

Counterfactual Causality for Reachability and Safety based on Distance Functions

Julie Parreaux¹

Jakob Piribauer^{2,3} Christel Baier²

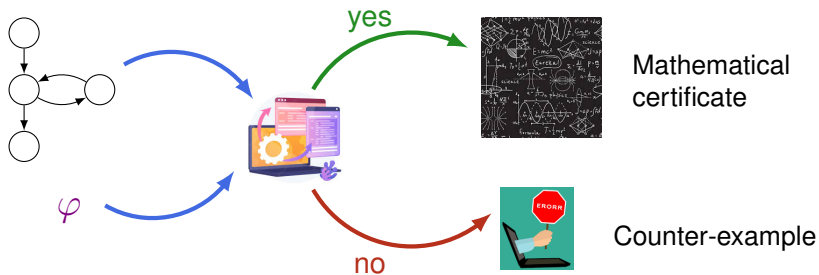
¹Aix-Marseille Université, France

²Technische Universität Dresden, Germany

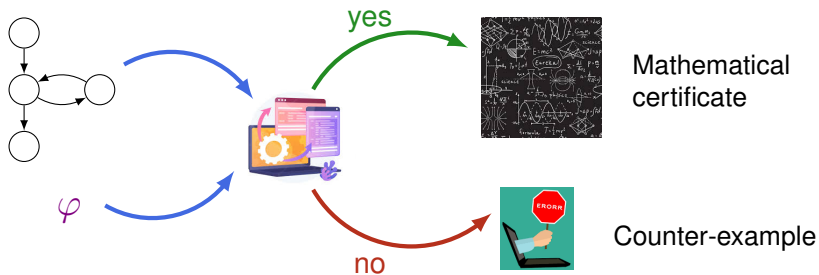
³Technische Universität München, Germany

GandALF 2023

From verification to causality

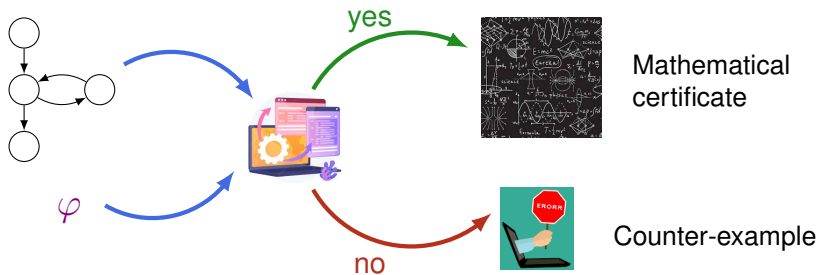


From verification to causality



Causality: explain why the property holds or not

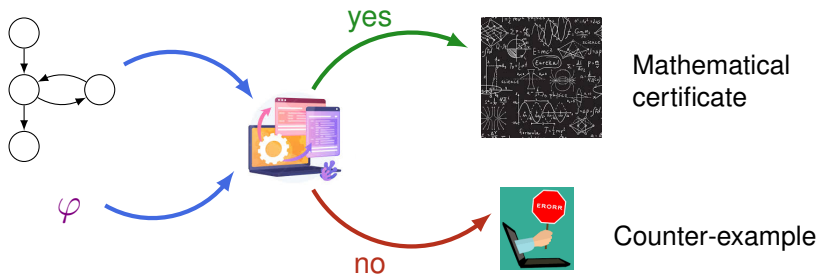
From verification to causality



Causality: explain why the property holds or not

- ▶ what causes the specification to hold for the full model?

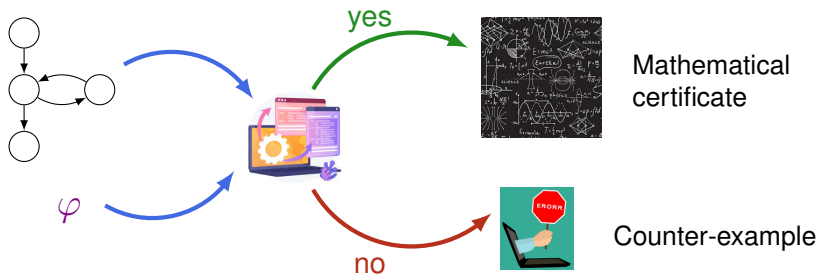
From verification to causality



Causality: explain why the property holds or not

- ▶ what causes the specification to hold for the full model?
- ▶ who is responsible for a requirement violation? and to which degree?

From verification to causality



Causality: explain why the property holds or not

- ▶ what causes the specification to hold for the full model?
- ▶ who is responsible for a requirement violation? and to which degree?
- ▶ if a bad behavior occurs, what has caused the violation of the specification?

Forward vs backward causality



Forward vs backward causality



Forward vs backward causality



Why?



Forward vs backward causality



Why?



Forward causality

Describes causes before the execution:

Forward vs backward causality



Why?



Forward causality

Describes causes before the execution:
what can cause an event in a given model?

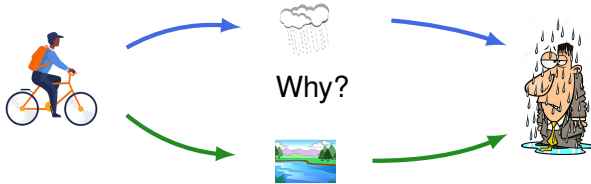
Forward vs backward causality



Forward causality

Describes causes before the execution:
what can cause an event in a given model?

Forward vs backward causality



Forward causality

Describes causes before the execution:
what can cause an event in a given model?

Forward vs backward causality

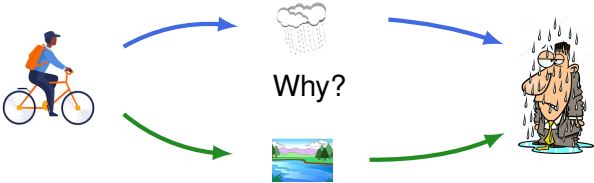


Forward causality

Describes causes before the execution:
what can cause an event in a given model?

