Log Grouping and Causality Analysis

Fateme Faraji Daneshgar
Research Associate

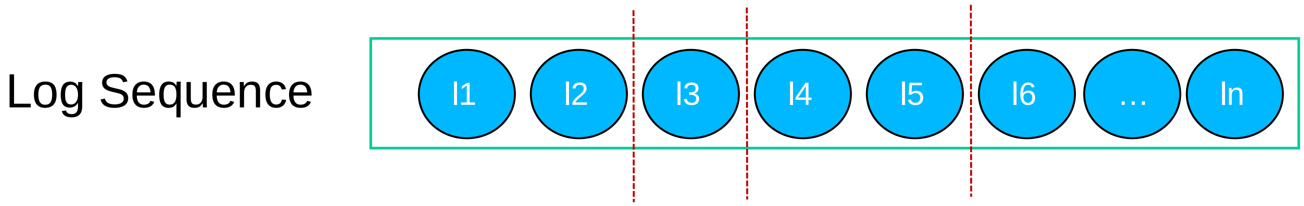
Polytechnique Montréal

DORSAL Laboratory

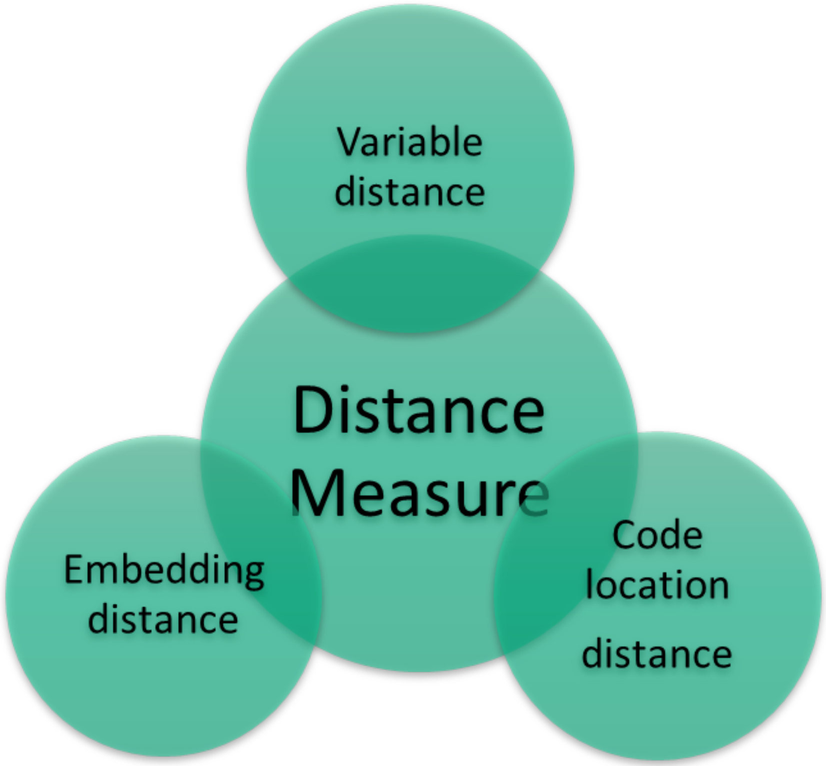
Agenda

- Log grouping
- Incremental Prefix Tree
- Co-occurrence Probability
- Causality Analysis

Log Grouping



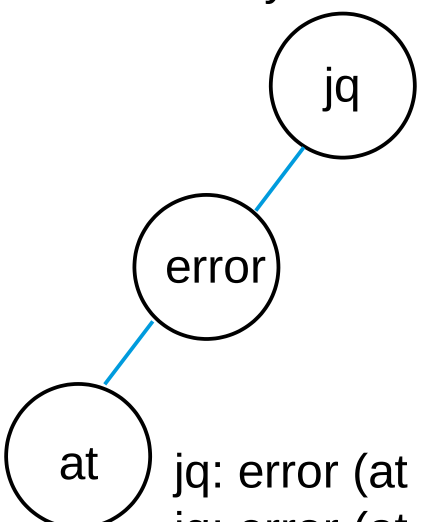
- Log
  - Template
    - Embedding distance
  - Variable
    - Cosine distance



