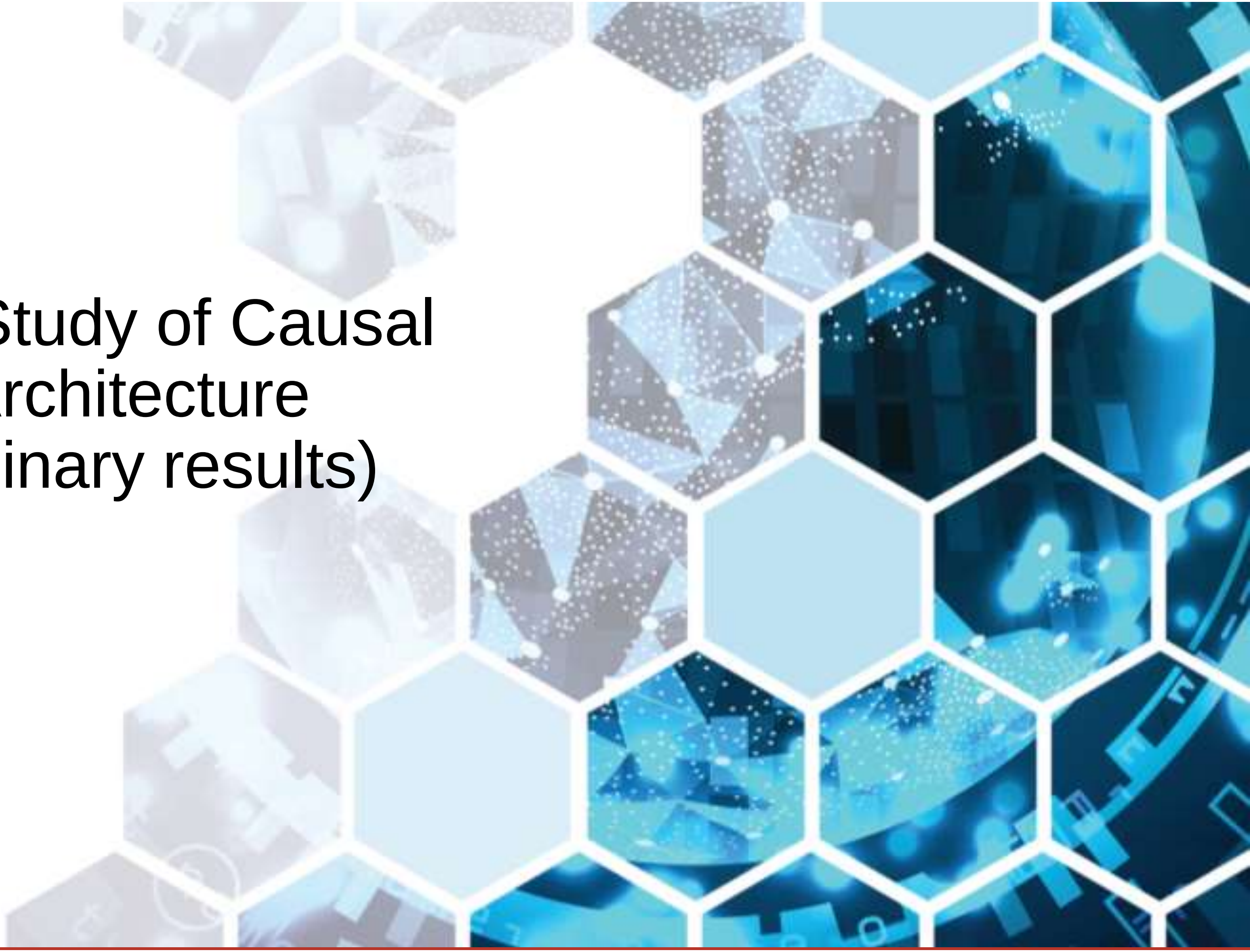
Extended Case Study of Causal Learning within Architecture Research (preliminary results)

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

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213



Goal of the Authors

Robert Stoddard
SEI

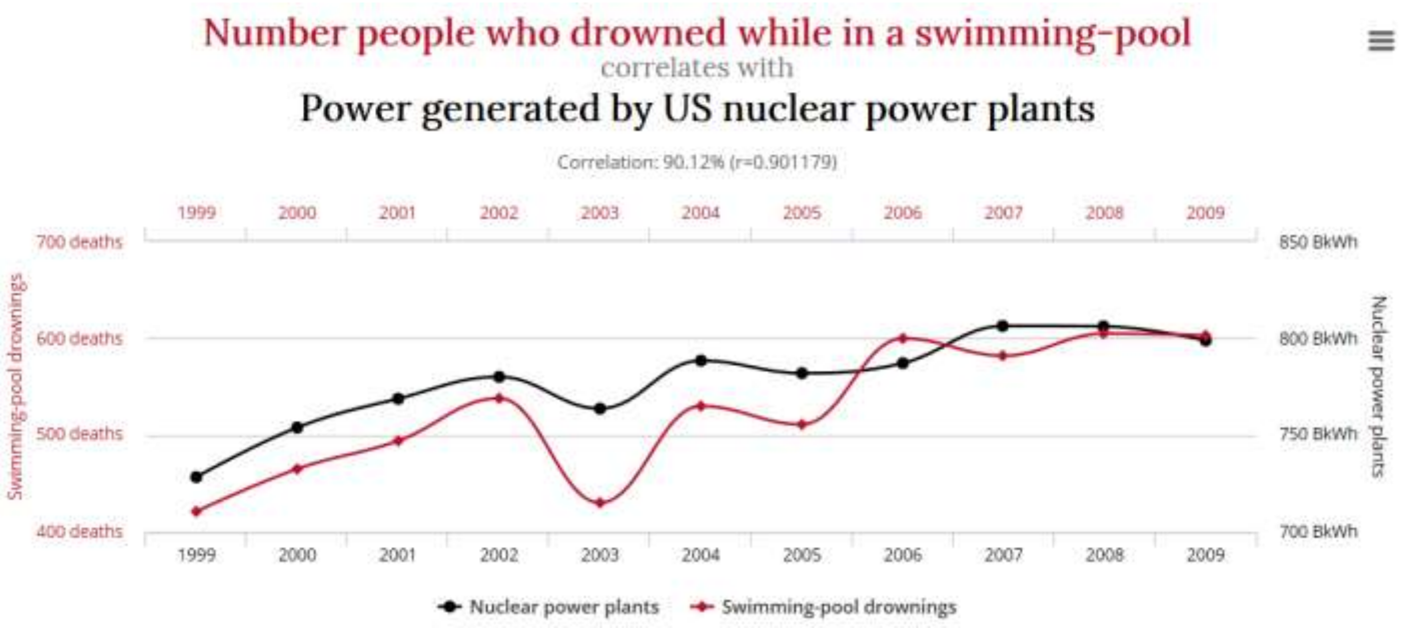
Dr. Mike Konrad
SEI

Dr. David Danks
CMU

Goal is to motivate adoption of Causal Learning within software measurement!

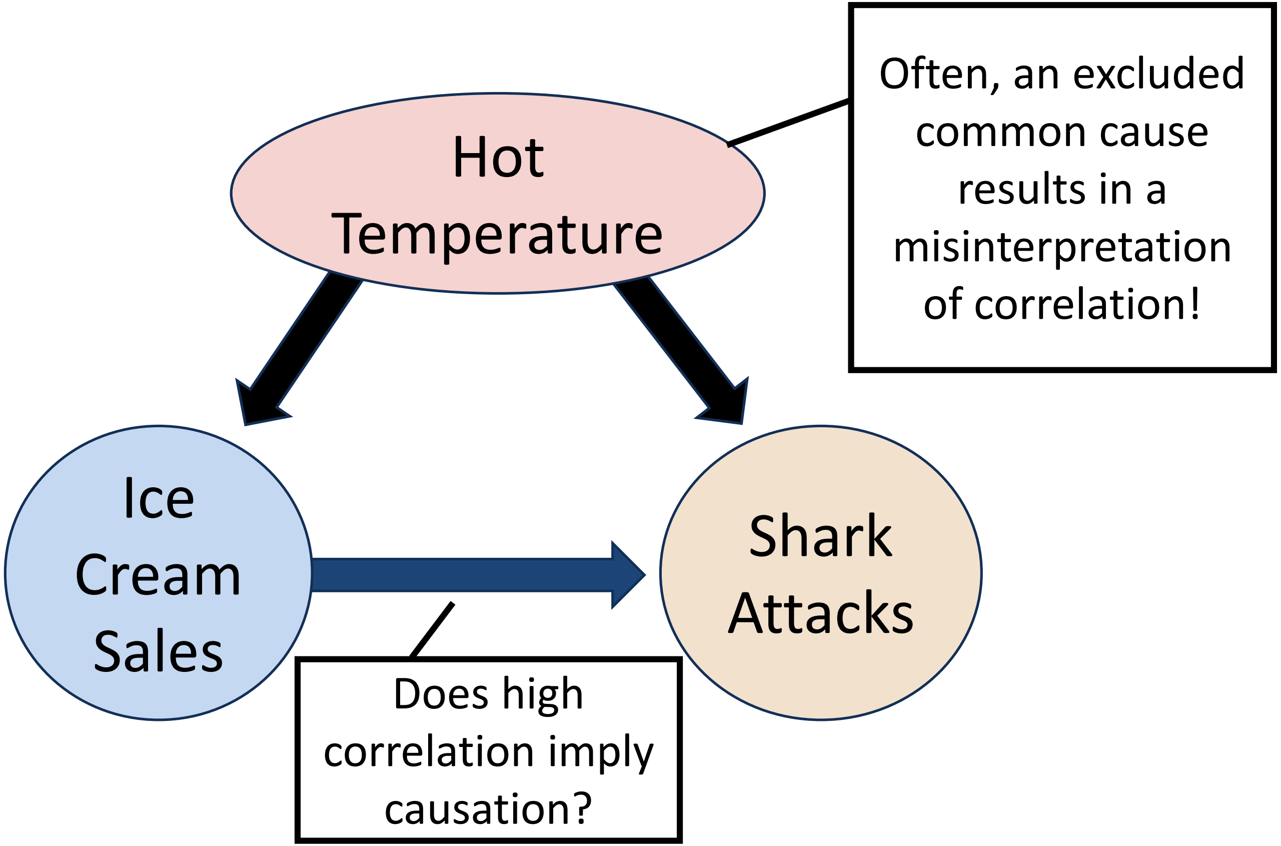
Dr. Rick Kazman
SEI

Why Do We Care about Causation?



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

More about Misinterpreting Correlation!



