Extended Case Study of Causal Learning within Architecture Research (preliminary results)

Robert Stoddard, SEI
Mike Konrad, SEI
Rick Kazman, SEI
David Danks, CMU
Goal of the Authors

Dr. Mike Konrad
SEI

Robert Stoddard
SEI

Dr. David Danks
CMU

Dr. Rick Kazman
SEI

Goal is to motivate adoption of Causal Learning within software measurement!
Why Do We Care about Causation?

http://www.tylervigen.com/spurious-correlations
More about Misinterpreting Correlation!

Often, an excluded common cause results in a misinterpretation of correlation!

Does high correlation imply causation?
Regression must be interpreted in context of a DAG!

Correlation, hence regression, may be fooled by spurious association!

Before jumping into regression, we need a Directed Acyclic Graph (DAG) representing our context

We then need to determine which paths are causal and which are spurious.

We then must block spurious correlation paths.

Lastly, we then conduct regression with the correct set of factors!

Remember, context of the DAG determines the suitability of the regression model!
The Causal Learning Landscape

Causal Discovery using CMU Tetrad which implements a variety of algorithms

Causal Directed Acyclic Graph Model

Prior Knowledge & Observational Data

Formulate Hypotheses using domain knowledge and prior scholarly publication

Estimated SEM Model

© 2018 Carnegie Mellon University

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.
Preliminary Architecture Research Causal Findings

Nine open source systems analyzed using static code analysis (> 9000 files)

Four architecture pattern violations studied for impact on quality

Each file had the following attributes measured:

- **Age** in Months
- **Number of Developers** touching each file
- **Size** in Lines of Code
- Number of times the file participated in a **pattern violation** of:
  - the cyclic dependency
  - Improper inheritance
  - Unstable interface
  - Lack of modularity
- **Quality** outcome of Number of Bugs associated with each file
- **Bug churn** associated with each file

Correlation Matrix of All Factors

<table>
<thead>
<tr>
<th></th>
<th>AgeMonths</th>
<th>NumDev</th>
<th>NumCommits</th>
<th>LOC</th>
<th>NumBugs</th>
<th>NumChanges</th>
<th>BugChurn</th>
<th>ChangeChurn</th>
<th>NumCyclicDepend</th>
<th>NumModularityVio</th>
<th>NumUnstableInter</th>
</tr>
</thead>
<tbody>
<tr>
<td>NumDev</td>
<td>0.1790</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>NumCommits</td>
<td>0.0930</td>
<td>0.6890</td>
<td>0.0000</td>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>LOC</td>
<td>0.0460</td>
<td>0.2640</td>
<td>0.2720</td>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>NumBugs</td>
<td>0.1160</td>
<td>0.6540</td>
<td>0.9330</td>
<td>0.2570</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>NumChanges</td>
<td>0.0960</td>
<td>0.6880</td>
<td>0.9990</td>
<td>0.2720</td>
<td>0.9340</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>BugChurn</td>
<td>0.0380</td>
<td>0.3920</td>
<td>0.5810</td>
<td>0.7270</td>
<td>0.6390</td>
<td>0.5820</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>ChangeChurn</td>
<td>0.0180</td>
<td>0.2980</td>
<td>0.4180</td>
<td>0.9400</td>
<td>0.4120</td>
<td>0.4180</td>
<td>0.8300</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>NumCyclicDepend</td>
<td>0.0340</td>
<td>0.1520</td>
<td>0.2920</td>
<td>0.1000</td>
<td>0.2430</td>
<td>0.2900</td>
<td>0.1240</td>
<td>0.1080</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>NumModularityVio</td>
<td>0.0490</td>
<td>0.3270</td>
<td>0.2100</td>
<td>0.1070</td>
<td>0.1590</td>
<td>0.2100</td>
<td>0.0980</td>
<td>0.1000</td>
<td>0.0130</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>NumUnstableInter</td>
<td>0.0390</td>
<td>0.5400</td>
<td>0.4820</td>
<td>0.1580</td>
<td>0.3940</td>
<td>0.4810</td>
<td>0.2220</td>
<td>0.2000</td>
<td>0.1420</td>
<td>0.2670</td>
<td>0.0000</td>
</tr>
<tr>
<td>NumImproperInher</td>
<td>0.1280</td>
<td>0.2060</td>
<td>0.2110</td>
<td>0.1040</td>
<td>0.1850</td>
<td>0.2120</td>
<td>0.1150</td>
<td>0.0740</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

NumBugs, NumChanges, and NumCommits are highly correlated; Will keep NumBugs only in the modeling; Likewise, ChangeChurn and LOC highly correlated, so kept only LOC in the modeling.
All Remaining Factors are Non-Normal - 01
All Remaining Factors are Non-Normal - 02
Eyeballing Bivariate Relationships
Best Subsets Regression

