



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

Robert Stoddard, SEI
Mike Konrad, SEI
Rick Kazman, SEI
David Danks, CMU

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Document Markings

Copyright 2018 Carnegie Mellon University. All Rights Reserved.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM18-1046

Goal of the Authors



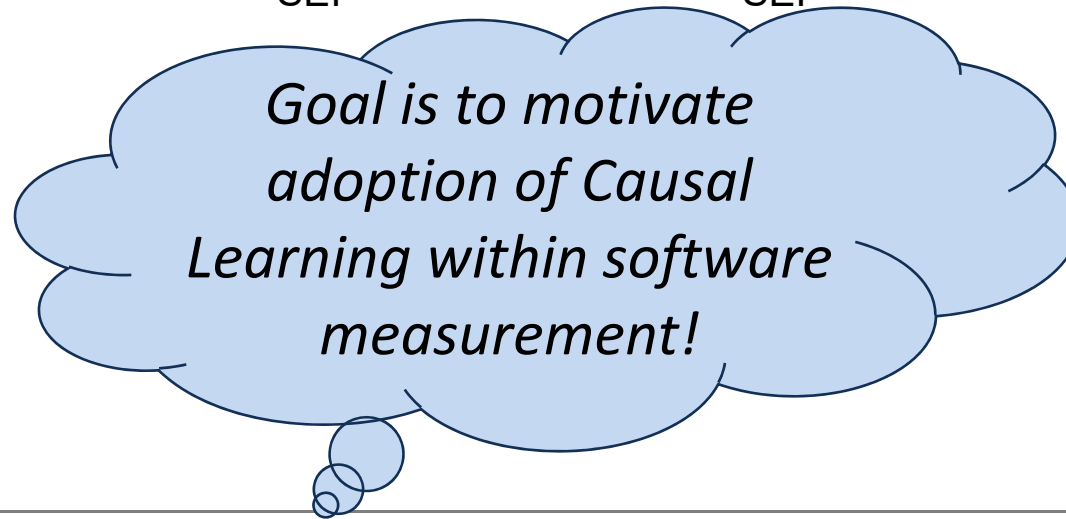
Robert Stoddard
SEI



Dr. Mike Konrad
SEI

