} \in E/\sim_{X|E}$. Then, we have $\phi([s']^p_{X|E}) = [s']^p_X = [s]^p_X$ And bam, we found a point that maps onto it. Surjectivity is established. It's a bijection.

The inverse of the bijection maps $[s]^p_X$ to $[s']^p_{X|E}$, where s' is an arbitrary element of E where $s' \sim_X s$ (which must exist because $[s]^p_X$ is taken from the image of E).

Now, let's show continuity of ϕ. Fix an arbitrary open set O in $\sim_X (E)$. By how the subspace topology works, $O = O' \cap \sim_X (E)$, where O' is some open in S/ $\sim_X$. We will attempt to show that $\sim^{-1}_{X|E} (\phi^{-1}(O)) = E \cap \sim^{-1}_X (O')$, establishing that the preimage of $\phi^{-1}(O)$ is open in E equipped with the subspace topology (intersection of E and the preimage of an open, ie, an open), which would show that $\phi^{-1}(O)$ is open in E/ $\sim_{X|E}$ by how the quotient topology works, establishing the continuity of ϕ.

So, our proof goal switches to proving $\sim^{-1}_{X|E} (\phi^{-1}(O)) = E \cap \sim^{-1}_X (O')$

Let a point s lie in the first set. Then it must be in E, and we must have that $\sim_{X|E} (s) \in \phi^{-1}(O)$. Rewrite this a little bit as $[s]^p_{X|E} \in \phi^{-1}(O)$, and then, by applying ϕ to both sides, we have $[s]^p_X \in O \subseteq O'$, establishing that $s \in \sim^{-1}_X (O')$. So, we have one subset inclusion direction, $\sim^{-1}_{X|E} (\phi^{-1}(O)) \subseteq E \cap \sim^{-1}_X (O')$

For the other subset inclusion direction, let a point s lie in E, and applying $\sim_X$, we have $[s]^p_X \in O'$. In particular, $[s]^p_X \in \sim_X (E)$, so we have $[s]^p_X \in O$.

Now, for this s, since it's in E, $\phi(\sim_{X|E} (s)) = \phi([s]^p_{X|E}) = [s]^p_X \in O$. So, it lies in the composition of preimages on the left side, and we have equality, which, by previous arguments, establishes that ϕ is continuous.

Now, we'll show that E/ $\sim_{X|E}$ is Hausdorff. Take two distinct points in it, shove them through ϕ to get two distinct points in $\sim_X$ (E), use Hausdorffness of$\sim_X$ (E) (which happens because E is closed and so compact, and $\sim_X$ is continuous, so $\sim_X$ (E) is compact, and S/ $\sim_X$ is compact metrizable, so $\sim_X$ (E) is a closed subset of a compact metrizable space and so is compact metrizable, and thus Hausdorff) to fit two disjoint open neighborhoods around your two points, then use continuity of ϕ to show that ϕ^{-1} pulls those disjoint open neighborhoods back to make disjoint open neighborhoods of your two original points in E/ $\sim_{X|E}$.

Hm, E/ $\sim_{X|E}$ is a Hausdorff quotient of a compact metrizable space (E) and so is compact metrizable. So, in particular, any closed set C must be compact, and using continuity of ϕ, ϕ(C) is compact, and by Hausdorffness of $\sim_X$ (E), ϕ(C) is closed. So, ϕ maps closed sets to closed sets. This is equivalent to the preimage of a closed set being closed, according to the function ϕ^{-1}. So, ϕ^{-1} is continuous too, and we have a homeomorphism. □

Alright, where to go from here? Well, things go perfectly fine up until you hit Proposition 21.5, the "conditional history exists" result. We'll need to strengthen it to arbitrary unions and intersections for things from here on out to work properly. In particular, it means that you can just intersect all the sets of coordinates C where $C \vdash X$, to get a unique minimal set of coordinates that generates X, and bam, that's a history, but for the infinite case. Perfectly well-defined, no problems whatsoever.

