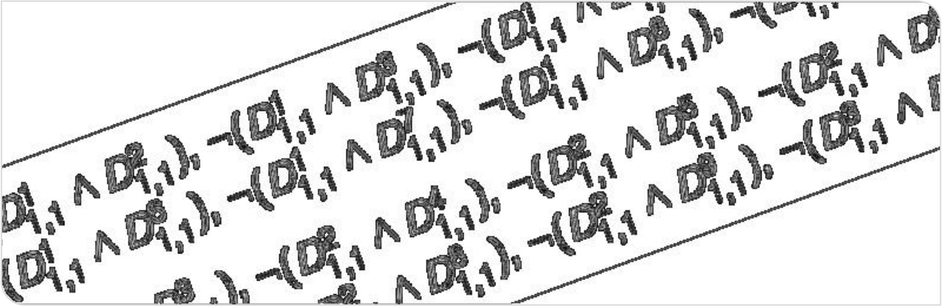
Karlsruhe Institute of Technology

An Information-Flow Perspective on Algorithmic Fairness

KeY Symposium 2023

Samuel Teuber, Bernhard Beckert | August 10, 2023



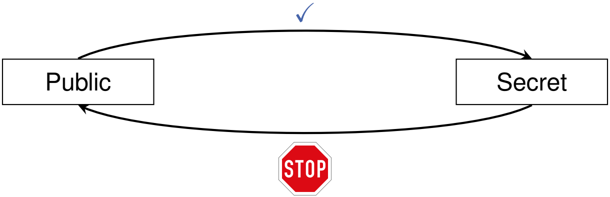
Motivation I: Information-Flow

- Established topic in **computer security**
- **Tools** available to analyze source code

Motivation I: Information-Flow

- Established topic in **computer security**
- **Tools** available to analyze source code

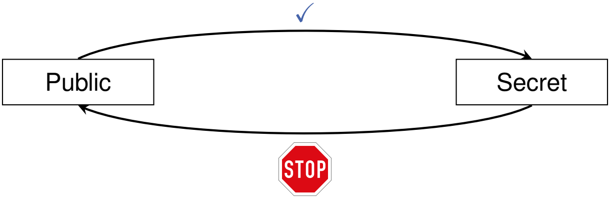
General Idea:



Motivation I: Information-Flow

- Established topic in **computer security**
- **Tools** available to analyze source code

General Idea:

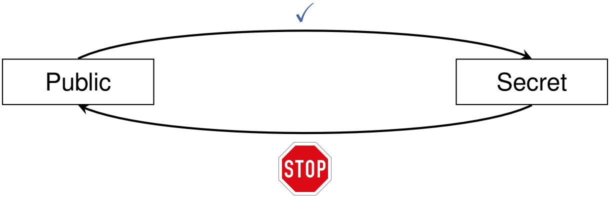



```
func f1(public, secret):
    result=public+secret
    return result
```

Motivation I: Information-Flow

- Established topic in **computer security**
- **Tools** available to analyze source code

General Idea:



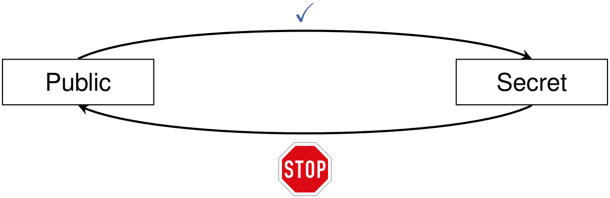

```
func f1(public, secret):
    result=public+secret
    return result
```

Insecure Information-Flow

Motivation I: Information-Flow

- Established topic in **computer security**
- **Tools** available to analyze source code

General Idea:




```
func f1(public, secret):
    result=public+secret
    return result
```

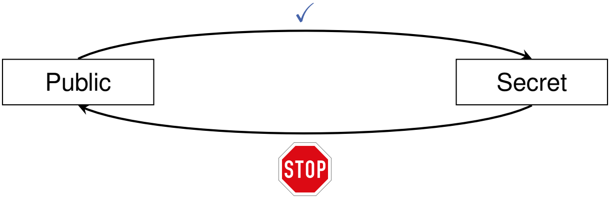
```
func f2(public, secret)
    result=public+1
    return result
```

Insecure Information-Flow

Motivation I: Information-Flow

- Established topic in **computer security**
- **Tools** available to analyze source code

General Idea:




```
func f1(public, secret):
    result=public+secret
    return result
```

Insecure Information-Flow

```
func f2(public, secret)
    result=public+1
    return result
```

Unconditional Noninterference

Institute of Information Security and Dependability (KASTEL)

Motivation II: Group Fairness

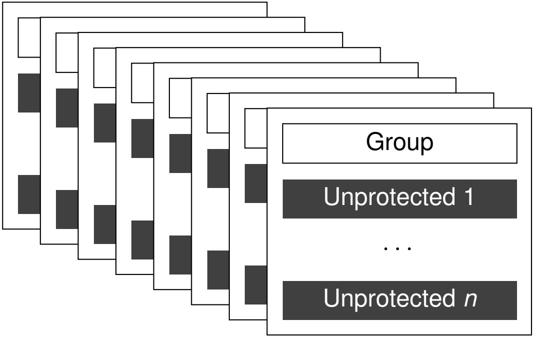
- Important class of Fairness definitions in **Algorithmic Fairness**
- Usually framed as **probabilistic properties**

General Idea:

Motivation II: Group Fairness

- Important class of Fairness definitions in **Algorithmic Fairness**
- Usually framed as **probabilistic properties**

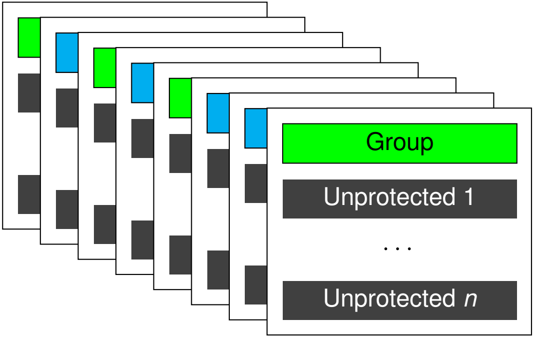
General Idea:



Motivation II: Group Fairness

- Important class of Fairness definitions in **Algorithmic Fairness**
- Usually framed as **probabilistic properties**

General Idea:

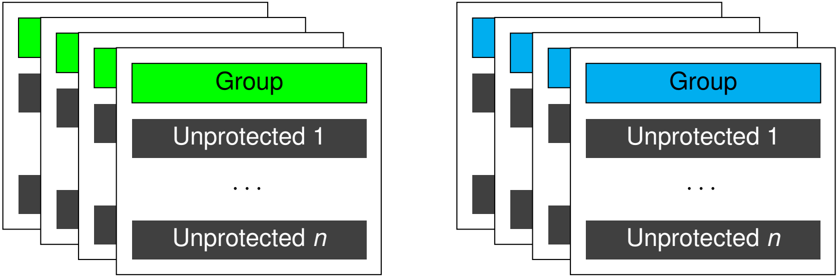


Motivation II: Group Fairness

Karlsruhe Institute of Technology

- Important class of Fairness definitions in **Algorithmic Fairness**
- Usually framed as **probabilistic properties**

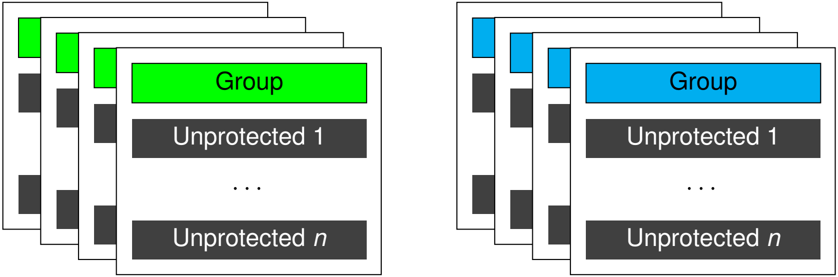
General Idea:



Motivation II: Group Fairness

- Important class of Fairness definitions in **Algorithmic Fairness**
- Usually framed as **probabilistic properties**

General Idea:



Does a decision procedure **disparately treat** individuals from different groups?

Motivation II: Group Fairness

Karlsruhe Institute of Technology

- Important class of Fairness definitions in **Algorithmic Fairness**
- Usually framed as **probabilistic properties**

General Idea:

- Group Attribute: Random Variable $G \in \mathcal{G}$
- Unprotected Attribute: Random Variable $U \in \mathcal{U}$
- Deterministic Decision Procedure: $P : \mathcal{G} \times \mathcal{U} \to \mathcal{D}$
- Finite domains

Does a decision procedure **disparately treat** individuals from different groups?

Examples

`age` given in decades

```
func credit1(age, score):
  return (age != 5)
```

Examples

`age` given in decades

```
func credit1(age, score):
  return (age != 5)
```

```
func credit2(age, score):
  return (score>8)
```

Examples

`age` given in decades

```
func credit1(age, score):
  return (age != 5)
```

```
func credit2(age, score):
  return (score>8)
```

```
func credit3(age, score):
  if (age >= 6):
    return (score >= 8)
  else:
    return (score >= 6)
```

This Work

Analyze **Decision Procedures** w.r.t Fairness Criteria by assigning **high security status** to a protected group attribute and performing **Information-Flow analyses**

Outline

Qualitative Information-Flow

Qualitative Information-Flow

Unconditional Noninterference

A program P satisfies *Unconditional Noninterference* iff **for all public** inputs $u \in \mathcal{U}$ and **all secret** inputs $g, g' \in \mathcal{G}$ it holds that

$$P(u, g) = P(u, g').$$

Qualitative Information-Flow

Unconditional Noninterference

A program P satisfies *Unconditional Noninterference* iff **for all public** inputs $u \in \mathcal{U}$ and **all secret** inputs $g, g' \in \mathcal{G}$ it holds that

$$P(u, g) = P(u, g').$$

Demographic Parity

A decision procedure satisfies demographic parity iff for all $d \in \mathcal{D}$ and $g_1, g_2 \in \mathcal{G}$

Qualitative Information-Flow

Unconditional Noninterference

A program P satisfies *Unconditional Noninterference* iff **for all public** inputs $u \in \mathcal{U}$ and **all secret** inputs $g, g' \in \mathcal{G}$ it holds that

$$P(u, g) = P(u, g') .$$

Demographic Parity

A decision procedure satisfies demographic parity iff for all $d \in \mathcal{D}$ and $g_1, g_2 \in \mathcal{G}$

$$\Pr\left[P(G, U) = d \mid G = g_1\right] = \Pr\left[P(G, U) = d \mid G = g_2\right]$$