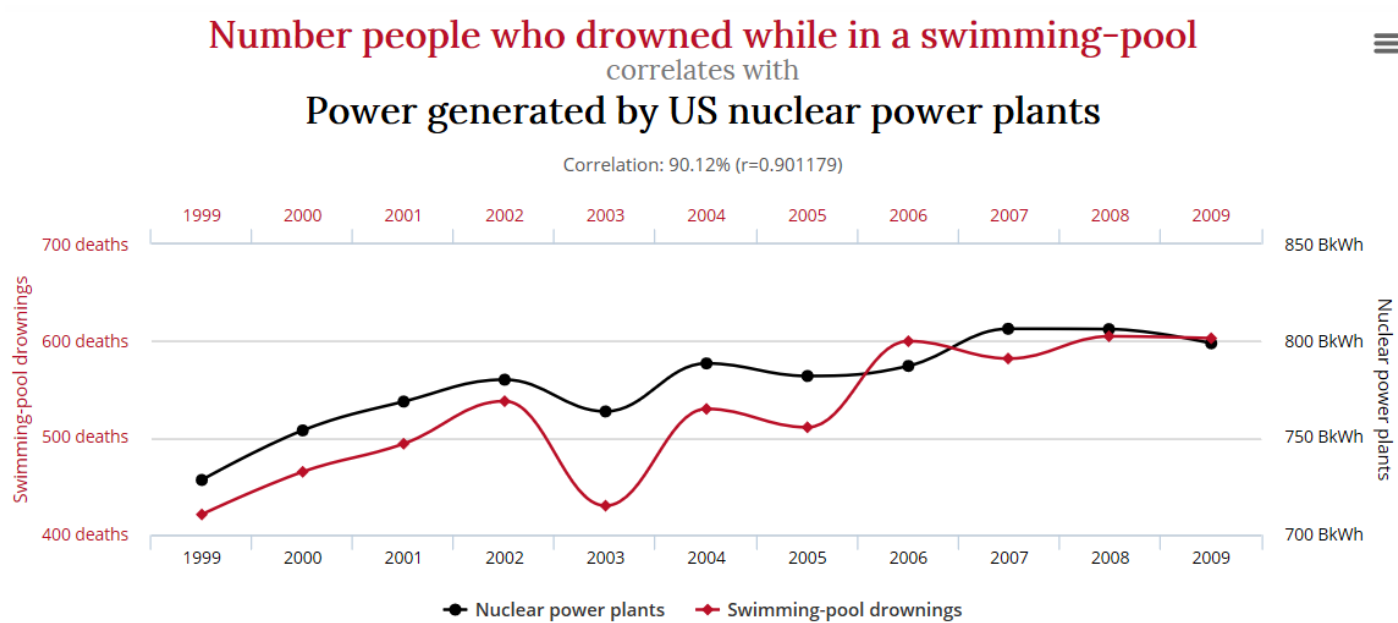


Dr. David Danks
CMU



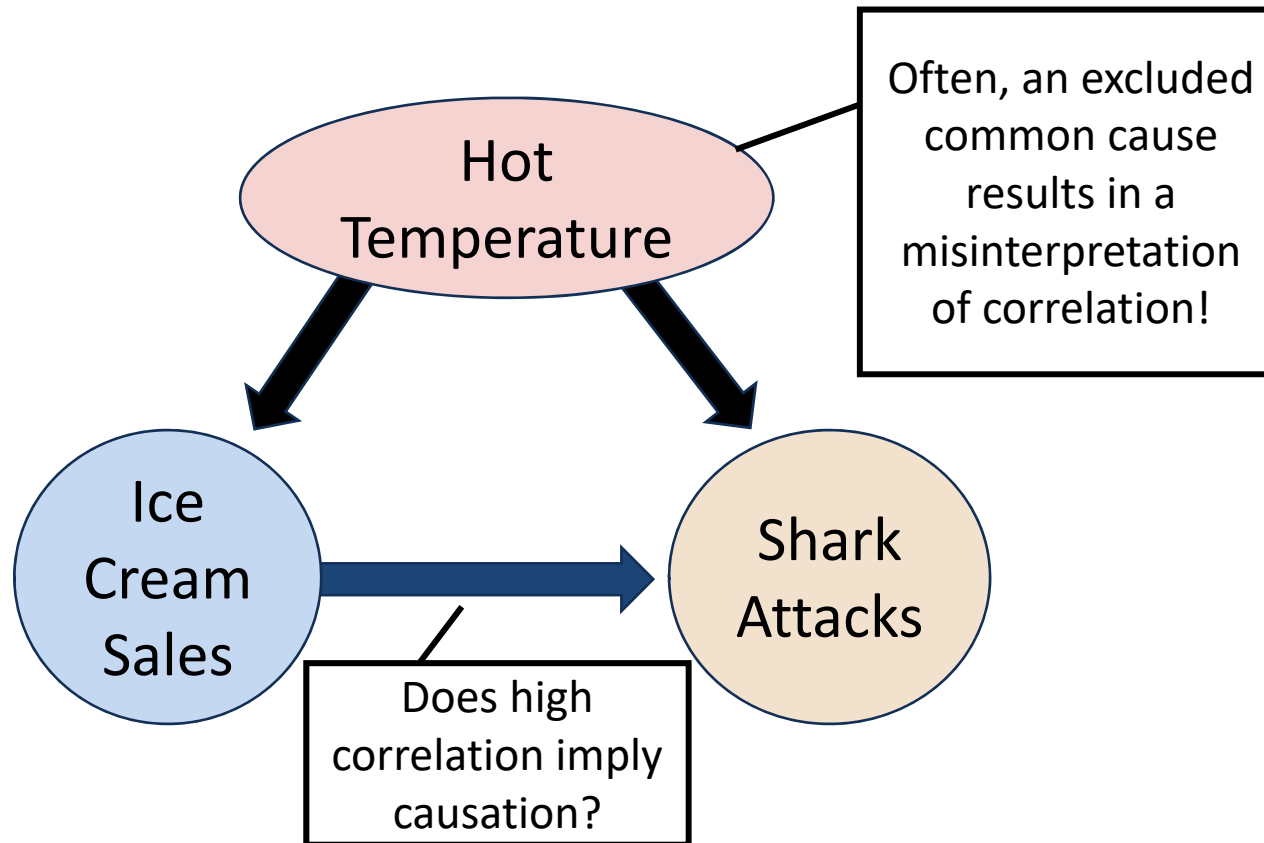
Dr. Rick Kazman
SEI

Why Do We Care about Causation?



<http://www.tylervigen.com/spurious-correlations>

More about Misinterpreting Correlation!



Regression Cannot be Trusted without 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

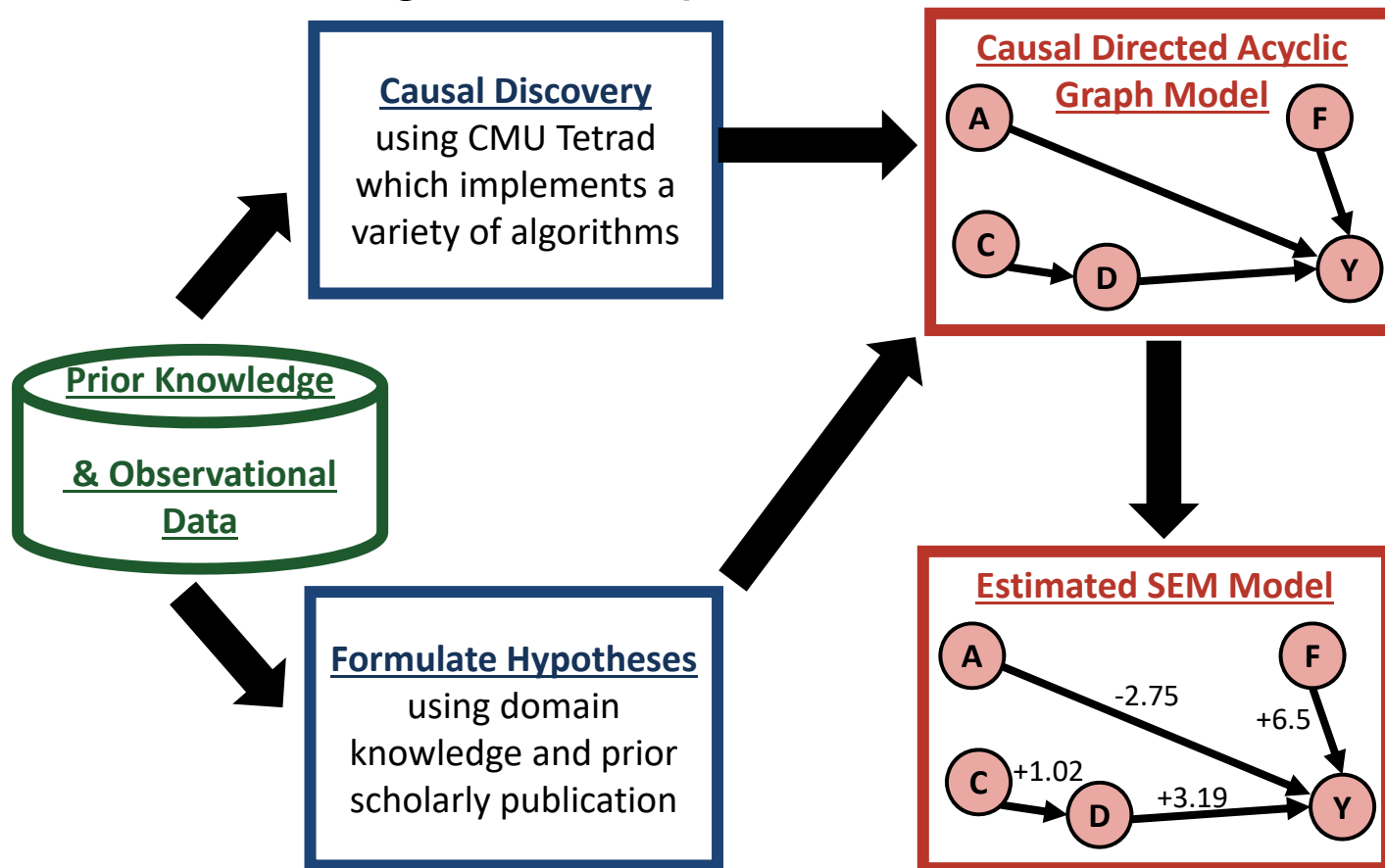
We then need to determine which paths are causal and which are spurious.

