A Quantitative Causal Analysis for Network Log Data

Richard Jarry[1], Satoru Kobayashi[2], Kensuke Fukuda[2]

richard.jarry@grenoble-inp.org, sat@nii.ac.jp, kensuke@nii.ac.jp

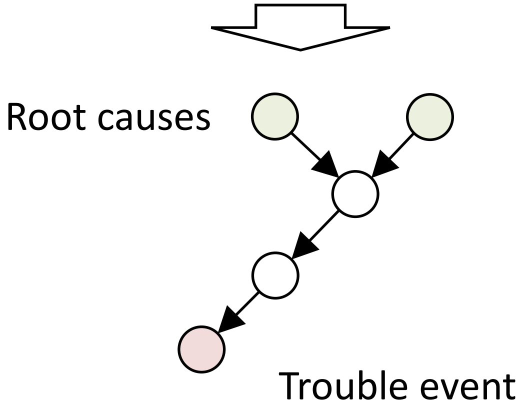
[1]Grenoble INP Ensimag, [2]National Institute of Informatics

ADMNET 2021

Log analysis for automated network operation

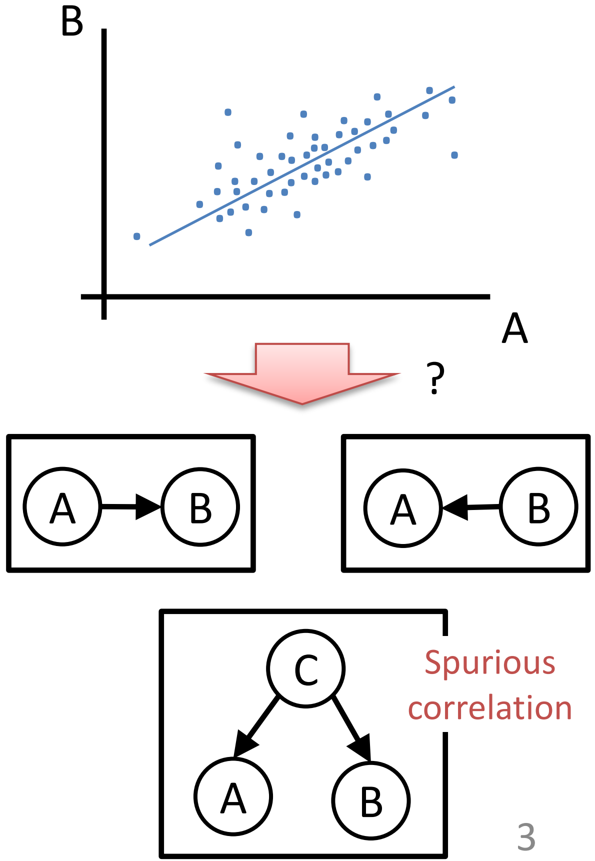
- Network log data
  - Important data source for operation
  - Too large, difficult to use manually
- Automated log analysis
  - Anomaly detection
  - Fault localization
  - <u>Root cause analysis</u>

```
Jul 12 13:00:25 sv1 interface eth1 down
Jul 12 13:00:26 rt2 connection failed to 192.168.1.4
Jul 12 13:02:16 sv1 user sat logged in from 192.168.1.15
Jul 12 13:02:29 sv1 su for root by sat
Jul 12 13:02:58 sv1 interface eth1 up
…
```