Hot Temperature

Ice Cream Sales

Shark Attacks

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

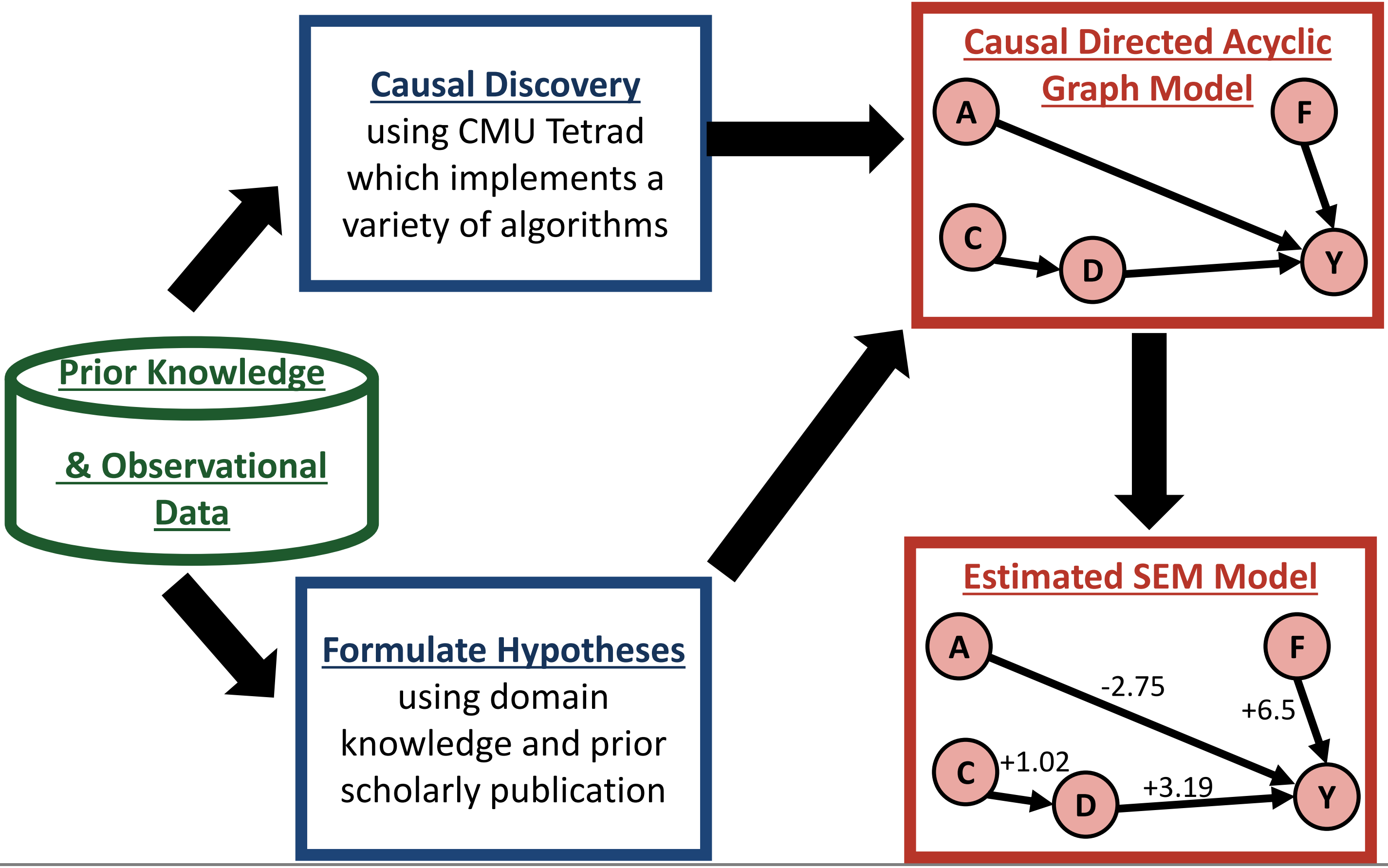
We then need to determine which paths are causal and which are spurious.

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



Prior Knowledge

& Observational Data

Causal Discovery using CMU Tetrad which implements a variety of algorithms

Formulate Hypotheses using domain knowledge and prior scholarly publication

Causal Directed Acyclic Graph Model

Estimated SEM Model

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

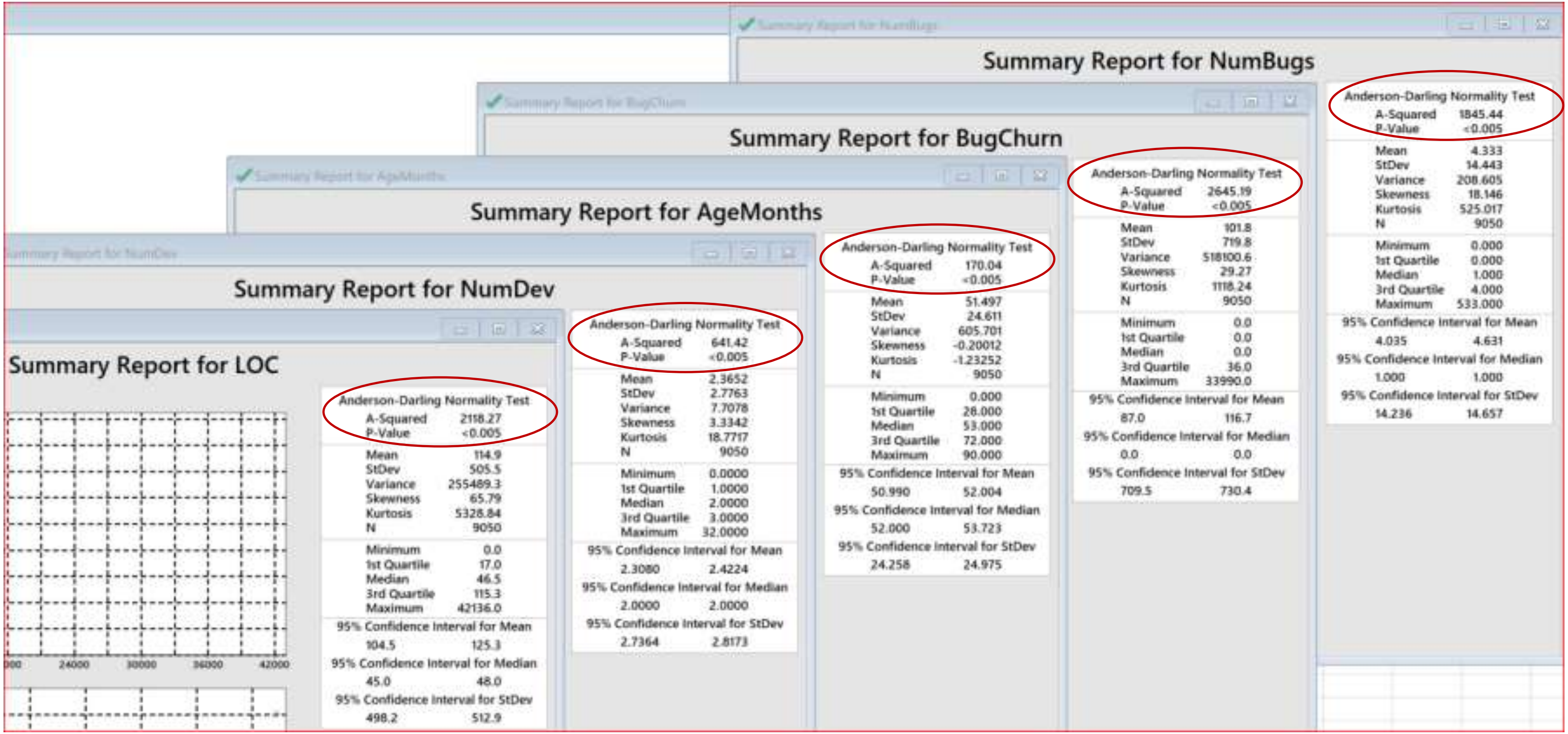
R. Mo, Y. Cai, R. Kazman and L. Xiao, "Hotspot Patterns: The Formal Definition and Automatic Detection of Architecture Smells," *2015 12th Working IEEE/IFIP Conference on Software Architecture*, Montreal, QC, 2015, pp. 51-60. doi: 10.1109/WICSA.2015.12

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



```
Summary Report for LOC

Anderson-Darling Normality Test
    A-Squared      2118.27
    P-Value        <0.005

    Mean             114.9
    StDev            505.5
    Variance      255489.3
    Skewness         65.79
    Kurtosis       5328.84
    N                 9050

    Minimum            0.0
    1st Quartile      17.0
    Median            46.5
    3rd Quartile     115.3
    Maximum        42136.0
95% Confidence Interval for Mean
    104.5            125.3
95% Confidence Interval for Median
    45.0              48.0
95% Confidence Interval for StDev
    498.2            512.9
```