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



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

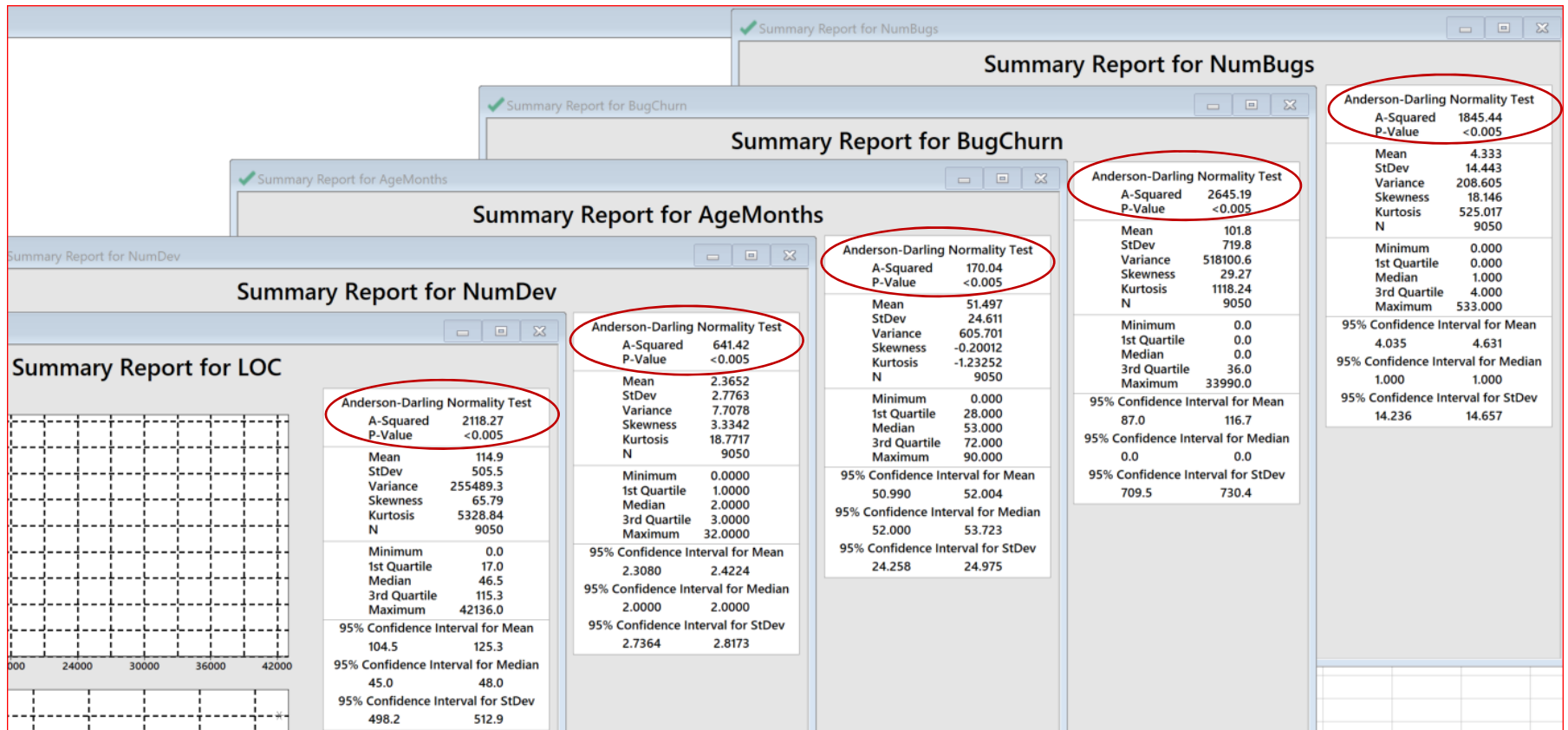
R. Mo, Y. Cai, R. Kazman and L. Xiao, "Hotspot Patterns: The Formal Definition and Automatic Detection of Architecture Smells," *2015 12th Working IEEE/IFIP Conference on Software Architecture*, Montreal, QC, 2015, pp. 51-60. doi: 10.1109/WICSA.2015.12

Correlation Matrix of All Factors

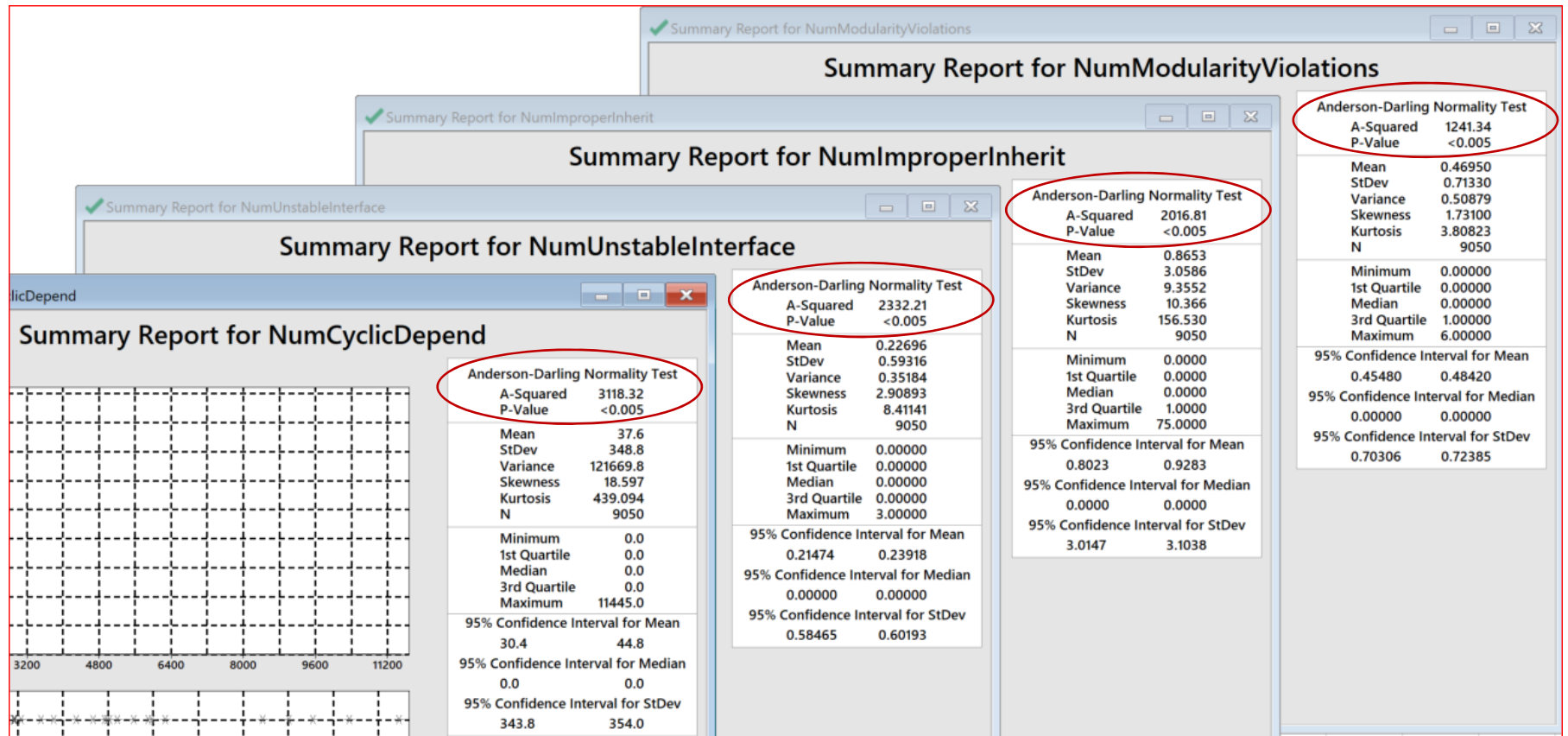
	AgeMonths	NumDev	NumCommits	LOC	NumBugs	NumChanges	BugChurn	ChangeChurn	NumCyclicDepend	NumModularityVio	NumUnstableInter
NumDev	0.1790										
	0.0000										
NumCommits	0.0930	0.6890									
	0.0000	0.0000									
LOC	0.0460	0.2640	0.2720								
	0.0000	0.0000	0.0000								
NumBugs	0.1160	0.6540	0.9330	0.2570							
	0.0000	0.0000	0.0000	0.0000							
NumChanges	0.0960	0.6880	0.9990	0.2720	0.9340						
	0.0000	0.0000	0.0000	0.0000	0.0000						
BugChurn	0.0380	0.3920	0.5810	0.7270	0.6390	0.5820					
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000					
ChangeChurn	0.0180	0.2980	0.4180	0.9400	0.4120	0.4180	0.8300				
	0.0880	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000				
NumCyclicDepend	0.0340	0.1520	0.2920	0.1000	0.2430	0.2900	0.1240	0.1080			
	0.0010	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000			
NumModularityVio	0.0490	0.3270	0.2100	0.1070	0.1590	0.2100	0.0980	0.1000	0.0130		
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2140		
NumUnstableInter	0.0390	0.5400	0.4820	0.1580	0.3940	0.4810	0.2220	0.2000	0.1420	0.2670	
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
NumImproperInher	0.1280	0.2060	0.2110	0.1040	0.1850	0.2120	0.1150	0.0740	0.1620	0.0020	0.1330
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8540	0.0000