Incremental Prefix Tree

- Parsing logs
  - Drain method
    - Prefix Tree

- jq: error (at <stdin>:45): null (null) has no keys
- jq: error (at <stdin>:21): null (null) has no keys

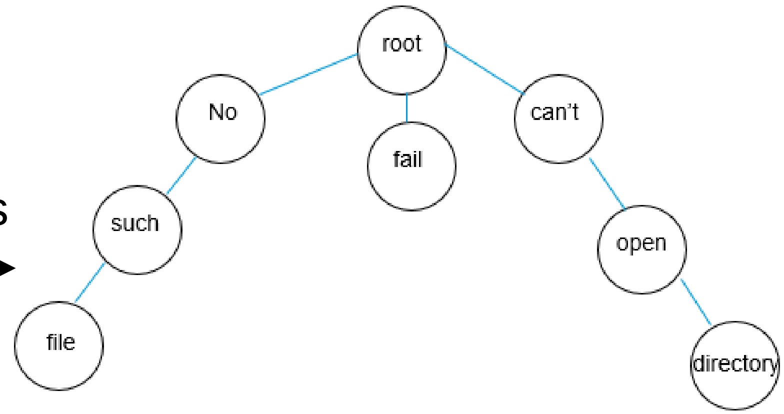


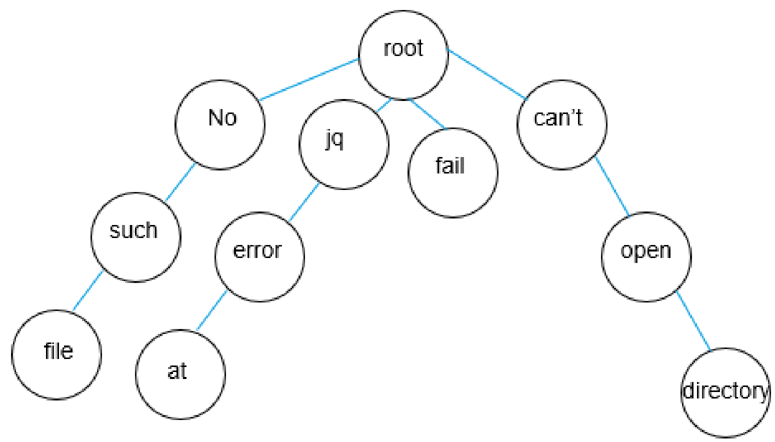
Prefix Tree

Template	Variables
jq: error (at <stdin>:<*>): null (null) has no keys	45
jq: error (at <stdin>:<*>): null (null) has no keys	21