Qualitative Information-Flow

Karlsruhe Institute of Technology

Unconditional Noninterference

A program P satisfies *Unconditional Noninterference* iff **for all public** inputs $u \in \mathcal{U}$ and **all secret** inputs $g, g' \in \mathcal{G}$ it holds that

$$P(u, g) = P(u, g') .$$

Demographic Parity

A decision procedure satisfies demographic parity iff for all $d \in \mathcal{D}$ and $g_1, g_2 \in \mathcal{G}$

$$\Pr[P(G, U) = d \mid G = g_1] = \Pr[P(G, U) = d \mid G = g_2]$$

For arbitrary but independent variables G, U:

$$\text{Unconditional Noninterference} \Rightarrow \text{Demographic Parity}$$

Qualitative Information-Flow

Karlsruhe Institute of Technology

Unconditional Noninterference

A program P satisfies *Unconditional Noninterference* iff **for all public** inputs $u \in \mathcal{U}$ and **all secret** inputs $g, g' \in \mathcal{G}$ it holds that

$$P(u, g) = P(u, g').$$

Demographic Parity

A decision procedure satisfies demographic parity iff for all $d \in \mathcal{D}$ and $g_1, g_2 \in \mathcal{G}$

$$\Pr[P(G, U) = d \mid G = g_1] = \Pr[P(G, U) = d \mid G = g_2]$$

For arbitrary but independent variables G, U:

$$\text{Unconditional Noninterference} \Rightarrow \text{Demographic Parity}$$
$$\text{Unconditional Noninterference} \not\Leftarrow \text{Demographic Parity}$$

Qualitative Information Flow (Refined)

Karlsruhe Institute of Technology

Instead of unconditional guarantee:

```
boolean credit3(int age, int score){
    if (age >= 6){
        return (score >= 8);
    } else {
        return (score >= 6);
    }
}
```

Qualitative Information Flow (Refined)

Instead of unconditional guarantee:

- Restrict guarantee to **parts of the input space**

```
//@ requires age < 6;
//@ determines \result \by score;
```

```
boolean credit3(int age, int score){
    if (age >= 6){
        return (score >= 8);
    } else {
        return (score >= 6);
    }
}
```

Qualitative Information Flow (Refined)

Instead of unconditional guarantee:

- Restrict guarantee to **parts of the input space**

```
//@ requires age < 6;
//@ determines \result \by score;
```

- Provide **classification** of inputs that shall be treated equally

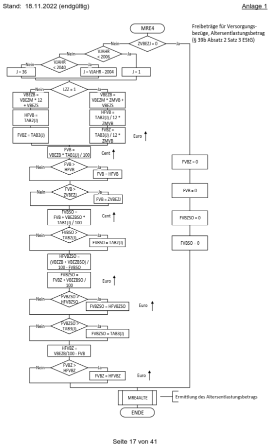
```
//@ determines \result \by score, (age >= 6);

boolean credit3(int age, int score){
    if (age >= 6){
        return (score >= 8);
    } else {
        return (score >= 6);
    }
}
```

German Wage Tax Code

35 Inputs
Yearly Wage
Tax category
...
Health Insurance

German Wage Tax Code

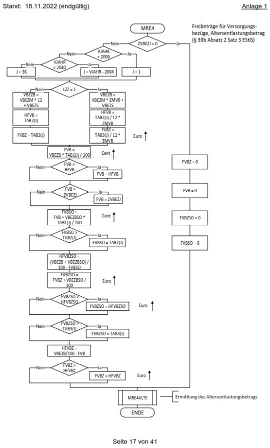
35 Inputs
Yearly Wage
Tax category
...
Health Insurance



26 pages of flow charts

German Wage Tax Code

35 Inputs
Yearly Wage
Tax category
...
Health Insurance

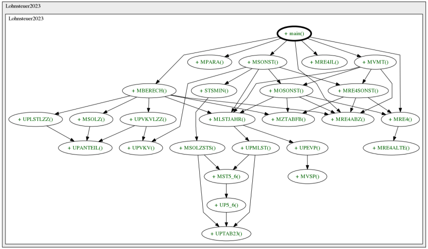


26 pages of flow charts

17 Output
Wage tax
Additional wage tax
...
Tax Exemption

German Wage Tax Code

35 Inputs
Yearly Wage
Tax category
...
Health Insurance

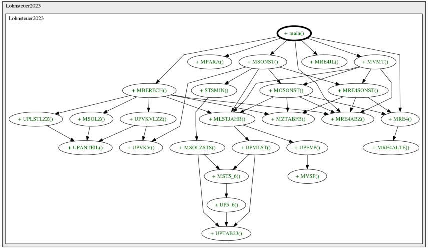


1500 lines of Java Code
26 methods
global state
intermediate variables

17 Output
Wage tax
Additional wage tax
...
Tax Exemption

German Wage Tax Code

35 Inputs
Yearly Wage
Tax category
...
Health Insurance
Religious Affiliation

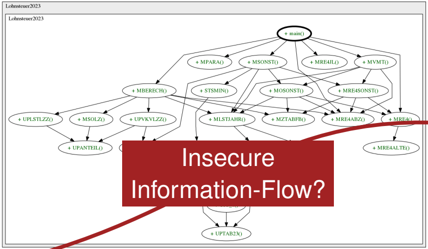


1500 lines of Java Code
26 methods
global state
intermediate variables

17 Output
Wage tax
Additional wage tax
...
Tax Exemption
Church Tax

German Wage Tax Code

35 Inputs
Yearly Wage
Tax category
...
Health Insurance
Religious
Affiliation



Insecure
Information-Flow?