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

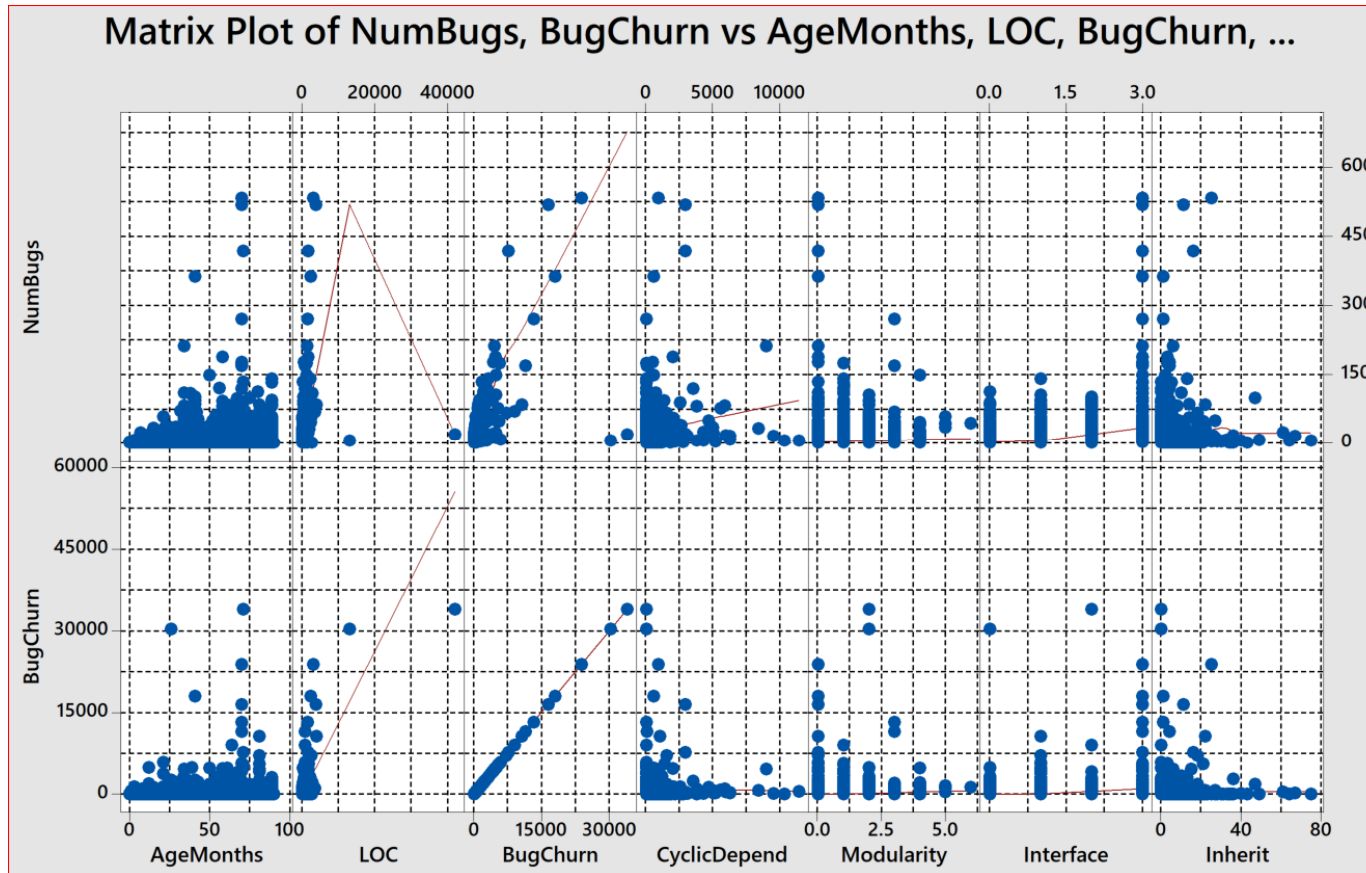
All Remaining Factors are Non-Normal - 01



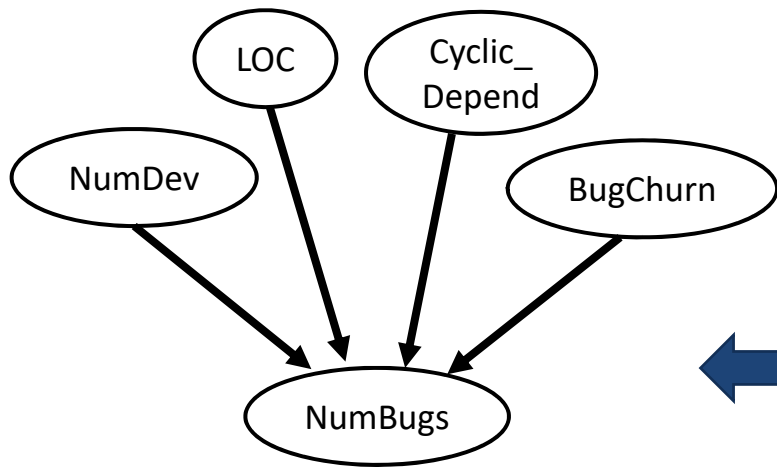
All Remaining Factors are Non-Normal - 02