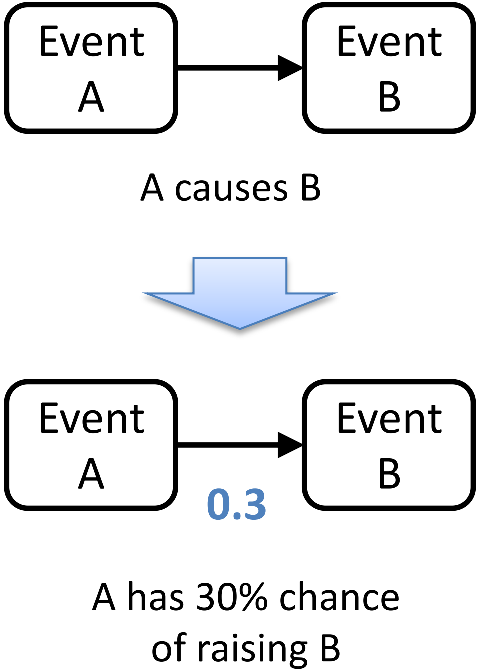
Relation mining for root cause analysis

- Traditional approach -> Correlation
  - Raise Spurious correlation
    - ➢Many False Positives
- Recent approach -> **Causal Inference**
  - Determine causal directions
    - ➢Help finding root causes
  - Remove spurious correlation by searching conditional independence
    - ➢Focus on important relations



Challenges in causal analysis of network logs

- Past literature: Use PC algorithm [1]
  - Basic causal discovery algorithm
  - Can determine only part of edge directions
  - No quantative weight of edges
- Proposed approach: Use MixedLiNGAM
  - Determine all edge directions
  - Determine weight value of edges



[1] S. Kobayashi, et al. “Mining Causality of Network Events in Log Data”, IEEE Transactions on Network and Service Management, pp.53-67, vol.15, no.1, 2018

Goal

- Quantitative causal analysis of network logs
  - Use MixedLiNGAM for causal discovery
    - To determine accurate causal direction
    - To determine quantitative weight of causal edges
- Evaluate proposed method
  - With synthetic data
    - For validation and comparison
  - With real network log data
    - For case study and performance measurement

