Finite Factored Sets

Scott Garrabrant

Machine Intelligence Research Institute

Some Context

For people who are **not** already familiar with my work:

- Reduce existential risk.
- Figure out how to align advanced AI.
- Become less confused about intelligence/optimization/agency.
- Develop a theory of agents embedded in the environment they are optimizing.
- Do a bunch of weird math/philosophy.

For people who **are** already familiar with my work:

- According to my own personal aesthetics, the subject of this talk is about as exciting as Logical Induction.

Factoring the Talk

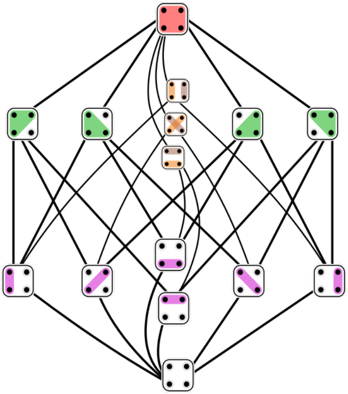
- This talk can be split into 2 parts:
  - Part 1: a short, pure math, combinatorics talk
  - Part 2: a more applied and philosophical main talk
- This talk can also be split into 5 parts, differentiated by color: Title Slides, Motivation, Table of Contents, Main Body, and Examples.
- This gives 10 distinct sections, labeled by the ordered pair on the bottom left.
- Slide numbers are given below:

	Part 1: Short Combinatorics Talk	Part 2: The Main Talk
Title Slides	1	7
Motivation	2	8
Table of Contents	3	9
Main Body	4-5	10, 12-15, 18
Examples	6	11, 16-17

Set Partitions

A **partition** of a set S is a set X of nonempty subsets of S, called **parts**, such that for each $s \in S$ there exists a unique part in X that contains s.

- A partition of S is a way to view S as a disjoint union.
- $\mathrm{Part}(S)$ is the set of all partitions of S.
- X is **trivial** if it has exactly one part.
- $[s]_X$ is the unique part in X containing s.
- $s \sim_X t$ if s and t are in the same part in X.
- $X \geq_S Y$ (X is finer than Y and Y is coarser than X) if for all $s, t \in S$, $s \sim_X t$ implies $s \sim_Y t$.
- $X \vee_S Y$ (The common refinement of X and Y) is the coarsest partition that is finer than both X and Y.



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.

- A factorization of S is a way to view S as a product.
- If $B = \{b_0, \ldots, b_n\}$ is a factorization of S, then there exists a bijection between S and $b_0 \times \cdots \times b_n$ given by $s \mapsto ([s]_{b_0}, \ldots, [s]_{b_n})$. Thus, $|S| = \prod_{b \in B} |b|$.
- A factor must be a partition into parts of equal size.
- $\text{Fact}(S)$ is the set of all factorizations of S.
- A **finite factored set** F is a pair (S, B), where S is a finite set and $B \in \text{Fact}(S)$.

Partition: 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.

Factorization: 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.

Enumerating Factorizations

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

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

0	1	2	3

$\left\{ \begin{array}{c} \{\{0,1\}, \{2,3\}\}, \\ \{\{0,2\}, \{1,3\}\} \end{array} \right\}$

0	1
2	3

$\left\{ \begin{array}{c} \{\{0,1\}, \{2,3\}\}, \\ \{\{0,3\}, \{1,2\}\} \end{array} \right\}$

0	1
3	2

$\left\{ \begin{array}{c} \{\{0,2\}, \{1,3\}\}, \\ \{\{0,3\}, \{1,2\}\} \end{array} \right\}$

0	2
3	1

$\|S\|$	$\|\text{Fact}(S)\|$	$\|S\|$	$\|\text{Fact}(S)\|$
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

This sequence was not on OEIS!

