# Mining causes of network logs in log data with causal inference

Satoru Kobayashi, Kensuke Fukuda, Hiroshi Esaki
The University of Tokyo / NII
IFIP/IEEE IM 2017, Lisbon, Portugal
May 9, 2017

# Difficulty of leveraging system log

- Huge dataset
  - Large scale and complicated systems
  - 120,000 lines / day in SINET4
    - An academic network in Japan
  - Automated analysis required
- Difficult to analyze automatically
  - Discrete and sparse
  - Loss contextual information with simple approach

# Related works

- Anomaly / change point detection
- Fault localization
- Root cause analysis
  - Heuristic-based [1]
  - Causal inference [2,3,4]

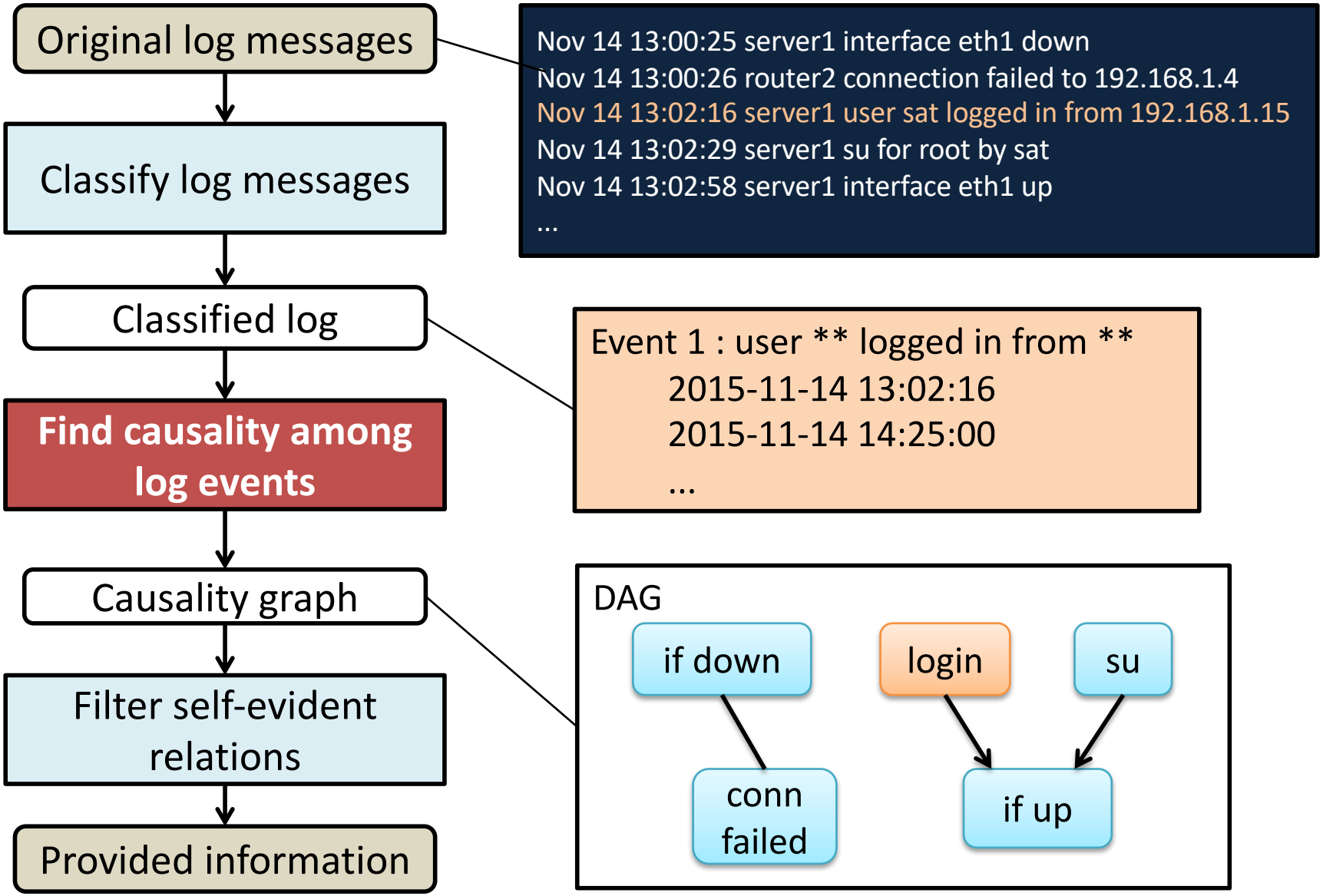
-> Causal graph approach based on causal inference