Overview of log causal analysis

Original log data

↓

Generate log time-series

↓

Causal analysis

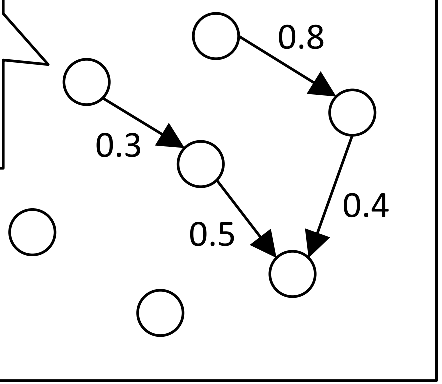
↓

Provided information

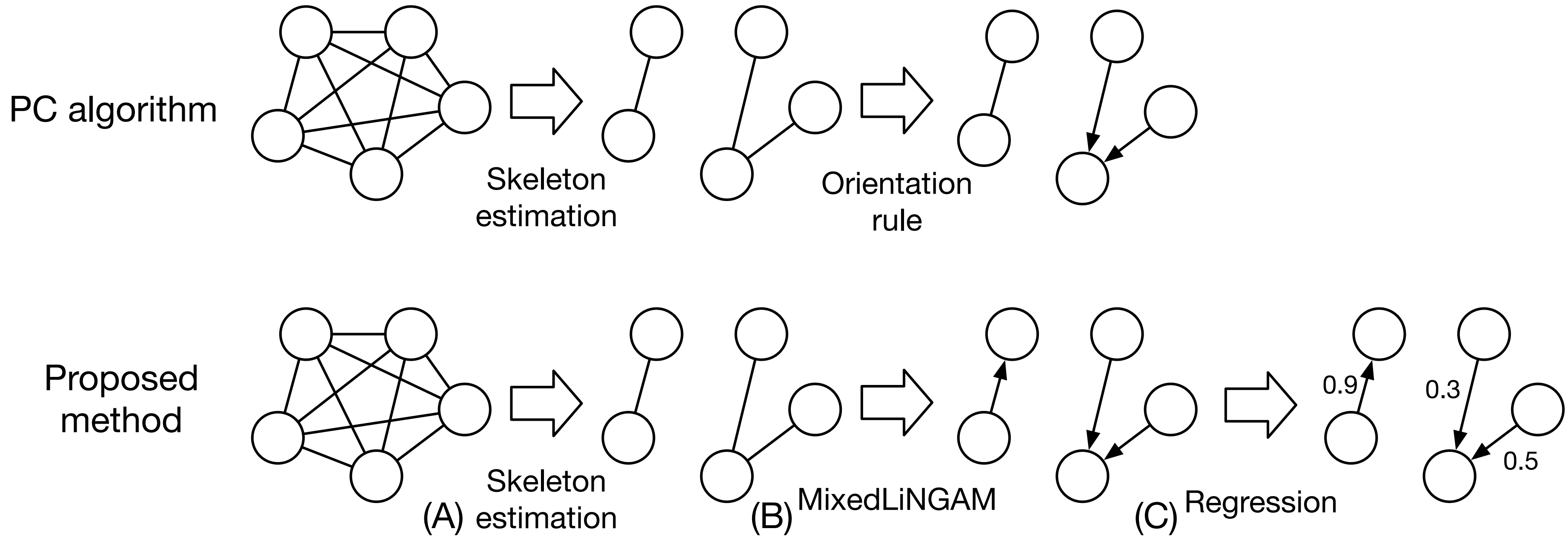
```
Jul 12 13:00:25 sv1 interface eth1 down
Jul 12 13:00:26 rt2 connection failed to 192.168.1.4
Jul 12 13:02:16 sv1 user sat logged in from 192.168.1.15
Jul 12 13:02:29 sv1 su for root by sat
Jul 12 13:02:58 sv1 interface eth1 up
...
```

```
E1: sv1, user * logged in from *
   2021-07-12 13:02:16
   2021-07-12 14:25:00
   ...
```

1 event node for
- 1 host device
- 1 log template



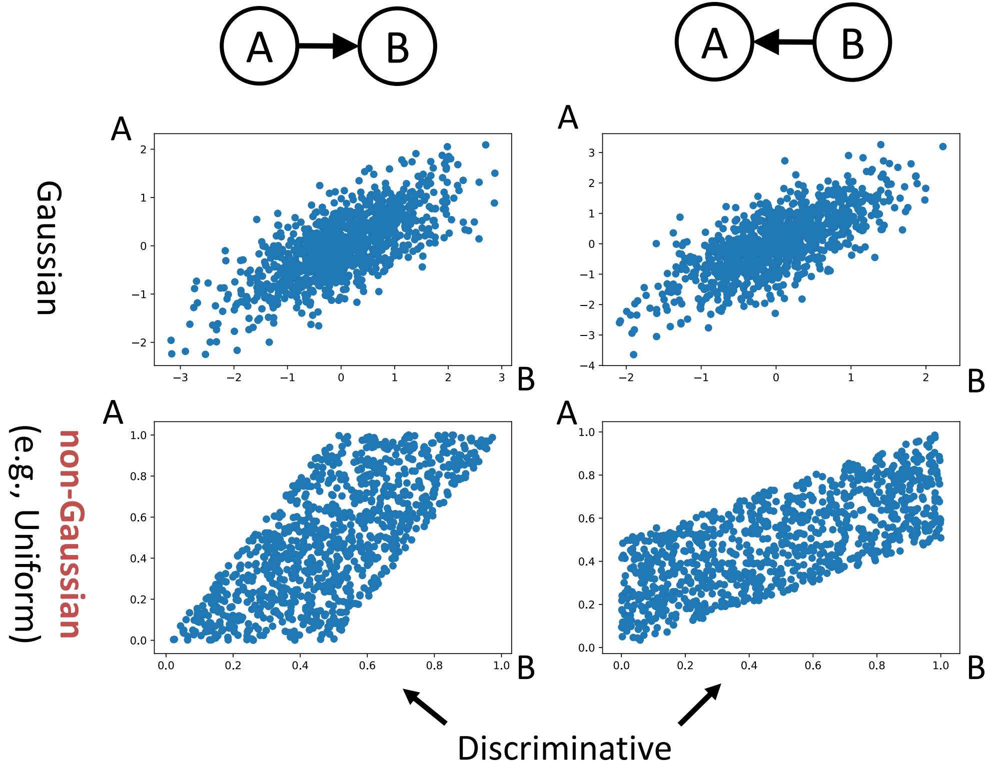
Causal Discovery with MixedLiNGAM



LiNGAM (Linear Non-Gaussian Acyclic Model)[3]