Necessary causes
Cause implies Effect

Forward vs backward causality



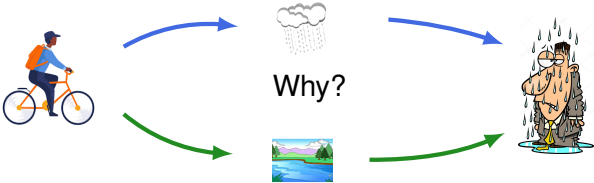
Forward causality

Describes causes before the execution:
what can cause an event in a given model?

Necessary causes
Cause implies Effect



Forward vs backward causality



Forward causality

Describes causes before the execution:
what can cause an event in a given model?

Necessary causes
Cause implies Effect



Backward causality

Describes causes after the execution:

Forward vs backward causality



Forward causality

Describes causes before the execution:
what can cause an event in a given model?

Necessary causes

Cause implies Effect



Backward causality

Describes causes after the execution:

Forward vs backward causality



Forward causality

Describes causes before the execution:
what can cause an event in a given model?

Necessary causes

Cause implies Effect



Backward causality

Describes causes after the execution:
what has caused an observed effect in a
given execution?

Forward vs backward causality



Forward causality

Describes causes before the execution:
what can cause an event in a given model?

Necessary causes

Cause implies Effect



Backward causality

Describes causes after the execution:
what has caused an observed effect in a
given execution?

Counterfactual causes

Fixed an execution:

\neg Cause implies \neg Effect

Forward vs backward causality



Forward causality

Describes causes before the execution:
what can cause an event in a given model?

Necessary causes

Cause implies Effect



Backward causality

Describes causes after the execution:
what has caused an observed effect in a
given execution?

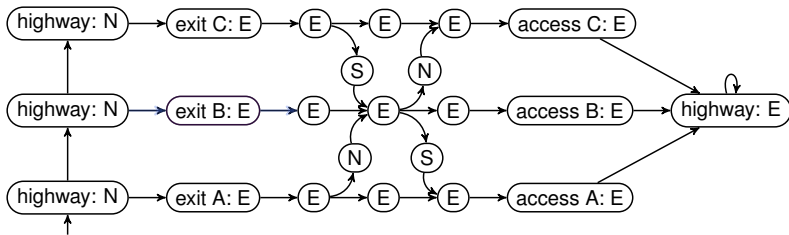
Counterfactual causes

Fixed an execution:

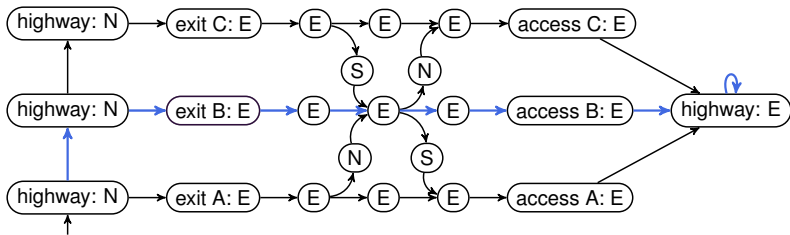
\neg Cause implies \neg Effect



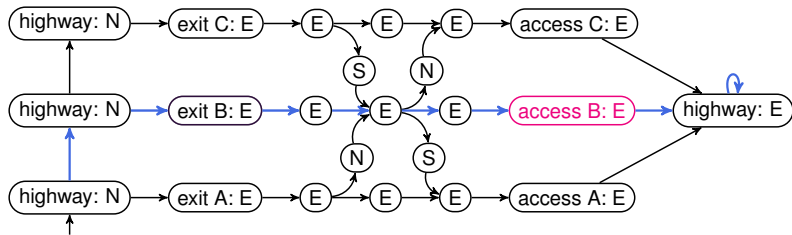
Counterfactual causality based on distance functions



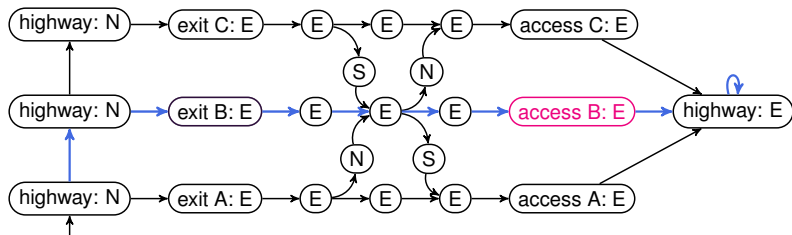
Counterfactual causality based on distance functions



Counterfactual causality based on distance functions



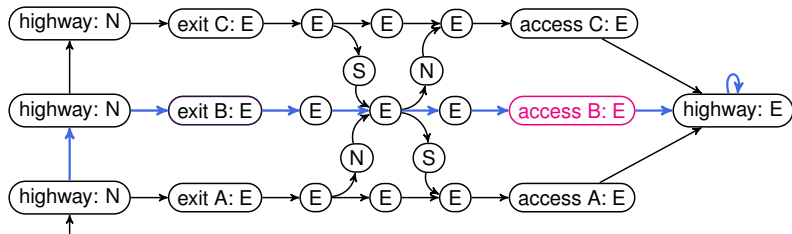
Counterfactual causality based on distance functions



Counterfactual cause

$\pi = NNE^\omega$ Effect = {access B}

Counterfactual causality based on distance functions



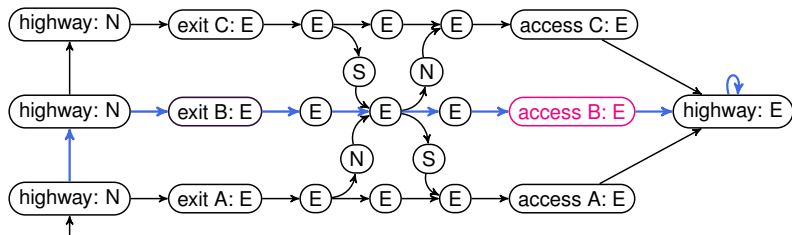
Stalnaker-Lewis-semantics

Counterfactual cause

$\pi = NNE^\omega$

Effect = {access B}

Counterfactual causality based on distance functions



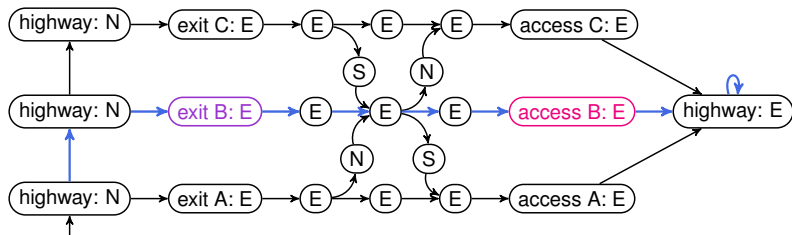
Stalnaker-Lewis-semantics

Counterfactual cause for the closest set of executions according to a similarity metric