Incremental Prefix Tree

jq: error (at <stdin>:<*>): null (null) has no keys

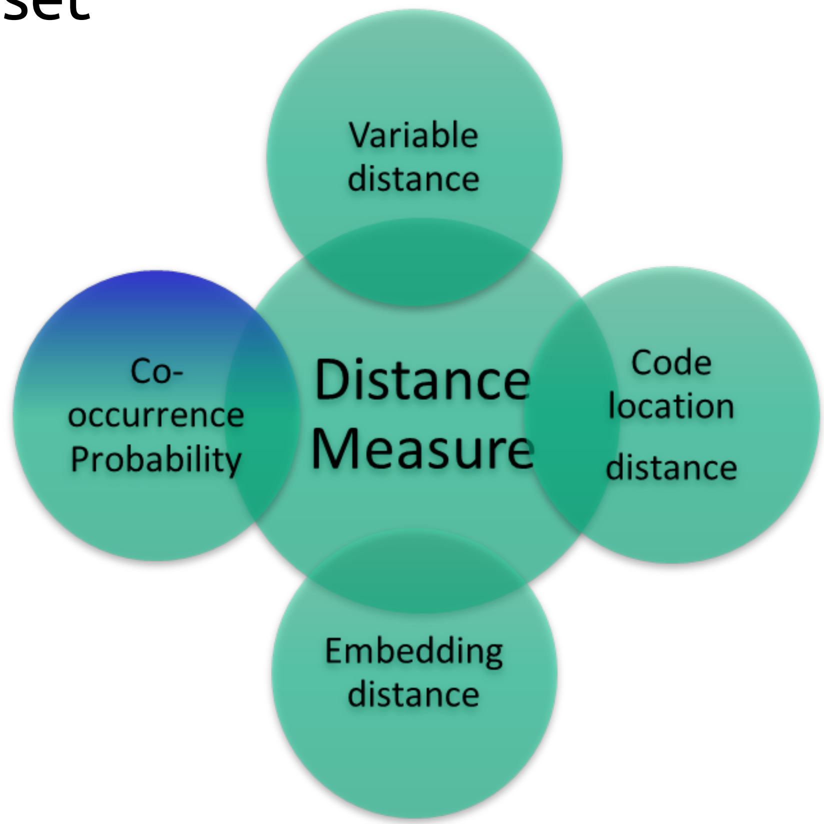




Template	Variables
jq: error (at <stdin>:<*>): null (null) has no keys	--

Co-occurrence Probability

- A log file is an ordered sequence of logs
  - Sequential information
- Representing logs as sequence set
  - Each sequence is a log file
  - Each event is a log