[1] B. Tak et al. "LOGAN: Problem Diagnosis in the Cloud Using Log-Based Reference Models," in IEEE IC2E, 2016, pp. 62-67.
[2] Z. Zheng et al. "3-Dimensional root cause diagnosis via co-analysis," in ACM ICAC, 2012, pp. 181.
[3] K. Nagaraj et al. "Structured Comparative Analysis of Systems Logs to Diagnose Performance Problems," in NSDI, 2012, pp. 1–14.
[4] A. Mahimkar et al. "Towards automated performance diagnosis in a large iptv network," in ACM SIGCOMM, 2009, pp. 231–242.

# Goal

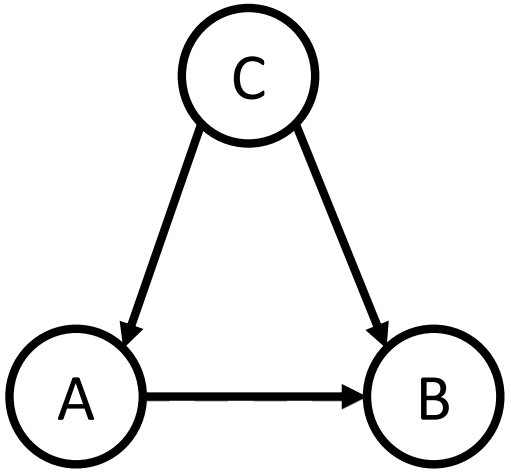
- Extract causality of events in system logs
  - Based on causal inference
  - Present as Directed Acyclic Graph (DAG) to pinpoint root causes
  - Across multiple devices
- Provide useful information for system management and troubleshooting
  - Enough lean for operators to read

# System architecture



# Causal inference

- Conditional independence
  - Remove redundant edges
  - A and B are conditionally independent given C
- DAG estimation
  - Recursive search of conditional independence
  - -> PC algorithm [5]