End of Part 1

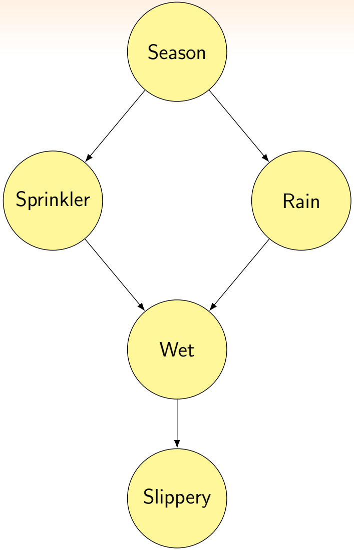
The Main Talk

(It's About Time)

Scott Garrabrant

Machine Intelligence Research Institute

The Pearlian Paradigm



The Pearlian causal inference paradigm is really awesome.

- Given a collection of variables and a joint probability distribution over those variables, Pearl can infer causal (i.e. temporal) relationships between the variables.
- Can infer temporal data (causation) from statistical data (correlation)!

However, I claim that the Pearlian paradigm is cheating.

- "Given a collection of variables" is actually hiding a lot of the work!
- It doesn't infer temporal data from statistical data alone. It uses statistical data **and factorization data**.
- This issue is also related to a failure to adequately handle abstraction and determinism.

We Can Do Better

We will introduce an alternative to the Pearlian paradigm that does not rely on being given factorization data, and works better with abstraction and determinism.
Our approach will be heavily inspired by Pearl, but will not involve graphical models.

Time and Orthogonality

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

History

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 $s \sim_X t$.

Time

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

Orthogonality

We say X and Y are **orthogonal**, written $X \perp^F Y$, if $h^F(X) \cap h^F(Y) = \{\}$.

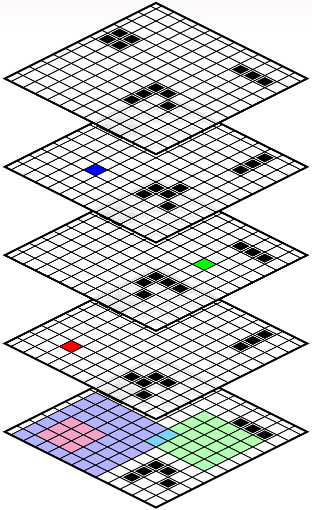
Game of Life

Let S be the set of all game of life computations starting from an $[-n, n] \times [-n, n]$ board. $|S| = 2^{(2n+1)^2}$, the number of initial board states.

- Let $R = \{(r, c, t) \in \mathbb{Z}^3 \mid 0 \leq t \leq n,\ |r| \leq n - t,\ |c| \leq n - t\}$ (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.
- 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 $B = \{L_{(r,c,0)} \mid -n \leq r, c \leq n\}$.

Fix $X = L_{(r_X, c_X, t_X)}$, $Y = L_{(r_Y, c_Y, t_Y)}$, where $(r_X, c_X, t_X), (r_Y, c_Y, t_Y) \in R$.

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



Conditional Orthogonality

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

Conditional History

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 $r \in E$.

Note: Without the second condition, conditional history would not even be well defined.

Conditional Orthogonality

We say X and Y are **orthogonal given** E, written $X \perp^F Y \mid E$, if $h^F(X|E) \cap h^F(Y|E) = \{\}$. We say X and Y are **orthogonal given** Z, written $X \perp^F Y \mid Z$, if $X \perp^F Y \mid z$ for all $z \in Z$.

Compositional Semigraphoid Axioms

Theorem (Compositional Semigraphiod Axioms)

Let $F = (S, B)$ be a finite factored set. Let $X, Y, Z, W \in Part(S)$ be partitions of S.

- *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 $X \perp^F W \mid Z$, then $X \perp^F (Y \vee_S W) \mid Z$. (composition)*

These are a standard set of axioms discussed in the graphical models literature, slightly modified to be in the language of partitions of S, rather than sets of variables.

The Fundamental Theorem

Probability Distribution on a Finite Factored Set

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

Theorem (The Fundamental Theorem of Finite Factored Sets)

Let $F = (S, B)$ be a finite factored set, and let $X, Y, Z \in Part(S)$ 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 $P(x \cap z) \cdot P(y \cap z) = P(x \cap y \cap z) \cdot P(z)$.