Sequence set

- Each log template is mapped to a unique number

```
1 0 1 2 2 3 3 3 4 5 6 7 8 9 2 2 2 10 2 2 2 2 2 10 10 10 10 10 10 11 12 10 10 10 13 13 13 13 13 13 13 13 13 13 13
2 14 14 14 14 15 14 15 15 15 15 15 14 15 15 15 16 16 17 18 19 15 20 19 14 14
3 0 1 2 2 3 3 3 4 5 6 7 8 9 2 2 2 10 2 2 2 2 2 10 10 10 10 10 10 21 10 10 10 13 13 13 13 13 13 13 13 13 13
4 14 22 14 22 22 12 23 14 14 14 14 14 14 14 14 22 22 22 22 22 22 24 25 25 25 25 25 25 25 25
5 26 26 26 26 27 28 26 26 29 30 13 13 13 13 13 13 13 13 13 13
6 31 31 32 32 32 32 32 32 32 32 32 32 33 33 33 33 33 33 33 33 33 34 35 36 34 34 37 34 34 37 34 34 34 34 34 37 38 39 38 39 38 39 40 41 42 43 44
7 14 14 14 14 14 14 14 14 14 14 45 45 46 47 48 16 16 49 47 48 16 16 49 47 48 16 16 49 50 50 25 25 25 25 25 25 25 25 25
8 26 26 27 27 27 26 26 26 26 51 51 38 39 52 53 38 39 40 41 42 43 13 13 13 13 54
9 26 26 26 26 27 28 26 26 55 56 29 57 30 13 13 13 13 13 13 13 13 13 13 13
10 0 58 59 60 61 62 39 59 60 63 64 65 66 40 41 67
11 14 14 68 14 14 69 14 14 14 70 14 14 71 72 14 70 70 70 16 16 73 74 75 25 25 25 25
12 14 14 14 14 76 14 14 14 14 14 14 77 78 77 78 47 48 16 16 79 80 47 48 16 16 79 47 48 16 16 79 23 50 50 25 25 25 25 25 25 25 25 25 25
13 0 1 2 2 3 3 3 4 5 6 7 8 9 2 2 2 10 2 2 2 2 2 10 10 10 10 10 11 81 10 10 10 13 13 13 13 13 13 13 13 13 13
14 14 14 14 14 14 14 14 14 14 14 82 82 79 82 82 79 16 16 79 83 16 16 79 47 48 16 16 79 47 48 16 16 79 47 48 16 16 79 79 79 25 25 25 25
15 14 14 23 14 14 84 14 14 14 14 14 14 16 16 79 16 16 79 47 48 16 16 79 16 16 79 47 48 16 16 79 47 48 79 79
16 85 86 85 86 85 86 85 86 85 86 85 86 85 86 85 86 85 86 85 86 45 45 87 45 45 45 45 45 45 45 45 88 89 90 13
17 26 26 27 27 27 26 26 26 26 91 92 93 93 93 94 95 95 95 95 95 57 95 95 95 95 95 57 57 30 13 13 13 13 13 13 13 13 13 13
18 26 26 26 26 27 28 26 26 57 96 96 97 38 39 52 53 38 39 40 41 42 43 13 13 13 13 13 13 13 13 13 13 54
19 14 22 14 22 22 14 14 14 14 14 14 14 14 22 22 22 22 22 22 22 82 82 49 82 82 49 16 16 49 98 24 16 16 49 47 48 16 16 49 47 48 16 16 49 47 48 16 16 49 50 50 25 25
20 99 99 0 100 40 41 13 13 13 13 13 13 13 13 13 13 101 54
21 26 26 27 27 27 26 26 26 26 102 102 102 102 102 102 102 102 102 102 103 104 38 39 52 53 38 39 40 41 42 43 44
22 26 26 27 27 27 26 26 26 26 105 106 106 104 38 39 52 53 38 39 40 41 42 43 44
23 107 107 107 107 107 107 107 107 107 107 108 109 110 23 86 23 86 87 86 85 86 23 86 23 86 85 86 85 86 85 86 85 86 85 85 85 85 85 111 112 113 13 13 13 13 13 13 13 13 13 13
24 0 107 114 115 65 40 41 13 13 13 13 13 13 13 13 13 13 54
25 0 114 58 59 116 117 118 118 118 118 118 118 118 118 118 119 119 119 119 119 119 119 119 39 39 119 119 39 39 59 63 64 116 120 121 121 121 121 121 121 39 63 64 116 117 122 123 124 63 64 59 116
   125 126 125 126 125 125 125 125 125 125 126 125 125 126 126 126 126 126 126 126 108 127 39 39 59 63 64 128 129 40 41 13 13 13 13 13 13 13 13 13 13 54
26 14 14 14 14 76 14 14 14 14 14 14 16 16 79 130 130 79 16 16 79