- Assumption
  - Linear causal model
  - **non-Gaussian** disturbance
  - DAG (Directed acyclic model)

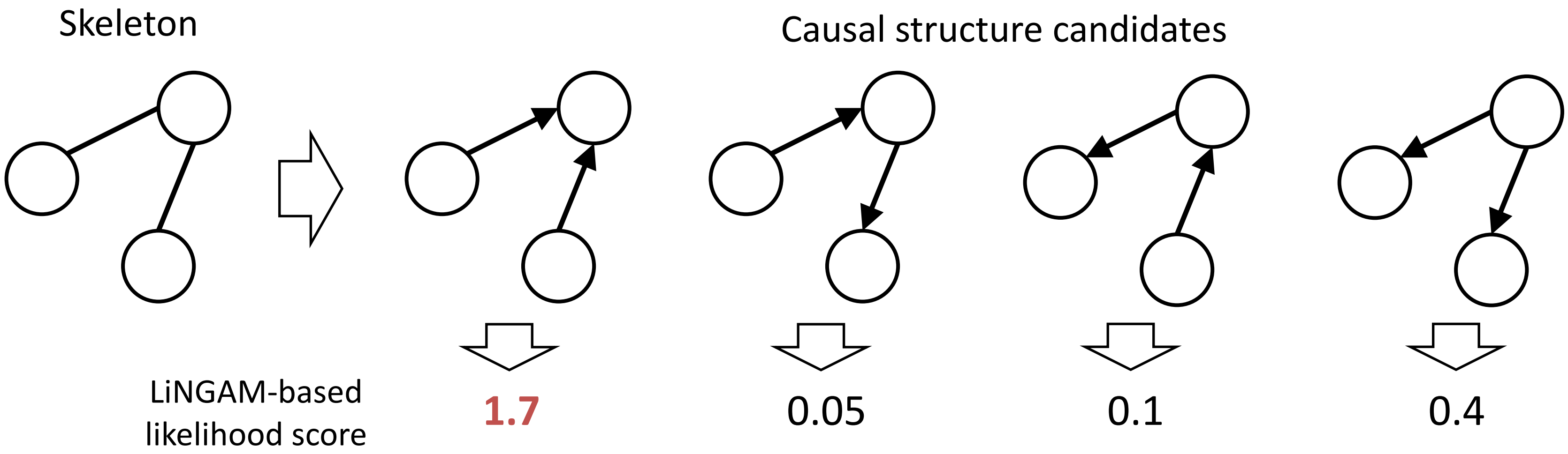
➢ Causal direction can be determined by the data distribution



[3] S. Shimizu, et al. “A Linear Non-Gaussian Acyclic Model for Causal Discovery,” Journal of Machine Learning Research, vol.7, pp.2003-2030, 2006.

(B) MixedLiNGAM[4]

1. Generate DAG candidates (corresponding to input skeleton)
2. Calculate LiNGAM-based likelihood score of each DAG
3. Select DAG with best score

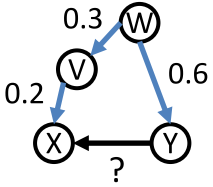


[4] C. Li, et al. "Combining Linear Non-Gaussian Acyclic Model with Logistic Regression Model for Estimating Causal Structure from Mixed Continuous and Discrete Data," arXiv, 2018.