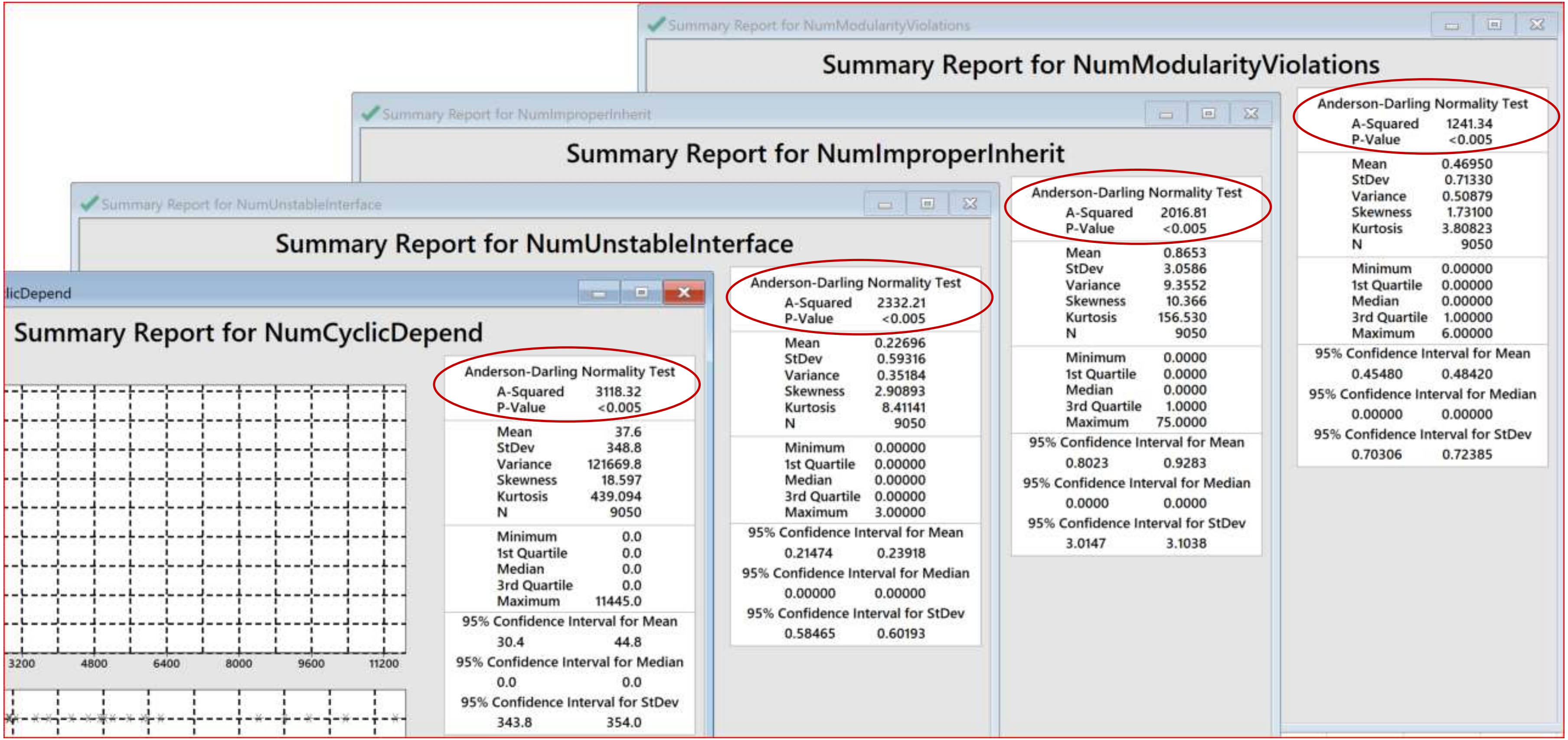
```
Summary Report for NumDev

Anderson-Darling Normality Test
    A-Squared       641.42
    P-Value         <0.005

    Mean            2.3652
    StDev           2.7763
    Variance        7.7078
    Skewness        3.3342
    Kurtosis       18.7717
    N                 9050

    Minimum         0.0000
    1st Quartile    1.0000
    Median          2.0000
    3rd Quartile    3.0000
    Maximum        32.0000
95% Confidence Interval for Mean
    2.3080          2.4224
95% Confidence Interval for Median
    2.0000          2.0000
95% Confidence Interval for StDev
    2.7364          2.8173
```

```
Summary Report for AgeMonths

Anderson-Darling Normality Test
    A-Squared       170.04
    P-Value         <0.005

    Mean            51.497
    StDev           24.611
    Variance       605.701
    Skewness      -0.20012
    Kurtosis      -1.23252
    N                 9050

    Minimum          0.000
    1st Quartile    28.000
    Median          53.000
    3rd Quartile    72.000
    Maximum         90.000
95% Confidence Interval for Mean
    50.990          52.004
95% Confidence Interval for Median
    52.000          53.723
95% Confidence Interval for StDev
    24.258          24.975
```

```
Summary Report for BugChurn

Anderson-Darling Normality Test
    A-Squared      2645.19
    P-Value         <0.005

    Mean             101.8
    StDev            719.8
    Variance      518100.6
    Skewness         29.27
    Kurtosis       1118.24
    N                 9050

    Minimum            0.0
    1st Quartile       0.0
    Median             0.0
    3rd Quartile      36.0
    Maximum        33990.0
95% Confidence Interval for Mean
    87.0             116.7
95% Confidence Interval for Median
    0.0                0.0
95% Confidence Interval for StDev
    709.5            730.4
```

```
Summary Report for NumBugs

Anderson-Darling Normality Test
    A-Squared      1845.44
    P-Value         <0.005

    Mean             4.333
    StDev           14.443
    Variance       208.605
    Skewness        18.146
    Kurtosis       525.017
    N                 9050

    Minimum          0.000
    1st Quartile     0.000
    Median           1.000
    3rd Quartile     4.000
    Maximum        533.000
95% Confidence Interval for Mean
    4.035            4.631
95% Confidence Interval for Median
    1.000            1.000
95% Confidence Interval for StDev
    14.236          14.657
```

All Remaining Factors are Non-Normal - 02


Summary Report for NumModularityViolations
Summary Report for NumImproperInherit
Summary Report for NumUnstableInterface
licDepend
3200
4800
6400
8000
9600
11200

Summary Report for NumCyclicDepend

Anderson-Darling Normality Test	
A-Squared	3118.32
P-Value	<0.005
Mean	37.6
StDev	348.8
Variance	121669.8
Skewness	18.597
Kurtosis	439.094
N	9050
Minimum	0.0
1st Quartile	0.0
Median	0.0
3rd Quartile	0.0
Maximum	11445.0
95% Confidence Interval for Mean	
30.4	44.8
95% Confidence Interval for Median	
0.0	0.0
95% Confidence Interval for StDev	
343.8	354.0

Summary Report for NumUnstableInterface

Anderson-Darling Normality Test	
A-Squared	2332.21
P-Value	<0.005
Mean	0.22696
StDev	0.59316
Variance	0.35184
Skewness	2.90893
Kurtosis	8.41141
N	9050
Minimum	0.00000
1st Quartile	0.00000
Median	0.00000
3rd Quartile	0.00000
Maximum	3.00000
95% Confidence Interval for Mean	
0.21474	0.23918
95% Confidence Interval for Median	
0.00000	0.00000
95% Confidence Interval for StDev	
0.58465	0.60193

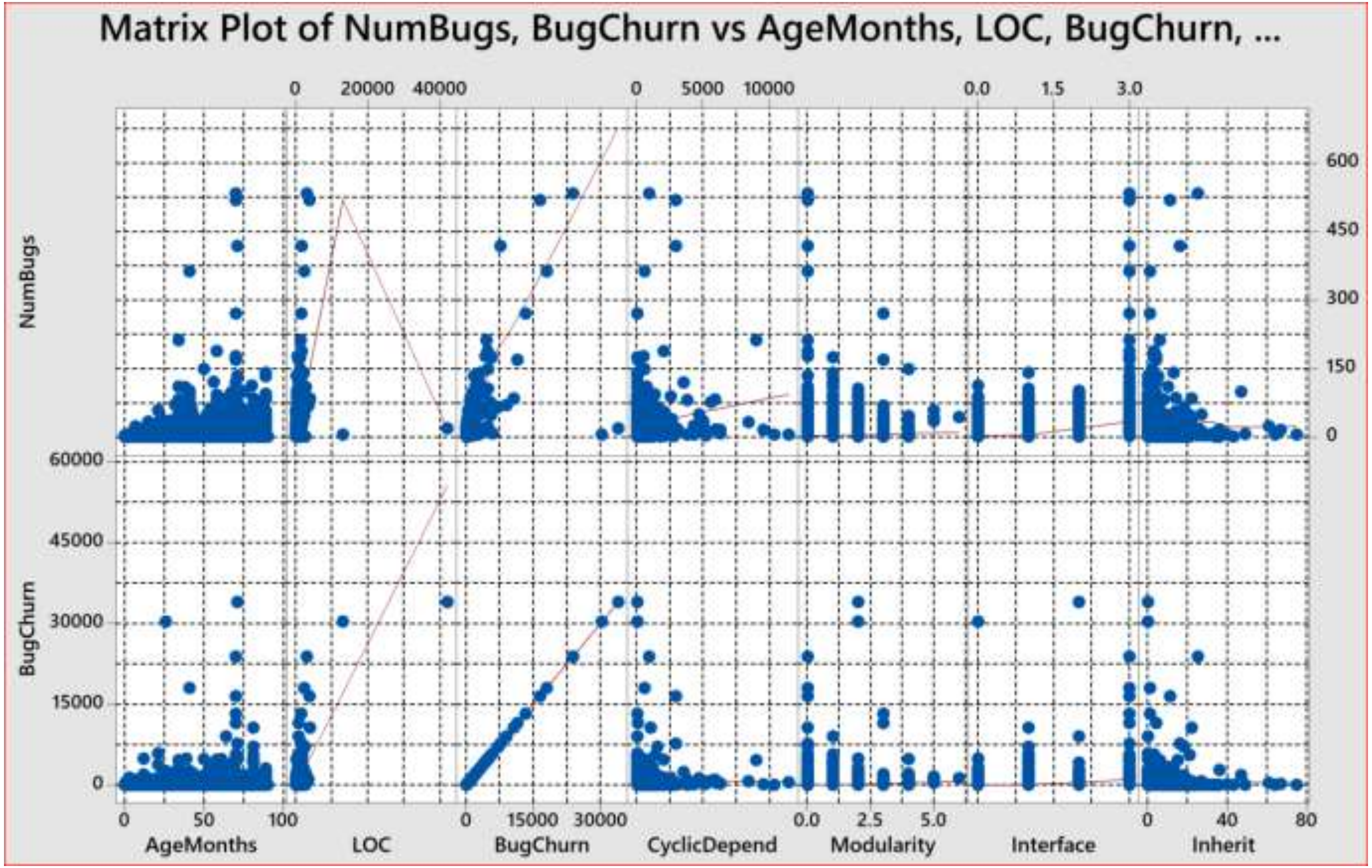
Summary Report for NumImproperInherit

Anderson-Darling Normality Test	
A-Squared	2016.81
P-Value	<0.005
Mean	0.8653
StDev	3.0586
Variance	9.3552
Skewness	10.366
Kurtosis	156.530
N	9050
Minimum	0.0000
1st Quartile	0.0000
Median	0.0000
3rd Quartile	1.0000
Maximum	75.0000
95% Confidence Interval for Mean	
0.8023	0.9283
95% Confidence Interval for Median	
0.0000	0.0000
95% Confidence Interval for StDev	
3.0147	3.1038

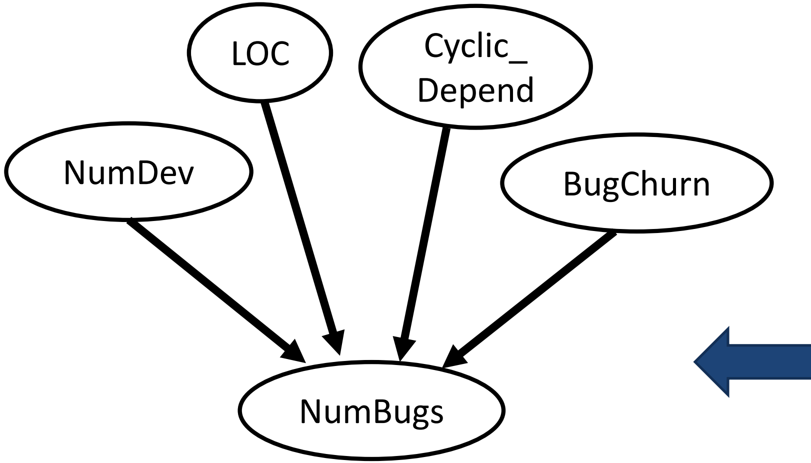
Summary Report for NumModularityViolations