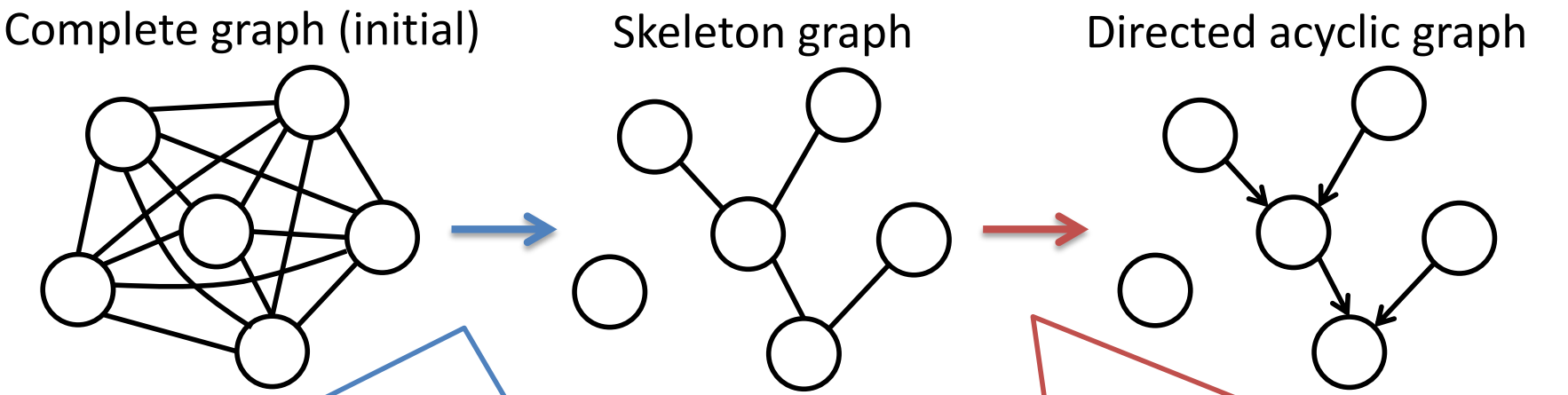



$$P(A|C)P(B|C) = P(A, B|C)$$

[5] P. Spirtes et al. "An algorithm for fast recovery of sparse causal graphs", Social science computer review, vol. 9, pp. 62–72, 1991.

# Causation mining

- PC algorithm [5]



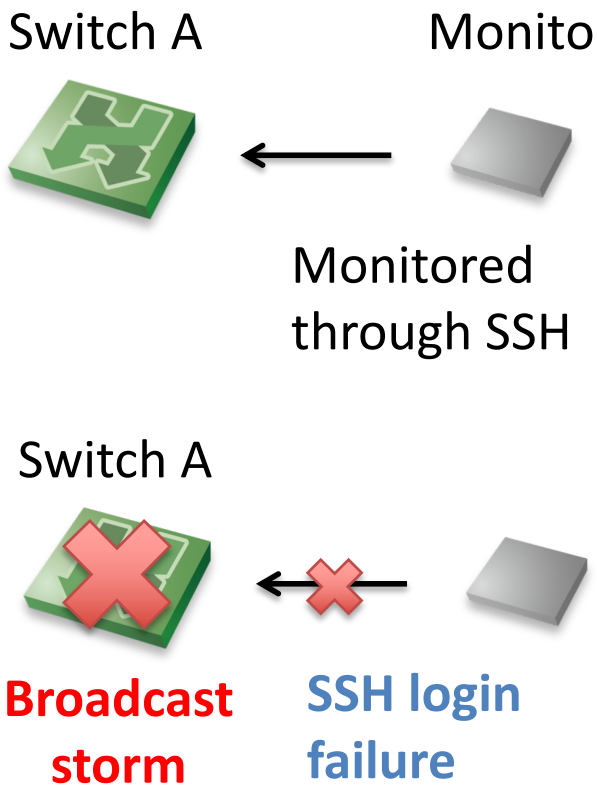

- Remove edges of conditional independence
- G square test [6]