Eyeballing Bivariate Relationships



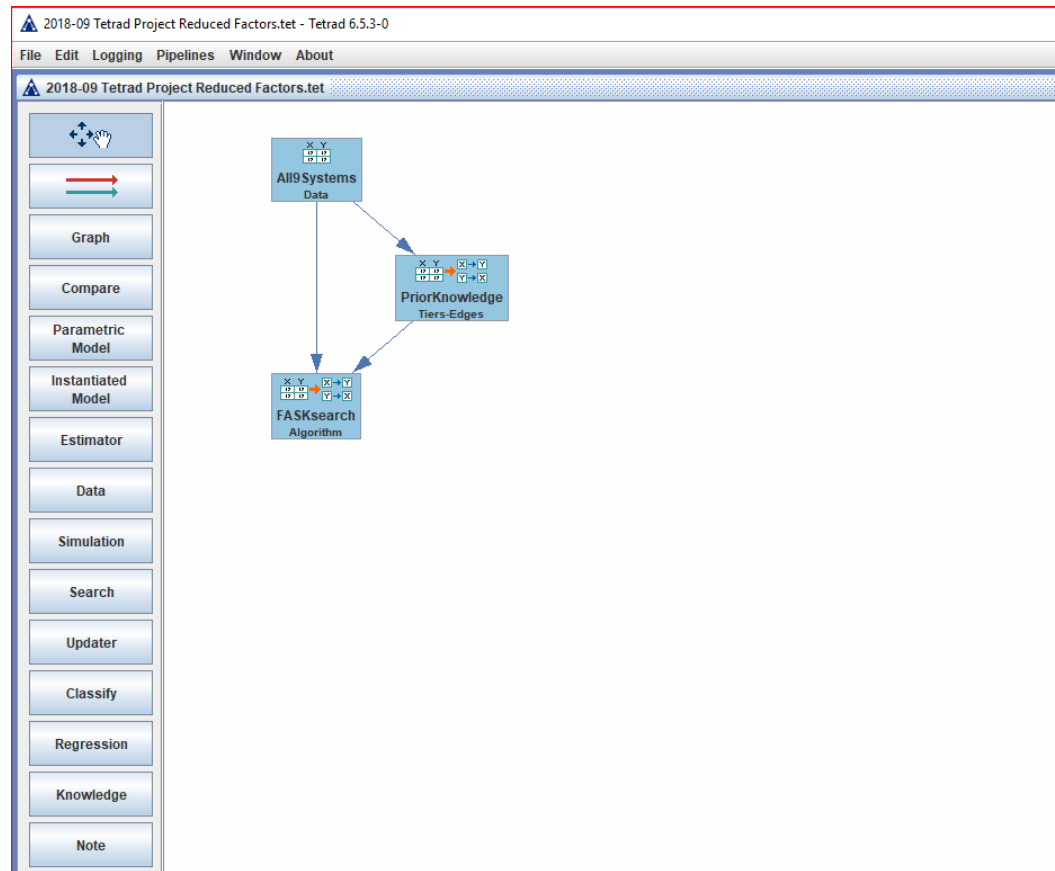
Best Subsets Regression



Response is NumBugs

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	Age Months	NumDev	LOC	BugChurn	Cyclic_Depend	Modularity	Interface	Inheritance
1	42.8	42.8	42.1	8301.1	10.921		X						
1	40.9	40.9	27.7	8891.2	11.106				X				
2	60.1	60.1	50.8	3051.3	9.1201		X		X				
2	50.0	50.0	27.0	6132.0	10.216			X	X				
3	68.4	68.4	63.0	544.7	8.1199		X	X	X				
3	61.5	61.5	52.1	2647.3	8.9662		X		X	X			
4	69.9	69.9	64.7	101.9	7.9297		X	X	X	X			
4	68.6	68.6	63.2	480.3	8.0921		X	X	X				X
5	70.0	70.0	64.8	59.3	7.9108		X	X	X	X			X
5	70.0	69.9	64.8	77.0	7.9184		X	X	X	X			X
6	70.1	70.1	64.9	35.5	7.9000		X	X	X	X			X
6	70.1	70.1	64.9	42.5	7.9030		X	X	X	X	X		X
7	70.2	70.1	65.0	21.6	7.8935		X	X	X	X	X	X	X
7	70.2	70.1	65.0	23.0	7.8941	X	X	X	X	X			X
8	70.2	70.2	65.0	9.0	7.8875	X	X	X	X	X	X	X	X

Conduct Causal Search using Tetrad



A View of the Data File Loaded into Tetrad

The screenshot shows a window titled "All9 Systems (Data)" with a menu bar containing "File", "Edit", and "Tools". Below the menu bar is a tab labeled "All 9 for Tetrad-v010.csv". The main area contains a data table with 15 rows and 9 columns. The columns are labeled C1 through C9, and the rows are numbered 1 through 15. The data values are as follows:

	C1	C2	C3	C4	C5	C6	C7	C8	C9
	AgeMonths	NumDev	LOC	NumBugs	BugChurn	NumCyclic...	NumModul...	NumUnsta...	NumImpro...
1	71.0000	8.0000	491.0000	18.0000	241.0000	8.0000	2.0000	3.0000	1.0000
2	35.0000	5.0000	270.0000	10.0000	329.0000	167.0000	1.0000	1.0000	4.0000
3	52.0000	2.0000	58.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	42.0000	1.0000	47.0000	2.0000	13.0000	0.0000	0.0000	0.0000	0.0000
5	49.0000	1.0000	10.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
6	36.0000	2.0000	103.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
7	54.0000	2.0000	29.0000	2.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	75.0000	8.0000	163.0000	13.0000	134.0000	0.0000	1.0000	3.0000	0.0000
9	74.0000	2.0000	15.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
10	57.0000	2.0000	26.0000	1.0000	16.0000	22.0000	0.0000	0.0000	0.0000
11	48.0000	4.0000	81.0000	2.0000	6.0000	0.0000	1.0000	0.0000	0.0000
12	39.0000	1.0000	30.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
13	49.0000	2.0000	46.0000	3.0000	36.0000	0.0000	0.0000	0.0000	0.0000
14	46.0000	3.0000	34.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
15	75.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

A "Done" button is located at the bottom center of the window.

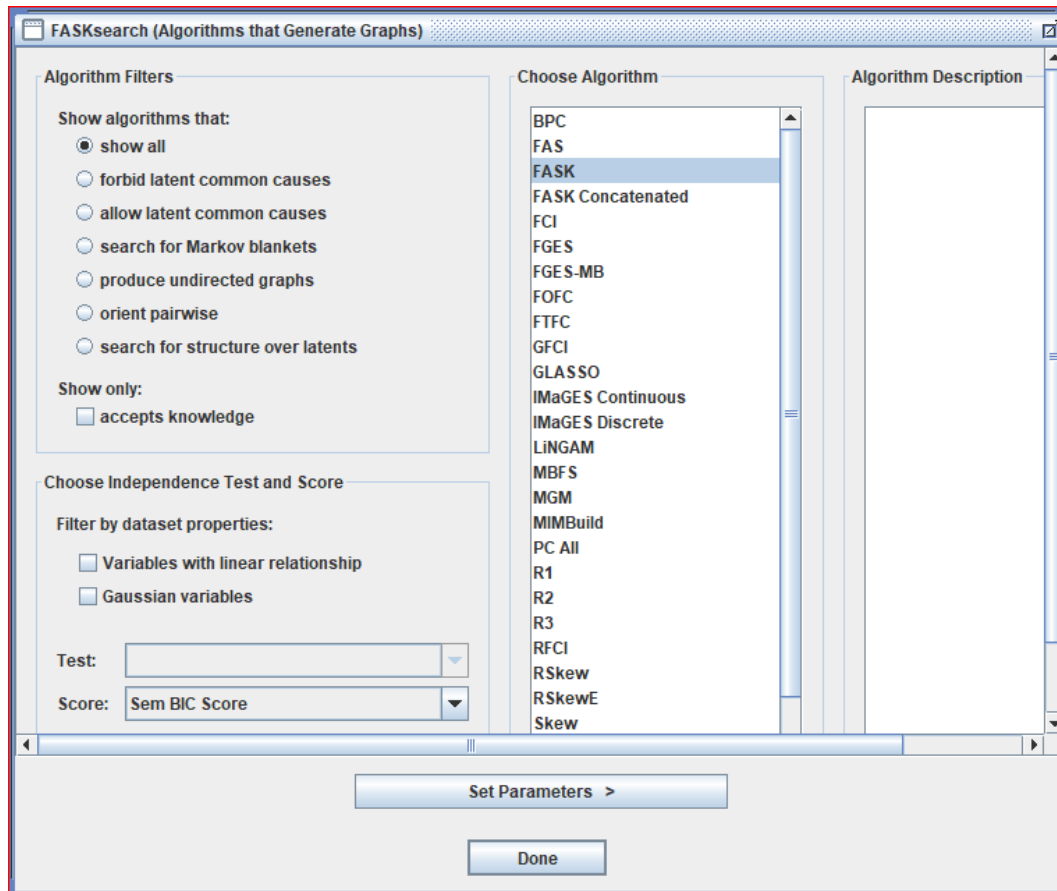
Prior Knowledge Entered into Tetrad

The screenshot shows the 'PriorKnowledge1 (Tiers and Edges)' window in the Tetrad software. The window has a 'File' menu and three tabs: 'Tiers', 'Other Groups', and 'Edges'. The 'Tiers' tab is active. At the top right, there is a label '# Tiers =' followed by a text box containing the number '3' and a small spinner control. Below this is a section labeled 'Not in tier:' with an empty text box. The main area is divided into three tiers:

- Tier 1:** Labeled 'Tier 1' on the left and has a checked checkbox for 'Forbid Within Tier'. It contains three buttons: 'AgeMonths', 'LOC', and 'NumDev'.
- Tier 2:** Labeled 'Tier 2' on the left and has a checked checkbox for 'Forbid Within Tier'. It contains three buttons: 'NumCyclicDepend', 'NumImproperInherit', and 'NumModularityViolations'. Below these is another button: 'NumUnstableInterface'.
- Tier 3:** Labeled 'Tier 3' on the left and has an unchecked checkbox for 'Forbid Within Tier'. It contains two buttons: 'BugChurn' and 'NumBugs'.

At the bottom of the window, there is a blue bar with the text 'Use shift key to select multiple items.' and a 'Done' button.

Using FASK Search with Associated Parameters



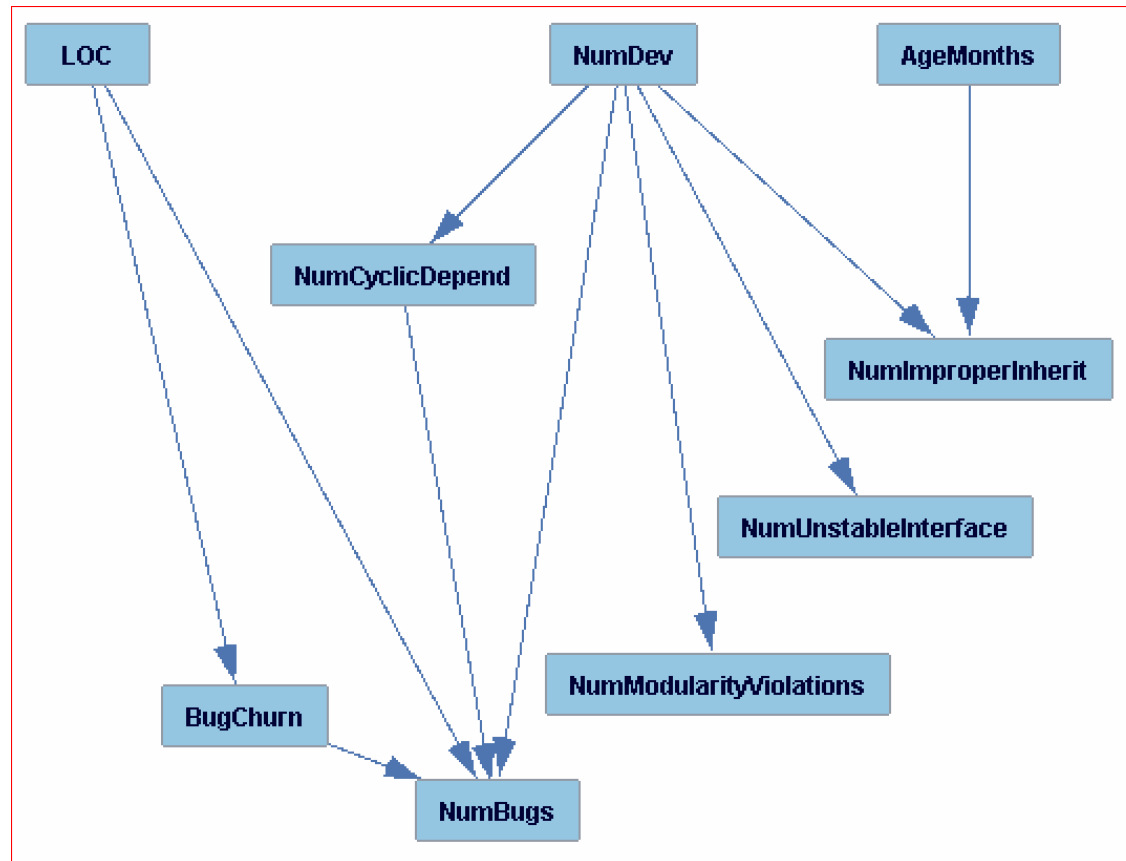
Additional FASK Search Parameter Settings

The screenshot shows a dialog box titled "FASKsearch (Algorithms that Generate Graphs)". It contains several parameters for the search algorithm, each with a corresponding input field or radio button. The parameters are:

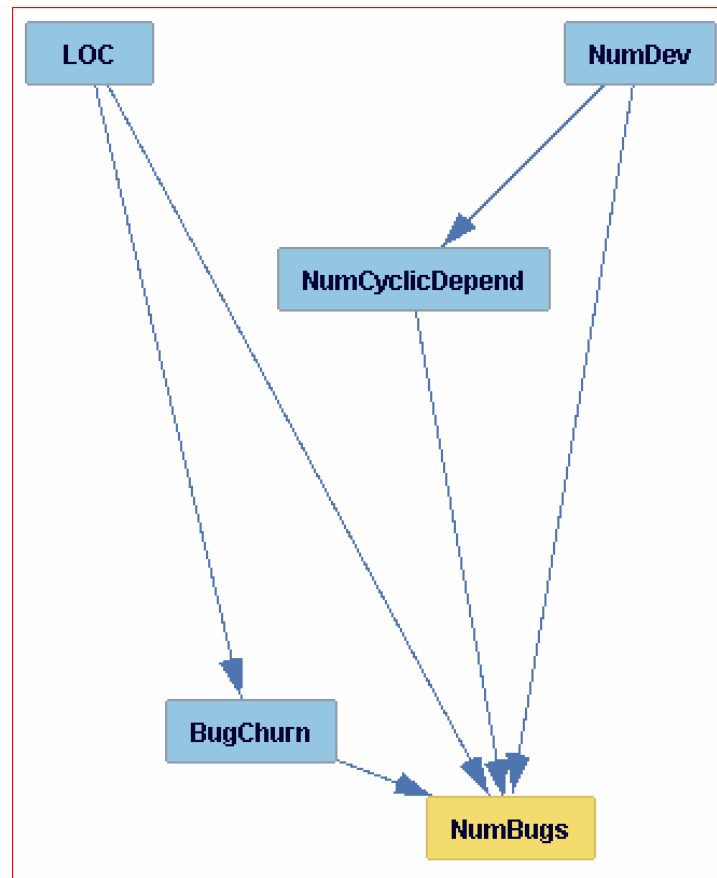
- Penalty discount (min = 0.0): Input field with value 2.
- Maximum size of conditioning set (unlimited = -1): Input field with value -1.
- Alpha orienting 2-cycles (min = 0.0): Input field with value 1.0E-6.
- Threshold for including extra edges: Input field with value 0.3.
- Threshold for judging negative coefficient edges as X->Y (range (-1, 0)): Input field with value -0.2.
- Yes if adjacencies from the FAS search should be used: Radio buttons for Yes (selected) and No.
- Yes if adjacencies from conditional correlation differences should be used: Radio buttons for Yes (selected) and No.
- The number of bootstraps (min = 0): Input field with value 0.
- Ensemble method: Preserved (0), Highest (1), Majority (2): Input field with value 1.
- Yes if verbose output should be printed or logged: Radio buttons for Yes and No (selected).