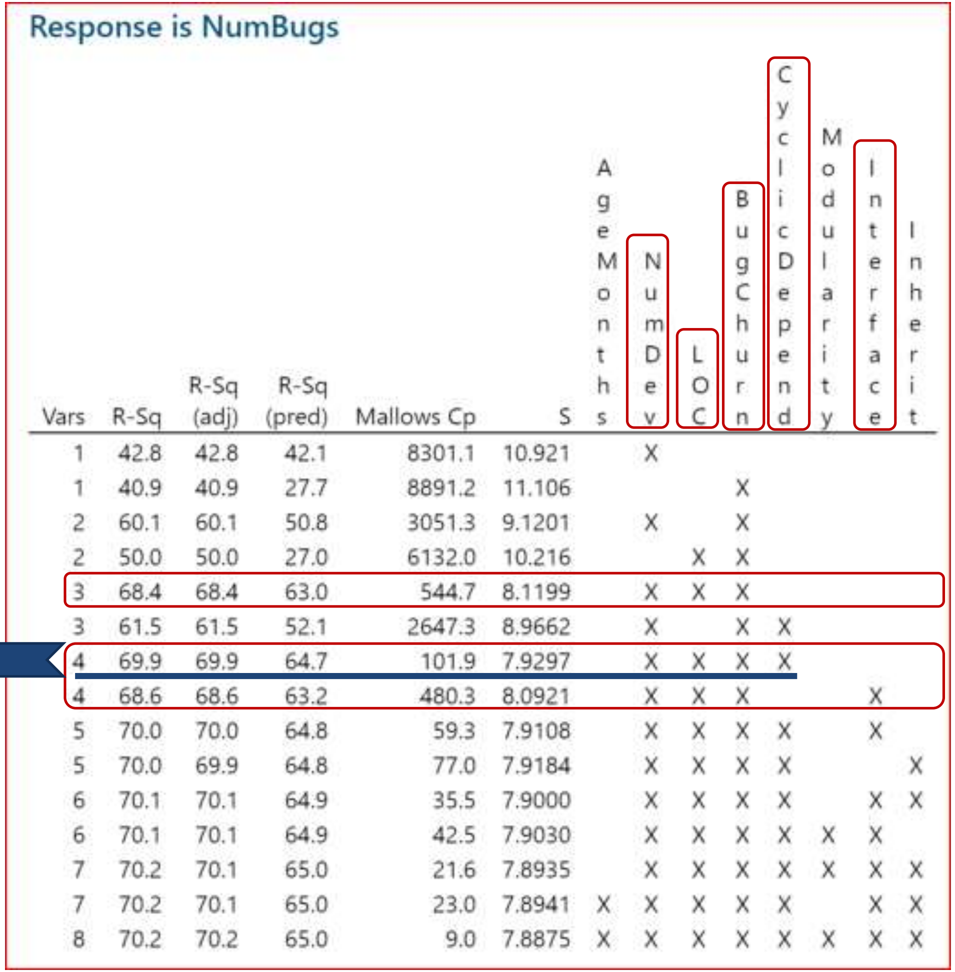
Anderson-Darling Normality Test	
A-Squared	1241.34
P-Value	<0.005
Mean	0.46950
StDev	0.71330
Variance	0.50879
Skewness	1.73100
Kurtosis	3.80823
N	9050
Minimum	0.00000
1st Quartile	0.00000
Median	0.00000
3rd Quartile	1.00000
Maximum	6.00000
95% Confidence Interval for Mean	
0.45480	0.48420
95% Confidence Interval for Median	
0.00000	0.00000
95% Confidence Interval for StDev	
0.70306	0.72385

Eyeballing Bivariate Relationships



Best Subsets Regression

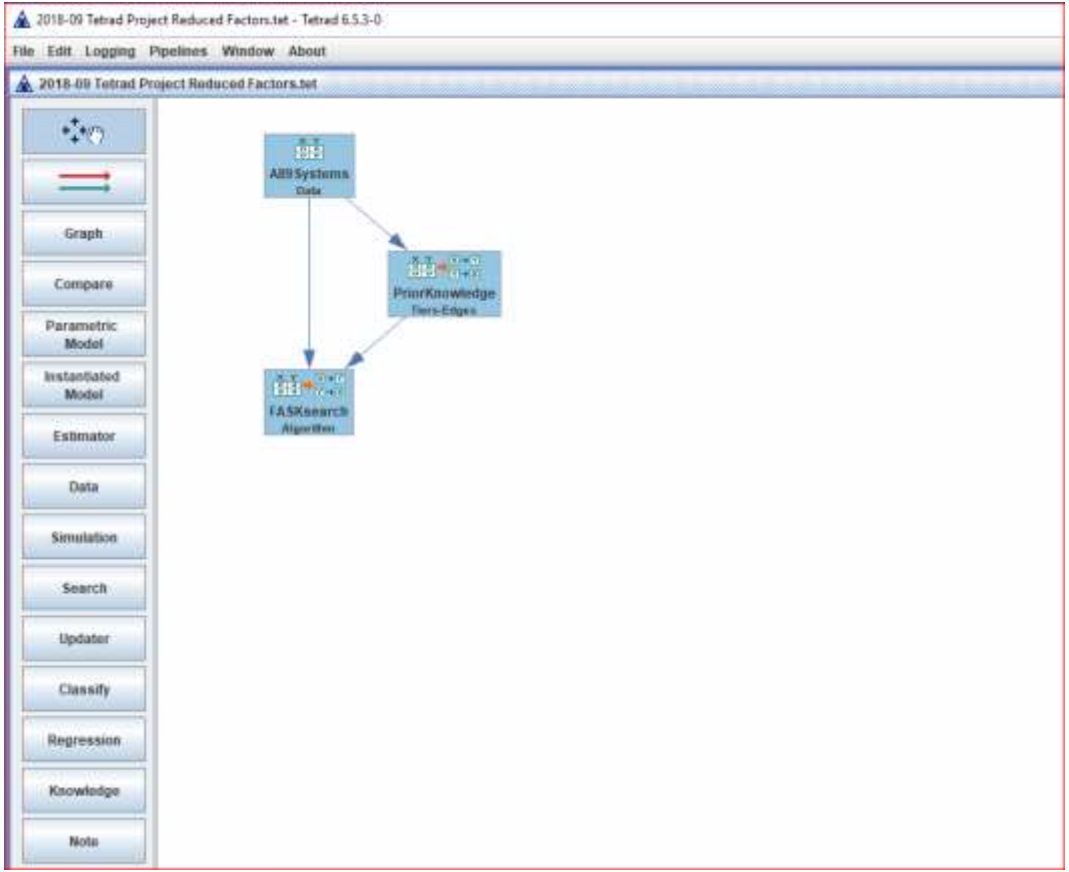




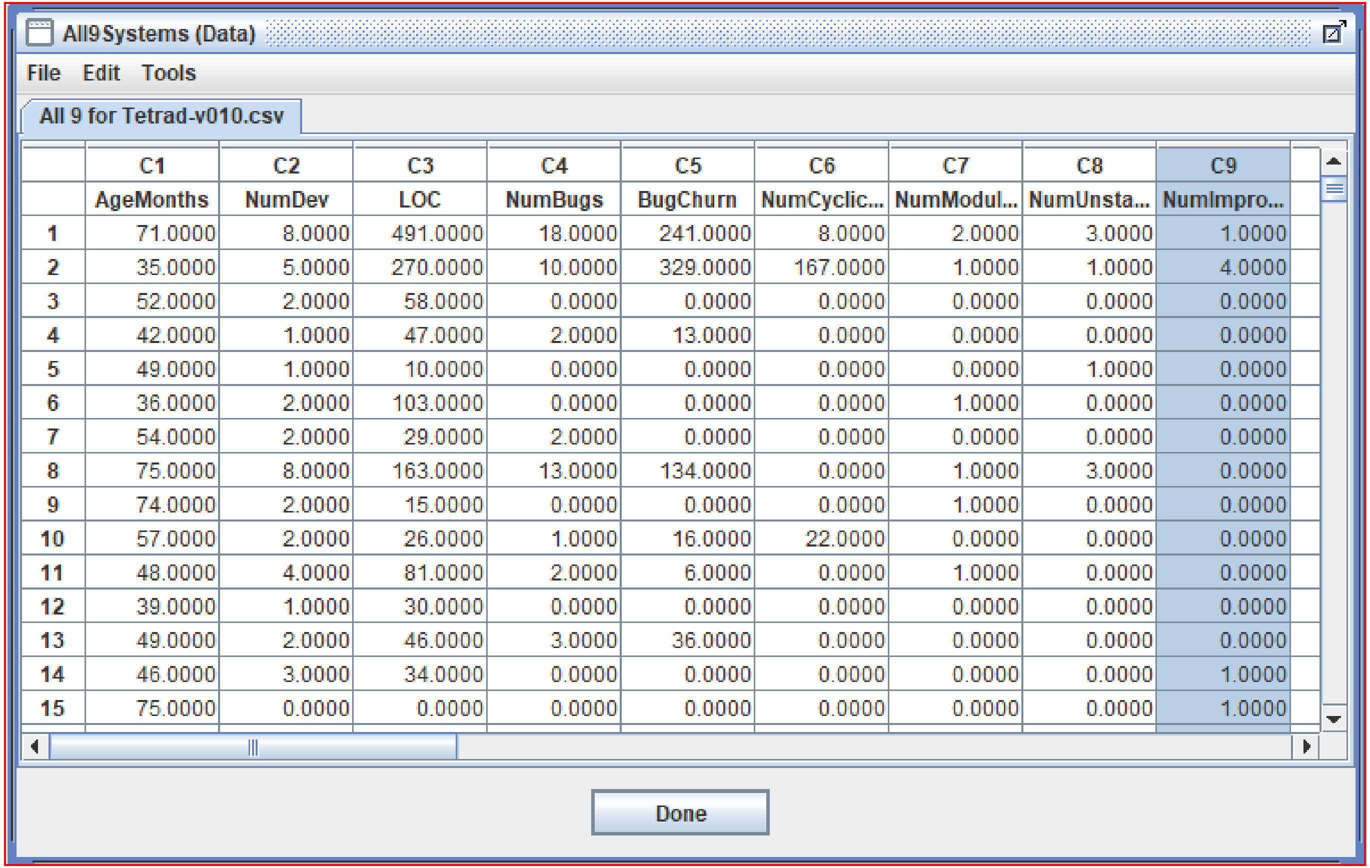
Response is NumBugs

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	AgeMonths	NumDev	LOC	BugChurn	CyclicDepend	Modularity	Interface	Inherit
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	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	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

