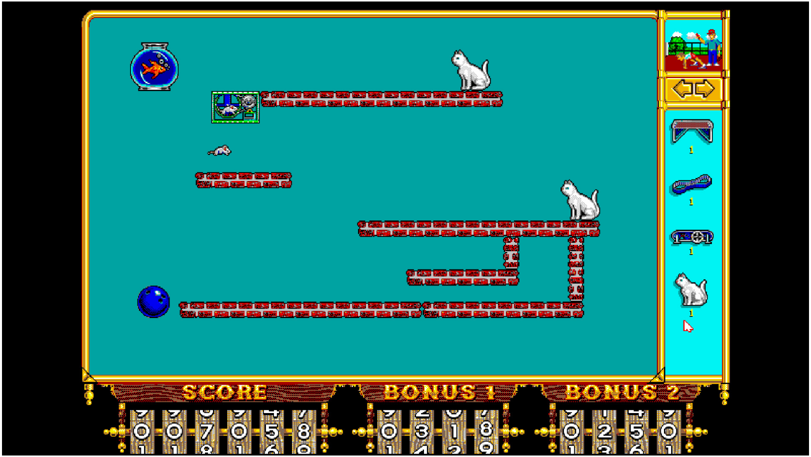
Some thoughts on inferring system structure

based on arXiv:1609.00672

(Elie Wolfe, Robert W. Spekkens, Tobias Fritz)

December 2016



Problem setting

Problem

Given the observed behaviour of some ‘black box’ system, what can be said about the internal structure of the black box?

Problem setting

Problem

Given the observed behaviour of some 'black box' system, what can be said about the internal structure of the black box?

Let's consider the simplest form of the problem: given some hypothesis about the internal structure of the black box, is it compatible with the observed behaviour?

Problem setting

Problem

Given the observed behaviour of some 'black box' system, what can be said about the internal structure of the black box?

Let's consider the simplest form of the problem: given some hypothesis about the internal structure of the black box, is it compatible with the observed behaviour?

This should remind you of:

Problem setting

Problem

Given the observed behaviour of some 'black box' system, what can be said about the internal structure of the black box?

Let's consider the simplest form of the problem: given some hypothesis about the internal structure of the black box, is it compatible with the observed behaviour?

This should remind you of:

- Hacking, reverse engineering.

Problem setting

Problem

Given the observed behaviour of some ‘black box’ system, what can be said about the internal structure of the black box?

Let’s consider the simplest form of the problem: given some hypothesis about the internal structure of the black box, is it compatible with the observed behaviour?

This should remind you of:

- Hacking, reverse engineering.
- Circuit Complexity! (Behaviour desired rather than observed.)

Problem setting

Problem

Given the observed behaviour of some 'black box' system, what can be said about the internal structure of the black box?

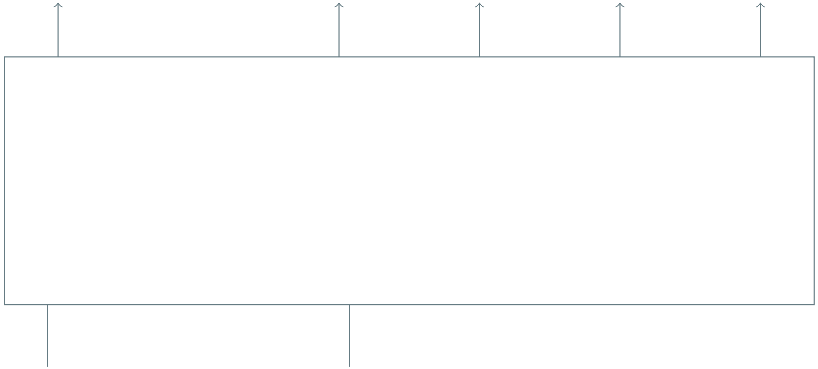
Let's consider the simplest form of the problem: given some hypothesis about the internal structure of the black box, is it compatible with the observed behaviour?

This should remind you of:

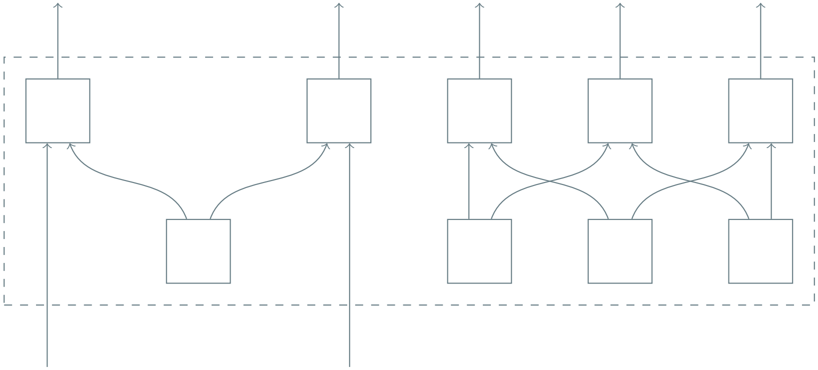
- Hacking, reverse engineering.
- Circuit Complexity! (Behaviour desired rather than observed.)