Counterfactual cause

$\pi = NNE^\omega$ *Effect* = {access B}

Counterfactual causality based on distance functions



Stalnaker-Lewis-semantics

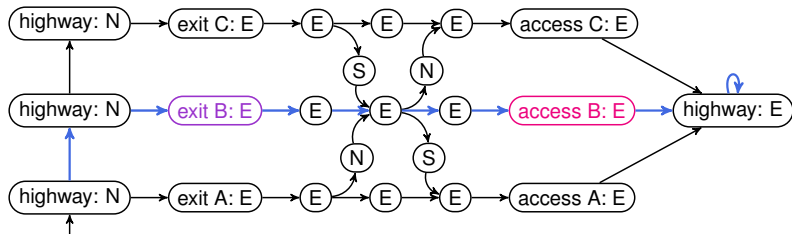
Counterfactual cause for the closest set of executions according to a similarity metric

Counterfactual cause

$\pi = NNE^\omega$ *Effect* = {access B}

Cause $\stackrel{?}{=} \{\text{exit B}\}$

Counterfactual causality based on distance functions



Stalnaker-Lewis-semantics

Counterfactual cause for the closest set of executions according to a similarity metric

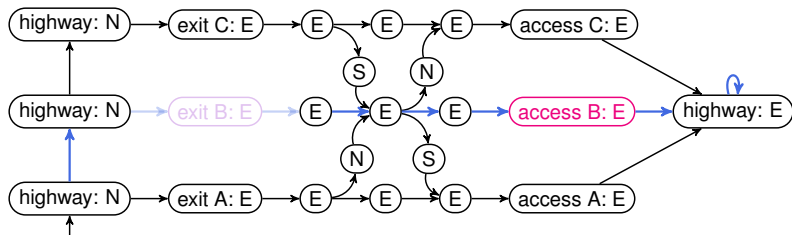
Counterfactual cause

$\pi = NNE^\omega$ *Effect* = {access B}

Cause $\stackrel{?}{=} \{\text{exit B}\}$

$$\zeta \in \{\zeta' \mid d(\pi, \zeta') = d_{\min} \text{ and } \zeta' \models \Box \neg \text{Cause}\}$$

Counterfactual causality based on distance functions



Stalnaker-Lewis-semantics

Counterfactual cause for the closest set of executions according to a similarity metric

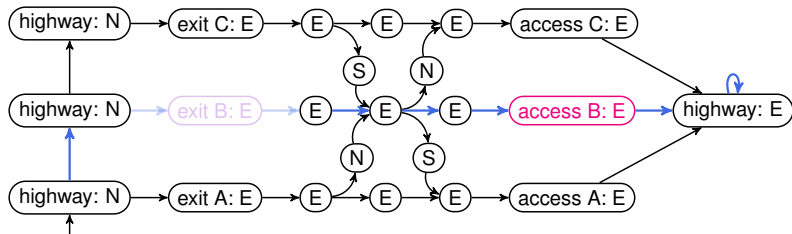
Counterfactual cause

$\pi = NNE^\omega$ *Effect* = {access B}

Cause $\stackrel{?}{=} \{\text{exit B}\}$

$$\zeta \in \{\zeta' \mid d(\pi, \zeta') = d_{\min} \text{ and } \zeta' \models \Box \neg \text{Cause}\}$$

Counterfactual causality based on distance functions



Stalnaker-Lewis-semantics

Counterfactual cause for the closest set of executions according to a similarity metric

Counterfactual cause

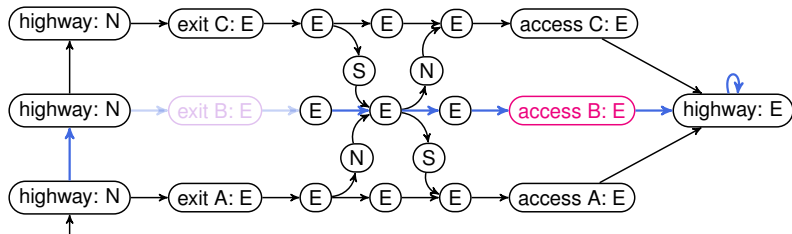
$\pi = NNE^\omega$ *Effect* = {access B}

Cause $\stackrel{?}{=} \{ \text{exit B} \}$

$\zeta \in \{ NE^\omega \}$

$$\zeta \in \{ \zeta' \mid d(\pi, \zeta') = d_{\min} \text{ and } \zeta' \models \Box \neg \textit{Cause} \}$$

Counterfactual causality based on distance functions



Stalnaker-Lewis-semantics

Counterfactual cause for the closest set of executions according to a similarity metric

Counterfactual cause

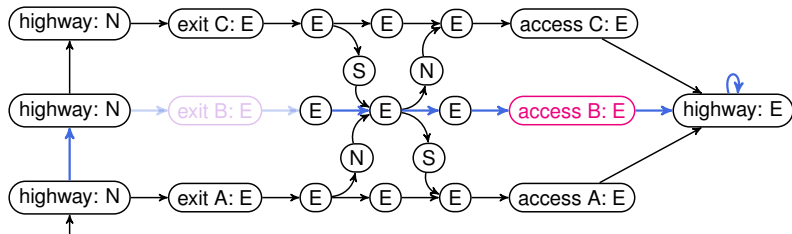
$\pi = NNE^\omega$ *Effect* = {access B}

