Comparative Causal Analysis of Network Log Data in Two Large ISPs

Satoru Kobayashi, Keiichi Shima, Kenjiro Cho, Osamu Akashi, Kensuke Fukuda

AnNet 2022
Toward automated network log analysis

• Automated log analysis
  – Necessary for recent large-scale networks and their logs
  – Especially important for root cause analysis of network troubles

• Machine learning approach for network root cause analysis
  – Lack of diversity in training data
  – Weak for unknown trouble cases

➢ Collaborative (inter-ISP) log analysis
Future collaborative log analysis

• Collaborative log analysis of multiple ISPs
  – Learn more (and diverse) troubles
  – Can be effective for unknown troubles (If appeared in other ISPs)
Difficulty in collaborative log analysis

• Does there really exist transferable knowledge?
  – If not, transfer learning loses accuracy and reliability
• We need to preliminarily compare multiple log datasets
  – To examine the transfer learning is effective or not in advance

Research goal: Propose a comparative log analysis technique of ISPs
Challenges for comparative log analysis

• Difference of environment (vendors and network topology)
  – Different system behavior
  – Different log formats and variables
  – Different logging behavior (when to log)
    ➢ Difficult to compare directly

• Data publication policy of ISPs
  – Network logs include sensitive information
    ➢ Need anonymization
Key idea

• Log messages -> Time-series event with log templates
  – Time-series event: same log template, same host device
  – Helpful for anonymization

• Focus on event causality
  – If there is a same network behavior, there can also be similar causal
    relations of log events [1]
  – Clear and direct relations without spurious correlation

  Management, pp.53-67, vol.15, no.1, March, 2018
Analysis flow overview

ISP A

Log → Log template generation → Log, Time-series → Preprocess → Causal discovery → 1 DAG for 1-day log data

ISP B

Log → Log template generation → Log, Time-series → Preprocess → Causal discovery

More comparable

1 time-series node: events with 1 log template from 1 host device
(1) Log template generation

1. Parse log messages into header information and statements
2. Generate log templates from log statements
3. Classify log messages with the templates

No variables
-> Anonymized

(2) Preprocessing of input time-series nodes

- Decrease processing time of causal discovery
  - Remove periodic component of log time-series [1]
    - Ignore daily or weekly (planned) system behavior
  - Prune causal edge candidates with prior knowledge [3]
    - Considering network topology and protocol layers of events
  - Merge completely synchronizing time-series nodes (new)
    - Decrease the number of input nodes

- These three methods can be used together

(3) Causal discovery

- **PC algorithm** [4]
  - Relatively fast causal discovery method (available for large dataset)
  - With G square test (for binary time-series)

Evaluation outline

- Datasets
- Validation of node-merging preprocessing
- Evaluation (Comparative causal analysis of ISPs)
  - Causal analysis results
  - Details of Circuit-related causal edges
  - Case study
Datasets

• ISP A
  – nation-wide ISP in Japan
  – 56,968,361 log lines
  – 92 days
  – 1,861 hostnames
  – 36 corresponding trouble tickets
  – 5,182 log templates
  – NOT using preprocessing of prior knowledge

• ISP B
  – nation-wide ISP in Japan
  – 34,722,785 log lines
  – 365 days
  – 131 hostnames
  – 88 corresponding trouble tickets
  – 1,789 log templates
  – Using preprocessing of prior knowledge
Validation of node-merging preprocessing

- Causal results of ISP B without/with node merging
  - (each value is the average of every 1-day log data)
  - Processing time: 76.0 sec -> 36.3 sec (52% decreased)
  - Number of nodes: 360.1 -> 279.4
  - Number of edges: 56.8 -> 49.8
  - Corresponding trouble tickets: 70/88 -> 71/88

Node merging enables:
- Faster calculation
- More reliable causal results
### Evaluation - Comparative causal analysis of two ISPs