Proposition 5 (Reproof of Proposition 21.5, Infinite History Remix:) *If* X *is a permissible subpartition, and there's a bunch of sets* $C_i \subseteq B$ *where* $\forall i : C_i \vdash X$*, then* $\bigcap_i C_i \vdash X$*, and* $\bigcup_i C_i \vdash X$*.*

The proofs of these two results are extremely similar, just flipped around, so we'll provide the general proof framework for both cases.

Step 1 is to show it for the intersection of two sets of coordinates, or the union of two sets of coordinates. Step 2 is to show it for the intersection of a descending sequence of sets of coordinates, or the union of an ascending sequence of sets of coordinates. Step 3 is to use steps 1 and 2 and "there's only countably many coordinates" to prove the whole thing.

The proof of Step 1 perfectly follows the way it works in Scott's post, there's no meaningful differences going on, the exact same argument works.

For the proof of step 2, we'll given the argument for intersection (and for union in parentheses). Fix a sequence of sets C_n s.t. $C_0 \supseteq C_1 \supseteq C_2 \ldots$ (or, for union, have $\subseteq$ instead).

Now, for any particular coordinate b, if $b \in \bigcap_n C_n$ (for union, $b \notin \bigcup_n C_n$), then it'll be in all the C_n (none of the C_n),

so for intersections we have $(\chi_{\bigcap_n C_n}(s,t))_b = s_b = \lim_{n\to\infty}(\chi_{C_n}(s,t))_b$

and for unions we have $(\chi_{\bigcup_n C_n}(s,t))_b = t_b = \lim_{n\to\infty}(\chi_{C_n}(s,t))_b$

now, if $b \notin \bigcap_n C_n$ (for union, $b \in \bigcup_n C_n$), then there's some finite n where C_n excludes b (includes b), and then it never returns after that (always is present after that) because the C_n get smaller (larger) as n increases. So, again, for b like that,

for intersections, we have $(\chi_{\bigcap_n C_n}(s,t))_b = t_b = \lim_{n\to\infty}(\chi_{C_n}(s,t))_b$

and for unions we have $(\chi_{\bigcup_n C_n}(s,t))_b = s_b = \lim_{n\to\infty}(\chi_{C_n}(s,t))_b$

So, since we have convergence in each individual coordinate, and there's only countably many coordinates, this means that $\chi_{\bigcap_n C_n}(s,t) = \lim_{n\to\infty} \chi_{C_n}(s,t)$ (for union, just switch the intersection to a union).

Now, since all the $C_n \vdash X$, this means that all the $\chi_{C_n}(s,t)$ will land in $[s]_X$, which is a closed set (single points are closed in Hausdorff spaces, the preimage of a closed set through a continuous function is closed, so $[s]_X$ is closed). Thus, the limit point will also land in $[s]_X$, establishing that $\bigcap_n C_n \vdash X$ (or $\bigcup_n C_n \vdash X$)

Now for step 3. Again, the union argument is in parentheses. Given your collection of sets C_i, index the coordinates by N, there's only countably many coordinates. For coordinate n, if $n \in \bigcap_i C_i$ (or $n \notin \bigcup_i C_i$), then let C_n be whatever set C_i you want. If $n \notin \bigcap_i C_i$ (or $n \in \bigcup_i C_i$), then let C_n be some C_i which excludes (includes) the coordinate n, which must exist. Now, we can go

$$\bigcap_{i \in I} C_i = \bigcap_{n \in N} C_n = \bigcap_{n \in N} \bigcap_{m \leq n} C_m$$

(or for union) $\bigcup_{i\in I} C_i = \bigcup_{n\in N} C_n = \bigcup_{n\in N} \bigcup_{m\leq n} C_m$

Basically, since our C_n were picked to exclude (include) every coordinate it's possible to exclude (include), we can rewrite our big intersection (union) as a countable intersection (union). Then just rewrite a bit.