Response is NumBugs

<table>
<thead>
<tr>
<th>Vars</th>
<th>R-Sq (adj)</th>
<th>R-Sq (pred)</th>
<th>Mallows Cp</th>
<th>S</th>
<th>Age Mth</th>
<th>NumDev</th>
<th>LOC</th>
<th>BugChurn</th>
<th>Cyclic Depend</th>
<th>Modularrit</th>
<th>Interfacit</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42.8</td>
<td>42.8</td>
<td>42.1</td>
<td>8301.1</td>
<td>10.921</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>40.9</td>
<td>40.9</td>
<td>27.7</td>
<td>8891.2</td>
<td>11.106</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>60.1</td>
<td>60.1</td>
<td>50.8</td>
<td>3051.3</td>
<td>9.1201</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>50.0</td>
<td>50.0</td>
<td>27.0</td>
<td>6132.0</td>
<td>10.216</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>68.4</td>
<td>68.4</td>
<td>63.0</td>
<td>544.7</td>
<td>8.1199</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>61.5</td>
<td>61.5</td>
<td>52.1</td>
<td>2647.3</td>
<td>8.9662</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>69.9</td>
<td>69.9</td>
<td>64.7</td>
<td>101.9</td>
<td>7.9297</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>68.6</td>
<td>68.6</td>
<td>63.2</td>
<td>480.3</td>
<td>8.0921</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>70.0</td>
<td>70.0</td>
<td>64.8</td>
<td>59.3</td>
<td>7.9108</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>70.0</td>
<td>69.9</td>
<td>64.8</td>
<td>77.0</td>
<td>7.9184</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>70.1</td>
<td>70.1</td>
<td>64.9</td>
<td>35.5</td>
<td>7.9000</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>70.1</td>
<td>70.1</td>
<td>64.9</td>
<td>42.5</td>
<td>7.9030</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>70.2</td>
<td>70.1</td>
<td>65.0</td>
<td>21.6</td>
<td>7.8935</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>70.2</td>
<td>70.1</td>
<td>65.0</td>
<td>23.0</td>
<td>7.8941</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>70.2</td>
<td>70.2</td>
<td>65.0</td>
<td>9.0</td>
<td>7.8875</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
Conduct Causal Search using Tetrad
A View of the Data File Loaded into Tetrad

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
<th>C8</th>
<th>C9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>71.0000</td>
<td>8.0000</td>
<td>431.0000</td>
<td>18.0000</td>
<td>241.0000</td>
<td>8.0000</td>
<td>2.0000</td>
<td>3.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>2</td>
<td>35.0000</td>
<td>5.0000</td>
<td>270.0000</td>
<td>10.0000</td>
<td>329.0000</td>
<td>167.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>4.0000</td>
</tr>
<tr>
<td>3</td>
<td>52.0000</td>
<td>2.0000</td>
<td>58.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>42.0000</td>
<td>1.0000</td>
<td>47.0000</td>
<td>2.0000</td>
<td>13.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>5</td>
<td>49.0000</td>
<td>1.0000</td>
<td>10.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>6</td>
<td>36.0000</td>
<td>2.0000</td>
<td>103.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>7</td>
<td>54.0000</td>
<td>2.0000</td>
<td>29.0000</td>
<td>2.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>8</td>
<td>75.0000</td>
<td>8.0000</td>
<td>133.0000</td>
<td>13.0000</td>
<td>134.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>3.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>9</td>
<td>74.0000</td>
<td>2.0000</td>
<td>15.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>10</td>
<td>57.0000</td>
<td>2.0000</td>
<td>26.0000</td>
<td>1.0000</td>
<td>16.0000</td>
<td>22.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>11</td>
<td>48.0000</td>
<td>4.0000</td>
<td>81.0000</td>
<td>2.0000</td>
<td>6.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>12</td>
<td>39.0000</td>
<td>1.0000</td>
<td>30.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>13</td>
<td>49.0000</td>
<td>2.0000</td>
<td>46.0000</td>
<td>3.0000</td>
<td>36.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>14</td>
<td>46.0000</td>
<td>3.0000</td>
<td>34.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>15</td>
<td>75.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Prior Knowledge Entered into Tetrad
Using FASK Search with Associated Parameters
Additional FASK Search Parameter Settings

FASK Parameters

- Penalty discount (min = 0.0)
- Maximum size of conditioning set (unlimited = -1)
- Alpha orienting 2-cycles (min = 0.0)
- Threshold for including extra edges
- Threshold for judging negative coefficient edges as X->Y (range (-1, 0))
- Yes if adjacencies from the FAS search should be used
- Yes if adjacencies from conditional correlation differences should be used
- The number of bootstraps (min = 0)
- Ensemble method: Preserved (0), Highest (1), Majority (2)
- Yes if verbose output should be printed or logged