Formally, a 'system' is a morphism in a suitable monoidal category C. ($\rightarrow$ Baez, Spivak)

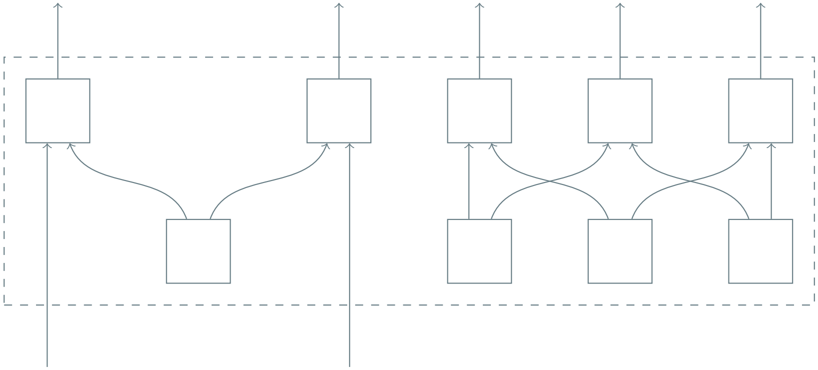
Example: a system with two inputs and five outputs,



Example: a system with two inputs and five outputs, with hypothesis about internal structure:

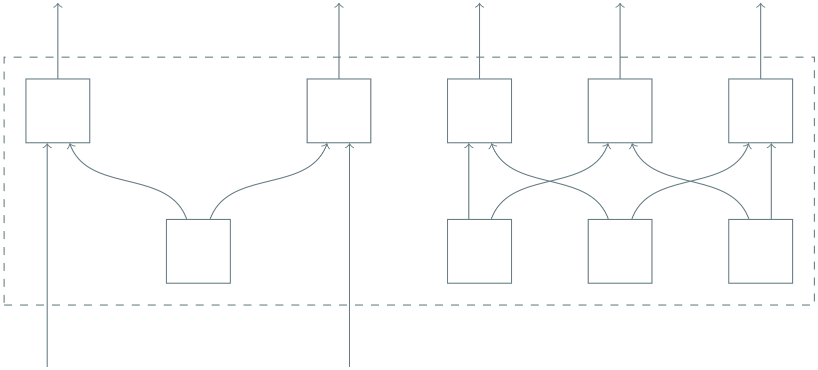


Example: a system with two inputs and five outputs, with hypothesis about internal structure:



Some necessary conditions for the hypothesis to be viable can be read off: the morphism describing the system must factor into two components.

Example: a system with two inputs and five outputs, with hypothesis about internal structure:



Some necessary conditions for the hypothesis to be viable can be read off: the morphism describing the system must factor into two components.

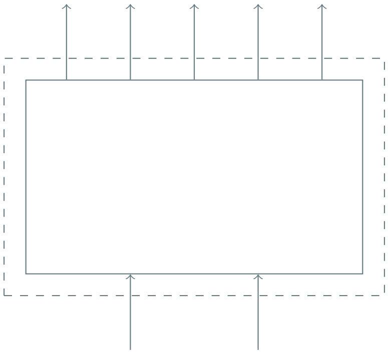
As we will see, this type of condition is far from sufficient in general.

But let's talk about some formal aspects first.

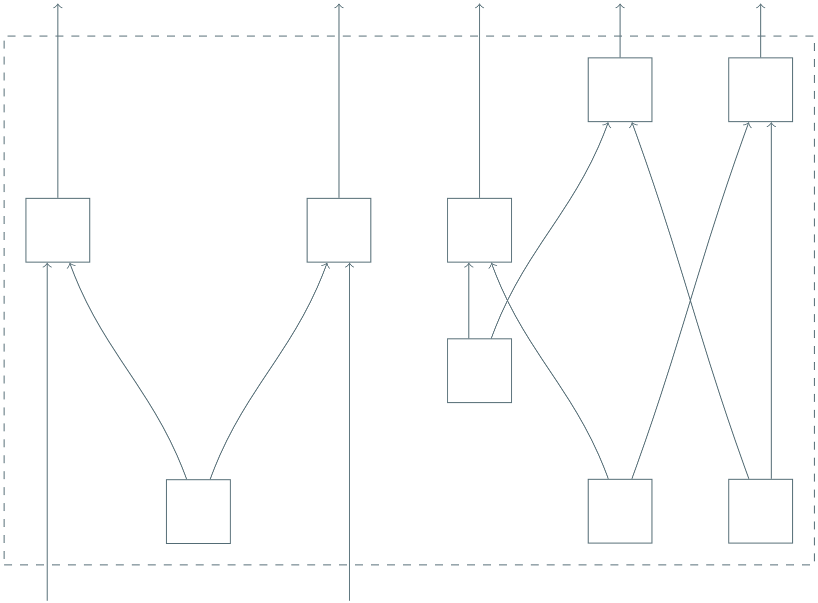
But let's talk about some formal aspects first. What is the structure of the set of feasible structure hypotheses?

But let's talk about some formal aspects first. What is the structure of the set of feasible structure hypotheses?