<table>
<thead>
<tr>
<th>Network</th>
<th>#Nodes</th>
<th>#Edges</th>
<th>#Tickets</th>
<th>% of Tickets</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISP A</td>
<td>2,758.3</td>
<td>349.8</td>
<td>18</td>
<td>42%</td>
</tr>
<tr>
<td>ISP B</td>
<td>279.4</td>
<td>49.8</td>
<td>71</td>
<td>81%</td>
</tr>
</tbody>
</table>

### Classification of tickets

<table>
<thead>
<tr>
<th>Network</th>
<th>Class</th>
<th>#All tickets</th>
<th>#Tickets with edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISP A</td>
<td>Circuit</td>
<td>22</td>
<td>15 (68%)</td>
</tr>
<tr>
<td></td>
<td>Connection</td>
<td>7</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>Device</td>
<td>7</td>
<td>3 (43%)</td>
</tr>
<tr>
<td>ISP B</td>
<td>Circuit</td>
<td>22</td>
<td>14 (63%)</td>
</tr>
<tr>
<td></td>
<td>Connection</td>
<td>55</td>
<td>50 (91%)</td>
</tr>
<tr>
<td></td>
<td>Device</td>
<td>7</td>
<td>4 (57%)</td>
</tr>
<tr>
<td></td>
<td>Blackout</td>
<td>4</td>
<td>3 (75%)</td>
</tr>
</tbody>
</table>
### Evaluation - Details of Circuit-related causal edges

Aggregated with adjacent nodes of causal edges

<table>
<thead>
<tr>
<th>Network</th>
<th>Node label</th>
<th>Days w/ logs</th>
<th>Days w/ edges</th>
<th>Days w/ tickets (edges/tickets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISP A</td>
<td>MPLS</td>
<td>88</td>
<td>69</td>
<td>12 (17%)</td>
</tr>
<tr>
<td></td>
<td>System</td>
<td>92</td>
<td>92</td>
<td>5 (5%)</td>
</tr>
<tr>
<td></td>
<td>Interface</td>
<td>92</td>
<td>92</td>
<td>5 (5%)</td>
</tr>
<tr>
<td></td>
<td>Monitor</td>
<td>90</td>
<td>53</td>
<td>4 (4%)</td>
</tr>
<tr>
<td></td>
<td>OSPF</td>
<td>61</td>
<td>5</td>
<td>1 (20%)</td>
</tr>
<tr>
<td>ISP B</td>
<td>Monitor</td>
<td>191</td>
<td>60</td>
<td>10 (17%)</td>
</tr>
<tr>
<td></td>
<td>MPLS</td>
<td>39</td>
<td>13</td>
<td>4 (31%)</td>
</tr>
<tr>
<td></td>
<td>BGP</td>
<td>315</td>
<td>291</td>
<td>4 (1%)</td>
</tr>
<tr>
<td></td>
<td>Interface</td>
<td>318</td>
<td>211</td>
<td>3 (1%)</td>
</tr>
<tr>
<td></td>
<td>OSPF</td>
<td>54</td>
<td>1</td>
<td>1(100%)</td>
</tr>
</tbody>
</table>