Now, for each n, we have $\bigcap_{m\leq n} C_m \vdash X$ (or $\bigcup_{m\leq n} C_m \vdash X$) because we proved the finite intersection (union) case and can use induction. And the sequence $C_0, C_0 \cap C_1, C_0 \cap C_1 \cap C_2 \ldots$ (same thing but with $\cup$) is a descending (ascending) sequence of sets, so we can apply our proof from that case to establish that $\bigcap_{i\in I} C_i = \bigcap_{n\in N} \bigcap_{m\leq n} C_m \vdash X$ (or $\bigcup_{i\in I} C_i = \bigcup_{n\in N} \bigcup_{m\leq n} C_m \vdash X$) and we're done! □

Ok, now that that's taken care of... is there anything else on the list that may be particularly difficult? Well, our next spot of mild trouble is in Proposition 23, specifically, the part about showing that the history of a supremum of partitions is the union of the histories of the component parts.

Proposition 6 (Reproof of Proposition 23.2): *Given X_i, a collection of subpartitions all with the same domain, we have $h(\bigvee X_i) = \bigcup_{i\in I} h(X_i)$.*

So, one direction of this, that $\forall j : h(\bigvee X_i) \vdash X_j$ to show that the history of the supremum is as larger or larger than the union of all the other histories, is pretty easy. We have that $\bigvee X_i \geq_E X_j$ for all j, so the history of $\bigvee X_i$ manages to generate X_j, for all j, so all the $h(X_i)$ are a subset of the history of $\bigvee X_i$, establishing that $h(\bigvee X_i) \supseteq \bigcup_{i\in I} h(X_i)$.

The other subset direction, which will be established by showing $\bigcup_{i\in I} h(X_i) \vdash \bigvee X_i$, requires taking a bit more care. Our first order of business is showing that $\bigvee X_i \leq_E \bigcup_{i\in I} h(X_i)|E$. Assume that two points, s and t, both in E, fulfill $s \sim_{\bigcup_{i\in I} h(X_i)} t$. Then, we have $\forall i \in I : s \sim_{h(X_i)} t$. And since $h(X_i) \vdash X_i$ for all i, and s, t are both in E, this implies $\forall i \in I : s \sim_{X_i} t$. Which is equivalent to $s \sim_{\bigvee X_i} t$. And so, we have established that inequality.

But that isn't enough, the last piece we need to conclude the argument is that $\chi_{\bigcup_{i\in I} h(X_i)}(E,E) = E$. We'll do this by showing that for arbitrary $s,t \in E$, that $\chi_{\bigcup_{i\in I} h(X_i)}(s,t) \in E$. First off, we can go

$$\chi_{\bigcup_{i\in I} h(X_i)}(s,t) = \chi_{\bigcup_n \bigcup_{m\leq n} h(X_m)}(s,t) = \lim_{n\to\infty} \chi_{\bigcup_{m\leq n} h(X_m)}(s,t)$$

The first inequality was the same sort of ascending/descending countable chain argument used to make a sequence of ever-larger sets in Proposition 6. Then, we just use the usual limit argument with that of "each coordinate individually converges, so we have overall convergence". Now, if we knew that each $\chi_{\bigcup_{m\leq n} h(X_m)}(s,t)$ was in E, then by closure of E, we'd have that our desired point lands in E. So, we just have to show these finite stages land in E. We do this with an induction proof. Clearly, $\chi_{h(X_0)}(s,t) \in E$, because $h(X_0) \vdash X_0$, so $\chi_{h(X_0)}(E,E) = E$. For the induction step, we go

$$\chi_{\bigcup_{m\leq n+1} h(X_m)}(s,t) = \chi_{h(X_{n+1})}(s, \chi_{\bigcup_{m\leq n} h(X_m)}(s,t)) \in E$$

Where that last inclusion is because, by induction assumption, $\chi_{\bigcup_{m\leq n} h(X_m)}(s,t) \in E$, and also $h(X_{n+1}) \vdash X_{n+1}$.

And so, we have that $\chi_{\bigcup_{i\in I} h(X_i)}(E,E) = E$, and so we've established that $\bigcup_{i\in I} h(X_i) \vdash \bigvee X_i$, establishing the other subset inclusion direction, and thus equality. □

Alright, where is this heading? Well, the semigraphoid axioms, of course. The only roadblock left is that we have to reprove Lemma 2 from scratch. The old proof no longer suffices due to using impermissible partitions, and the old proof path cannot be repaired.

Proposition 7 (reproof of Lemma 2): *Let* X, Y *be subpartitions, and* E *be their domain. Then*

$h(X \vee Y) = h(X) \cup \bigcup_{x\in X} h(Y|x)$.