Choose Algorithm  Run Search & Generate Graph  Done
Causal Search Algorithms

**Constraint-based:** Calculate independences in the data and do “backwards inference”; used to minimize the degree of false negative edges

**Score-based (Bayesian):** Calculate the likelihood of different DAGs given the data; used to minimize the degree of false positive edges

**Hybrid:** Use constraint-based to get “close,” then Bayesian search around neighborhood

- A → B  No evidence of a causal link
- A → B  Evidence of a causal link from A to B
- A ← B  Evidence of a causal link from B to A
- A ↔ B  Evidence of an unmeasured confounder
Some Algorithms Exploit Non-Gaussianity

Linear Gaussian

Linear non-Gaussian

\[
\begin{align*}
X & \rightarrow Y \\
X & \leftarrow Y
\end{align*}
\]
Causal Search Capable with Small Data

**Challenge:** Which genes regulate flowering time in Arabidopsis thaliana?

Using only 47 observations, causal search identified 9 out of 21,326 genes as causal on gene activation.

Subsequent greenhouse study, that used knockout variants, confirmed that 4 of the 9 were actual regulators.

Taken from Dave Danks, 2016 Summer Causal Workshop
Causal Structure Graph Result
Markov Blanket of the NumBugs Factor

- LOC
- NumDev
- NumCyclicDepend
- BugChurn
- NumBugs
Motivation to Look at Multi-Level SEM Models (MSEM)

Within schools, students with better Spanish skills had higher academic achievement. Yet, schools with highest proportion of Spanish speakers performed poorest.

Also called Simpson’s Paradox and the Ecological Fallacy
Mplus Code

```
TITLE: Basic Model of NumBugs Markov Blanket;

DATA: FILE IS All9forMplus.csv;

VARIABLE: NAMES ARE AgeMos NumDev LOC Cycles Inherit Interfac Modular BugChurn NumBugs System;

USEVARIABLES ARE NumDev LOC Cycles BugChurn NumBugs System;

CLUSTER IS System;

ANALYSIS: TYPE IS TWOLEVEL;

MODEL:

%BETWEEN%
NumBugs ON BugChurn LOC NumDev Cycles;
NumBugs; BugChurn; LOC; NumDev; Cycles;
[NumBugs]; [BugChurn]; [LOC]; [NumDev]; [Cycles];

%WITHIN%
NumBugs ON BugChurn LOC NumDev Cycles;

OUTPUT: SAMPSTAT STDXY;
```
### SUMMARY OF DATA

**Number of clusters**: 9  
**Average cluster size**: 1005.556

**Estimated Intraclass Correlations for the Y Variables**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Intraclass Correlation</th>
<th>Variable</th>
<th>Intraclass Correlation</th>
<th>Variable</th>
<th>Intraclass Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMBUGS</td>
<td>0.052</td>
<td>NUMDEV</td>
<td>0.084</td>
<td>LOC</td>
<td>0.008</td>
</tr>
<tr>
<td>CYCLES</td>
<td>0.039</td>
<td>BUGCHURN</td>
<td>0.026</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Traditional SEM Results from Tetrad

File Parameters Layout

Graphical Editor Tabular Editor

Degrees of Freedom = 4
Chi Square = 2358.0099
P Value = 0.0000E0
BIC Score = 2321.5678
CFI = 0.9907
RMSEA = 0.2550
Conclusions

1. We attempted MSEM modeling to be sensitive to the “between” and “within” variation components of all the factors

2. We also wanted to guard against Simpson’s paradox

3. The Mplus MSEM analysis, via the Intraclass Correlation measures, showed that in this data situation, we do not need to perform MSEM with two levels

4. We then conducted a single level, univariate SEM within Tetrad

5. We achieved regression coefficients that take into account the mediation effects occurring on the outcome, NumBugs

6. Traditional regression would have been ignorant of the above
Next Steps

Perform more causal searches
- Additional algorithms
- Sensitivity analysis of algorithm parameters
- Using bootstrapping to get confidence intervals on causal edges

Perform additional multilevel structural equation models:
- Investigate more factors associated with attributes of the open source system
- Evaluate whether a latent factor representing the “voice” of any architecture pattern might be helpful

Publish results:
- Comparison of different models
- Distinguish the causal influence of factors at both the file level and within a system

Convince others in the community to adopt Causal Learning and MSEM
Questions?

Robert Stoddard, SEI
rws@sei.cmu.edu

Mike Konrad, SEI
mdk@sei.cmu.edu

Rick Kazman, SEI
rkazman@sei.cmu.edu

David Danks, CMU
ddanks@cmu.edu