Causal fault localisation in dataflow systems

Andrei Paleyes and Neil D. Lawrence, University of Cambridge
ap2169@cam.ac.uk
arXiv: 2304.11987 || GitHub: https://github.com/apaleyes/causality-fbp

The system that is designed with **dataflow architecture** provides full dataflow graph natively. This graph can be treated as **complete causal graph**. We demonstrate its utility by using **causal inference for fault localisation** in different dataflow frameworks and applications.

Intro

The paper **"Dataflow graphs as complete causal graphs"** (arxiv:2303.09552, CAIN'23) suggested that dataflow architecture s benefits of dataflow graphs that can be seen as complete causal graphs of an entire system. We aim to provide the first realistic demonstration of the idea, with a focus on fault localisation.

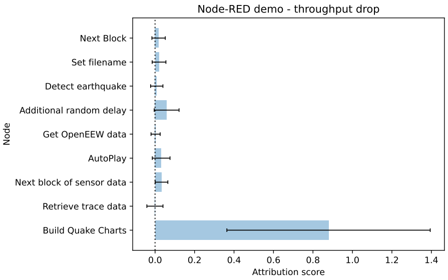
Methods

We build a series of demonstration projects with various modern dataflow frameworks, and conduct experiments on these demos. In each experiment we intervene on of the nodes inflicting a shift in the overall system's output distribution. We then use causal inference to localise the faulty node.

- 3 demo projects
- 3 dataflow frameworks
- 2 types of interventions: code bug, input data shift
- 5 experiments in total

Results

- Correct node identified in all experiments
- T-test gives $p < 0.01$ for identifying the correct node
- Each node assigned attribution score - native UQ (example below)



Future work

Despite convincing performance in all experiments, this idea is far from being industry-ready yet. Some further steps are:

- Scale up graph size
- Estimate data storage cost
- Convert demos to fully automated plugins in corresponding frameworks

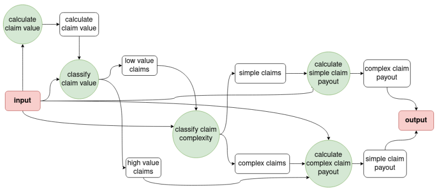
Dataflow frameworks we use

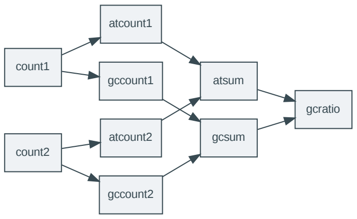
SciPipe

SELDON CORE

Node-RED

Dataflow graph examples





← Paper

SCAN ME

Code →

SCAN ME

UNIVERSITY OF CAMBRIDGE