Cause $\stackrel{?}{=} \{\text{exit B}\}$

$\zeta \in \{NE^\omega, NNE^\omega\}$

$$\zeta \in \{\zeta' \mid d(\pi, \zeta') = d_{\min} \text{ and } \zeta' \models \Box \neg \text{Cause}\}$$

Counterfactual causality based on distance functions



Stalnaker-Lewis-semantics

Counterfactual cause for the closest set of executions according to a similarity metric

Counterfactual cause

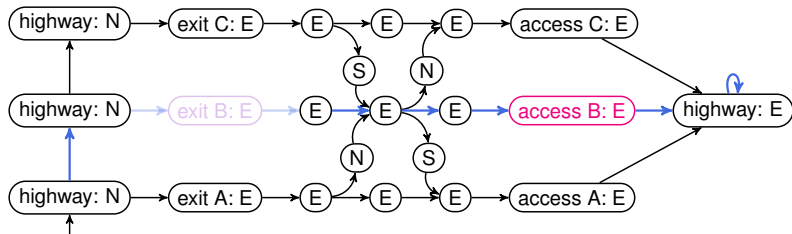
$\pi = NNE^\omega$ *Effect* = {access B}

Cause $\stackrel{?}{=} \{\text{exit B}\}$

$\zeta \in \{NE^\omega, NNNE^\omega\}$

Do all $\zeta \in \{\zeta' \mid d(\pi, \zeta') = d_{\min} \text{ and } \zeta' \models \Box \neg \text{Cause}\}$ satisfy $\Box \neg \text{Effect}$?

Counterfactual causality based on distance functions



Stalnaker-Lewis-semantics

Counterfactual cause for the closest set of executions according to a similarity metric

Counterfactual cause

$\pi = NNE^\omega$ *Effect* = {access B}

Cause = {exit B}

$\zeta \in \{NE^\omega, NNNE^\omega\}$

Do all $\zeta \in \{\zeta' \mid d(\pi, \zeta') = d_{\min} \text{ and } \zeta' \models \Box \neg \textit{Cause}\}$ satisfy $\Box \neg \textit{Effect}$?

Contributions on transition systems

Checking counterfactual cause problem in transition systems

Contributions on transition systems

Checking counterfactual cause problem in transition systems

Given a distance over executions, check if **Cause** is a cause for **Effect**

Contributions on transition systems

Checking counterfactual cause problem in transition systems

Given a distance over executions, check if **Cause** is a cause for **Effect**

distance	causality

Contributions on transition systems

Checking counterfactual cause problem in transition systems

Given a distance over executions, check if **Cause** is a cause for **Effect**

size of the longest common prefix



distance	causality
prefix	

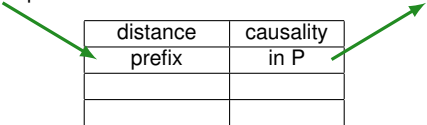
Contributions on transition systems

Checking counterfactual cause problem in transition systems

Given a distance over executions, check if **Cause** is a cause for **Effect**

size of the longest common prefix

CTL satisfiability



distance prefix	causality in P

Contributions on transition systems

Checking counterfactual cause problem in transition systems

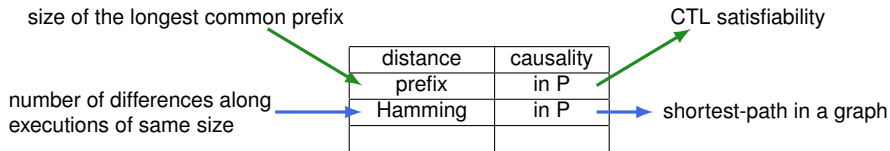
Given a distance over executions, check if **Cause** is a cause for **Effect**



Contributions on transition systems

Checking counterfactual cause problem in transition systems

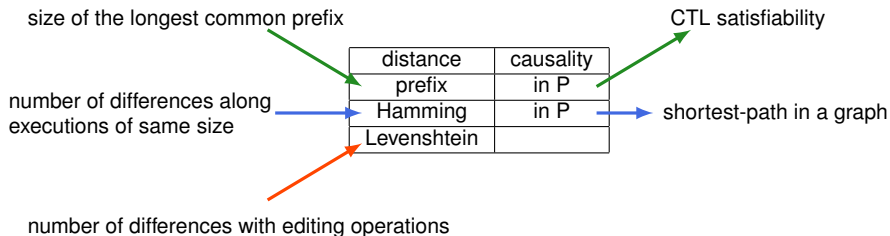
Given a distance over executions, check if **Cause** is a cause for **Effect**



Contributions on transition systems

Checking counterfactual cause problem in transition systems

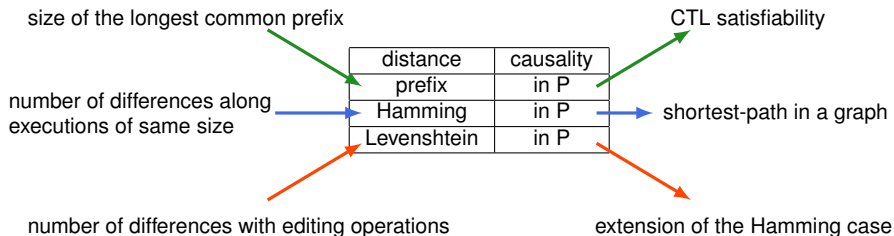
Given a distance over executions, check if **Cause** is a cause for **Effect**



Contributions on transition systems

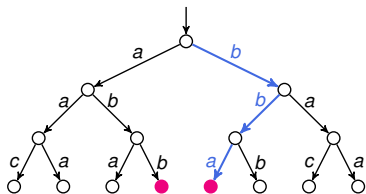
Checking counterfactual cause problem in transition systems

Given a distance over executions, check if **Cause** is a cause for **Effect**



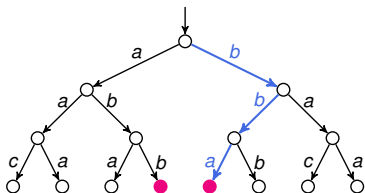
The case of Hamming distance

Reduction to a shortest-path problem



The case of Hamming distance

Reduction to a shortest-path problem

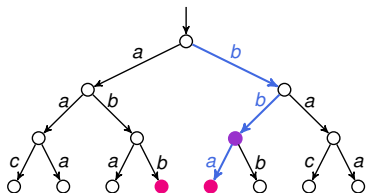


Hypothesis

Transition system where all executions have the same size.