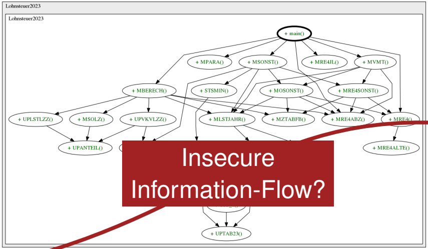
17 Output
Wage tax
Additional wage
tax
...
Tax Exemption
Church Tax

1500 lines of Java Code
26 methods
global state
intermediate variables

Institute of Information Security
and Dependability (KASTEL)

German Wage Tax Code

35 Inputs
Yearly Wage
Tax category
...
Health Insurance
Religious Affiliation



Insecure Information-Flow?

1500 lines of Java Code
26 methods
global state
intermediate variables

17 Output
Wage tax
Additional wage tax
...
Tax Exemption
Church Tax

Analysis of Java Code 2015-2023 using the tool Joana

No insecure Information-Flow!

Graf et al. 2013; Snelting et al. 2014

Quantitative Information-Flow

Conditional Vulnerability

Karlsruhe Institute of Technology

Intuition:

You observe a randomly sampled $u \in \mathcal{U}$ and P's outcome $d \in \mathcal{D}$.
With what probability can you guess G?

Conditional Vulnerability

Intuition:
You observe a randomly sampled $u \in \mathcal{U}$ and P's outcome $d \in \mathcal{D}$.
With what probability can you guess G?

Conditional Vulnerability

For a program P and random independent variables G, U, we define the *Conditional Vulnerabiliy* $V(G|P, U)$ as follows:

$$\sum_{(u,d)\in\mathcal{U}\times\mathcal{D}} \Pr\left[P(G, U) = d, U = u\right] \cdot \max_{g\in\mathcal{G}} \Pr\left[G = g | P(G, U) = d, U = u\right]$$

see e.g. Smith 2009

Conditional Vulnerability

Intuition:
You observe a randomly sampled $u \in \mathcal{U}$ and P's outcome $d \in \mathcal{D}$.
With what probability can you guess G?

Conditional Vulnerability

For a program P and random independent variables G, U, we define the *Conditional Vulnerabiliy* $V(G|P,U)$ as follows:

$$\sum_{(u,d)\in\mathcal{U}\times\mathcal{D}} \Pr\left[P(G,U)=d, U=u\right] \cdot \max_{g\in\mathcal{G}} \Pr\left[G=g \middle| P(G,U)=d, U=u\right]$$

see e.g. Smith 2009

Can we use this as a Fairness Metric?

Conditional Vulnerability

Karlsruhe Institute of Technology

Intuition:
You observe a randomly sampled $u \in \mathcal{U}$ and P's outcome $d \in \mathcal{D}$.
With what probability can you guess G?

Conditional Vulnerability

For a program P and random independent variables G, U, we define the *Conditional Vulnerabiliy* $V(G|P, U)$ as follows:

$$\sum_{(u,d) \in \mathcal{U} \times \mathcal{D}} \Pr\left[P(G, U) = d, U = u\right] \cdot \max_{g \in \mathcal{G}} \Pr\left[G = g | P(G, U) = d, U = u\right]$$

see e.g. Smith 2009

Can we use this as a Fairness Metric?
...for binary decisions? ($|\mathcal{D}| = 2$)

A naive approach

Karlsruhe Institute of Technology

Given known distributions of G and U:
Compute $V\left(G \middle| P, U\right)$

A naive approach

Karlsruhe Institute of Technology

Given known distributions of G and U:
Compute $V(G|P, U)$

Problem: Vulnerability Measures two things at the same time:

- How easy is it to guess G?
- How much of G is revealed by P?

A naive approach

Karlsruhe Institute of Technology

Given known distributions of G and U:
Compute $V\left(G \middle| P, U\right)$

Problem: Vulnerability Measures two things at the same time:

- How easy is it to guess G?
- How much of G is revealed by P?

$\Rightarrow$ If $G = g_1$ is *extremely* likely, P does not matter

A naive approach

Karlsruhe Institute of Technology

Given known distributions of G and U:
Compute $V(G|P, U)$

Problem: Vulnerability Measures two things at the same time:

- How easy is it to guess G?
- How much of G is revealed by P?

$\Rightarrow$ If $G = g_1$ is *extremely* likely, P does not matter

$\Rightarrow$ Independence of P is an undesirable property for a metric evaluating P

Measuring for uniformly distributed G

Fairness Spread

We define the *Fairness Spread* $S(G, U, P)$ as follows:

$$\sum_{u \in \mathcal{U}} \Pr[U = u] \cdot \max_{g_1, g_2 \in \mathcal{G}} \left(\Pr[P(g_1, u) = 1] - \Pr[P(g_2, u) = 1]\right)$$

Measuring for uniformly distributed G

Fairness Spread

We define the *Fairness Spread* $S(G, U, P)$ as follows:

$$\sum_{u \in \mathcal{U}} \Pr[U = u] \cdot \max_{g_1, g_2 \in \mathcal{G}} \left(\Pr[P(g_1, u) = 1] - \Pr[P(g_2, u) = 1]\right)$$

Theorem

*Assume G **is distributed uniformly** and U is independent of G, then:*

$$S(G, U, P) = |\mathcal{G}| \cdot V(G|P, U) - 1$$

Measuring for uniformly distributed G

Fairness Spread

We define the *Fairness Spread* $S(G, U, P)$ as follows:

$$\sum_{u \in \mathcal{U}} \Pr[U = u] \cdot \max_{g_1, g_2 \in \mathcal{G}} \left(\Pr[P(g_1, u) = 1] - \Pr[P(g_2, u) = 1]\right)$$

Theorem

*Assume G **is distributed uniformly** and U is independent of G, then:*

$$S(G, U, P) = |\mathcal{G}| \cdot V(G|P, U) - 1$$

$\Rightarrow$ $S(G, U, P)$ is **independent** of G's distribution!