Proof. Since $X \leq_E X \vee_E Y$, we have $h(X) \subseteq h(X \vee Y)$. Symmetrically, for all $x \in X$, since $Y|x \subseteq X \vee Y$, we have $h(Y|x) \subseteq h(X \vee Y)$ by Proposition 23 in Scott's paper. Thus, $h(X \vee Y) \supseteq h(X) \cup \bigcup_{x\in X} h(Y|x)$. One direction down, one to go.

In order to begin attacking the reverse direction, our first order of business is taking a detour to establish that $h(X) \cup \bigcup_{x\in X} h(Y|x) \vdash \mathrm{Ind}_E$, the indiscrete partition of E. In order to do this, we'll show that for any particular $x \in X$, $h(X) \cup h(Y|x) \vdash \mathrm{Ind}_E$. Then, just yeet Proposition 5 (from this post) at that to show our desired result, that you can union everything together and it'll still generate Ind_E.

So, let x be arbitrary in X, and r be an arbitrary point in x, and s,t be arbitrary points in E.

$$\chi_{h(X)\cup h(Y|x)}(s,t) = \chi_{h(X)}(s, \chi_{h(Y|x)}(s,t)) = \chi_{h(X)}(s, \chi_{h(X)}(r, \chi_{h(Y|x)}(s,t)))$$

$$= \chi_{h(X)}(s, \chi_{h(Y|x)}(\chi_{h(X)}(r,s), \chi_{h(X)}(r,t)))$$

And both $\chi_{h(X)}(r,s)$ and $\chi_{h(X)}(r,t)$ land in $[r]_X = x$, because $h(X) \vdash X$ and r,s,t are all in E.

Accordingly, $\chi_{h(Y|x)}(\chi_{h(X)}(r,s), \chi_{h(X)}(r,t)) \in [\chi_{h(X)}(r,s)]_Y \subseteq E$, because $h(Y|x) \vdash Y|x$, and both of the chi thingies land in x. And then, since both s and the long $\chi_{h(Y|x)}$ term land in E, $\chi_{h(X)}(s, \text{stuff})$ lands in E. And so, since s and t were arbitrary in E, we have established that $\chi_{h(X)\cup h(Y|x)}(E,E) = E$, which is sufficient by itself to establish that $h(X) \cup h(Y|x) \vdash \mathrm{Ind}_E$ for our arbitrary x. And then, by Proposition 5, $h(X) \cup \bigcup_{x\in X} h(Y|x) \vdash \mathrm{Ind}_E$.

At this point, we can back out of the lemma and resume our work of showing that this batch of histories is capable of generating $X \vee_E Y$.

We'll be using the definition of generation where we try to show that if s and t lie in the appropriate set (E in this case), then chimera-ing them together lands in the same equivalence class as s. We have

$$\chi_{h(X) \cup \bigcup_{x \in X} h(Y|x)}(s, t) = \chi_{h(Y|[s]_X)}(s, \chi_{h(X)}(s, \chi_{h(X) \cup \bigcup_{x \in X} h(Y|x)}(s, t)))$$

Now, since we had previously derived that $h(X) \cup \bigcup_{x \in X} h(Y|x) \vdash \text{Ind}_E$, that means that $\chi_{h(X) \cup \bigcup_{x \in X} h(Y|x)}(s, t) \in E$. Moving out one layer, $\chi_{h(X)}(s, \text{stuff})$ has both components landing in E, and $h(X) \vdash X$, so that result will land in $[s]_X$.

Moving out another layer from that, $\chi_{h(Y, [s]_X)}(s, \text{stuff})$ has both components landing in $[s]_X$, and $h(Y|[s]_X) \vdash Y|[s]_X$, so the result will land in $[s]_{Y|[s]_X}$, aka $[s]_Y \cap [s]_X$, aka $[s]_{X \vee Y}$. And we're done! Our collection of histories generates $X \vee Y$.

□

But of course, all of this was just a warmup for showing the semigraphoid axioms, which... go suprisingly smoothly. Most of the aggrivation was concentrated in reproving Lemma 2.