The case of Hamming distance

Reduction to a shortest-path problem

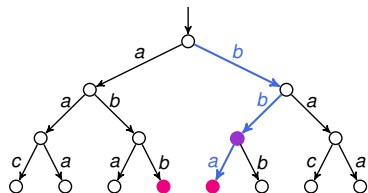


Hypothesis

Transition system where all executions have the same size.

The case of Hamming distance

Reduction to a shortest-path problem



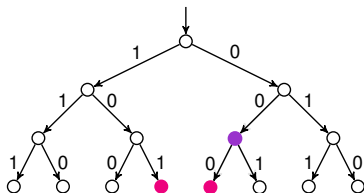
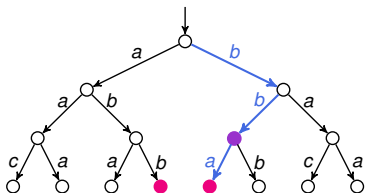
Hypothesis

Transition system where all executions have the same size.

Algorithm to check a potential cause

The case of Hamming distance

Reduction to a shortest-path problem



Hypothesis

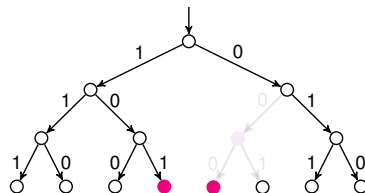
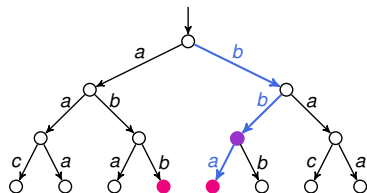
Transition system where all executions have the same size.

Algorithm to check a potential cause

- ▶ Defining the weighted graph such that $w(u, v) = 0$ iff label of (u, v) is the same than in the execution

The case of Hamming distance

Reduction to a shortest-path problem



Hypothesis

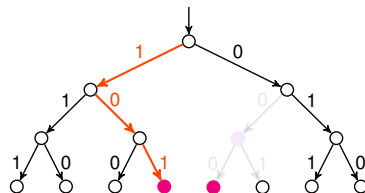
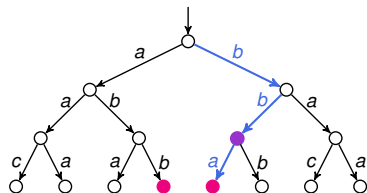
Transition system where all executions have the same size.

Algorithm to check a potential cause

- ▶ Defining the weighted graph such that $w(u, v) = 0$ iff label of (u, v) is the same than in the execution
- ▶ Removing the potential cause **Cause**

The case of Hamming distance

Reduction to a shortest-path problem



Hypothesis

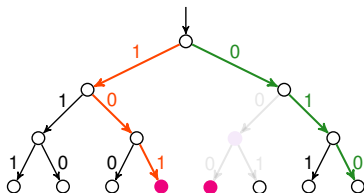
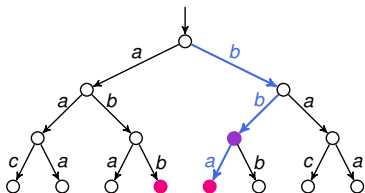
Transition system where all executions have the same size.

Algorithm to check a potential cause

- ▶ Defining the weighted graph such that $w(u, v) = 0$ iff label of (u, v) is the same than in the execution
- ▶ Removing the potential cause Cause
- ▶ Computing the shortest path to reach Effect: ζ

The case of Hamming distance

Reduction to a shortest-path problem



Hypothesis

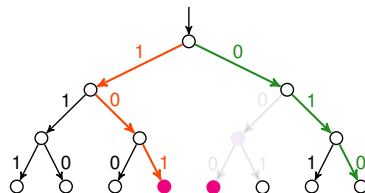
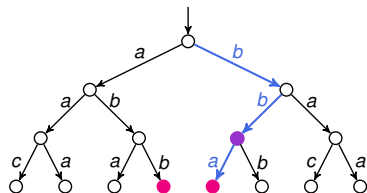
Transition system where all executions have the same size.

Algorithm to check a potential cause

- ▶ Defining the weighted graph such that $w(u, v) = 0$ iff label of (u, v) is the same than in the execution
- ▶ Removing the potential cause Cause
- ▶ Computing the shortest path to reach Effect: ζ
- ▶ Computing the shortest path to reach \neg Effect: ζ'

The case of Hamming distance

Reduction to a shortest-path problem



Hypothesis

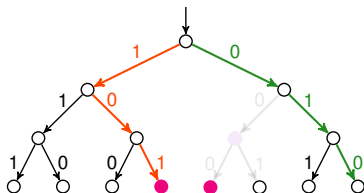
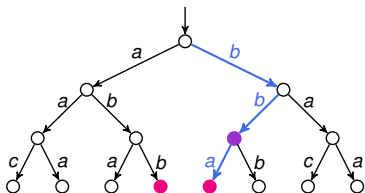
Transition system where all executions have the same size.

Algorithm to check a potential cause

- ▶ Defining the weighted graph such that $w(u, v) = 0$ iff label of (u, v) is the same than in the execution
- ▶ Removing the potential cause Cause
- ▶ Computing the shortest path to reach Effect: ζ
- ▶ Computing the shortest path to reach \neg Effect: ζ'
- ▶ Test $\text{weight}(\zeta') < \text{weight}(\zeta)$

The case of Hamming distance

Reduction to a shortest-path problem



Hypothesis

Transition system where all executions have the same size.