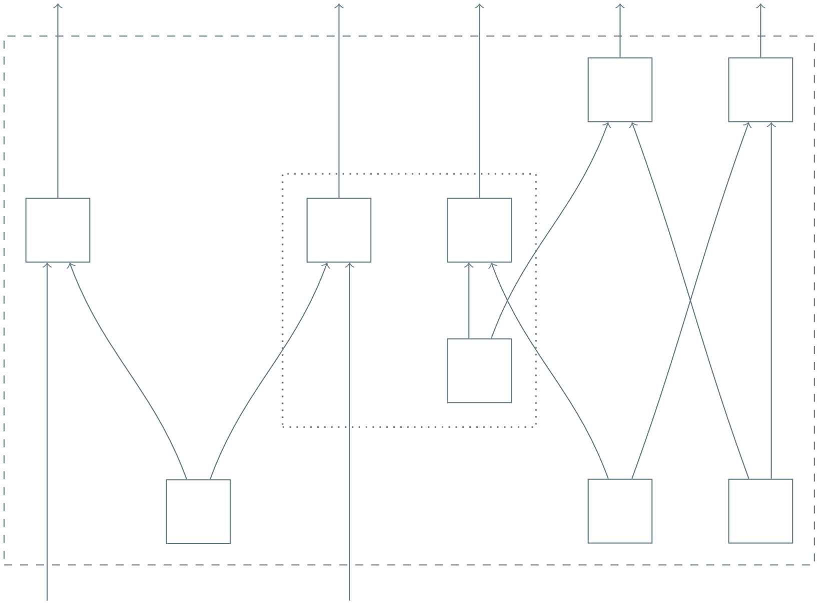
1) There is a trivial hypothesis that always works:



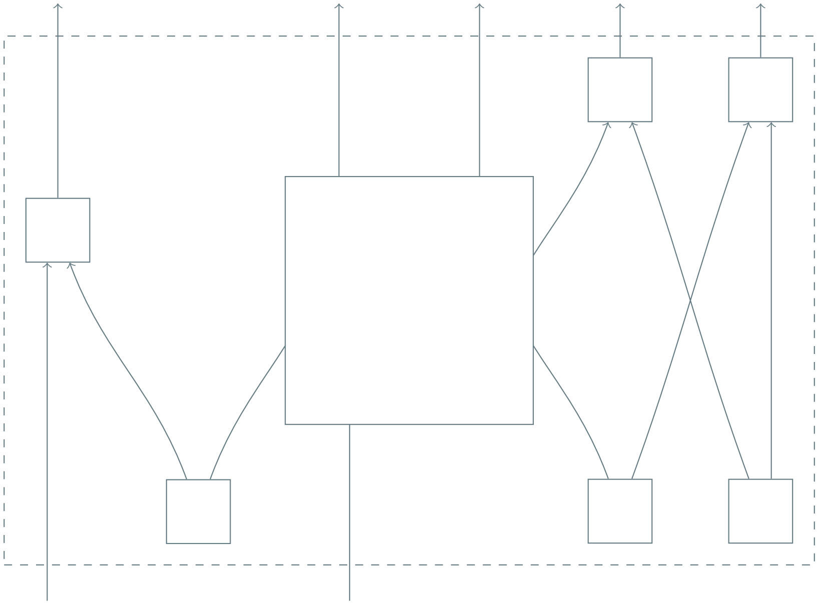
2) If some hypothesis works, then so does every ‘black boxing’ of it.



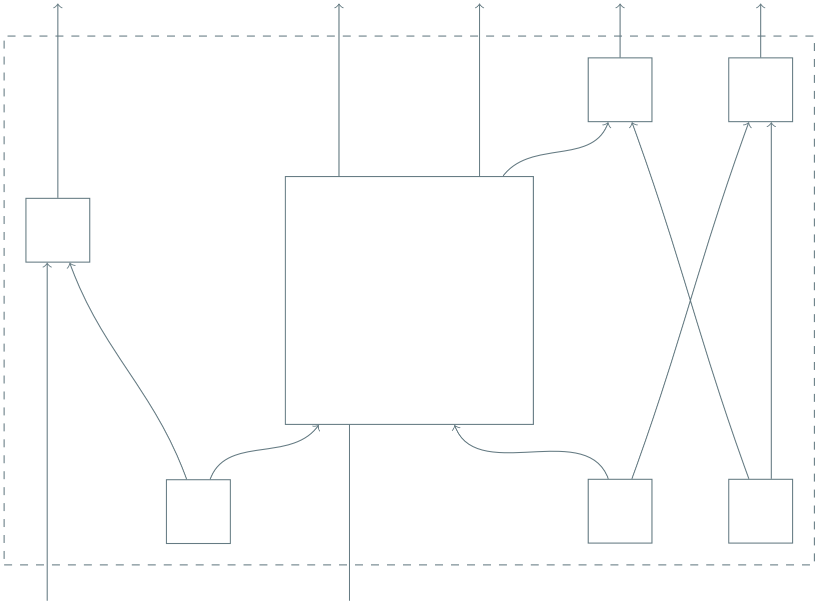
2) If some hypothesis works, then so does every 'black boxing' of it.