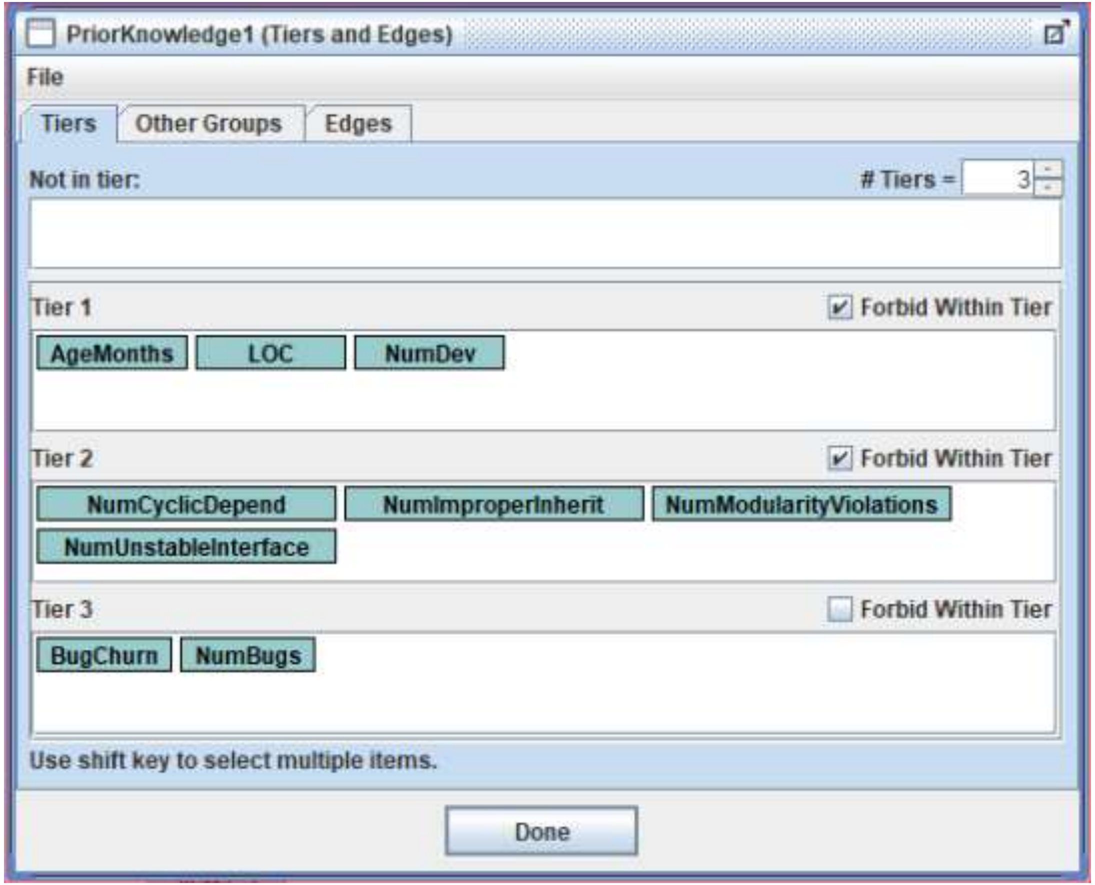


A View of the Data File Loaded into Tetrad



	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

Prior Knowledge Entered into Tetrad



PriorKnowledge1 (Tiers and Edges)

Not in tier:

# Tiers = 3

Tier 1 ☑ Forbid Within Tier

AgeMonths, LOC, NumDev

Tier 2 ☑ Forbid Within Tier

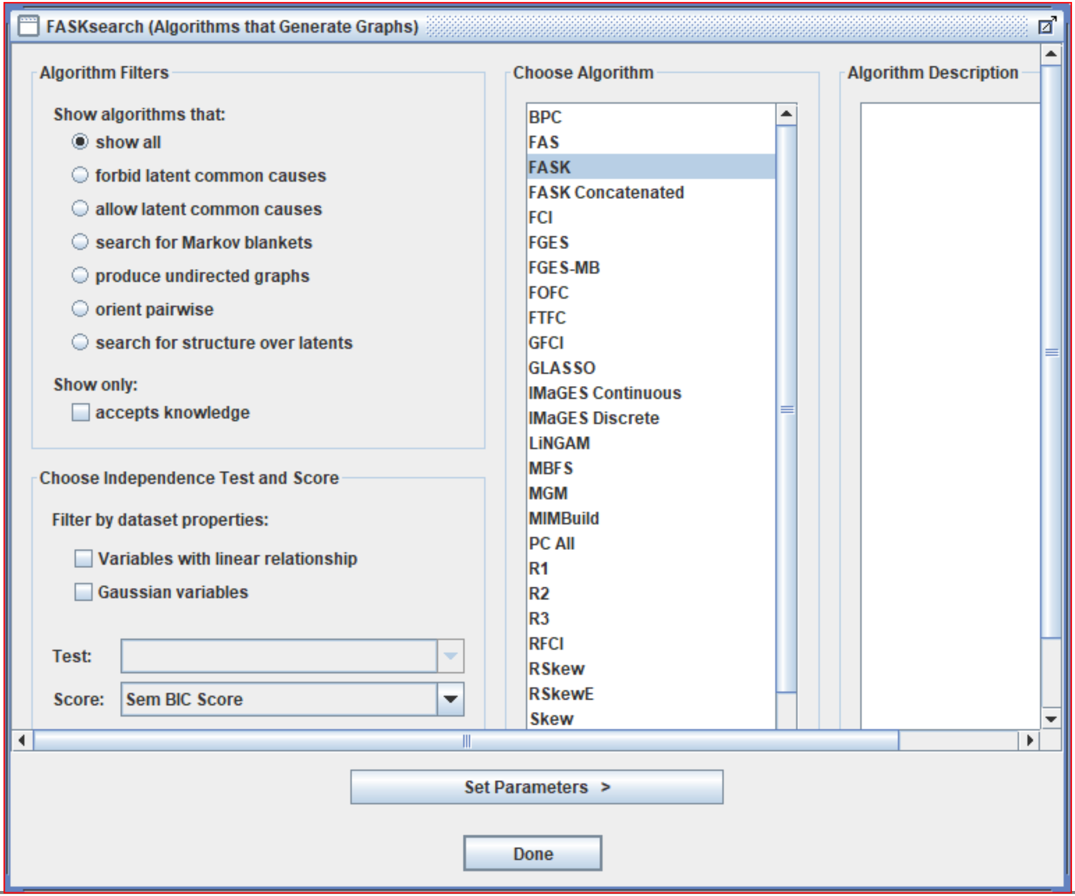
NumCyclicDepend, NumImproperInherit, NumModularityViolations, NumUnstableInterface

Tier 3 □ Forbid Within Tier

BugChurn, NumBugs

Use shift key to select multiple items.

Using FASK Search with Associated Parameters



FASKsearch (Algorithms that Generate Graphs)

Algorithm Filters

Show algorithms that:

- ● show all
- ○ forbid latent common causes
- ○ allow latent common causes
- ○ search for Markov blankets
- ○ produce undirected graphs
- ○ orient pairwise
- ○ search for structure over latents

Show only:

- □ accepts knowledge

Choose Independence Test and Score

Filter by dataset properties:

- □ Variables with linear relationship
- □ Gaussian variables

Test:

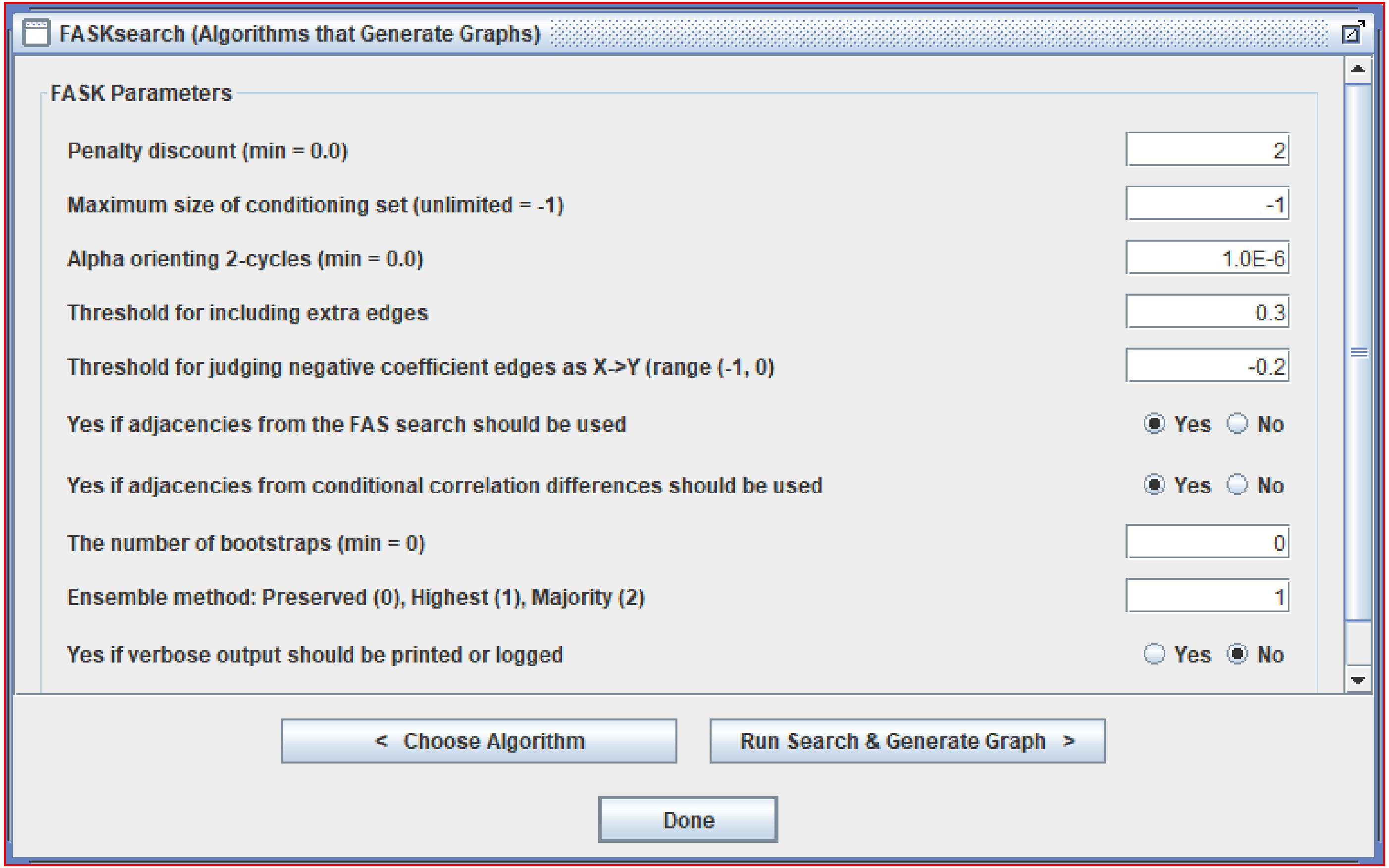
Score: Sem BIC Score

Choose Algorithm

BPC
FAS
FASK
FASK Concatenated
FCI
FGES
FGES-MB
FOFC
FTFC
GFCI
GLASSO
IMaGES Continuous
IMaGES Discrete
LiNGAM
MBFS
MGM
MIMBuild
PC All
R1
R2
R3
RFCI
RSkew
RSkewE
Skew