At the bottom of the dialog box, there are three buttons: "< Choose Algorithm", "Run Search & Generate Graph >", and "Done".

Causal Structure Graph Result



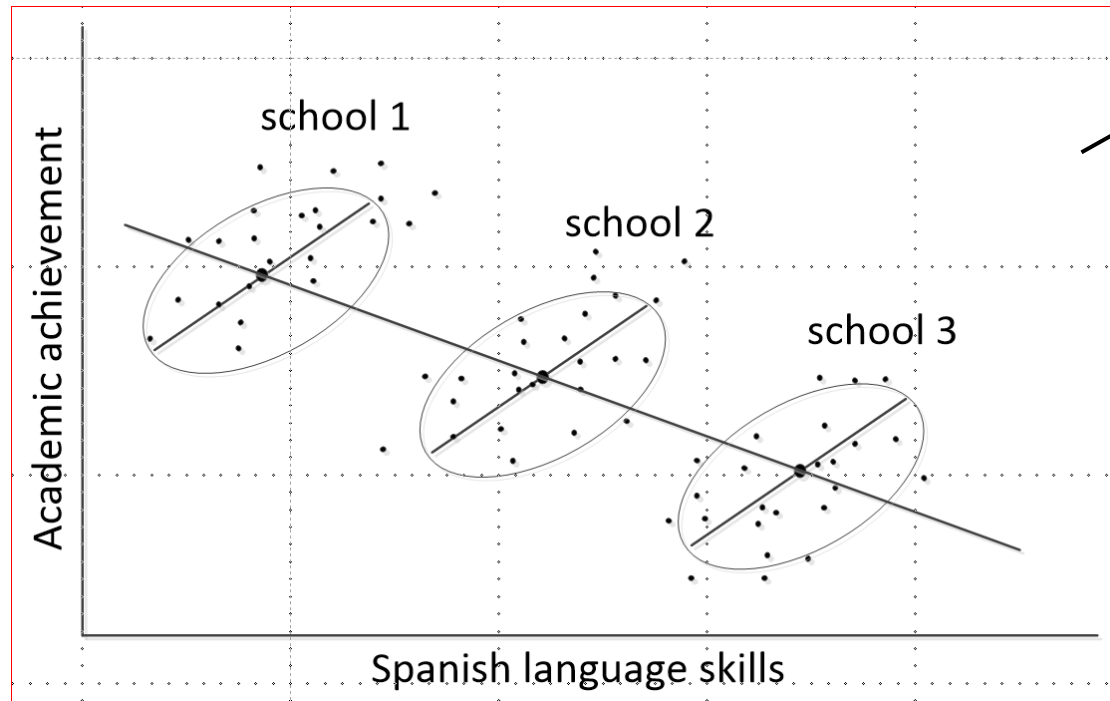
Markov Blanket of the NumBugs Factor



Motivation to Look at Multi-Level SEM Models (MSEM)

Within schools, students with better Spanish skills had higher academic achievement.

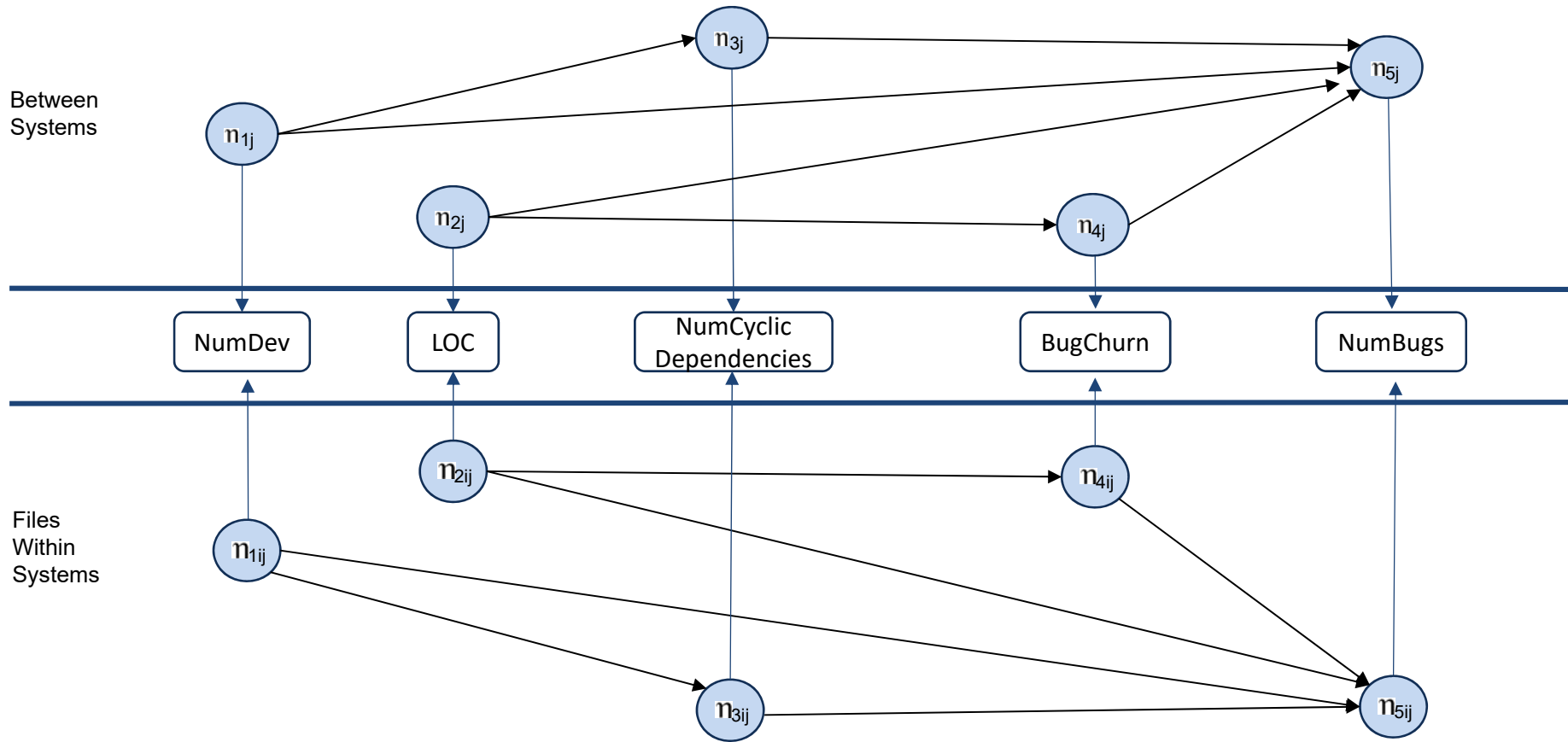
Yet, schools with highest proportion of Spanish speakers performed poorest.