27 26 26 27 27 27 26 26 26 26 131 131 131 131 132 131 132 131 132 132 131 132 132 131 132 132 131 132 132 103 38 39 131 133 134 135 136 135 136 135 136 135 136 135 136 135 136 135 103 38 39 133 134
   136 135 136 135 136 136 103 38 39 52 53 38 39 40 41 42 43 44
28 0 137 138 139 117 63 140 65 141 142 143 144 145 13 13 13 13 13 13 13 13 13 13 101 54
29 26 26 26 26 27 28 26 26 55 56 57 51 51 51 51 51 97 38 39 52 53 38 39 40 41 42 43 13 13 13 13 13 13 13 13 13 13 54
30 14 14 14 14 14 14 14 14 14 14 45 45 47 48 16 16 79 47 48 16 16 79 47 48 16 16 79 23 146 146
31 14 14 14 14 14 14 14 14 14 14 16 16 79 147 16 16 79 148 47 48 16 16 79 47 48 16 16 79 47 48 16 16 79 79 79 79 25 25
32 0 149 150 13 13 13 13 13 13 13 13 13 13
33 14 22 14 22 22 151 151 14 14 14 14 14 14 14 14 22 22 22 22 22 22 22 16 16 49 47 48 16 16 49 47 48 16 16 49 47 48 16 16 49 16 16 49 24 49 73 50 50 25 25
34 0 107 114 152 117 153 39 154 155 63 156 152 117 153 39 63 156 128 129 40 41 13 13 13 13 13 13 13 13 13 13 54
35 157 39 158 159 40 41 13 13 13 13 13 13 13 13 13 13 160 54
36 14 14 14 14 14 14 14 14 14 14 16 16 79 161 47 48 16 16 79 47 48 16 16 79 16 16 79 47 48 16 16 79 79 79 25 25 50 50 25 25 25 25 25 25 25 25
37 162 0 107 114 13 13 13 13 13 13 13 13 13 13
38 0 107 114 59 60 117 163 164 163 39 39 39 39 59 60 63 64 65 66 40 41 13 13 13 13 13 13 13 13 13 13 54
39 14 22 12 12 12 12 12 12 12 12 165 14 22 22 14 22
40 14 22 14 22 22 151 14 14 14 14 14 14 14 14 22 22 22 22 22 22 22 166 47 48 16 16 49 47 48 16 16 49 146 146
41 14 14 14 14 76 14 14 14 14 14 14
42 14 14 14 14 14 14 14 14 14 14 82 82 79 82 82 79 167 16 16 79 16 16 79 168 168 47 48 16 16 79 47 48 16 16 79 16 16 79 47 48 79 79 79
43 169 169 169 170 44
44 14 14 14 14 14 14 14 14 14 14 171 47 48 16 16 79 47 48 16 16 79 47 48 16 16 79 50 50 25 25 25 25 25 25 25 25 25 25
45 26 26 27 27 27 26 26 26 26 91 92 93 93 93 95 95 95 95 95 57 95 95 95 95 95 57 57 57 172 51 97 38 39 52 53 38 39 40 41 42 43 13 13 13 13 13 13 13 13 13 13 54
46 26 26 27 27 27 26 26 26 26 136 173 38 39 136 174 38 39 175 175 174 38 39 52 53 38 39 40 41 42 43 44
47 176 176 176 176 176 176 176 176 176 176 177 178 177 177 179 177 178 177 177 179 177 178 177 177 179 177 178 180 181 145 27 13 13 13 13 13 13 13 13 13 13 182 54
48 0 114 58 59 60 117 183 39 39 39 59 60 63 64 65 66 40 41 13 13 13 13 13 13 13 13 13 13 54
49 14 22 23 14 22 76 22 23 14 14 14 14 14 14 14 14 22 22 22 22 22 22 22 184 47 48 16 16 49 47 48 16 16 49 47 48 16 16 49 24 185 23 50 50 25 25 25 25 25 25 25 25 25 25
50 14 14 186 16 16 79 187 14 14 14 14
51 14 14 23 14 14 76 14 14 14 14 14 14 188 189 190 191 192 189 190 192 189 190 192 47 48 16 16 79 47 48 16 16 79 47 48 16 16 79
52 14 14 14 14 14 14 14 14 14 14 188 189 193 191 192 188 189 193 191 192 188 189 193 191 192 189 193 192 189 193 192 189 193 192 189 193 192 47 48 16 16 79 189 193 192 189 193 192 189 193 192 47 48 16 16 79 47
   16 79 23
53 194 195
54 0 107 114 196 117 107 63 115 65 141 40 41 13 13 13 13 13 13 13 13 13 13 54
55 197 198
```