Algorithm Description

Additional FASK Search Parameter Settings



FASKsearch (Algorithms that Generate Graphs)

FASK Parameters

Penalty discount (min = 0.0) 2

Maximum size of conditioning set (unlimited = -1) -1

Alpha orienting 2-cycles (min = 0.0) 1.0E-6

Threshold for including extra edges 0.3

Threshold for judging negative coefficient edges as X->Y (range (-1, 0) -0.2

Yes if adjacencies from the FAS search should be used ● Yes ○ No

Yes if adjacencies from conditional correlation differences should be used ● Yes ○ No

The number of bootstraps (min = 0) 0

Ensemble method: Preserved (0), Highest (1), Majority (2) 1

Yes if verbose output should be printed or logged ○ Yes ● No

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

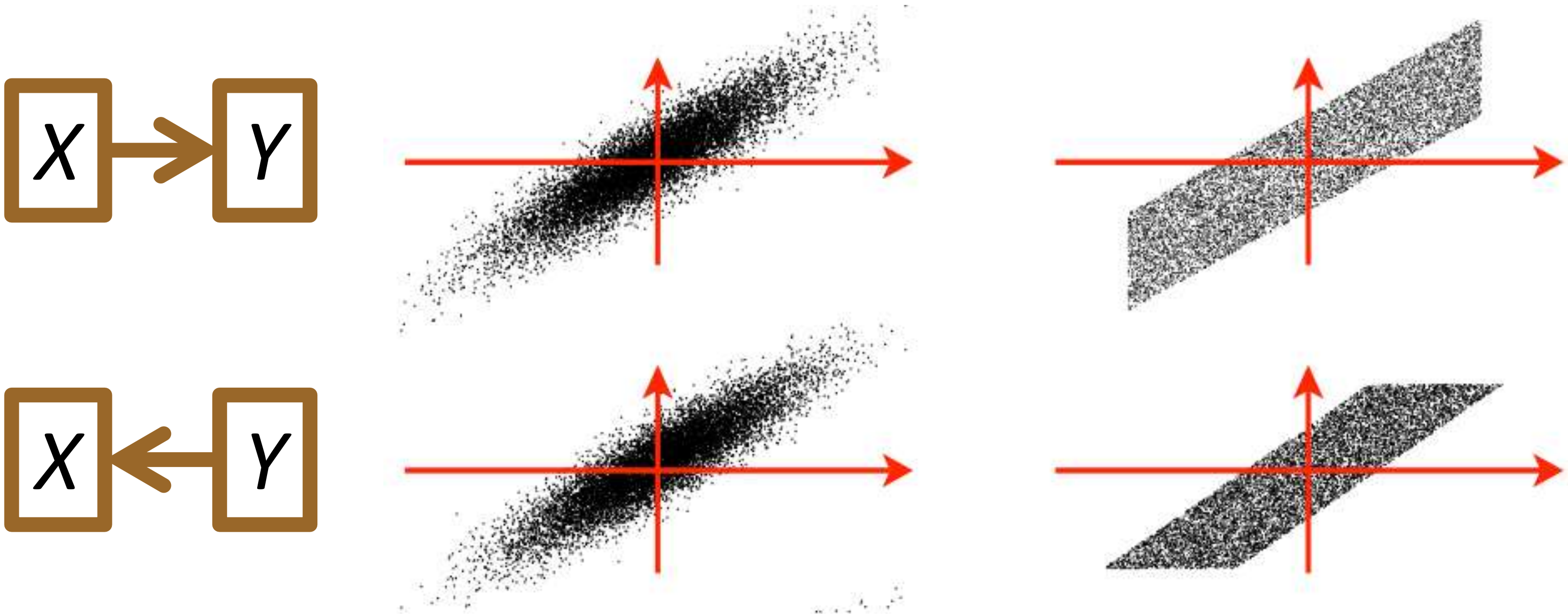
Hybrid: Use constraint-based to get “close,” then Bayesian search around neighborhood

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-Gaussianality

Linear Gaussian

Linear non-Gaussian



Causal Search Capable with Small Data

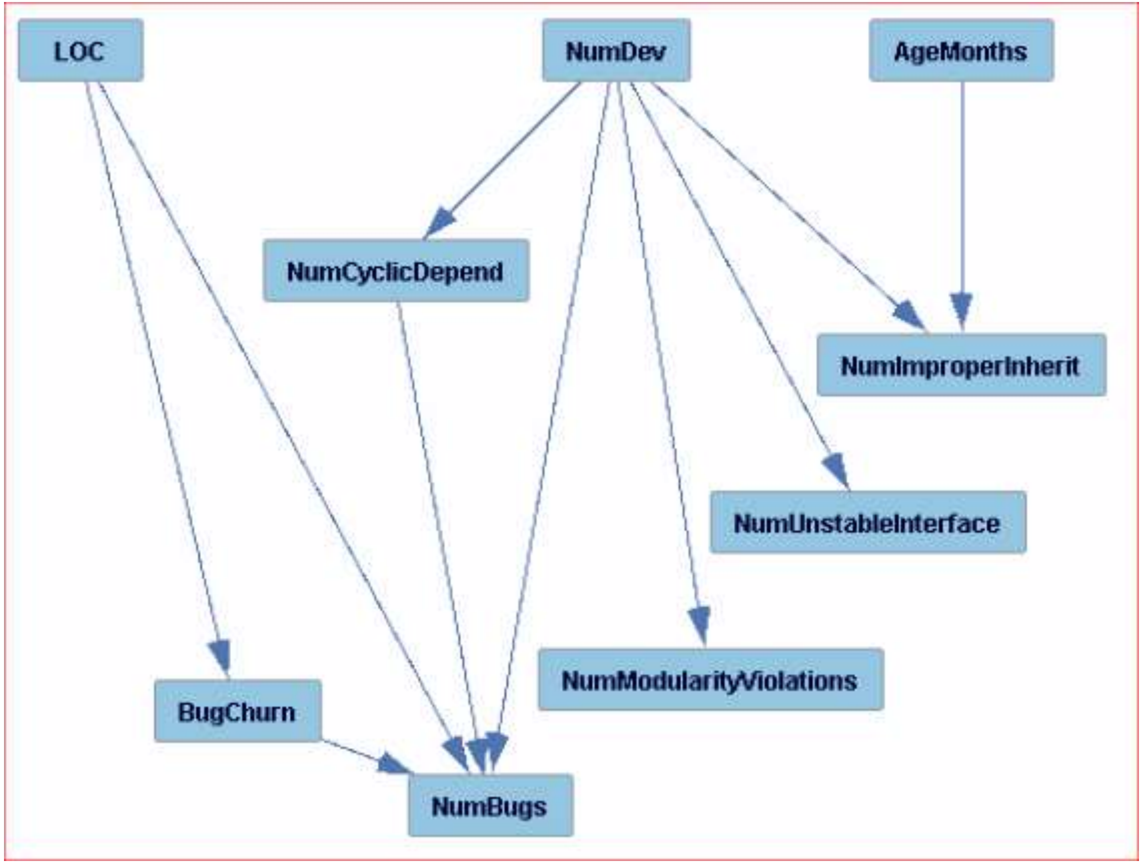
Challenge: Which genes regulate flowering time in Arabidopsis thaliana?

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

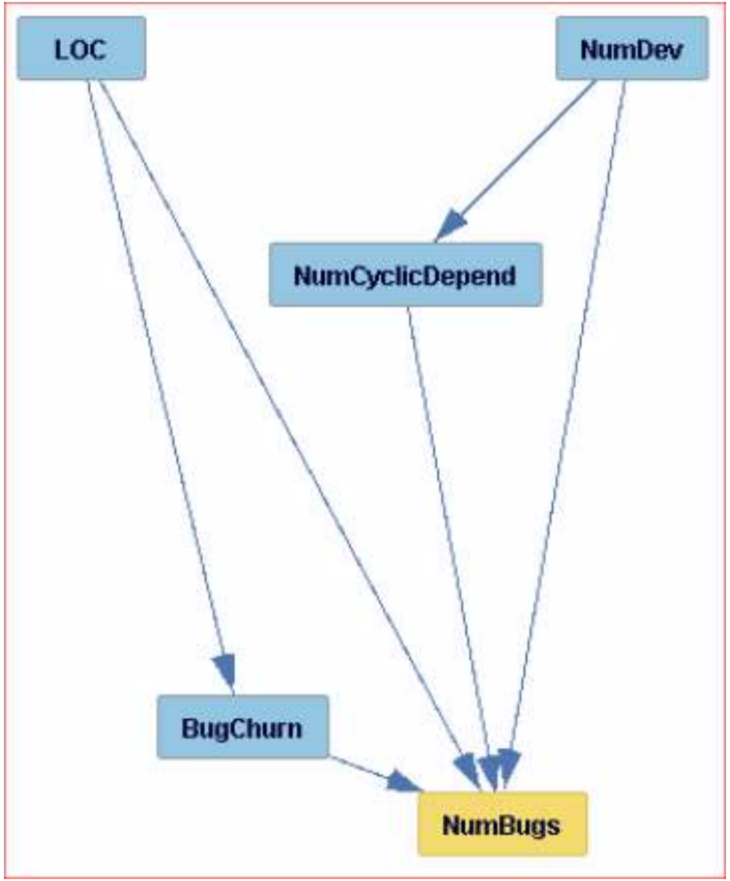
Subsequent greenhouse study, that used knockout variants, confirmed that 4 of the 9 were actual regulators