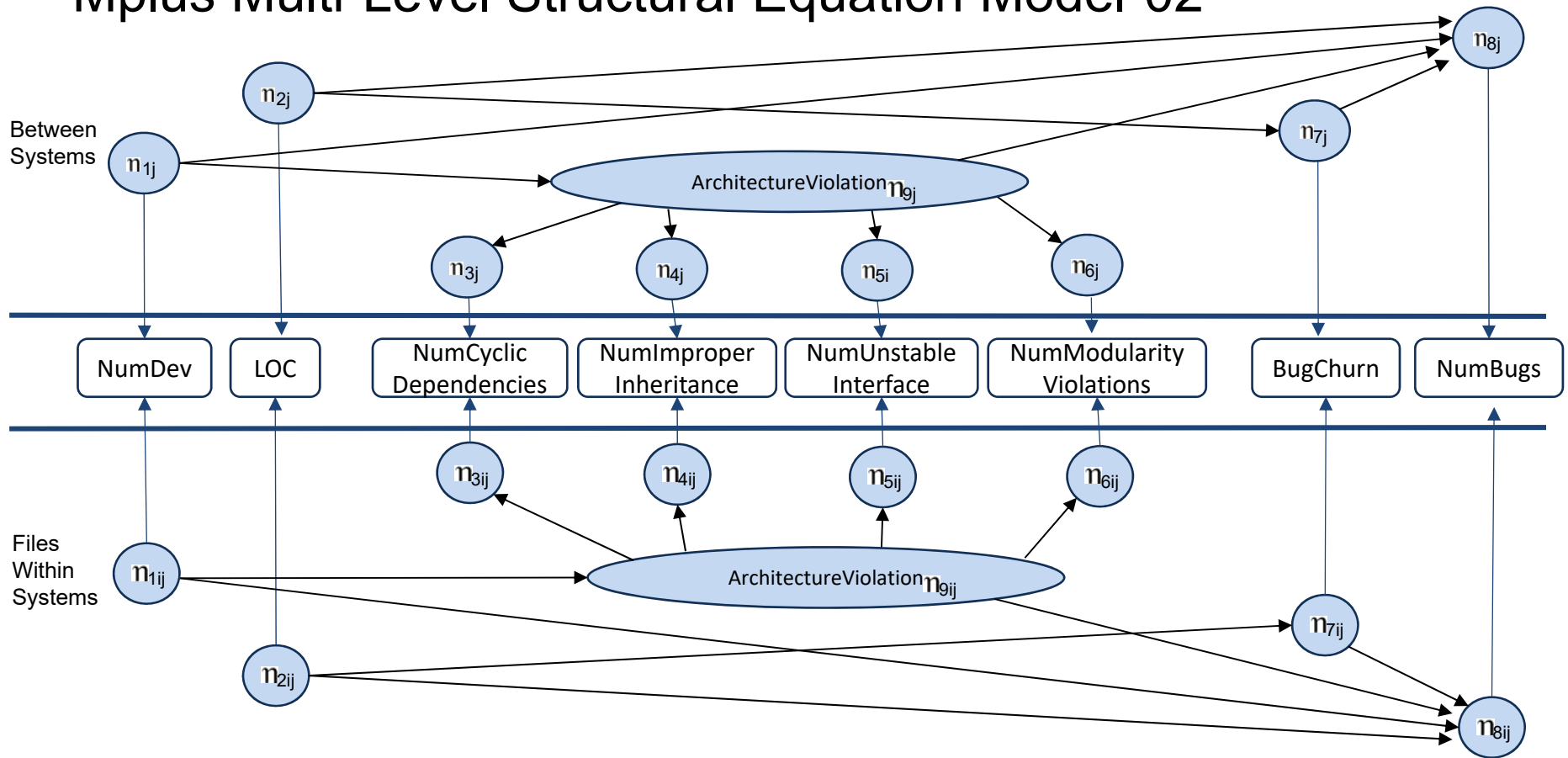
Also called
Simpson's
Paradox
and the
Ecological
Fallacy

Kris Preacher, 2018

Mplus Multi-Level Structural Equation Model-01



Mplus Multi-Level Structural Equation Model-02



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 STDYX;
```


Mplus MSEM Results

SUMMARY OF DATA

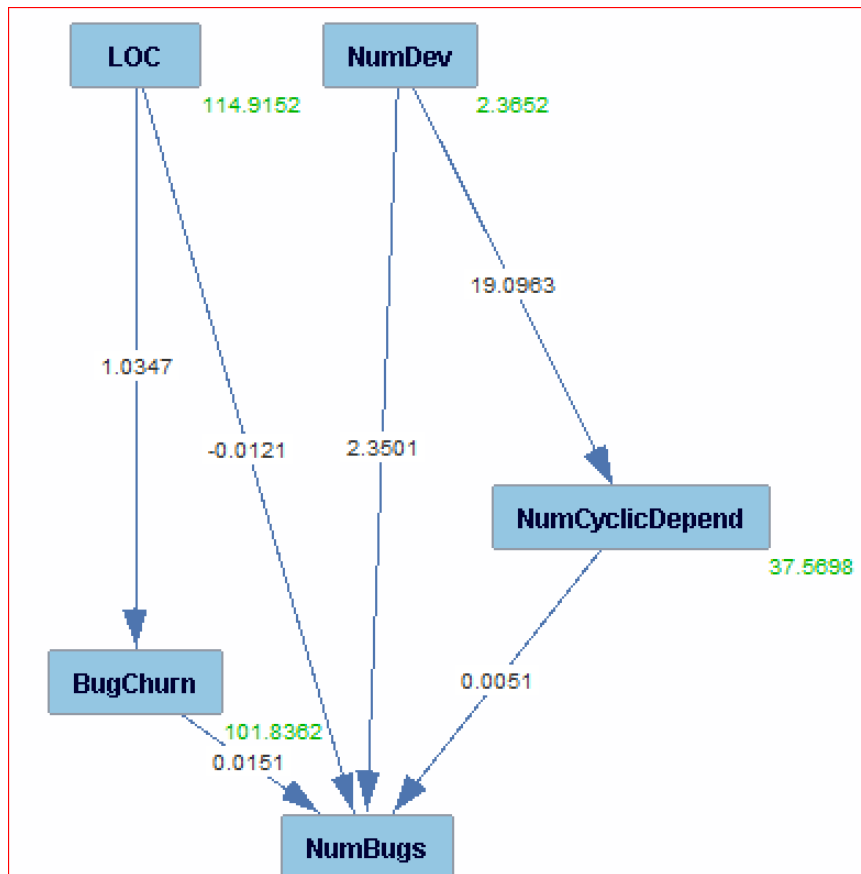
Number of clusters 9

Average cluster size 1005.556

Estimated Intraclass Correlations for the Y Variables

Variable	Intraclass Correlation	Variable	Intraclass Correlation	Variable	Intraclass Correlation
NUMBUGS	0.052	NUMDEV	0.084	LOC	0.008
CYCLES	0.039	BUGCHURN	0.026		

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

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Carnegie Mellon University
Software Engineering Institute



[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.]