Countably Factored Spaces: Polynomials and Probability

And then things go poorly right around here. Or at least, there's a big difficulty spike if you're trying to rescue all results. Apparently, for Daniel's "countable sets" case, since there can only be finitely many coordinates, you only have to deal with power series (generalization of polynomials) involving sums of countably many terms of finitely many variables each, which is a little tricky, but doable, and TurnTrout did it.

But here, when I tried, I wound up dealing with sums of uncountably many terms, consisting of countably many variables each. Which, apparently, don't work well at all. Now, an uncountable sum over infinitesimally small things is kinda like an integral, or taking the union of a bunch of points to make a set, and as it turns out, a whole bunch of the results in this section have analogues if you swap out the polynomial of a set for the set itself.

With this reframing, Proposition 26 becomes trivial, and Proposition 27 turns into the statement that chimera-ing sets together can be viewed as projecting them down to the appropriate coordinates, and then taking the product of those.

Let's use S_C as an abbreviation for $\prod_{b \in C} S/\sim_b$, basically, the space you get when you project down to the factors in C.

Put another way, Proposition 27 is basically saying that, when C_0 and C_1 are disjoint sets of coordinates, then

$$pr_{S_{C_0 \cup C_1}}(\chi_{C_0}(E_0, E_1)) = pr_{S_{C_0}}(E_0) \times pr_{S_{C_1}}(E_1)$$

This can be trivially verified from understanding what the chimera function does. $pr_{S_{C_0}}(E_0)$ are the possible coordinate values for the coordinates in C_0 which are available for use by the chimera function, and $pr_{B/S_{C_0}}(E_1)$ are the possible coordinate values for the coordinates in B/S_{C_0} which are available for use by the chimera function, and the chimera function can put these together however it wishes, making the set $pr_{S_{C_0}}(E_0) \times pr_{B/S_{C_0}}(E_1)$. Project out some extra coordinates, and you get the result.

Proposition 28 turns into the statement that, given a set $E \subseteq S$ that factorizes as $F \times G$, you can just let C be the set of relevant coordinates for the set F, and write F and G as $pr_{S_C}(E)$ and $pr_{S_{B/C}}(E)$.

Propositions 29 and 30 turn into the statement that given some set E, there's a way to split it into as many factors as possible, until eventually you hit a batch of "irreducible pieces", sets of coordinates C where $pr_C(E)$ just can't be factorized any further, it's an odd shape, and that these sets of coordinates partition B, and multiplying these things together makes E again. The rough proof path for this is showing that given any way of factorizing E in two different ways using coordinates $(C_0, B/C_0)$ in one way, and $(C_1, B/C_1)$ in another way, then $(C_0 \cap C_1, B/(C_0 \cap C_1))$ is also a factorization of the set. Then we do our usual sort of argument with the countable descending chains to get that for any particular coordinate b, there's a minimal set of coordinates including b that don't let you factor the set any more.

It's slightly tricky to show, but Lemma 3 carries over and shows an equivalence between $X \perp Y|Z$, and the statement "for all x, y, and z, if you decompose $x \cap z$ and $y \cap z$ into irreducible pieces (be very sure to label all pieces with the coordinates they're pinning down, to make the pieces distinguishable), and throw them into a multiset, it'll be the same as decomposing z and $x \cap y \cap z$ into irreducible pieces and throwing those pieces into a multiset"

But, sadly, when we go to the Fundamental Theorem of Finite Factored Sets, it's just too hard (for me personally, you may be different) to show that probabilistic independence implies two multisets of irreducible pieces (labeled with coordinates) are identical.

But, it *is* pretty easy to get the other direction, where conditional orthogonality implies conditional independence. Well, kinda. I'd ideally like to strengthen it to talk about conditional probabilities (if it's very improbable to select any particular event z), but I only proved the unconditional probabilities version, though a slightly stronger version than was originally stated.