Taken from Dave Danks, 2016 Summer Causal Workshop

Causal Structure Graph Result



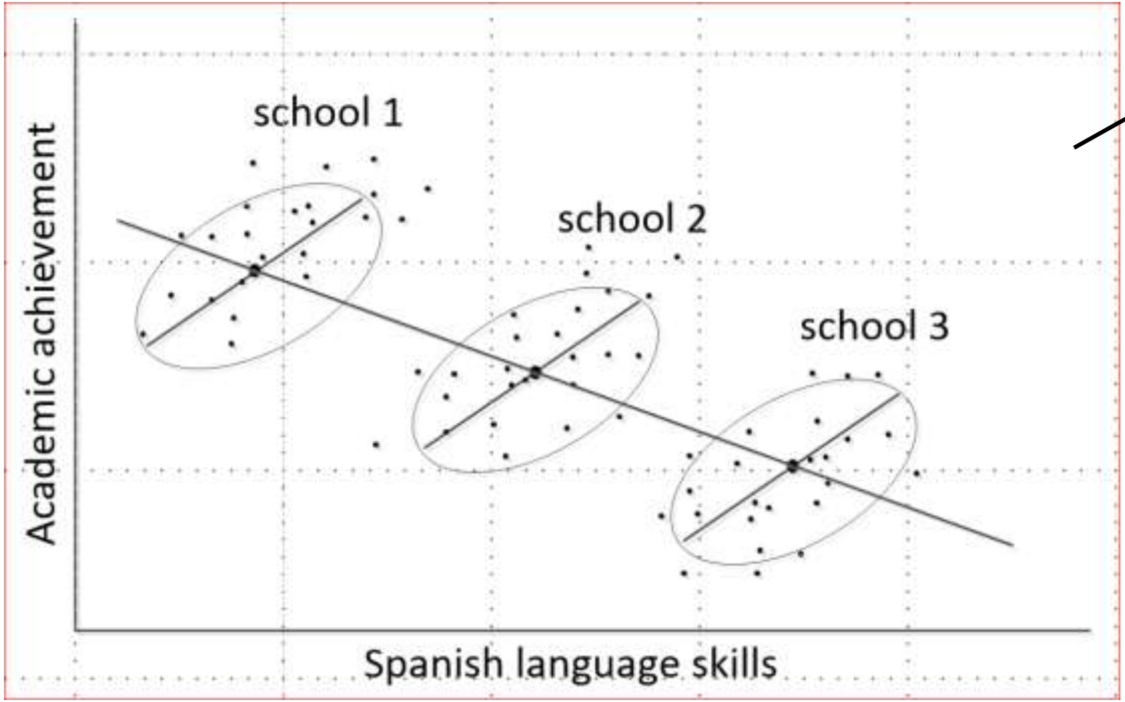
Markov Blanket of the NumBugs Factor



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

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

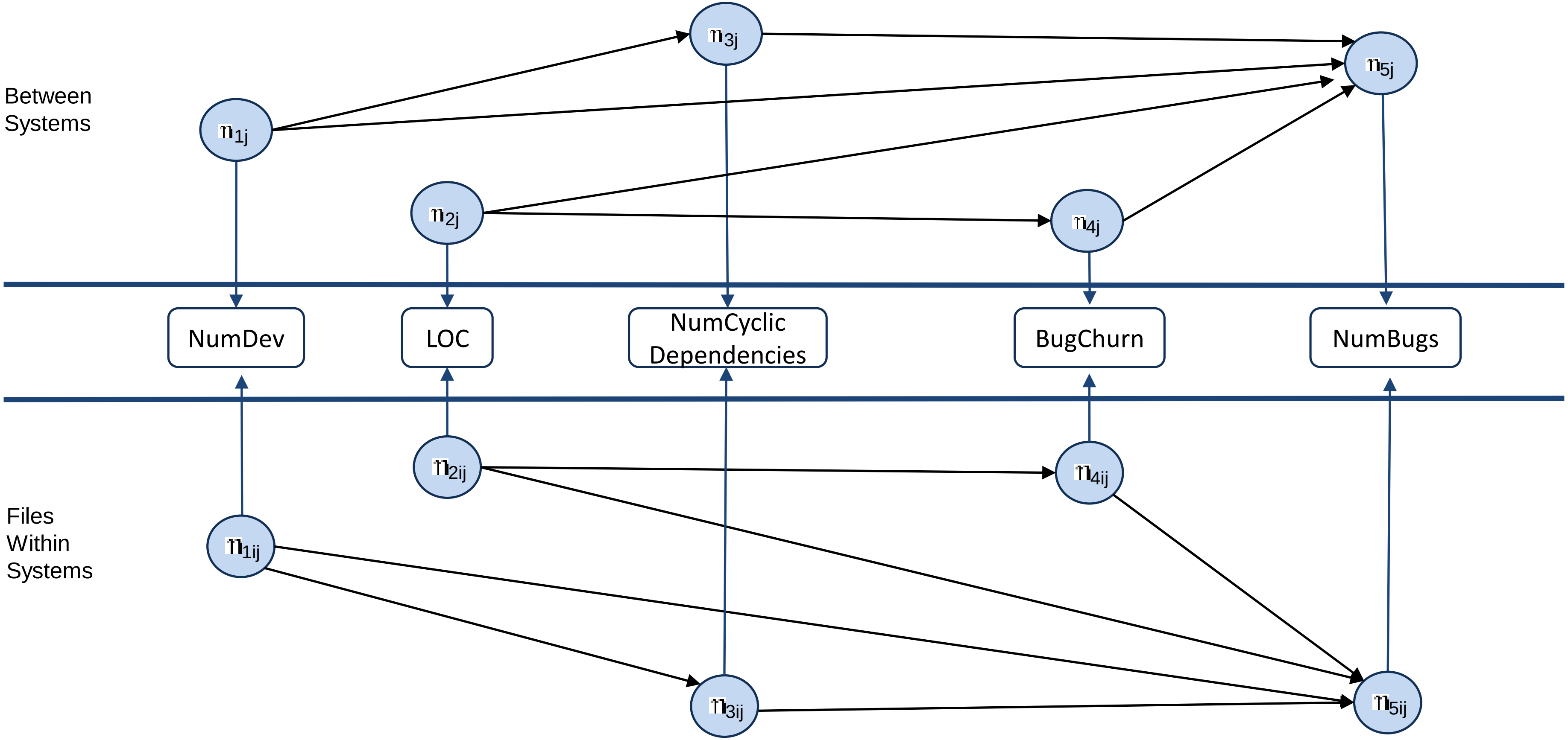
Yet, schools with highest proportion of Spanish speakers performed poorest.



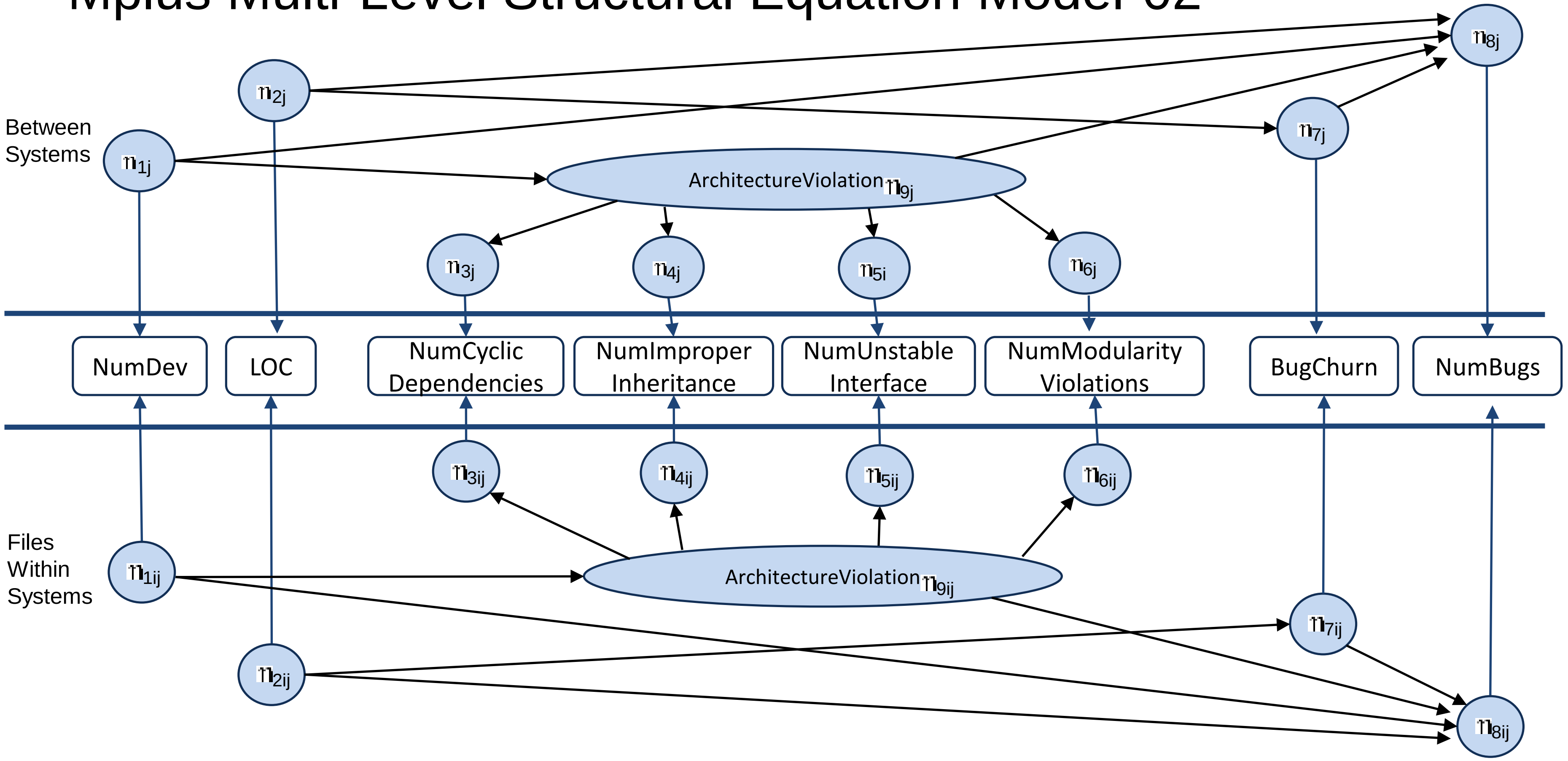
Also called Simpson's Paradox and the Ecological Fallacy