Extension

Same algorithm with a generalisation of Hamming distance

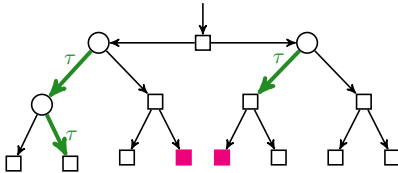
Algorithm to check a potential cause

- ▶ Defining the weighted graph such that $w(u, v) = 0$ iff label of (u, v) is the same than in the execution
- ▶ Removing the potential cause Cause
- ▶ Computing the shortest path to reach Effect: ζ
- ▶ Computing the shortest path to reach \neg Effect: ζ'
- ▶ Test $\text{weight}(\zeta') < \text{weight}(\zeta)$

Extension into reachability/safety games

Counterfactual causality in games

Winning player follows a non-winning strategy τ

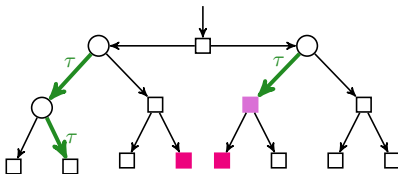


Extension into reachability/safety games

Counterfactual causality in games

Winning player follows a non-winning strategy τ

Cause= set of vertices that a winning strategy needs to avoid



Extension into reachability/safety games

Checking counterfactual cause problem in games

Given a distance over strategies, check if **Cause** is a cause for **Effect**

Extension into reachability/safety games

Checking counterfactual cause problem in games

Given a distance over strategies, check if **Cause** is a cause for **Effect**

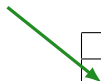
distance	causality

Extension into reachability/safety games

Checking counterfactual cause problem in games

Given a distance over strategies, check if **Cause** is a cause for **Effect**

used to extend distance as
functions applied on sets



distance	causality
Hausdorff lifting of the prefix distance	

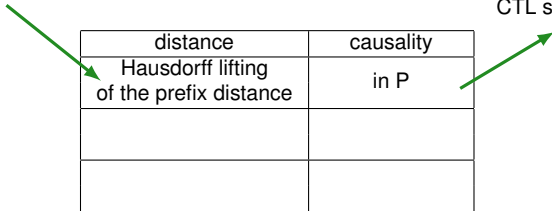
Extension into reachability/safety games

Checking counterfactual cause problem in games

Given a distance over strategies, check if **Cause** is a cause for **Effect**

used to extend distance as
functions applied on sets

attractor theory +
CTL satisfiability



distance	causality
Hausdorff lifting of the prefix distance	in P

Extension into reachability/safety games

Checking counterfactual cause problem in games

Given a distance over strategies, check if **Cause** is a cause for **Effect**

used to extend distance as
functions applied on sets

number of vertices
where strategies make
different choices

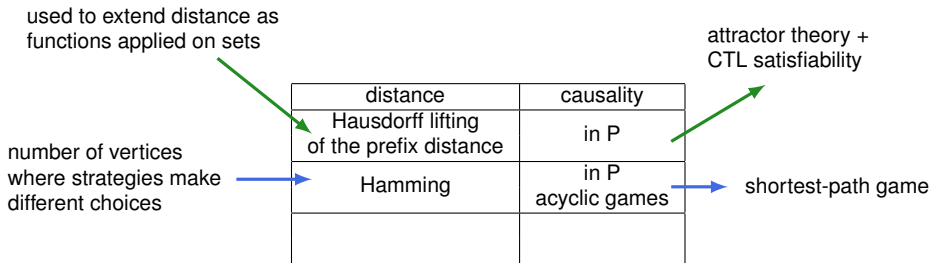
distance	causality
Hausdorff lifting of the prefix distance	in P
Hamming	

attractor theory +
CTL satisfiability

Extension into reachability/safety games

Checking counterfactual cause problem in games

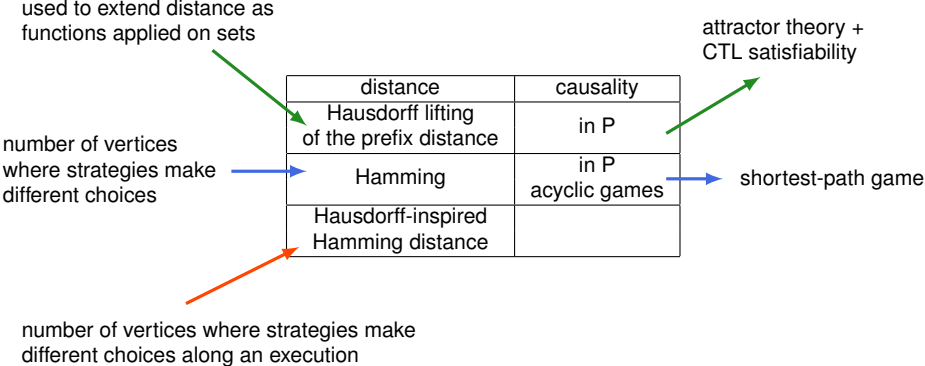
Given a distance over strategies, check if **Cause** is a cause for **Effect**



Extension into reachability/safety games

Checking counterfactual cause problem in games

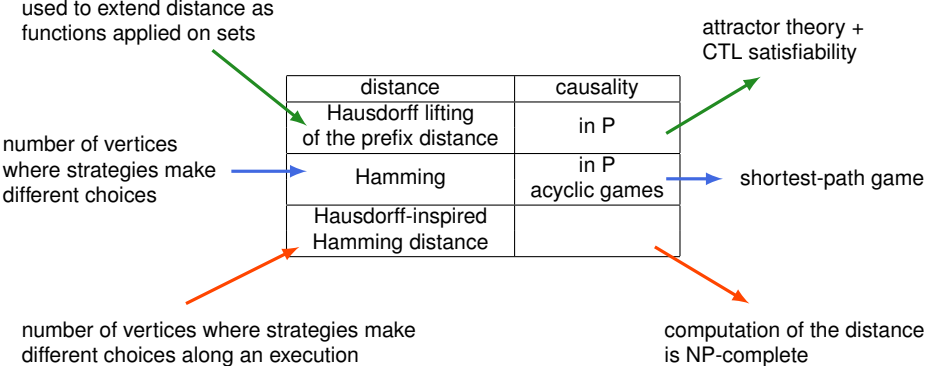
Given a distance over strategies, check if Cause is a cause for Effect



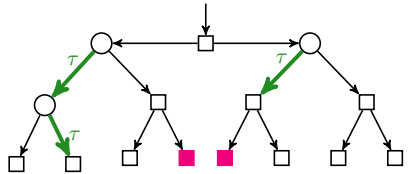
Extension into reachability/safety games