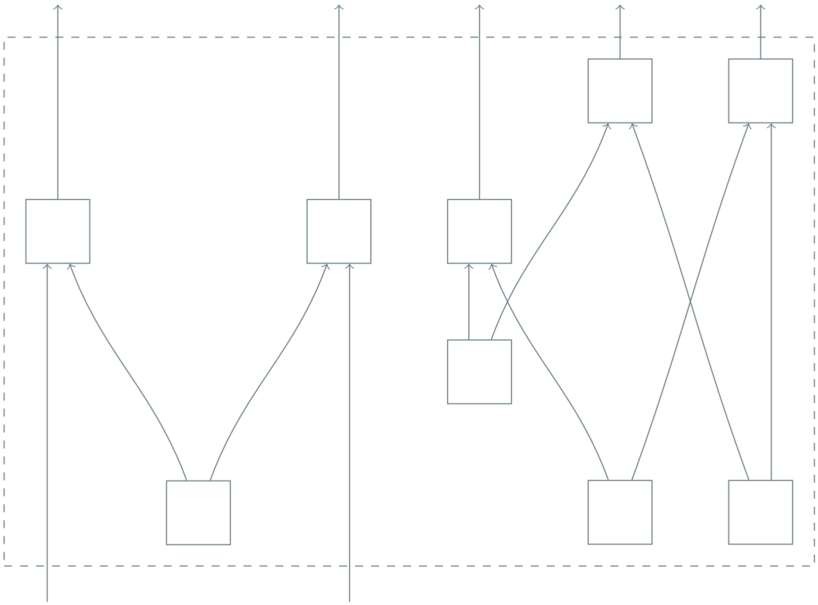
2) If some hypothesis works, then so does every ‘black boxing’ of it.



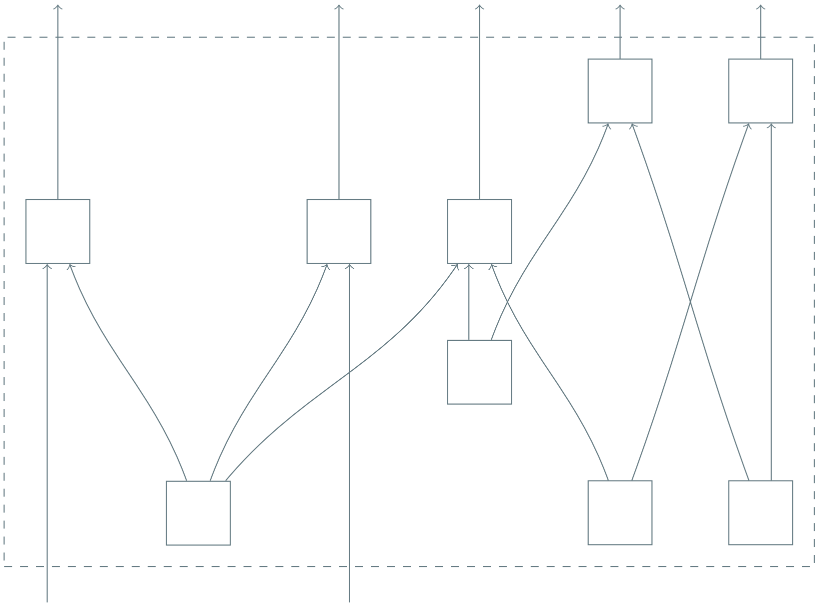
2) If some hypothesis works, then so does every ‘black boxing’ of it.



3) Adding internal wires also preserves feasibility:



3) Adding internal wires also preserves feasibility:



The poset of wiring diagrams

The structure hypotheses are wiring diagrams ($\rightarrow$ Spivak) with the right number of input and output wires.

The poset of wiring diagrams

The structure hypotheses are wiring diagrams ($\rightarrow$ Spivak) with the right number of input and output wires.

Black boxing and adding additional wires defines a partial order on the set of these wiring diagrams.

The poset of wiring diagrams

The structure hypotheses are wiring diagrams ($\rightarrow$ Spivak) with the right number of input and output wires.

Black boxing and adding additional wires defines a partial order on the set of these wiring diagrams. This poset depends on:

The poset of wiring diagrams

The structure hypotheses are wiring diagrams ($\to$ Spivak) with the right number of input and output wires.

Black boxing and adding additional wires defines a partial order on the set of these wiring diagrams. This poset depends on:

- The number of inputs and outputs

The poset of wiring diagrams

The structure hypotheses are wiring diagrams ($\rightarrow$ Spivak) with the right number of input and output wires.

Black boxing and adding additional wires defines a partial order on the set of these wiring diagrams. This poset depends on:

- The number of inputs and outputs
- The particular flavour of monoidal category considered (symmetric, traced, compact, etc.)

The poset of wiring diagrams

The structure hypotheses are wiring diagrams ($\rightarrow$ Spivak) with the right number of input and output wires.

Black boxing and adding additional wires defines a partial order on the set of these wiring diagrams. This poset depends on:

- The number of inputs and outputs
- The particular flavour of monoidal category considered (symmetric, traced, compact, etc.)
- If desired: a set of types for the wires

The poset of wiring diagrams

The structure hypotheses are wiring diagrams ($\rightarrow$ Spivak) with the right number of input and output wires.

Black boxing and adding additional wires defines a partial order on the set of these wiring diagrams. This poset depends on:

- The number of inputs and outputs
- The particular flavour of monoidal category considered (symmetric, traced, compact, etc.)
- If desired: a set of types for the wires

The feasible hypotheses for a given behaviour form a lower set in this poset.