Kris Preacher, 2018

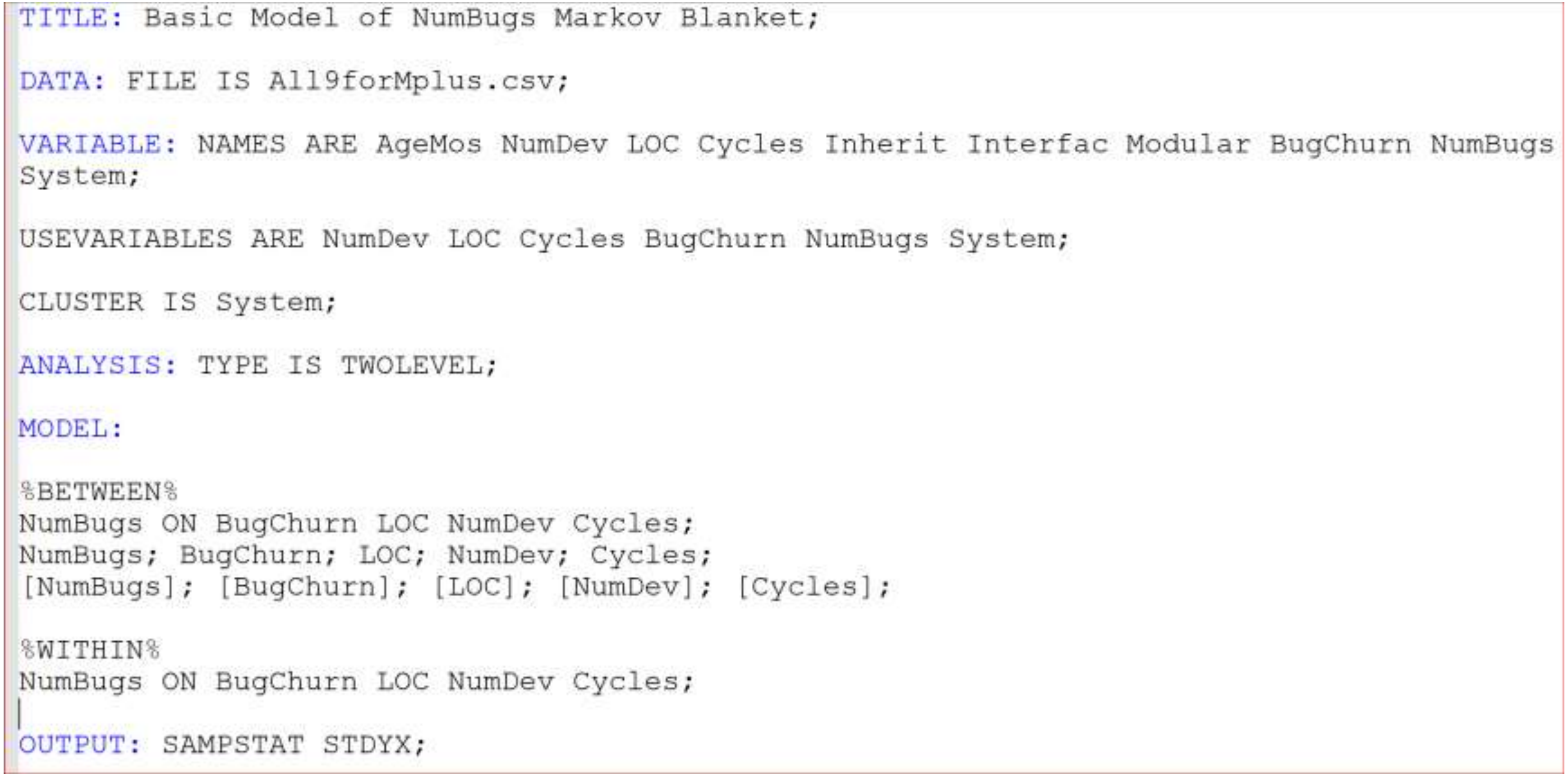
Mplus Multi-Level Structural Equation Model-01



Mplus Multi-Level Structural Equation Model-02

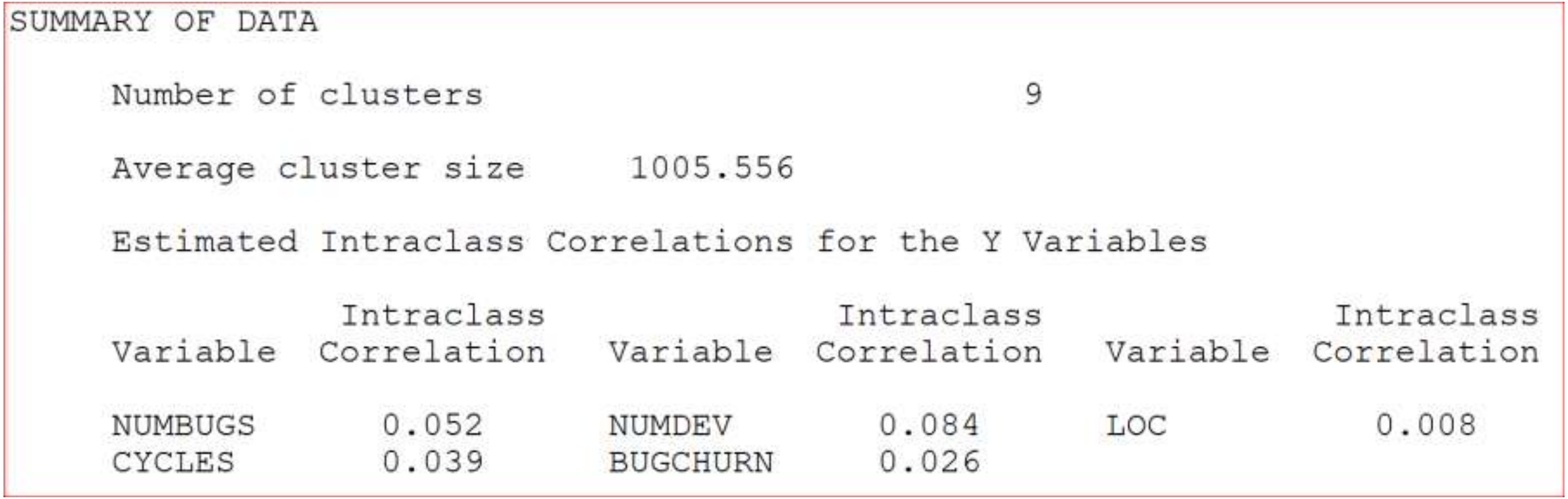


Mplus Code



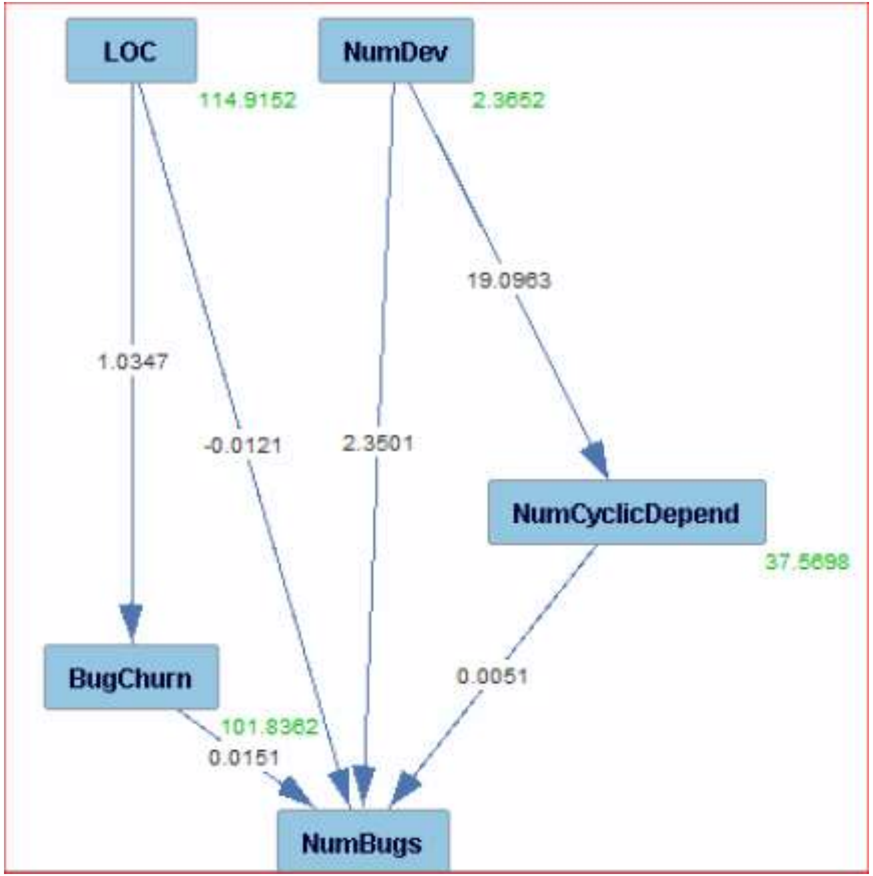
```
TITLE: Basic Model of NumBugs Markov Blanket;

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;
```

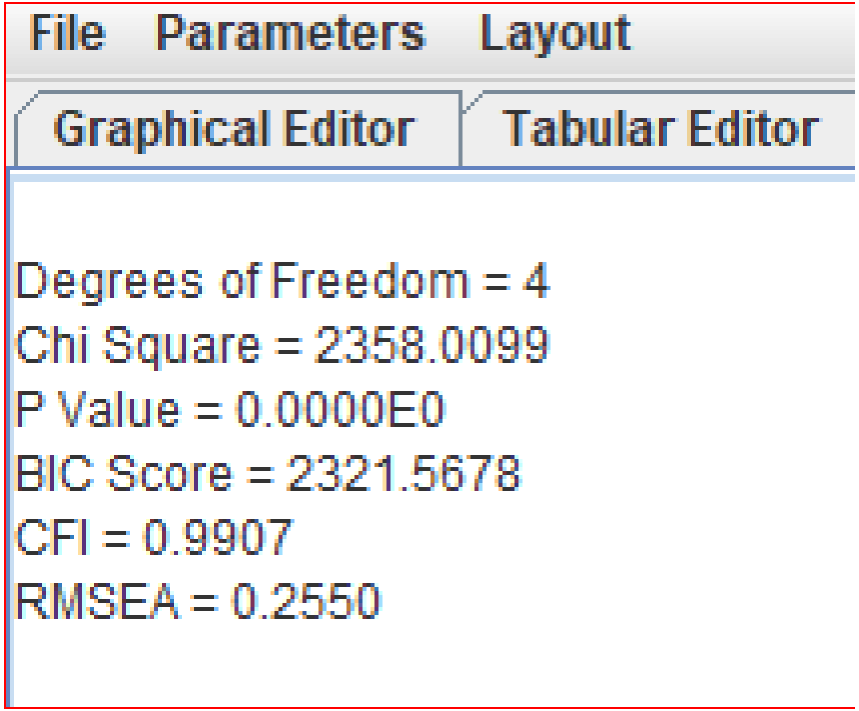
Mplus MSEM Results



```
SUMMARY OF DATA

     Number of clusters                               9

     Average cluster size     1005.556

     Estimated Intraclass Correlations for the Y Variables

                Intraclass                 Intraclass                 Intraclass
     Variable  Correlation   Variable     Correlation   Variable     Correlation

     NUMBUGS      0.052      NUMDEV          0.084      LOC             0.008
     CYCLES       0.039      BUGCHURN        0.026
```

Traditional SEM Results from Tetrad





```
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