Examples

```
func credit1(age, score):
  return (age != 5)
```

$S(G,U,P)$ uniform distribution	$S(G,U,P)$ $U \in [6,7]$ more likely
1.0	1.0

Institute of Information Security
and Dependability (KASTEL)

Examples

	$S(G, U, P)$ uniform distribution	$S(G, U, P)$ $U \in [6, 7]$ more likely
`func credit1(age, score):` `  return (age != 5)`	1.0	1.0
`func credit2(age, score):` `  return (score>8)`	0.0	0.0

Examples

	$S(G, U, P)$ uniform distribution	$S(G, U, P)$ $U \in [6, 7]$ more likely
`func credit1(age, score):` `  return (age != 5)`	1.0	1.0
`func credit2(age, score):` `  return (score>8)`	0.0	0.0

`func credit3(age, score):`
`  if (age >= 6):`
`    return (score >= 8)`
`  else:`
`    return (score >= 6)`
	0.2	0.3

The Meaning of Fairness Spread

Karlsruhe Institute of Technology

$$\underbrace{\sum_{u \in \mathcal{U}} \Pr\left[U = u\right]}_{\text{Weighted by } U} \cdot \underbrace{\max_{g_1, g_2 \in \mathcal{G}} \left(\Pr\left[P\left(g_1, u\right) = 1\right] - \Pr\left[P\left(g_2, u\right) = 1\right]\right)}_{\text{Maximal disparity between groups}}$$

The Meaning of Fairness Spread

Karlsruhe Institute of Technology

$$\underbrace{\sum_{u \in \mathcal{U}} \Pr\left[U = u\right]}_{\text{Weighted by } U} \cdot \underbrace{\max_{g_1, g_2 \in \mathcal{G}} \left(\Pr\left[P\left(g_1, u\right) = 1\right] - \Pr\left[P\left(g_2, u\right) = 1\right]\right)}_{\text{Maximal disparity between groups}}$$

Handwavy Explanation:

The higher the fairness spread the more group-based disparities.

The Meaning of Fairness Spread

Karlsruhe Institute of Technology

$$\underbrace{\sum_{u \in \mathcal{U}} \Pr\left[U = u\right]}_{\text{Weighted by } U} \cdot \underbrace{\max_{g_1, g_2 \in \mathcal{G}} \left(\Pr\left[P\left(g_1, u\right) = 1\right] - \Pr\left[P\left(g_2, u\right) = 1\right]\right)}_{\text{Maximal disparity between groups}}$$

Handwavy Explanation:

The higher the fairness spread the more group-based disparities.

- Is there a **more formal** but also intuitive explanation?
- Ability to handle dependent variables?

The Meaning of Fairness Spread

Karlsruhe Institute of Technology

$$\underbrace{\sum_{u \in \mathcal{U}} \Pr\left[U = u\right]}_{\text{Weighted by } U} \cdot \underbrace{\max_{g_1, g_2 \in \mathcal{G}} \left(\Pr\left[P\left(g_1, u\right) = 1\right] - \Pr\left[P\left(g_2, u\right) = 1\right]\right)}_{\text{Maximal disparity between groups}}$$

Handwavy Explanation:

The higher the fairness spread the more group-based disparities.

- Is there a **more formal** but also intuitive explanation?
- Ability to handle dependent variables?

$\Rightarrow$ Causal Analysis to the rescue

Information Flow and Causal Analysis

Structural Caual Models

Karlsruhe Institute of Technology

A rich framework for the (statistical) analysis of causal relationships

Three components

- Background Variables $B = \{B_1, \ldots, B_k\}$
- Modeled Variables $V = \{V_1, \ldots, V_n\}$
- Set of functions $f_i\left(\mathrm{pa}_i, B_{\mathrm{pa}_i}\right)$:
  How is V_i computed based on $\mathrm{pa}_i \subseteq V$ and $B_{\mathrm{pa}_i} \subseteq B$?

Example: Red Cars pay higher car insurance premiums Kusner et al. 2017

Structural Caual Models

A rich framework for the (statistical) analysis of causal relationships

Three components

- Background Variables $B = \{B_1, \ldots, B_k\}$
- Modeled Variables $V = \{V_1, \ldots, V_n\}$
- Set of functions $f_i\left(\mathrm{pa}_i, B_{\mathrm{pa}_i}\right)$:
  How is V_i computed based on $\mathrm{pa}_i \subseteq V$ and $B_{\mathrm{pa}_i} \subseteq B$?