  - Statistical test for conditional independence
  - Use binary or multi-level data as input

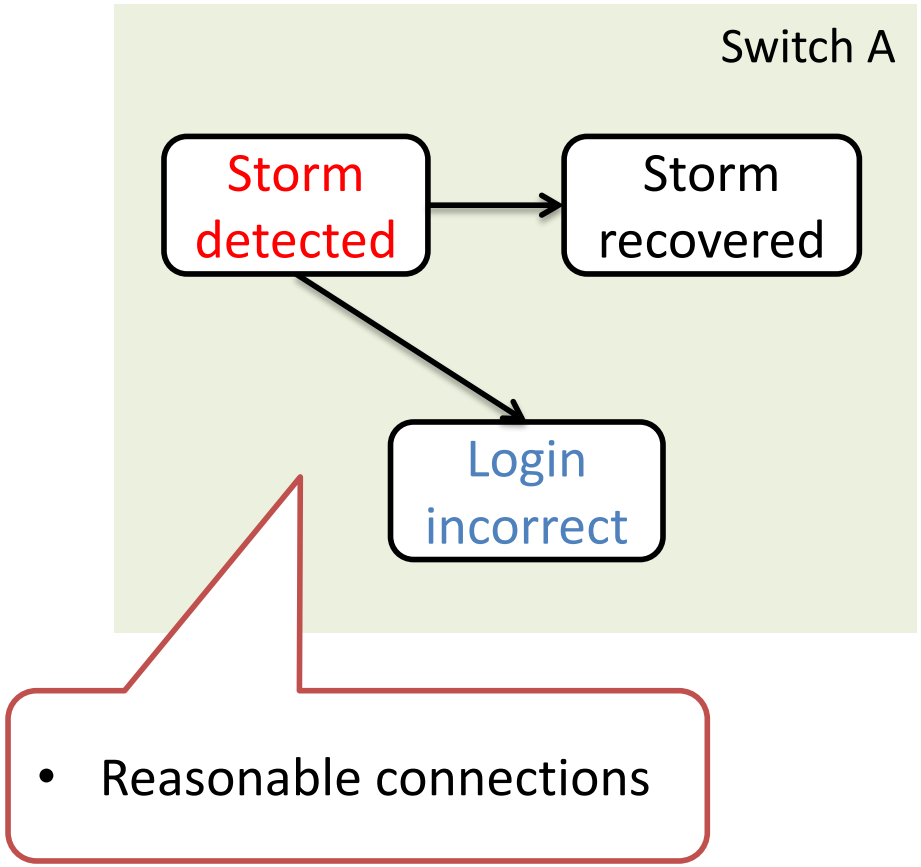
[5] P. Spirtes et al. "An algorithm for fast recovery of sparse causal graphs", Social science computer review, vol. 9, pp. 62–72, 1991.
[6] R. E. Neapolitan. "Learning Bayesian Networks." Prentice Hall Upper Saddle River, 2004.