Co-occurrence Probability

- Considering the time window T:
  - Linear probability
  - Non linear probability
- Suppose
  - M_{ij}: *the number of co occurrence time intervals,*
  - N_i: *the number of time intervals with logj*
  - $CP[i][i] = 1$
  - $CP[i][j] = CP[j][i] = Max(CP[i][j], CP[j][i])$

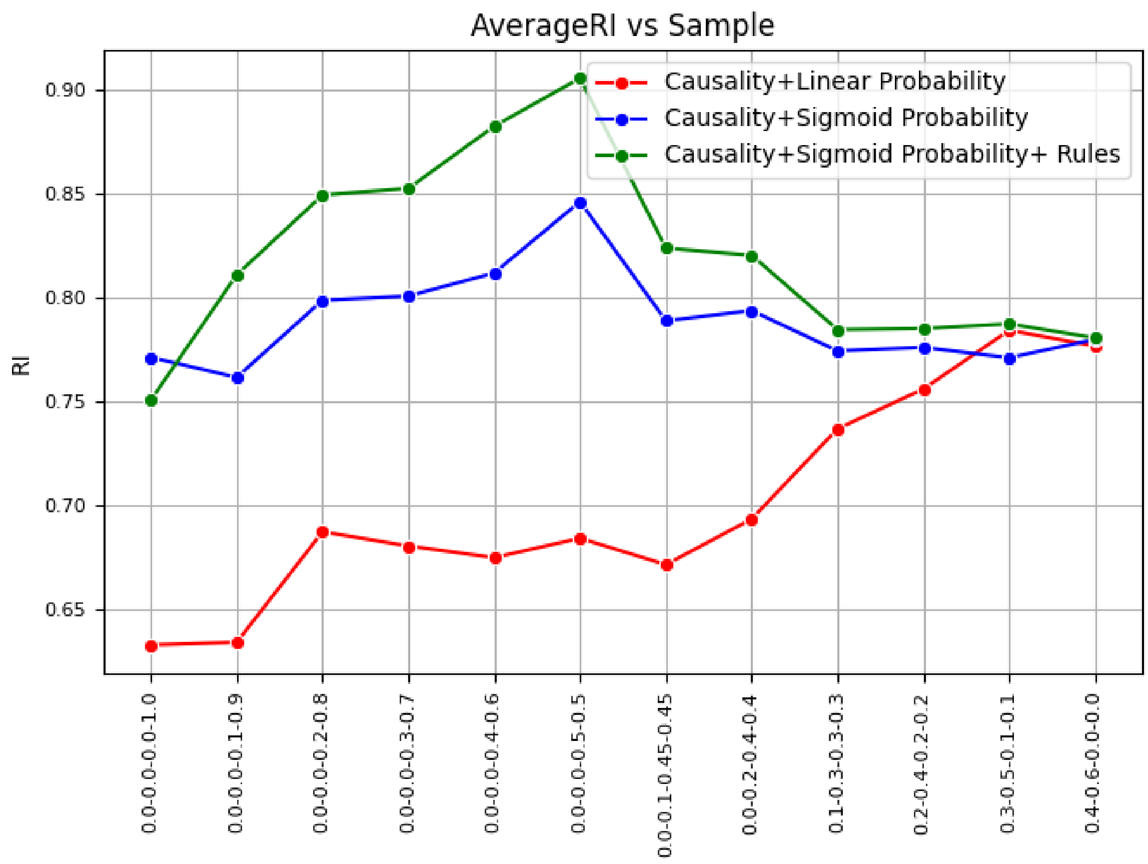
- Linear : $CP[i][j] = \frac{M_{ij}}{N_j}$,
- Non linear: $CP[i][j] = \frac{1}{1+e^{-X}}$ $, x = k\left(M_{ij} - \frac{N_i}{2}\right), k = 1$
-

Co-occurrence Probability Matrix

1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.8	0.4	0.12	0.84	0	0	0	0	0	0	0	0.24	0	0	0	0	0	0	0
0	0.75	0	0.25	0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.0286	0	0.8	0.0571	0.4286	0	0.0286	0	0	0	0	0	0.2571	0.0286	0	0	0.2	0.1143	0.142
0	0	0	0	0.8966	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.0055	0	0.8866	0.0066	0.4262	0.0407	0.0352	0.0286	0.0011	0	0.1828	0.0441	0.011	0.0242	0	0	0
0	0	0	0	0	0.9	0.9	0.8	0.7	0.6	0.6	0.4	0	0	0	0	0	0	0	0
0	0	0	0.0032	0	0.0317	0.0032	0.8508	0.1032	0.0921	0.1016	0.0032	0	0.0063	0.0952	0.054	0.1746	0	0	0
0	0	0	0	0	0.1053	0.0263	0.6053	0.5789	0.9474	0.8947	0.0263	0	0	0.1053	0.0263	0	0	0	0
0	0	0	0	0	0.1111	0.0278	0.6667	0.6111	0.5833	0.8889	0.0278	0	0	0.1111	0.0278	0	0	0	0
0	0	0	0	0	0.0652	0.0217	0.6087	0.5217	0.4783	0.6957	0.0217	0	0	0.1957	0.0217	0	0	0	0
0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.0238	0	0.3095	0	0.4643	0.0417	0.0417	0.0357	0	0	0.869	0.1369	0.25	0.0417	0	0	0
0	0	0	0	0.0822	0.3014	0	0.0411	0	0	0	0	0	0.137	0.3014	0	0.1233	0	0	0
0	0	0	0	0	0	0	0.6364	0	0	0	0	0	0	0.0909	0.2273	0.2273	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0988	0	0.8663	0	0	0
0	0	0	0.0294	0.1397	0	0	0	0	0	0	0	0	0	0.0699	0	0	0.8346	0.4007	0.264
0	0	0	0.0122	0.0854	0	0	0	0	0	0	0	0	0	0.0244	0	0	0.939	0.5122	0.207
0	0	0	0.0556	0.2778	0	0	0	0	0	0	0	0	0	0.1111	0	0	1	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.9091	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Experimental Results