**MPLS, Interface, Monitor:** Found in many days
- Regular behavior

**OSPF:**
- Logs regularly appear, but causality is rare
- Anomalous if OSPF has causality with others
### AnNet 2022

#### Another aggregation of causal edges related to Circuit tickets

<table>
<thead>
<tr>
<th>Network</th>
<th>Label 1</th>
<th>Label 2</th>
<th>Same host w/ edges</th>
<th>Days w/ tickets (edges/tickets)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ISP A</strong> (92 days)</td>
<td>MPLS</td>
<td>MPLS</td>
<td>✓</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>System</td>
<td>System</td>
<td>✓</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>MPLS</td>
<td>MPLS</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Monitor</td>
<td>Monitor</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>System</td>
<td>System</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Interface</td>
<td>Interface</td>
<td></td>
<td>59</td>
</tr>
<tr>
<td>Monitor</td>
<td>OSPF</td>
<td>✓</td>
<td>1</td>
<td>1 (100%)</td>
</tr>
<tr>
<td>Interface</td>
<td>OSPF</td>
<td>✓</td>
<td>1</td>
<td>1 (100%)</td>
</tr>
<tr>
<td>Interface</td>
<td>Monitor</td>
<td>✓</td>
<td>1</td>
<td>1 (100%)</td>
</tr>
<tr>
<td>Monitor</td>
<td>OSPF</td>
<td>✓</td>
<td>1</td>
<td>1 (100%)</td>
</tr>
<tr>
<td>Interface</td>
<td>OSPF</td>
<td>✓</td>
<td>1</td>
<td>1 (100%)</td>
</tr>
<tr>
<td>Interface</td>
<td>Monitor</td>
<td>✓</td>
<td>1</td>
<td>1 (100%)</td>
</tr>
<tr>
<td>System</td>
<td>MPLS</td>
<td>2</td>
<td>2 (100%)</td>
<td></td>
</tr>
<tr>
<td>System</td>
<td>Interface</td>
<td>3</td>
<td>1 (33%)</td>
<td></td>
</tr>
<tr>
<td>Interface</td>
<td>Monitor</td>
<td>✓</td>
<td>1</td>
<td>1 (100%)</td>
</tr>
<tr>
<td>Monitor</td>
<td>Monitor</td>
<td>✓</td>
<td>1</td>
<td>1 (100%)</td>
</tr>
<tr>
<td>Monitor</td>
<td>MPLS</td>
<td>1</td>
<td>1 (100%)</td>
<td></td>
</tr>
<tr>
<td><strong>ISP B</strong> (365 days)</td>
<td>Monitor</td>
<td>Monitor</td>
<td>✓</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>BGP</td>
<td>BGP</td>
<td>✓</td>
<td>215</td>
</tr>
<tr>
<td>Monitor</td>
<td>MPLS</td>
<td>✓</td>
<td>5</td>
<td>4 (80%)</td>
</tr>
<tr>
<td>Interface</td>
<td>Interface</td>
<td>✓</td>
<td>166</td>
<td>3 (2%)</td>
</tr>
<tr>
<td>Monitor</td>
<td>BGP</td>
<td>✓</td>
<td>3</td>
<td>1 (33%)</td>
</tr>
<tr>
<td>Monitor</td>
<td>MPLS</td>
<td>✓</td>
<td>1</td>
<td>1 (100%)</td>
</tr>
<tr>
<td>MPLS</td>
<td>MPLS</td>
<td>✓</td>
<td>1</td>
<td>1 (100%)</td>
</tr>
<tr>
<td>MPLS</td>
<td>MPLS</td>
<td>✓</td>
<td>3</td>
<td>1 (33%)</td>
</tr>
<tr>
<td>OSPF</td>
<td>OSPF</td>
<td>✓</td>
<td>1</td>
<td>1 (100%)</td>
</tr>
</tbody>
</table>

#### Aggregated with node pairs of causal edges

- **Edges between same labels**
  - (Within a protocol function)
  - ➔ Relatively frequent (regular) but sometimes related to troubles

- **Edges between different labels**
  - (Communication between protocols)
  - ➔ Anomalous and related to troubles

Mainly adjacent to subordinate functions *(Interface and Monitor)*
Evaluation - Case study

One of Circuit troubles in ISP B

Edges across devices between different labels (Rare and large behavior)

Found similar edge in ISP A too

-> Similar behavior in different ISPs
Discussion

• Causal approach is effective for dataset comparison
  – Logs appear regularly in any classes -> Which to focus?
  – Log causality can reveal large and relational behaviors

• How about other tickets (Connection and Device)?
  – Difficult to compare at least between these ISPs
  – Connection: Depends largely on used network protocols
  – Device: Depends largely on device vendors and models
Conclusion

• Goal: Comparative log analysis between different ISPs
• Approach: Causal discovery of time-series events classified with log templates and host devices
• Performance: Improved with node-merging by decreasing 52% of the processing time
• Result: Contribute to finding similar behaviors in two ISPs (especially on Circuit-related troubles)
• https://github.com/amulog/logdag