The poset of wiring diagrams

The structure hypotheses are wiring diagrams ($\rightarrow$ Spivak) with the right number of input and output wires.

Black boxing and adding additional wires defines a partial order on the set of these wiring diagrams. This poset depends on:

- The number of inputs and outputs
- The particular flavour of monoidal category considered (symmetric, traced, compact, etc.)
- If desired: a set of types for the wires

The feasible hypotheses for a given behaviour form a lower set in this poset.

Question

To apply Occam's razor:

The poset of wiring diagrams

The structure hypotheses are wiring diagrams ($\to$ Spivak) with the right number of input and output wires.

Black boxing and adding additional wires defines a partial order on the set of these wiring diagrams. This poset depends on:

- The number of inputs and outputs
- The particular flavour of monoidal category considered (symmetric, traced, compact, etc.)
- If desired: a set of types for the wires

The feasible hypotheses for a given behaviour form a lower set in this poset.

Question

To apply Occam's razor: under what conditions does this lower set have a maximal element?

The inflation technique

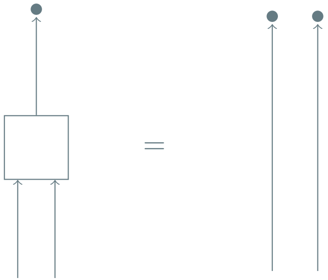
There is a general method for approaching the feasibility problem for those monoidal categories in which the unit object is terminal,

---

1

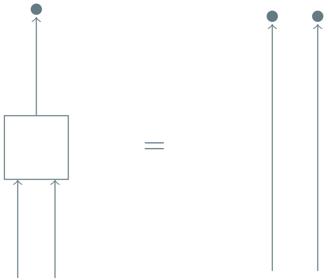
The inflation technique

There is a general method for approaching the feasibility problem for those monoidal categories in which the unit object is terminal,



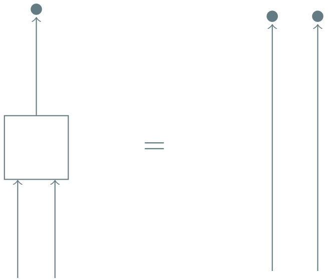
The inflation technique

There is a general method for approaching the feasibility problem for those monoidal categories in which the unit object is terminal,



The inflation technique

There is a general method for approaching the feasibility problem for those monoidal categories in which the unit object is terminal,

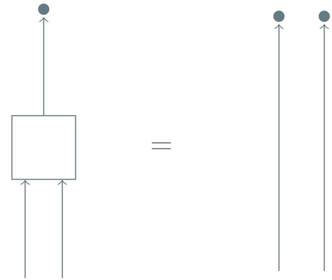


which means that systems can be discarded.

1

The inflation technique

There is a general method for approaching the feasibility problem for those monoidal categories in which the unit object is terminal,



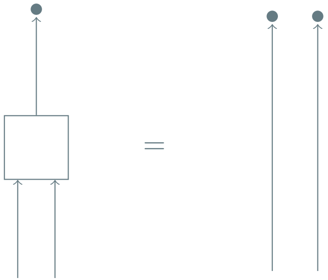
which means that systems can be discarded.

Let's take C to be the category of stochastic matrices. Then string diagrams in C are the same thing as **Bayesian networks** .

1

The inflation technique

There is a general method for approaching the feasibility problem for those monoidal categories in which the unit object is terminal,



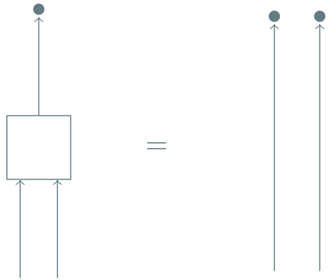
which means that systems can be discarded.

Let's take C to be the category of stochastic matrices. Then string diagrams in C are the same thing as **Bayesian networks**[1].

[1]Where the string diagrams may contain the comonoid structures of below.

The inflation technique

There is a general method for approaching the feasibility problem for those monoidal categories in which the unit object is terminal,



which means that systems can be discarded.

Let's take C to be the category of stochastic matrices. Then string diagrams in C are the same thing as **Bayesian networks**[1]. I will showcase the method with two examples.

[1] Where the string diagrams may contain the comonoid structures of below.

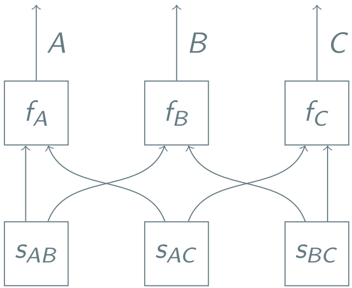
Example: three unbiased binary variables with perfect correlation:

$$P_{ABC} = \frac{[000] + [111]}{2}$$

Example: three unbiased binary variables with perfect correlation:

$$P_{ABC} = \frac{[000] + [111]}{2}$$

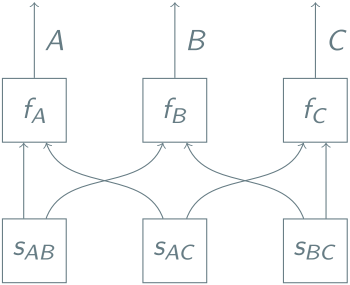
Hypothesis: at most pairwise common causes,