# Results

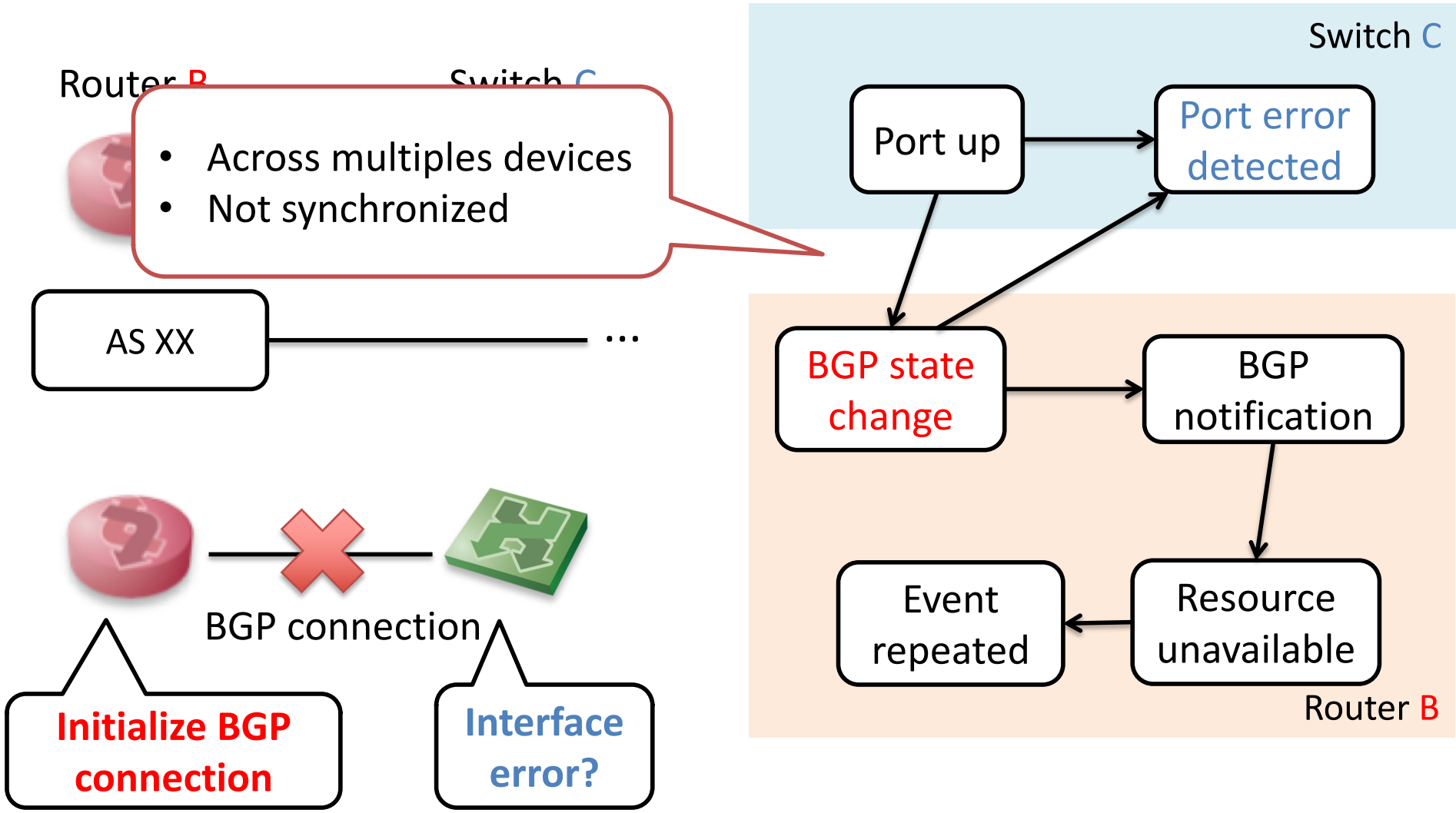
- Dataset
  - Syslog data of SINET 4
    - More than 100 network devices (switches and routers)
  - 35 million lines (15 months)
    - 4% remain after preprocessing
  - Classified with 1,414 log templates
- Results
  - 8,613 edges (causal relations) detected
  - **1,548 edges identified as important**
    - 3.4 edges / day