- Co-occurrence probability
  - Linear
  - Non linear
- removing noise



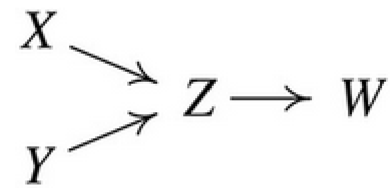
Conditional Independence

- Co-occurrence does not necessarily mean causality
- Buying cold water
- Buying ice cream
- Hot weather

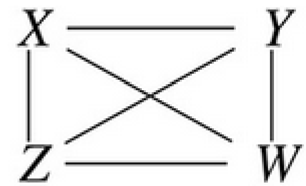
- Conditional Independency ($X \perp Y | Z$)
- Statistical test, chi-square

PC Algorithm

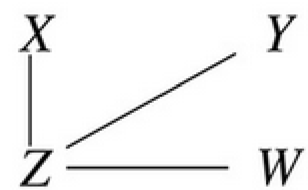
- Suppose we have 4 variable {X,Y,Z,W}



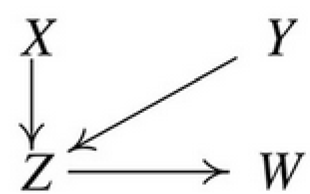
- Initialize a fully connected undirected graph



- Remove edges based on conditional independence tests
  - If X ⊥ Y | Z, remove the edge between X and Y



- Orient Edges to Form a Directed Graph



Chi-square Test

- Step1: create the contingency tables
- Suppose A⊥B|C

Group 1: $C = \text{Present}$

Event A	Event B	Frequency
Present	Present	1
Present	Absent	1
Absent	Present	1
Absent	Absent	1

Group 2: $C = \text{Absent}$

Event A	Event B	Frequency
Present	Present	1
Present	Absent	1
Absent	Present	1
Absent	Absent	2

Chi-square Test

- Calculate Expected Frequencies

$$\text{Expected Frequency} = \frac{(\text{Marginal Total of A}) \times (\text{Marginal Total of B})}{\text{Total Number of Observations}} = \frac{2 \times 2}{4} = 1$$

- Step 3: Compute Chi-Square Statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- If the chi-square statistic exceeds the critical value, we reject the null hypothesis and conclude that A and B are not independent given C.

Data Representation

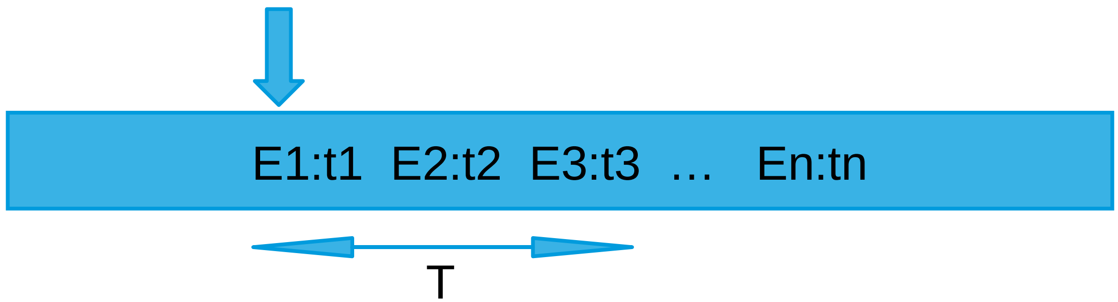
- Considering each log template as a variable to form a data frame

Seq	A	B	C	D
ABC	1	1	1	0
DC	0	0	1	1
ACD	1	0	1	1

- Time series:
  - For each log in each log file
  - The order and time of events are critical
  - Highly sparse
  - Statistical approaches rely on sufficient overlap and density in the data

Hybrid Representation

- The order is considered but not the absolute time



- Pair: (E1,E2) , (E1,E3), (E2,E3), $(E1 \perp E3 \mid E2)$
- Conditional independence test
- 92% accuracy with Ciena log files