Example: three unbiased binary variables with perfect correlation:

$$P_{ABC} = \frac{[000] + [111]}{2}$$

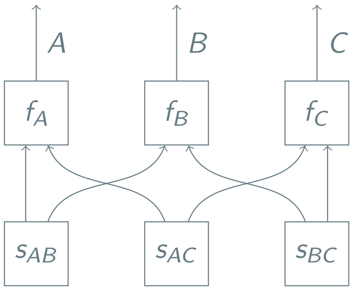
Hypothesis: at most pairwise common causes,



Example: three unbiased binary variables with perfect correlation:

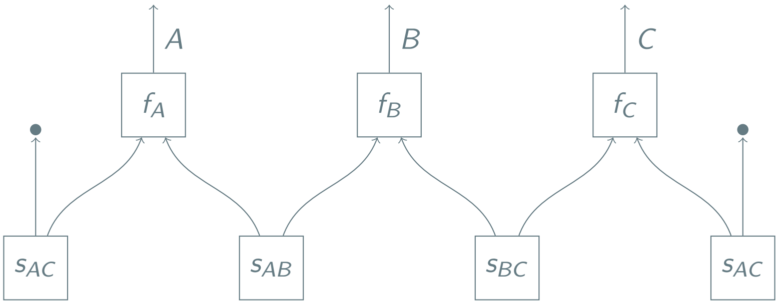
$$P_{ABC} = \frac{[000] + [111]}{2}$$

Hypothesis: at most pairwise common causes,



We will show that this hypothesis is not feasible.

Let's consider a slightly different network, **built out of copies of the same components**:



We call this an **inflated network**.

Crucial observations:

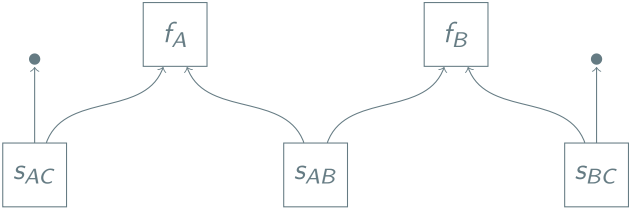
Crucial observations:

- Discarding C in the inflated network results in the same network as discarding C in the original one,

  and similarly for A.

Crucial observations:

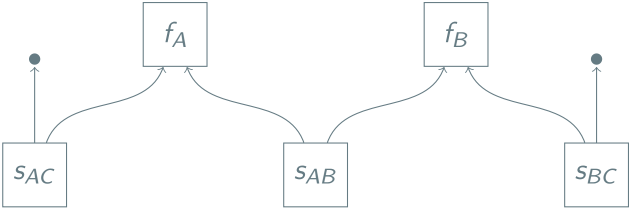
- Discarding C in the inflated network results in the same network as discarding C in the original one,



and similarly for A.

Crucial observations:

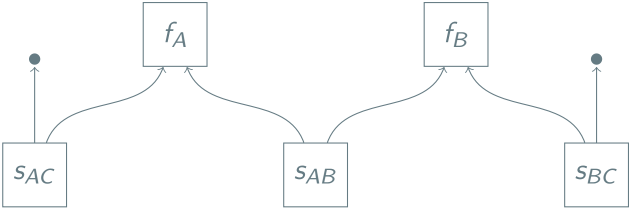
- Discarding C in the inflated network results in the same network as discarding C in the original one,



and similarly for A.

Crucial observations:

- Discarding C in the inflated network results in the same network as discarding C in the original one,



and similarly for A.

- Discarding B disconnects the network.

This results in conditions on the joint distribution P_{ABC} produced by the inflated network:

This results in conditions on the joint distribution P_{ABC} produced by the inflated network:

- A and B are as in the original network: unbiased and perfectly correlated.

This results in conditions on the joint distribution P_{ABC} produced by the inflated network:

- A and B are as in the original network: unbiased and perfectly correlated.
- Likewise for B and C.

This results in conditions on the joint distribution P_{ABC} produced by the inflated network:

- A and B are as in the original network: unbiased and perfectly correlated.
- Likewise for B and C.
- A and C are independent.

This results in conditions on the joint distribution P_{ABC} produced by the inflated network:

- A and B are as in the original network: unbiased and perfectly correlated.
- Likewise for B and C.
- A and C are independent.

There is no joint distribution with these properties at all! The constraints that we have inferred about the inflation network are so strong that they are inconsistent.

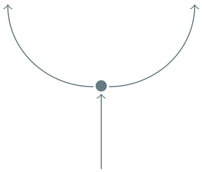
This results in conditions on the joint distribution P_{ABC} produced by the inflated network:

- A and B are as in the original network: unbiased and perfectly correlated.
- Likewise for B and C.
- A and C are independent.

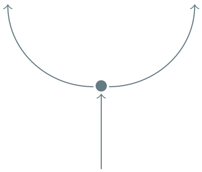
There is no joint distribution with these properties at all! The constraints that we have inferred about the inflation network are so strong that they are inconsistent.

$\Rightarrow$ The network structure hypothesis is not feasible.