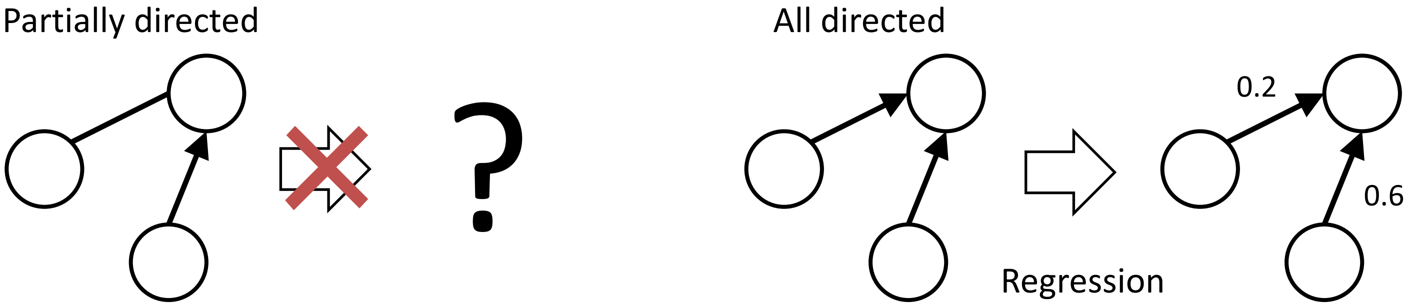
(C) Regression to determine causal weight

- Backdoor criterion[5]: We need to consider all backdoor path to determine the causal effect

Commonly incoming causal flow to X and Y



➢ If all edges are directed, edge weight can be calculated
  - Continuous data input -> Linear regression
  - Discrete (or binary) data input -> Logistic regression



[5] J. Pearl, "Causality: Models, Reasoning, and Inference", 2nd ed. Cambridge Press, 2009.

Analysis overview

Available in GitHub
https://github.com/cpflat/causaltestdata

A) Validation with synthetic data
- Randomly generated time-series data of Poisson Process
- Compare PC algorithm and MixedLiNGAM

B) Evaluation with real network log data
- Use log data of nation-wide academic network
- 8 core routers, over 100 L2 switches
- 35M lines in 456 days (of which 30 days used in evaluation)

Validation with synthetic data

Method	Data model Size	λ	Skeleton accuracy	Direction ratio	Weight diff.
PC algorithm	1,440	10	0.878	0.170	–
	1,440	100	0.980	0.272	–
	1,440	1,000	0.993	0.211	–
	10,800	10	0.973	0.271	–
	10,800	100	0.993	0.270	–
	10,800	1,000	0.957	0.283	–
MixedLiNGAM	1,440	10	0.878	0.704	0.198
	1,440	100	0.980	0.651	0.124
	1,440	1,000	0.993	0.296	0.080
	10,800	10	0.973	0.768	0.087
	10,800	100	0.993	0.682	0.097
	10,800	1,000	0.957	0.240	0.242

Time-series length
(1-day or 7-days)

Average appearance
per 1 day

Validation with synthetic data

Method	Data model Size	λ	Skeleton accuracy	Direction ratio	Weight diff.
PC algorithm	1,440	10	0.878	0.170	–
	1,440	100	0.980	0.272	–
	1,440	1,000	0.993	0.211	–
	10,800	10	0.973	0.271	–
	10,800	100	0.993	0.270	–
	10,800	1,000	0.957	0.283	–
MixedLiNGAM	1,440	10	0.878	0.704	0.198
	1,440	100	0.980	0.651	0.124
	1,440	1,000	0.993	0.296	0.080
	10,800	10	0.973	0.768	0.087
	10,800	100	0.993	0.682	0.097
	10,800	1,000	0.957	0.240	0.242