Example: Red Cars pay higher car insurance premiums Kusner et al. 2017

$\texttt{Group} \coloneqq \varepsilon_1 \sim \mathcal{U}_d\,(0, 1)$

G

Structural Caual Models

A rich framework for the (statistical) analysis of causal relationships

Three components

- Background Variables $B = \{B_1, \ldots, B_k\}$
- Modeled Variables $V = \{V_1, \ldots, V_n\}$
- Set of functions $f_i\left(\mathrm{pa}_i, B_{\mathrm{pa}_i}\right)$:
  How is V_i computed based on $\mathrm{pa}_i \subseteq V$ and $B_{\mathrm{pa}_i} \subseteq B$?

Example: Red Cars pay higher car insurance premiums   Kusner et al. 2017

$$\texttt{Group} := \varepsilon_1 \sim \mathcal{U}_d\,(0, 1)$$
$$\texttt{Aggressive} := \varepsilon_2 \sim \mathcal{U}\,(0, 1)$$

Structural Caual Models

Karlsruhe Institute of Technology

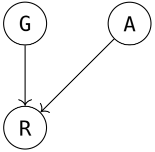
A rich framework for the (statistical) analysis of causal relationships

Three components

- Background Variables $B = \{B_1, \ldots, B_k\}$
- Modeled Variables $V = \{V_1, \ldots, V_n\}$
- Set of functions $f_i\left(\mathrm{pa}_i, B_{\mathrm{pa}_i}\right)$:
  How is V_i computed based on $\mathrm{pa}_i \subseteq V$ and $B_{\mathrm{pa}_i} \subseteq B$?

Example: Red Cars pay higher car insurance premiums Kusner et al. 2017

$$\begin{aligned}\texttt{Group} &:= \varepsilon_1 \sim \mathcal{U}_d\,(0,1)\\ \texttt{Aggressive} &:= \varepsilon_2 \sim \mathcal{U}\,(0,1)\\ \texttt{Red Car} &:= (0.5 \cdot \texttt{Group} + \texttt{Aggressive}) > 0.8\end{aligned}$$



Structural Caual Models

A rich framework for the (statistical) analysis of causal relationships

Three components

- Background Variables $B = \{B_1, \dots, B_k\}$
- Modeled Variables $V = \{V_1, \dots, V_n\}$
- Set of functions $f_i\left(\mathrm{pa}_i, B_{\mathrm{pa}_i}\right)$:
  How is V_i computed based on $\mathrm{pa}_i \subseteq V$ and $B_{\mathrm{pa}_i} \subseteq B$?

Example: Red Cars pay higher car insurance premiums Kusner et al. 2017

$$\begin{aligned} \texttt{Group} &:= \varepsilon_1 \sim \mathcal{U}_d\,(0,1) \\ \texttt{Aggressive} &:= \varepsilon_2 \sim \mathcal{U}\,(0,1) \\ \texttt{Red Car} &:= (0.5 \cdot \texttt{Group} + \texttt{Aggressive}) > 0.8 \\ \texttt{High P.} &:= \texttt{Red Car} \end{aligned}$$

$$\begin{tikzcd} G \arrow[d] & A \arrow[dl] \\ R \arrow[r] & H \end{tikzcd}$$

Institute of Information Security and Dependability (KASTEL)

Structural Caual Models

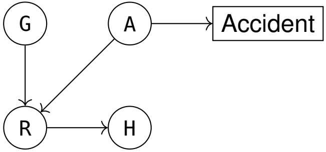
A rich framework for the (statistical) analysis of causal relationships

Three components

- Background Variables $B = \{B_1, \ldots, B_k\}$
- Modeled Variables $V = \{V_1, \ldots, V_n\}$
- Set of functions $f_i\left(\mathrm{pa}_i, B_{\mathrm{pa}_i}\right)$:
  How is V_i computed based on $\mathrm{pa}_i \subseteq V$ and $B_{\mathrm{pa}_i} \subseteq B$?

Example: Red Cars pay higher car insurance premiums — Kusner et al. 2017

$$\begin{aligned}
\texttt{Group} &:= \varepsilon_1 \sim \mathcal{U}_d(0,1) \\
\texttt{Aggressive} &:= \varepsilon_2 \sim \mathcal{U}(0,1) \\
\texttt{Red Car} &:= (0.5 \cdot \texttt{Group} + \texttt{Aggressive}) > 0.8 \\
\texttt{High P.} &:= \texttt{Red Car}
\end{aligned}$$



Interventions

Given a structural causal model and a concrete observation:
How would the observation be different for a modified variable?

Interventions

Given a structural causal model and a concrete observation:
How would the observation be different for a modified variable?

$$\begin{aligned}
\texttt{Group} &\coloneqq \varepsilon_1 \sim \mathcal{U}_d\,(0,1) \\
\texttt{Aggressive} &\coloneqq \varepsilon_2 \sim \mathcal{U}\,(0,1) \\
\texttt{Red Car} &\coloneqq \left(0.5 \cdot \texttt{Group} + \texttt{Aggressive}\right) > 0.8 \\
\texttt{High P.} &\coloneqq \texttt{Red Car}
\end{aligned}$$

Interventions

Karlsruhe Institute of Technology

Given a structural causal model and a concrete observation:
How would the observation be different for a modified variable?

$$\begin{aligned}
\texttt{Group} &:= \varepsilon_1 \sim \mathcal{U}_d\,(0,1) \\
\texttt{Aggressive} &:= \varepsilon_2 \sim \mathcal{U}\,(0,1) \\
\texttt{Red Car} &:= (0.5 \cdot \texttt{Group} + \texttt{Aggressive}) > 0.8 \\
\texttt{High P.} &:= \texttt{Red Car}
\end{aligned}$$