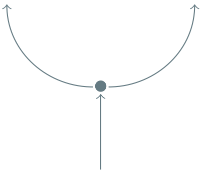
In some categories, such as stochastic matrices, it is also possible to make copies of systems,



In some categories, such as stochastic matrices, it is also possible to make copies of systems,

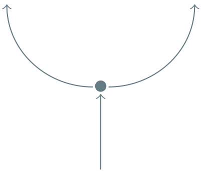


In some categories, such as stochastic matrices, it is also possible to make copies of systems,



equipping every object with a comonoid structure (not necessarily natural).

In some categories, such as stochastic matrices, it is also possible to make copies of systems,



equipping every object with a comonoid structure (not necessarily natural).

This can be leveraged to build inflation networks which witness more infeasbilities. Let's see an example!

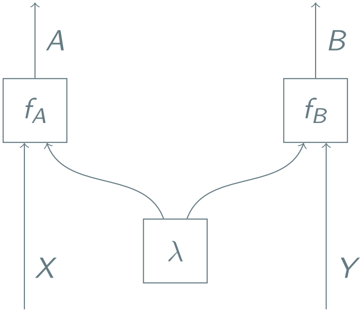
Again in stochastic matrices, consider a two-input and two-output morphism with binary variables:

$$P_{AB|XY}(ab|xy) = \begin{cases} \frac{1}{2} & \text{if } a \oplus b = xy, \\ 0 & \text{otherwise.} \end{cases}$$

Again in stochastic matrices, consider a two-input and two-output morphism with binary variables:

$$P_{AB|XY}(ab|xy) = \begin{cases} \frac{1}{2} & \text{if } a \oplus b = xy, \\ 0 & \text{otherwise.} \end{cases}$$

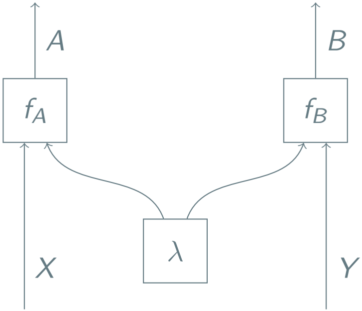
Structure hypothesis:



Again in stochastic matrices, consider a two-input and two-output morphism with binary variables:

$$P_{AB|XY}(ab|xy) = \begin{cases} \frac{1}{2} & \text{if } a \oplus b = xy, \\ 0 & \text{otherwise.} \end{cases}$$

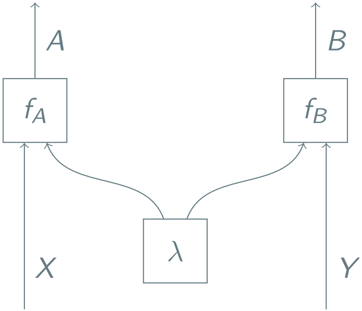
Structure hypothesis:



Again in stochastic matrices, consider a two-input and two-output morphism with binary variables:

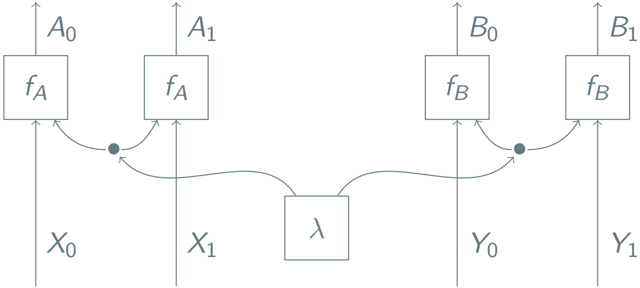
$$P_{AB|XY}(ab|xy) = \begin{cases} \frac{1}{2} & \text{if } a \oplus b = xy, \\ 0 & \text{otherwise.} \end{cases}$$

Structure hypothesis:

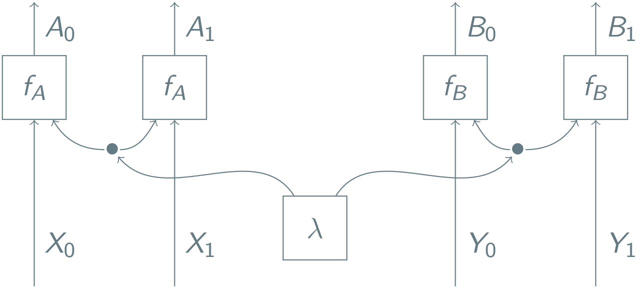


This looks promising: discarding A shows that B is only a function of Y, which is consistent with $P_{AB|XY}$.