Same method, same result

MixedLiNGAM is better in direction part

Evaluation with real network logs

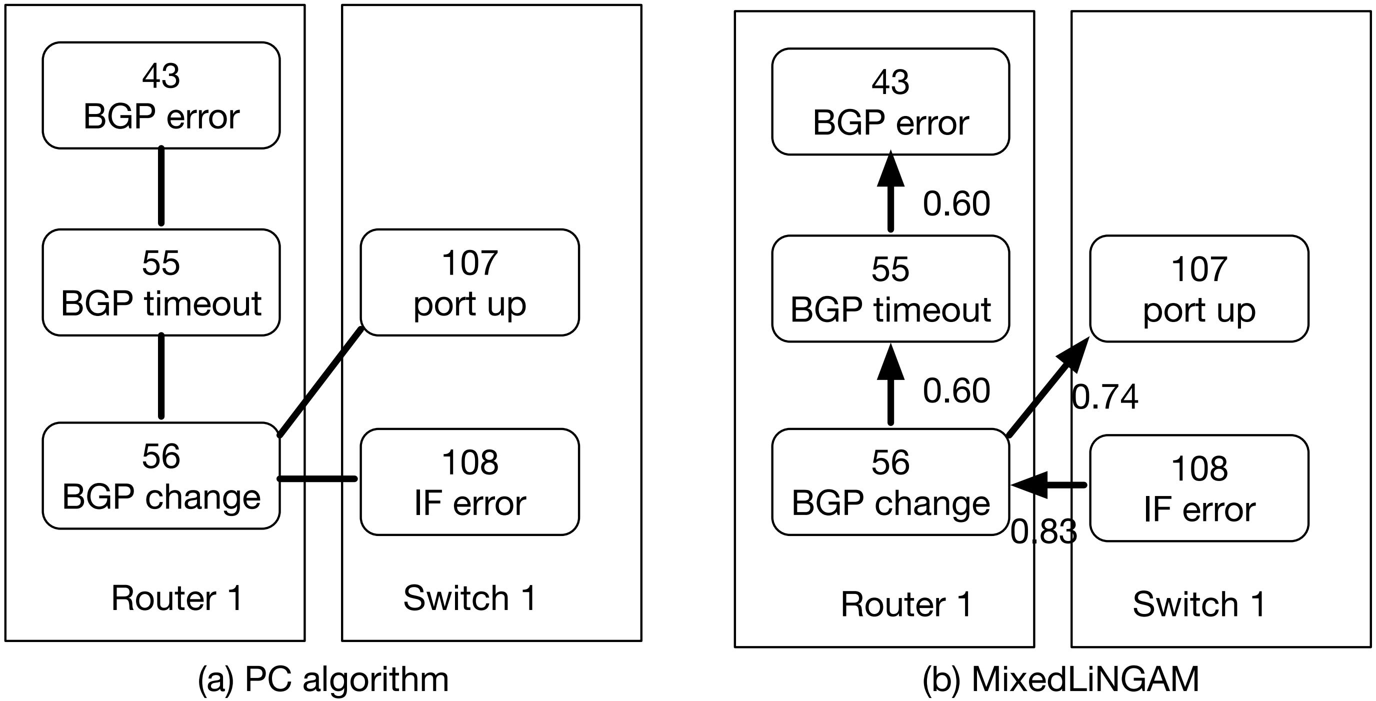
- Macroscopic analysis
  - – Causal analysis per day (1 DAG for 1 day data)
  - – Use 30-days logs (8,605 nodes in total)

Algorithm	**#edges**	**#directed edges**	**ave. weight**	**stdev**
Original PC	1289	121	–	–
MixedLingam	1289	1240	0.856	0.248