# Case study: Broadcast storm

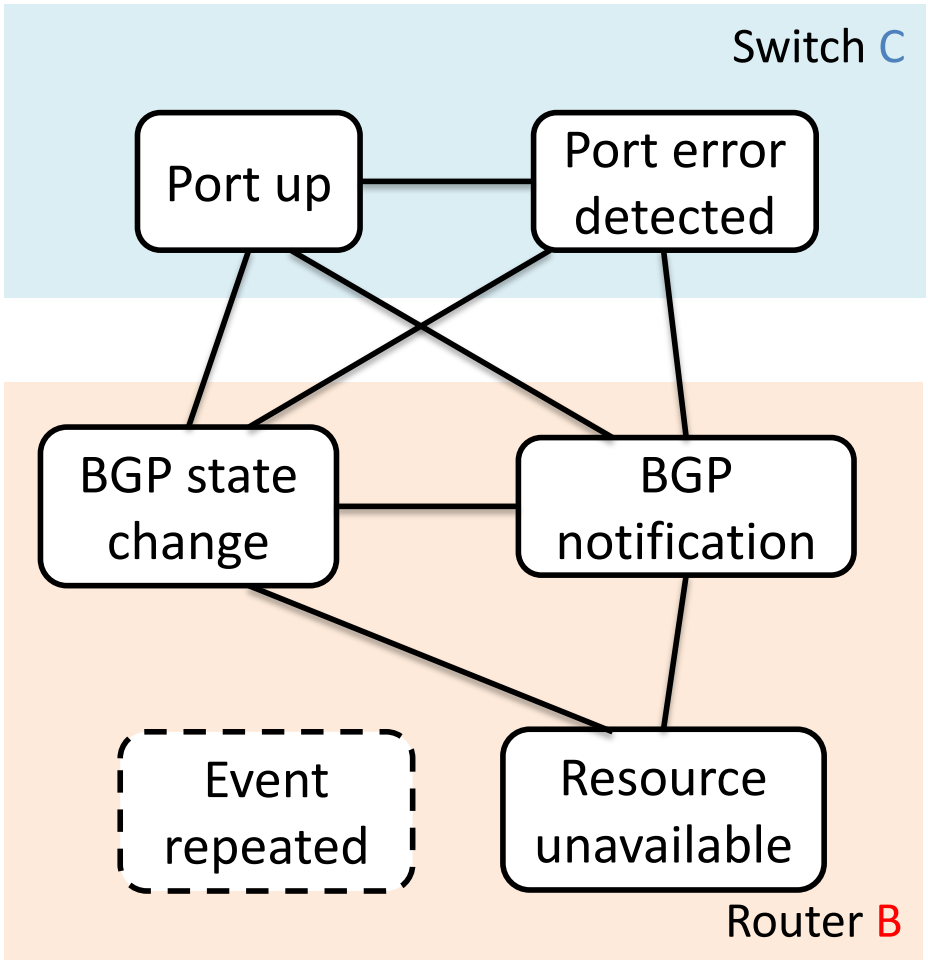


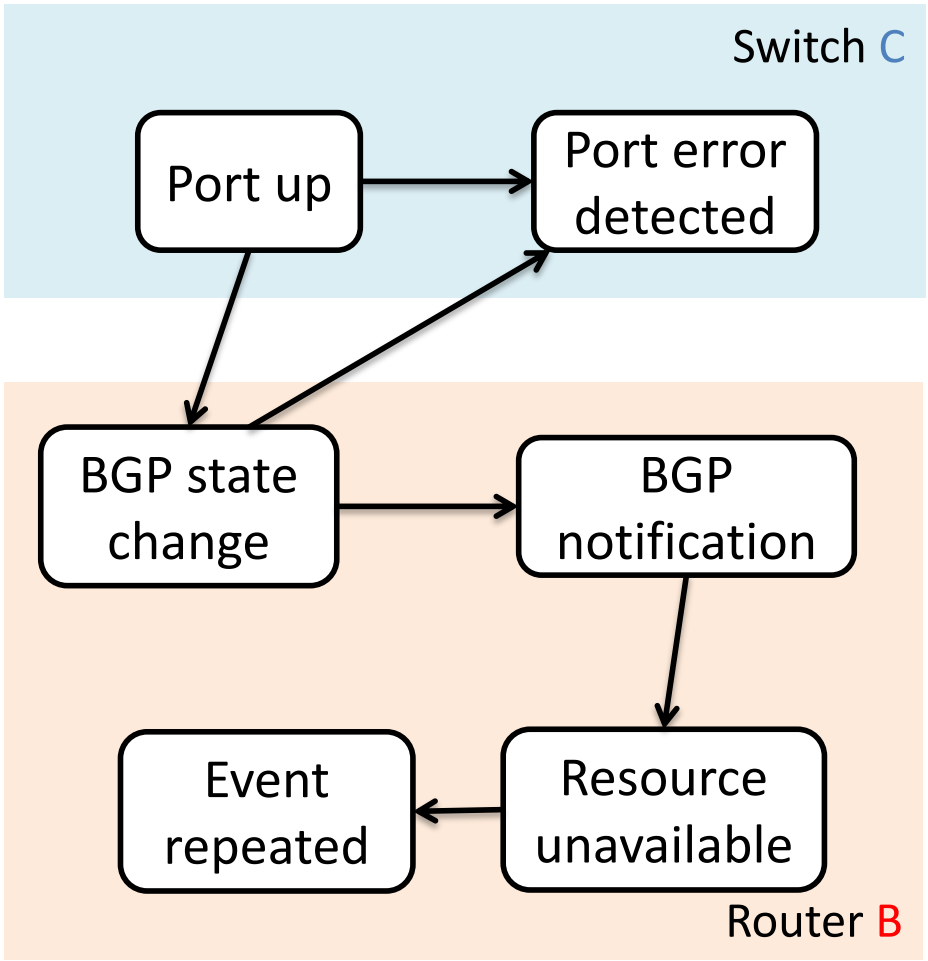
# Case study: BGP state initialization



# Case study: BGP state initialization




Correlation-based method




PC algorithm

# Comparison with trouble tickets

- Trouble tickets as a ground truth
  - 4 tickets with related messages in 1 month
  - Some descriptive edges found

| Ticket | Description | Detected edges |
|---|---|---|
| T1 | Disconnected device | 3 |
| T2 | Disconnected device | 3 |
| T3 | Hardware module failure | 2 |
| T4 | Polled I/O error | 0 |

Events of BGP connection state changes

Events appear only once

# Concluding remarks

- Extract causality of events in network logs
  - Use PC algorithm
- Evaluate with log data of actual large-scale network
  - Provide useful information for actual troubles
  - Effective for reported tickets

**Thank you for listening!**