G	$\Pr[\texttt{Red Car} = 1]$
0	0.2
1	0.7

Interventions

Karlsruhe Institute of Technology

Given a structural causal model and a concrete observation:
How would the observation be different for a modified variable?

$$\texttt{Group} \coloneqq \varepsilon_1 \sim \mathcal{U}_d(0, 1)$$
$$\texttt{Aggressive} \coloneqq \varepsilon_2 \sim \mathcal{U}(0, 1)$$
$$\texttt{Red Car} \coloneqq (0.5 \cdot \texttt{Group} + \texttt{Aggressive}) > 0.8$$
$$\texttt{High P.} \coloneqq \texttt{Red Car}$$

G	$\Pr[\texttt{Red Car} = 1]$
0	0.2
1	0.7

Observation:

$\texttt{Group} = 0$

$\texttt{Red Car} = 0$

Interventions

Given a structural causal model and a concrete observation:
How would the observation be different for a modified variable?

$$\texttt{Group} := \varepsilon_1 \sim \mathcal{U}_d(0,1)$$
$$\texttt{Aggressive} := \varepsilon_2 \sim \mathcal{U}(0,1)$$
$$\texttt{Red Car} := (0.5 \cdot \texttt{Group} + \texttt{Aggressive}) > 0.8$$
$$\texttt{High P.} := \texttt{Red Car}$$

G	$\Pr[\texttt{Red Car} = 1]$
0	0.2
1	0.7

Observation:

$\texttt{Group} = 0$

$\texttt{Red Car} = 0$

Intervention:

$\texttt{Group} \leftarrow 1$

Interventions

Given a structural causal model and a concrete observation:
How would the observation be different for a modified variable?

$$\begin{aligned}
\texttt{Group} &:= \varepsilon_1 \sim \mathcal{U}_d(0,1) \\
\texttt{Aggressive} &:= \varepsilon_2 \sim \mathcal{U}(0,1) \\
\texttt{Red Car} &:= (0.5 \cdot \texttt{Group} + \texttt{Aggressive}) > 0.8 \\
\texttt{High P.} &:= \texttt{Red Car}
\end{aligned}$$

G	$\Pr[\texttt{Red Car} = 1]$
0	0.2
1	0.7

Observation:

$\texttt{Group} = 0$

$\texttt{Red Car} = 0$

Intervention:

$\texttt{Group} \leftarrow 1$

Possible Outcomes:

$\Pr[\texttt{Red Car} = 1] = 0.625$

Interventions

Given a structural causal model and a concrete observation:
How would the observation be different for a modified variable?

$$\begin{aligned}
\texttt{Group} &:= \varepsilon_1 \sim \mathcal{U}_d\,(0,1)\\
\texttt{Aggressive} &:= \varepsilon_2 \sim \mathcal{U}\,(0,1)\\
\texttt{Red Car} &:= \left(0.5 \cdot \texttt{Group} + \texttt{Aggressive}\right) > 0.8\\
\texttt{High P.} &:= \texttt{Red Car}
\end{aligned}$$

G	$\Pr\left[\texttt{Red Car} = 1\right]$
0	0.2
1	0.7

Observation:

$\texttt{Group} = 0$

$\texttt{Red Car} = 0$

Intervention:

$\texttt{Group} \leftarrow 1$

Possible Outcomes:

$\Pr\left[\texttt{Red Car} = 1\right] = 0.625$

Interventions provide us with information on counterfactual events:
What if the applicant had been older?

Counterfactual Fairness

Karlsruhe Institute of Technology

For a program P and a causal model C we define $\hat{P}_C\,(b)$:

- Compue G, U from C with background variable assignements b
- Return $P\,(G, U)$

Counterfactual Fairness

Karlsruhe Institute of Technology

For a program P and a causal model C we define $\hat{P}_C(b)$:

- Compue G, U from C with background variable assignements b
- Return $P(G, U)$

Counterfactual Version: $\hat{P}_C(g, b)$ (intervenes for G)

Counterfactual Fairness

A program P with inputs G and U is *counterfactually fair* with respect to a causal model C

Counterfactual Fairness

Karlsruhe Institute of Technology

For a program P and a causal model C we define $\hat{P}_C(b)$:

- Compue G, U from C with background variable assignements b
- Return $P(G, U)$

Counterfactual Version: $\hat{P}_C(g, b)$ (intervenes for G)

Counterfactual Fairness

A program P with inputs G and U is *counterfactually fair* with respect to a causal model C iff for any $g_1, g_2 \in \mathcal{G}$, $u \in \mathcal{U}$, $d \in \mathcal{D}$ it holds that:

Counterfactual Fairness

Karlsruhe Institute of Technology

For a program P and a causal model C we define $\hat{P}_C(b)$:

- Compue G, U from C with background variable assignements b
- Return $P(G, U)$

Counterfactual Version: $\hat{P}_C(g, b)$ (intervenes for G)

Counterfactual Fairness

A program P with inputs G and U is *counterfactually fair* with respect to a causal model C iff for any $g_1, g_2 \in \mathcal{G}$, $u \in \mathcal{U}$, $d \in \mathcal{D}$ it holds that:

$$\Pr\left[\hat{P}_C(g_1, B) = d \middle| U = u, G = g_1\right] = \Pr\left[\hat{P}_C(g_2, B) = d \middle| U = u, G = g_1\right]$$

Causality and Fairness Spread

Fairness Spread is a bound on **the probability of having a deviating counterfactual**.