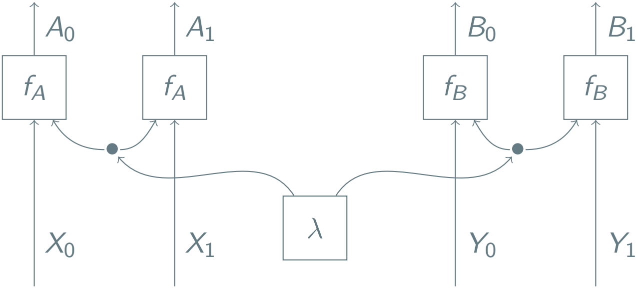
Inflation network:



Inflation network:

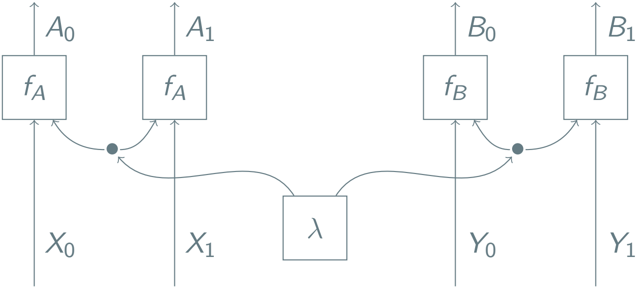


Inflation network:



Using inputs $X_0 = Y_0 = 0$ and $X_1 = Y_1 = 1$ hypothetically results in a distribution $P_{A_0 A_1 B_0 B_1}$ where:

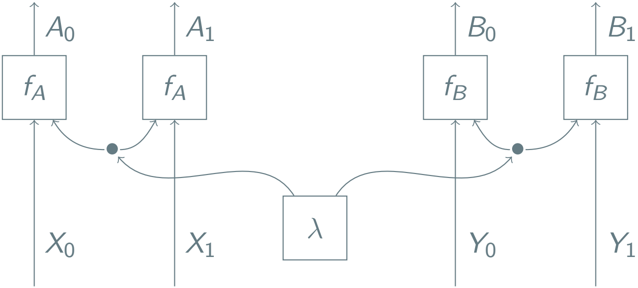
Inflation network:



Using inputs $X_0 = Y_0 = 0$ and $X_1 = Y_1 = 1$ hypothetically results in a distribution $P_{A_0 A_1 B_0 B_1}$ where:

- A_1 and B_0 are perfectly correlated,

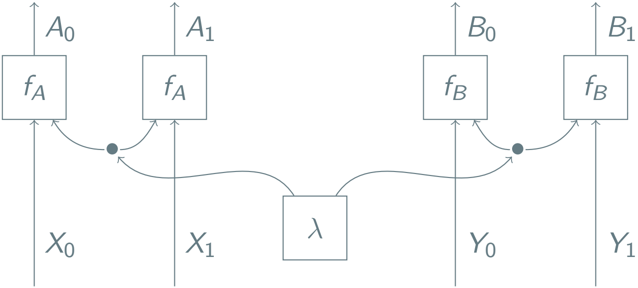
Inflation network:



Using inputs $X_0 = Y_0 = 0$ and $X_1 = Y_1 = 1$ hypothetically results in a distribution $P_{A_0 A_1 B_0 B_1}$ where:

- A_1 and B_0 are perfectly correlated,
- B_0 and A_0 are perfectly correlated,

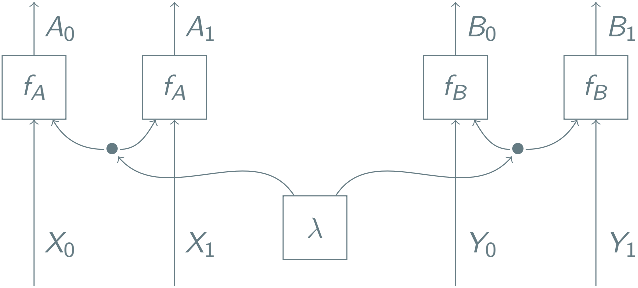
Inflation network:



Using inputs $X_0 = Y_0 = 0$ and $X_1 = Y_1 = 1$ hypothetically results in a distribution $P_{A_0 A_1 B_0 B_1}$ where:

- A_1 and B_0 are perfectly correlated,
- B_0 and A_0 are perfectly correlated,
- A_0 and B_1 are perfectly correlated,

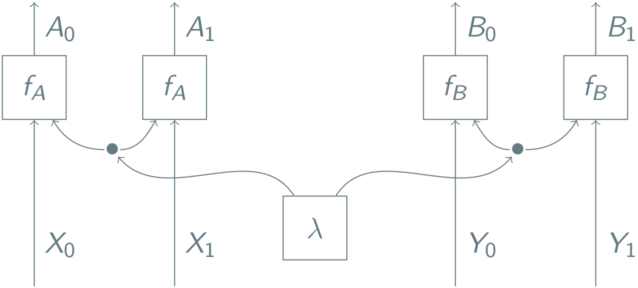
Inflation network:



Using inputs $X_0 = Y_0 = 0$ and $X_1 = Y_1 = 1$ hypothetically results in a distribution $P_{A_0 A_1 B_0 B_1}$ where:

- A_1 and B_0 are perfectly correlated,
- B_0 and A_0 are perfectly correlated,
- A_0 and B_1 are perfectly correlated,
- B_1 and A_1 are perfectly **anti**correlated.

Inflation network:



Using inputs $X_0 = Y_0 = 0$ and $X_1 = Y_1 = 1$ hypothetically results in a distribution $P_{A_0 A_1 B_0 B_1}$ where:

- A_1 and B_0 are perfectly correlated,
- B_0 and A_0 are perfectly correlated,
- A_0 and B_1 are perfectly correlated,
- B_1 and A_1 are perfectly **anti**correlated.

There is no distribution with these properties! $\Rightarrow$ Infeasible hypothesis.