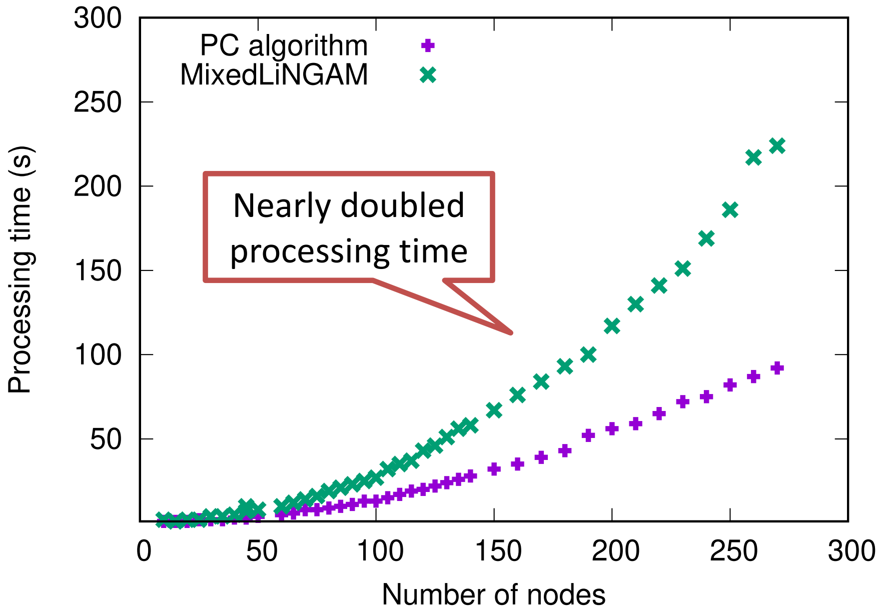
- 40 edges undirected?
➢ Edges with too small weight (nearly 0)

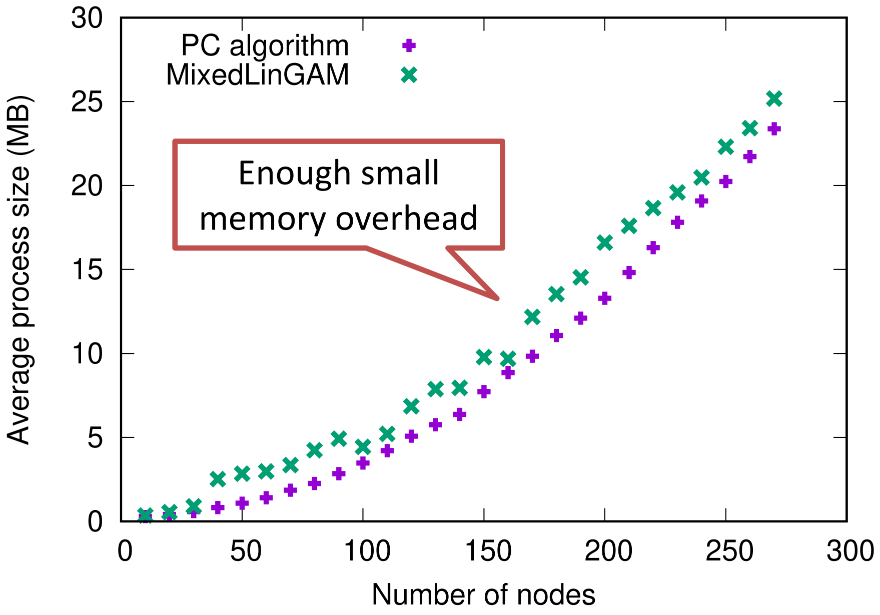
Most edges are weighted nearly 1.0

Case study


43
BGP error
55
BGP timeout
56
BGP change
107
port up
108
IF error
Router 1
Switch 1
(a) PC algorithm
43
BGP error
0.60
55
BGP timeout
0.60
56
BGP change
0.74
107
port up
108
IF error
0.83
Router 1
Switch 1
(b) MixedLiNGAM

Performance measurement





Concluding remarks

- We proposed a quantative causal analysis method
  - Based on MixedLiNGAM
- We demonstrated effectiveness of the proposed method
  - Validation with synthetic data -> Improved edge directions
  - Evaluation with network logs -> Appropriate results
- Future works
  - Improve performance for analysis with larger dataset
  - Automated root cause analysis based on obtained weighted DAGs

This work is supported by MIC/SCOPE #191603009.