The fundamental theorem allows us to infer orthogonality data from probabilistic data. Next, we will show how to infer temporal data from orthogonality data.

Temporal Inference

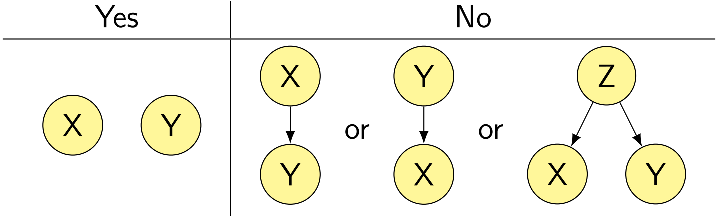
- Ω is a finite set, which is our sample space.
- A **model** of Ω is a pair (F, f), where $F = (S, B)$ is a finite factored set, and $f : S \to \Omega$. (f need not be injective or surjective.)
- If $X \in \text{Parts}(\Omega)$, $f^{-1}(X) \in \text{Parts}(S)$ is given by $s \sim_{f^{-1}(X)} t \Leftrightarrow f(s) \sim_X f(t)$.
- An **orthogonality database** is a pair $D = (O, N)$, where O and N are each sets of triples of partitions of Ω.
- (F, f) satisfies D if:
  - $f^{-1}(X) \perp^F f^{-1}(Y) \mid f^{-1}(Z)$ whenever $(X, Y, Z) \in O$, and
  - $\neg(f^{-1}(X) \perp^F f^{-1}(Y) \mid f^{-1}(Z))$ whenever $(X, Y, Z) \in N$.
- $X <_D Y$ if $f^{-1}(X) <^F f^{-1}(Y)$ for all models (F, f) that satisfy D.

But how does this compare to Pearlian temporal inference?

Two Binary Variables (Pearl)

Let X and Y be two binary variables. Pearl asks:

"Are X and Y independent?"



In either case, no temporal relationship can be concluded.
The Pearlian ontology blinds us from the natural next question:

"Are X and (X XOR Y) independent?"

If yes, the finite factored set paradigm can actually conclude that X is before Y!

Two Binary Variables (Factored Sets)

- Let $\Omega = \{00, 01, 10, 11\}$.
  - Let $X = \{\{00, 01\}, \{10, 11\}\}$. (What is the first bit?)
  - Let $Y = \{\{00, 10\}, \{01, 11\}\}$. (What is the second bit?)
  - Let $Z = \{\{00, 11\}, \{01, 10\}\}$. (Do the bits match?)
- Let $D = (O, N)$, where $O = \{(X, Z, \{\Omega\})\}$ and $N = \{(Z, Z, \{\Omega\})\}$.

Theorem

$X <_D Y$.

Proof. Let (F, f) satisfy D. Let $H_X = h^F(f^{-1}(X))$, $H_Y = h^F(f^{-1}(Y))$, and $H_Z = h^F(f^{-1}(Z))$.
Since $(X, Z, \{\Omega\}) \in O$ and $(Z, Z, \{\Omega\}) \in N$, we have $H_X \cap H_Z = \{\}$ and $H_Z \neq \{\}$.
Since $X \leq_\Omega Y \vee_\Omega Z$, $H_X \subseteq H_Y \cup H_Z$. Since $H_X \cap H_Z = \{\}$, this implies $H_X \subseteq H_Y$.
Similarly, since $Z \leq_\Omega X \vee_\Omega Y$, $H_Z \subseteq H_X \cup H_Y$.
If $H_X = H_Y$, then $\{\} \neq H_Z = (H_X \cup H_Y) \cap H_Z = H_X \cap H_Z = \{\}$, a contradiction.
Thus $H_X \neq H_Y$, so $H_X \subset H_Y$, so $f^{-1}(X) <^F f^{-1}(Y)$, so $X <_D Y$. $\square$

Applications/Future Work/Speculation

Inference:

- Decidablity 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

Infinity:

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

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

The End