Proposition 7 (reproof of one direction of the fundamental theorem): *Given any factorized probability distribution* μ*, permissible partitions* X, Y, Z *s.t.* $X \perp Y \mid Z$*, subsets* $X' \subseteq X$ *and* $Y' \subseteq Y$*, and* $z \in Z$*, then*

$$P_\mu(\bigcup_{x \in X'} x \cap z) \cdot P_\mu(\bigcup_{y \in Y'} y \cap z) = P_\mu(\bigcup_{x \in X'} x \cap \bigcup_{y \in Y'} y \cap z) \cdot P_\mu(z)$$

Proof: Since $X \perp Y|Z$, this means that for all $z \in Z$, we have $h(X|z) \cap h(Y|z) = \emptyset$. Use C to abbreviate $h(X|z)$. Now, clearly, by the definition of $\vdash$, we have $\chi_C(z, z) = z$, so we also have $\chi_{B/C}(z, z) = z$. Also, $h(Y|z) \subseteq B/C$, so $Y|z \leq_z h(Y|z)|z \leq_z \bigvee(B/C)|z$ and so this means that $B/C \vdash Y|z$.

Since $C \vdash X|z$ (remember, it's $h(X|z)$), we have that for all $x \in X$, $\chi_C(x \cap z, z) = x \cap z$. Rephrasing this somewhat, it's saying $pr_{S_C}(x \cap z) \times pr_{S_{B/C}}(z) = x \cap z$ for all x. Accordingly, we have

$$\bigcup_{x \in X'} x \cap z = \bigcup_{x \in X'} (pr_{S_C}(x \cap z) \times pr_{S_{B/C}}(z))$$

$$= (\bigcup_{x \in X'} (pr_{S_C}(x \cap z)) \times pr_{S_{B/C}}(z) = pr_{S_C}(\bigcup_{x \in X'} x \cap z) \times pr_{S_{B/C}}(z)$$

Also, since $B/C \vdash Y|z$, we can run through similar arguments to get that

$$\bigcup_{y \in Y'} y \cap z = pr_{S_{B/C}}(\bigcup_{y \in Y'} y \cap z) \times pr_{S_C}(z)$$

And, since we have

$$\chi_C(\bigcup_{x \in X'} x \cap z, \bigcup_{y \in X'} y \cap z) \subseteq \chi_C(\bigcup_{x \in X'} x \cap z, z) = \bigcup_{x \in X'} x \cap z$$

and

$$\chi_C(\bigcup_{x \in X'} x \cap z, \bigcup_{y \in X'} y \cap z) \subseteq \chi_C(z, \bigcup_{y \in X'} y \cap z) = \chi_{B/C}(\bigcup_{y \in X'} y \cap z, z) = \bigcup_{y \in Y'} y \cap z$$

This means that we have

$$\chi_C(\bigcup_{x\in X'} x\cap z, \bigcup_{y\in Y'} y\cap z)\subseteq \bigcup_{x\in X'} x\cap \bigcup_{y\in Y'} y\cap z$$

But, hang on, those sets we're chimera-ing together are projections of sets as large or larger than the intersection of X', Y', and z, so we must have equality there.

Accordingly, we can use that factorization and get

$$\bigcup_{x\in X'} x\cap \bigcup_{y\in Y'} y\cap z = pr_{S_C}(\bigcup_{x\in X'} x\cap z)\times pr_{S_{B/C}}(\bigcup_{y\in Y'} y\cap z)$$

and also, since we had $\chi_C(z,z)=z$, this means that $z = pr_{S_C}(z)\times pr_{S_{B/C}}(z)$

And so, throwing probabilities in the mix, we have

$$P_\mu(\bigcup_{x\in X'} x\cap z)\cdot P_\mu(\bigcup_{y\in Y'} y\cap z)$$

$$= P_{\mu_C}(pr_{S_C}(\bigcup_{x\in X'} x\cap z))\cdot P_{\mu_{B/C}}(pr_{S_{B/C}}(z)) + P_{\mu_{B/C}}(pr_{S_{B/C}}(\bigcup_{y\in Y'} y\cap z))\cdot P_{\mu_C}(pr_{S_C}(z))$$

$$= P_\mu(\bigcup_{x\in X'} x\cap \bigcup_{y\in Y'} y\cap z)\cdot P_\mu(z)$$

And we're done! □

Best of luck on trying to show the reverse direction!