Institute of Information Security and Dependability (KASTEL)

Causality and Fairness Spread

Fairness Spread is a bound on **the probability of having a deviating counterfactual**.

For two groups this bound is precise

Institute of Information Security and Dependability (KASTEL)

Causality and Fairness Spread

Fairness Spread is a bound on **the probability of having a deviating counterfactual**.
For two groups this bound is precise

Can be formally shown using the notion of a *difference function*:

$$\mathrm{Diff}_C\left(P, b\right) = \max_{g \in \mathcal{G}} \left| \hat{P}_C\left(b\right) - \hat{P}_C\left(g, b\right) \right|$$

Causality and Fairness Spread

Fairness Spread is a bound on **the probability of having a deviating counterfactual**.
For two groups this bound is precise

Can be formally shown using the notion of a *difference function*:

$$\mathrm{Diff}_C(P, b) = \max_{g \in \mathcal{G}} \left| \hat{P}_C(b) - \hat{P}_C(g, b) \right|$$

Consequences:

- Machinery for *Qualitative* Information Flow is applicable to $\hat{P}_C$
- Quantitative Information Flow Analyses can provide bounds for counterfactual unfairness

Example

Causal Model for Credit Example:

`score` provided by external entity with questionable methodology:

Example

Causal Model for Credit Example:

`score` provided by external entity with questionable methodology:

$$\texttt{group} \coloneqq \varepsilon_1 \sim \mathcal{U}_d\,(0, 9)$$

Example

Causal Model for Credit Example:

`score` provided by external entity with questionable methodology:

$$\texttt{group} \coloneqq \varepsilon_1 \sim \mathcal{U}_d\,(0, 9)$$
$$\texttt{income} \coloneqq \varepsilon_2 \sim \mathcal{U}\,(0, 9)$$

Example

Causal Model for Credit Example:

`score` provided by external entity with questionable methodology:

$$\begin{aligned}
\texttt{group} &\coloneqq \varepsilon_1 \sim \mathcal{U}_d\,(0,9)\\
\texttt{income} &\coloneqq \varepsilon_2 \sim \mathcal{U}\,(0,9)\\
\texttt{zipCode} &\coloneqq \textbf{if} \quad (\texttt{group} \geq 6) \quad \varepsilon_3 \sim \mathcal{U}\,(-1,5) \quad \textbf{else} \quad \varepsilon_4 \sim \mathcal{U}\,(-3,3)
\end{aligned}$$

Example

Causal Model for Credit Example:

`score` provided by external entity with questionable methodology:

$$
\begin{aligned}
\texttt{group} &\coloneqq \varepsilon_1 \sim \mathcal{U}_d\,(0,9) \\
\texttt{income} &\coloneqq \varepsilon_2 \sim \mathcal{U}\,(0,9) \\
\texttt{zipCode} &\coloneqq \textbf{if} \quad (\texttt{group} \geq 6) \quad \varepsilon_3 \sim \mathcal{U}\,(-1,5) \quad \textbf{else} \quad \varepsilon_4 \sim \mathcal{U}\,(-3,3) \\
\texttt{score} &\coloneqq \texttt{income} + \texttt{zipCode}
\end{aligned}
$$

Example

Causal Model for Credit Example:

`score` provided by external entity with questionable methodology:

$$\begin{aligned} \texttt{group} &:= \varepsilon_1 \sim \mathcal{U}_d\,(0,9) \\ \texttt{income} &:= \varepsilon_2 \sim \mathcal{U}\,(0,9) \\ \texttt{zipCode} &:= \textbf{if} \quad (\texttt{group} \geq 6) \quad \varepsilon_3 \sim \mathcal{U}\,(-1,5) \quad \textbf{else} \quad \varepsilon_4 \sim \mathcal{U}\,(-3,3) \\ \texttt{score} &:= \texttt{income} + \texttt{zipCode} \end{aligned}$$

```
func credit2(age, score):
    return (score>8)
```

```
func credit3(age, score):
    if (age >= 6):
        return (score >= 8)
    else:
        return (score >= 6)
```

Fairness Spread of $\hat{P}_C$: 0.27

Fairness Spread of $\hat{P}_C$: 0.23

Conclusion

We can use Information-Flow tools
to analyze fairness questions

Future Work:

- Machine Learning Systems
- Beyond binary decisions?
- Synthesizing restriced classifications?

References I

[1] Jürgen Graf, Martin Hecker, and Martin Mohr. "Using JOANA for Information Flow Control in Java Programs - A Practical Guide". In: *Proceedings of the 6th Working Conference on Programming Languages (ATPS'13)*. Lecture Notes in Informatics (LNI) 215. Springer Berlin / Heidelberg, Feb. 2013, pp. 123–138.

[2] Matt J. Kusner et al. "Counterfactual Fairness". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon et al. 2017, pp. 4066–4076.

References II

[3] Geoffrey Smith. "On the Foundations of Quantitative Information Flow". In: *Foundations of Software Science and Computational Structures, 12th International Conference, FOSSACS 2009, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2009, York, UK, March 22-29, 2009. Proceedings*. Ed. by Luca de Alfaro. Vol. 5504. LNCS. Cham: Springer, 2009, pp. 288–302. DOI: `10.1007/978-3-642-00596-1\_21`.

[4] Gregor Snelting et al. "Checking Probabilistic Noninterference Using JOANA". In: *it - Information Technology* 56 (Nov. 2014), pp. 280–287. DOI: `10.1515/itit-2014-1051`.