Karlsruhe Institute of Technology

Application-oriented Formal Verification Research Group
Institute of Information Security and Dependability (KASTEL)
Am Fasanengarten 5
76131 Karlsruhe
formal.kastel.kit.edu

An Information-Flow Perspective on Algorithmic Fairness

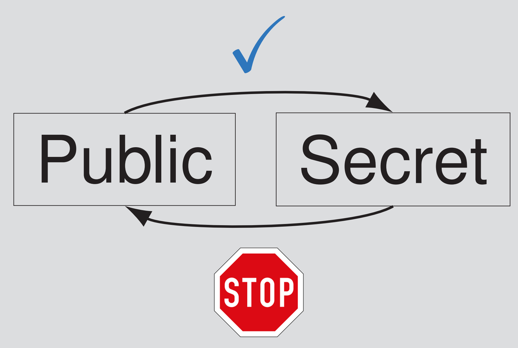
Samuel Teuber, Bernhard Beckert

Overview

(Java) Source Code
↓
White-Box Information-Flow Analysis
↓
Verdict on Fairness: Qualitative/Quantitative

Information-Flow Analysis

- Established topic in **computer security**
- Core Idea: **Public** and **Private** Information
- Private Information must not be leaked
- Many Tools [1, 2, 4]
- **Exhaustive** analysis of **source code**



Unconditional Noninterference

P satisfies Uncond. Noninterference iff for any $u \in \mathcal{U}$ and $g_1, g_2 \in \mathcal{G}$:

$$P(g_1, u) = P(g_2, u)$$

We can analyze Decision Making **Software** w.r.t Fairness Criteria by assigning **high security status** to a **protected group attribute** and performing **Information-Flow analyses**

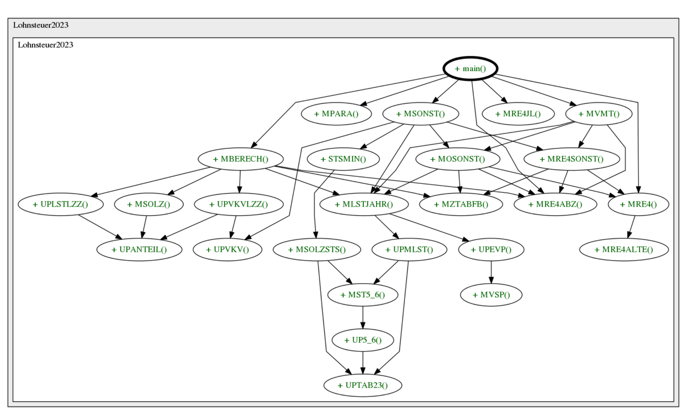
Unconditional Nonintereference $\Rightarrow$ Demographic Parity

If:

- Group $G \in \mathcal{G}$ and Unprotected Attribute $U \in \mathcal{U}$ independent
- Program P satisfies **unconditional noninterference**

Then:

$\Longrightarrow$ Outcome of P satisfies **demographic parity**



Wage Tax Software:

- **1.5k** LOC in **Java**
- 35 Input Variables including **Religion**
- Approx. 2^{153} possible input values!

Analysis using Joana:

Religious Affiliation has **no influence** on wage tax

Restricted Information-Flow

Define **restricted classification** $R : \mathcal{G} \times \mathcal{U} \rightarrow \mathcal{R}$

No Information-Flow within each class $r \in \mathcal{R}$

$\Longrightarrow$ **Auditable characterization** of limitations

$\Longrightarrow$ Conditional Demographic Parity

Quantitative Information-Flow: Conditional Vulnerability [3]

Intuition:
Observation of random $U \in \mathcal{U}$ and outcome $P(G, U)$.

$V(G|P, U) =$ Probability of correctly guessing G

Conditional Vulnerability measures **Fairness Spread**:

Fairness Spread $S(G, U, P)$

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

Handwavy Explanation:

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

Relation to Causal Analysis

Fairness Spread provides an **upper bound** on the probability that a random individual has a **counterfactual with a deviating outcome** for P

$\Longrightarrow$ Information-Flow Analysis is **compatible with Causal Graphs**

[1] Wolfgang Ahrendt et al., eds. *Deductive Software Verification - The KeY Book - From Theory to Practice.* Vol. 10001. LNCS. Cham: Springer, 2016. ISBN: 978-3-319-49811-9. DOI: 10.1007/978-3-319-49812-6.

[2] Jürgen Graf et al. "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.

[3] Geoffrey Smith. "On the Foundations of Quantitative Information Flow". In: *Foundations of Software Science and Computational Structures, 12th International Conference, FOSSACS 2009.* 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.