Checking counterfactual cause problem in games

Given a distance over strategies, check if **Cause** is a cause for **Effect**



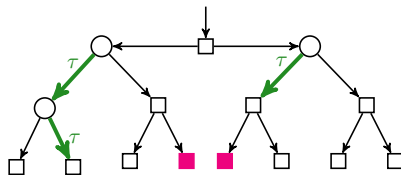
Counterfactual explanation



Counterfactual explanation

Explanation E

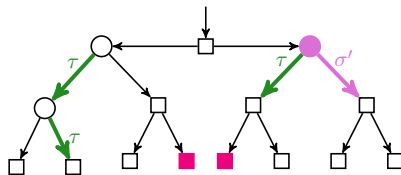
Given a non-winning strategy τ ,
check if there exists a winning strategy σ
such that $\tau(v) \neq \sigma(v)$ iff $v \in E$



Counterfactual explanation

Explanation E

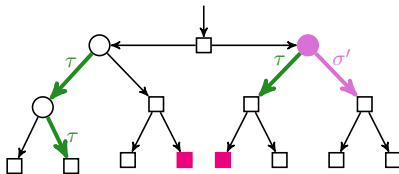
Given a non-winning strategy τ ,
check if there exists a winning strategy σ
such that $\tau(v) \neq \sigma(v)$ iff $v \in E$



Counterfactual explanation

Explanation E

Given a non-winning strategy τ ,
check if there exists a winning strategy σ
such that $\tau(v) \neq \sigma(v)$ iff $v \in E$



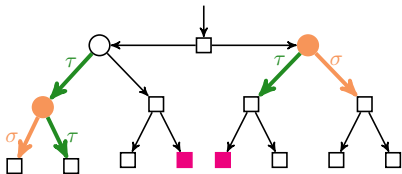
Contribution

- ▶ finding an explanation from a cause is in P

Counterfactual explanation

Explanation E

Given a non-winning strategy τ ,
check if there exists a winning strategy σ
such that $\tau(v) \neq \sigma(v)$ iff $v \in E$



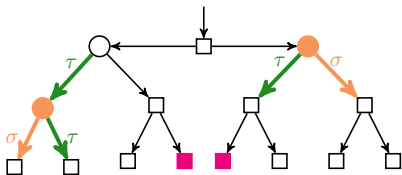
Contribution

- finding an explanation from a cause is in P

Counterfactual explanation

Explanation E

Given a non-winning strategy τ ,
check if there exists a winning strategy σ
such that $\tau(v) \neq \sigma(v)$ iff $v \in E$



Minimal explanation E

E is an explanation such that
 $d(\sigma, \tau) = d_{\min}^{\text{winning}}$

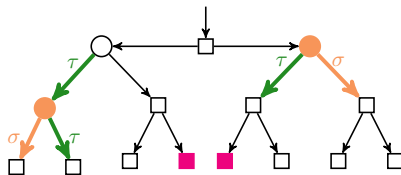
Contribution

- ▶ finding an explanation from a cause is in P

Counterfactual explanation

Explanation E

Given a non-winning strategy τ ,
check if there exists a winning strategy σ
such that $\tau(v) \neq \sigma(v)$ iff $v \in E$



Minimal explanation E

E is an explanation such that

$$d(\sigma, \tau) = d_{\min}^{\text{winning}}$$

Hamming distance

$$d(\sigma, \tau) = 2$$

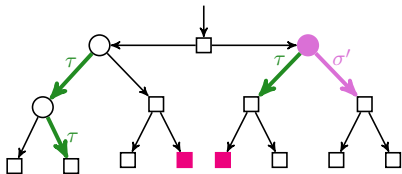
Contribution

- ▶ finding an explanation from a cause is in P

Counterfactual explanation

Explanation E

Given a non-winning strategy τ ,
check if there exists a winning strategy σ
such that $\tau(v) \neq \sigma(v)$ iff $v \in E$



Minimal explanation E

E is an explanation such that

$$d(\sigma, \tau) = d_{\min}^{\text{winning}}$$

Hamming distance

$$d(\sigma, \tau) = 2 > 1 = d(\sigma', \tau)$$

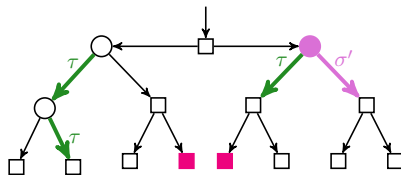
Contribution

- ▶ finding an explanation from a cause is in P

Counterfactual explanation

Explanation E

Given a non-winning strategy τ ,
check if there exists a winning strategy σ
such that $\tau(v) \neq \sigma(v)$ iff $v \in E$



Minimal explanation E

E is an explanation such that

$$d(\sigma, \tau) = d_{\min}^{\text{winning}}$$

Hamming distance

$$d(\sigma, \tau) = 2 > 1 = d(\sigma', \tau)$$

Minimal explanation problem

Check if E is a minimal explanation

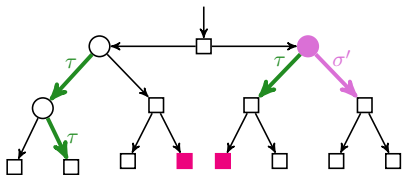
Contribution

- ▶ finding an explanation from a cause is in P

Counterfactual explanation

Explanation E

Given a non-winning strategy τ ,
check if there exists a winning strategy σ
such that $\tau(v) \neq \sigma(v)$ iff $v \in E$



Minimal explanation E

E is an explanation such that

$$d(\sigma, \tau) = d_{\min}^{\text{winning}}$$

Hamming distance

$$d(\sigma, \tau) = 2 > 1 = d(\sigma', \tau)$$

Minimal explanation problem

Check if E is a minimal explanation

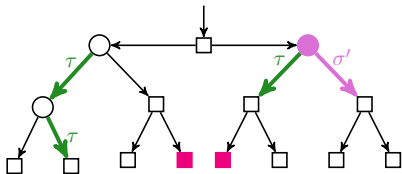
Contribution

- ▶ finding an explanation from a cause is in P
- ▶ coNP-complete problem for Hamming

