Noname manuscript No.
(will be inserted by the editor)

Shaky Structures: The Wobbly World of Causal Graphs in Software Analytics

Jeremy Hulse · Nasir Eisty · Tim Menzies

Received: date / Accepted: date

Abstract Causal graphs are a type of data structure useful for observing causal relationships (and drawing conclusions and observations about the causes of events or attributes). These are widely used in the software engineering literature.

But while they are widely used, they may also be wildly misleading. Causal structures generated from SE data can be highly variable. This instability is so large that conclusions drawn from one graph may be totally reversed in another, even when both graphs are learned from the same or very similar software project.

This paper tests for causal graph instability in graphs learned from standard software defect data sets (Ivy, Ant, Velocity, Camel, Xerces). Measured in terms of the Jaccard index of the number of edges shared by two different graphs, many of our treatments resulted in Jaccard counts less than 50% and often less than 20%.

These results reveal a major threat to the validity of causal research in SE. Causal graph conclusions may not generalize since those conclusions could be reversed by minor changes to the tuning or training data. Hence, before researchers can report supposedly general conclusions from causal graphs (e.g., "long functions cause more defects"), they should test that such conclusions hold over the numerous causal graphs that might be generated from the same data.

To allow for the repetition and/or refinement and/or refutation of our results, all our scripts and data are online at `https://github.com/jphulse/Stability_Of_Causal_Graphs_Public`

Keywords Empirical SE, Data Mining, Analytics

J.Hulse, T. Menzies
Computer Science, NC State, USA
E-mail: jphulse@ncsu.edu , E-mail: timm@ieeee.org

N. Eisty
Boise State University, USA
E-mail: nasireisty@boisestate.edu