Counterfactual explanation

Explanation E

Given a non-winning strategy τ ,
check if there exists a winning strategy σ
such that $\tau(v) \neq \sigma(v)$ iff $v \in E$



Minimal explanation E

E is an explanation such that

$$d(\sigma, \tau) = d_{\min}^{\text{winning}}$$

Hamming distance

$$d(\sigma, \tau) = 2 > 1 = d(\sigma', \tau)$$

Minimal explanation problem

Check if E is a minimal explanation

Contribution

- ▶ finding an explanation from a cause is in P
- ▶ coNP-complete problem for Hamming
- ▶ NP-hardness Hausdorff-inspired Hamming distances

Summary

Summary

In transition systems

Summary

In transition systems

- ▶ Check the counterfactual causality

distance	causality
prefix	in P
Hamming	in P
Levenshtein	in P

Summary

In transition systems

- ▶ Check the counterfactual causality
- ▶ Counterfactual causality for the Hamming distance is consistent with Halpern and Pearl's but-for causes

distance	causality
prefix	in P
Hamming	in P
Levenshtein	in P

Summary

In transition systems

- ▶ Check the counterfactual causality
- ▶ Counterfactual causality for the Hamming distance is consistent with Halpern and Pearl's but-for causes

distance	causality
prefix	in P
Hamming	in P
Levenshtein	in P

In reachability/safety games

Summary

In transition systems

- ▶ Check the counterfactual causality
- ▶ Counterfactual causality for the Hamming distance is consistent with Halpern and Pearl's but-for causes

distance	causality
prefix	in P
Hamming	in P
Levenshtein	in P

In reachability/safety games

- ▶ Generalisation of counterfactual causality with distances over strategies

Summary

In transition systems

- ▶ Check the counterfactual causality
- ▶ Counterfactual causality for the Hamming distance is consistent with Halpern and Pearl's but-for causes

distance	causality
prefix	in P
Hamming	in P
Levenshtein	in P

In reachability/safety games

- ▶ Generalisation of counterfactual causality with distances over strategies
- ▶ Check the counterfactual causality

distance	causality	
Hausdorff lifting of the prefix distance	in P	
Hamming strategy distance	in P acyclic games	
Hausdorff-inspired distance		

Summary

In transition systems

- ▶ Check the counterfactual causality
- ▶ Counterfactual causality for the Hamming distance is consistent with Halpern and Pearl's but-for causes

distance	causality
prefix	in P
Hamming	in P
Levenshtein	in P

In reachability/safety games

- ▶ Generalisation of counterfactual causality with distances over strategies
- ▶ Check the counterfactual causality
- ▶ Introduction of the notion of counterfactual explanation

distance	causality	
Hausdorff lifting of the prefix distance	in P	
Hamming strategy distance	in P acyclic games	
Hausdorff-inspired distance		

Summary

In transition systems

- ▶ Check the counterfactual causality
- ▶ Counterfactual causality for the Hamming distance is consistent with Halpern and Pearl's but-for causes

distance	causality
prefix	in P
Hamming	in P
Levenshtein	in P

In reachability/safety games

- ▶ Generalisation of counterfactual causality with distances over strategies
- ▶ Check the counterfactual causality
- ▶ Introduction of the notion of counterfactual explanation
- ▶ Check the minimal counterfactual explanation

distance	causality	explanations
Hausdorff lifting of the prefix distance	in P	
Hamming strategy distance	in P acyclic games	coNP-complete
Hausdorff-inspired distance		NP-hardness

Summary

In transition systems

- ▶ Check the counterfactual causality
- ▶ Counterfactual causality for the Hamming distance is consistent with Halpern and Pearl's but-for causes

In reachability/safety games

- ▶ Generalisation of counterfactual causality with distances over strategies
- ▶ Check the counterfactual causality
- ▶ Introduction of the notion of counterfactual explanation
- ▶ Check the minimal counterfactual explanation

Perspectives

Summary

In transition systems

- ▶ Check the counterfactual causality
- ▶ Counterfactual causality for the Hamming distance is consistent with Halpern and Pearl's but-for causes

In reachability/safety games

- ▶ Generalisation of counterfactual causality with distances over strategies
- ▶ Check the counterfactual causality
- ▶ Introduction of the notion of counterfactual explanation
- ▶ Check the minimal counterfactual explanation

Perspectives

- ▶ Check counterfactual causes in all reachability/safety games

Summary

In transition systems

- ▶ Check the counterfactual causality
- ▶ Counterfactual causality for the Hamming distance is consistent with Halpern and Pearl's but-for causes

In reachability/safety games

- ▶ Generalisation of counterfactual causality with distances over strategies
- ▶ Check the counterfactual causality
- ▶ Introduction of the notion of counterfactual explanation
- ▶ Check the minimal counterfactual explanation

Perspectives

- ▶ Check counterfactual causes in all reachability/safety games
- ▶ Finding a (good) counterfactual cause

Summary

In transition systems

- ▶ Check the counterfactual causality
- ▶ Counterfactual causality for the Hamming distance is consistent with Halpern and Pearl's but-for causes

In reachability/safety games

- ▶ Generalisation of counterfactual causality with distances over strategies
- ▶ Check the counterfactual causality
- ▶ Introduction of the notion of counterfactual explanation
- ▶ Check the minimal counterfactual explanation

Perspectives

- ▶ Check counterfactual causes in all reachability/safety games
- ▶ Finding a (good) counterfactual cause
- ▶ Study the impact of the distance over causes

Summary

In transition systems

- ▶ Check the counterfactual causality
- ▶ Counterfactual causality for the Hamming distance is consistent with Halpern and Pearl's but-for causes

In reachability/safety games

- ▶ Generalisation of counterfactual causality with distances over strategies
- ▶ Check the counterfactual causality
- ▶ Introduction of the notion of counterfactual explanation
- ▶ Check the minimal counterfactual explanation

Perspectives

- ▶ Check counterfactual causes in all reachability/safety games
- ▶ Finding a (good) counterfactual cause
- ▶ Study the impact of the distance over causes

Thank you! Questions?