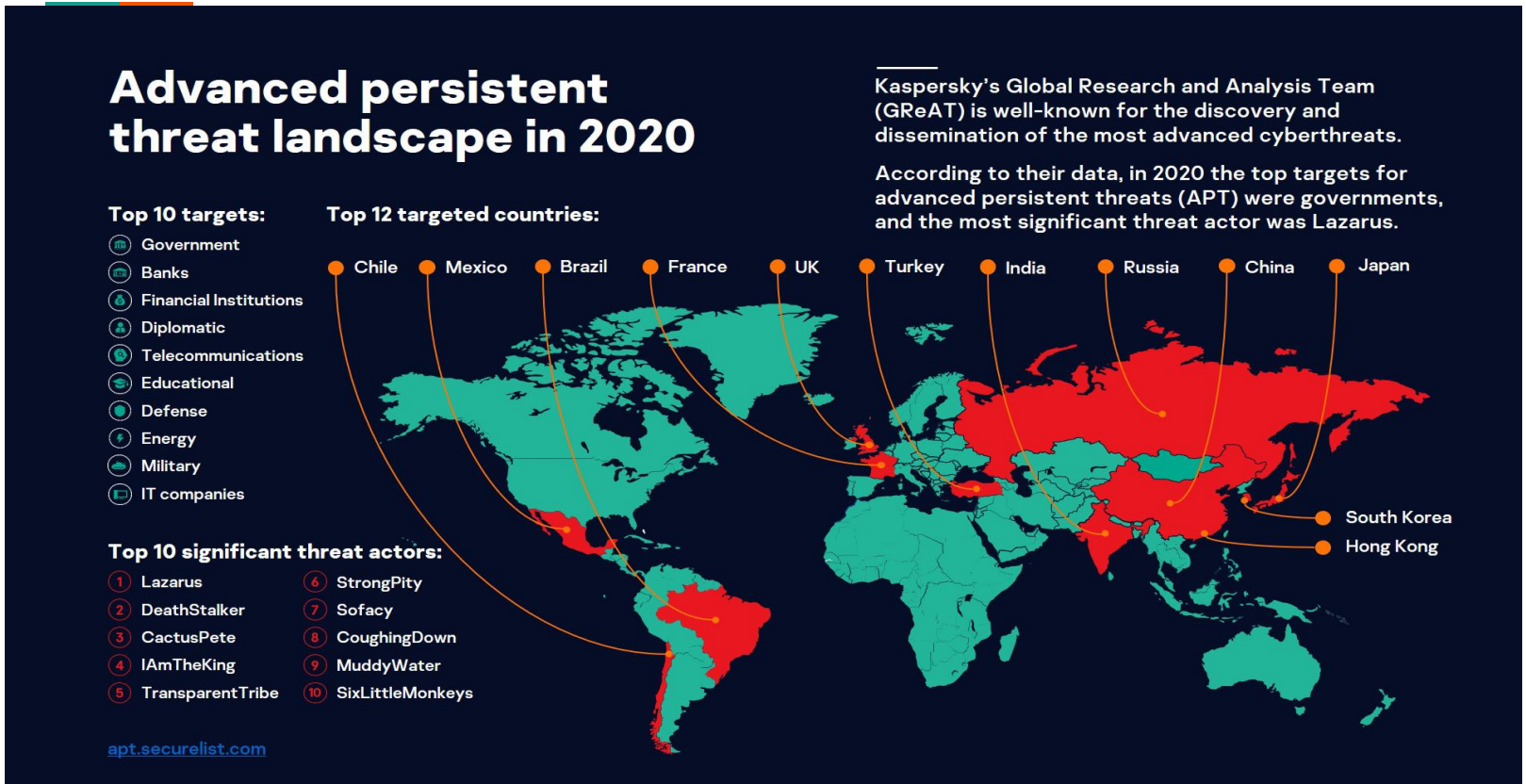
SEAL: Storage-efficient Causality Analysis on Enterprise Logs with Query-friendly Compression

Peng Fei, Zhou Li, Zhiying Wang, Xiao Yu, Ding Li, Kangkook Jee

Background and Motivation

Advanced persistent threat (**APT**):

Long-term presence on a network in order to mine highly sensitive data.



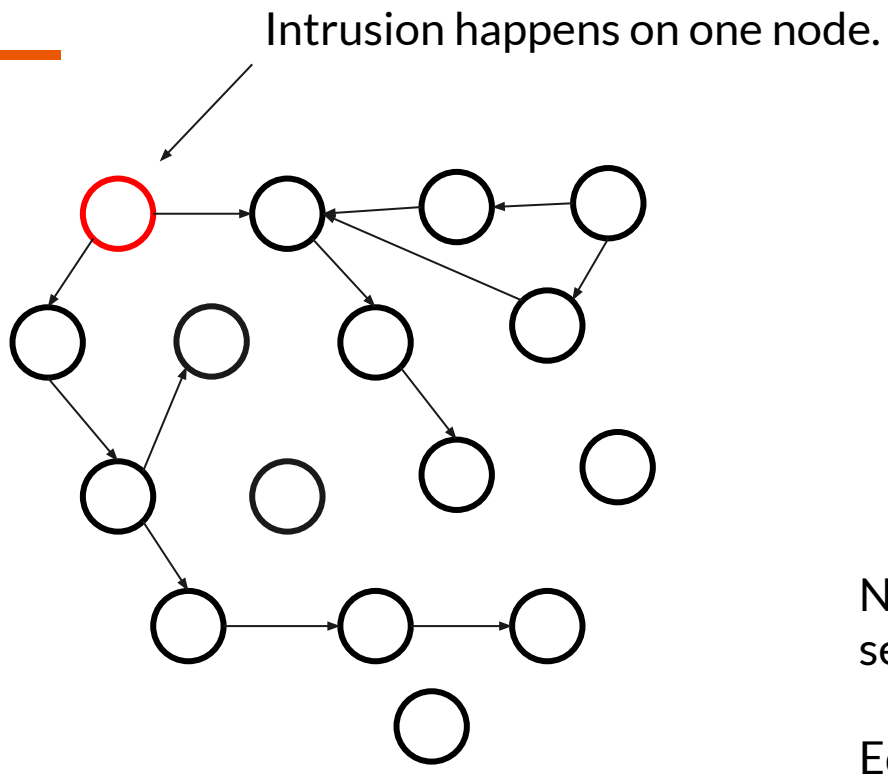
Background and Motivation

An advanced persistent threat (**APT**):

- Long-term presence.
  - The latest intrusion prolongs over 188 days before detection. (TrustWave global report, 2015)

- highly sensitive data.
  - Government data.
  - Banks.
  - ...

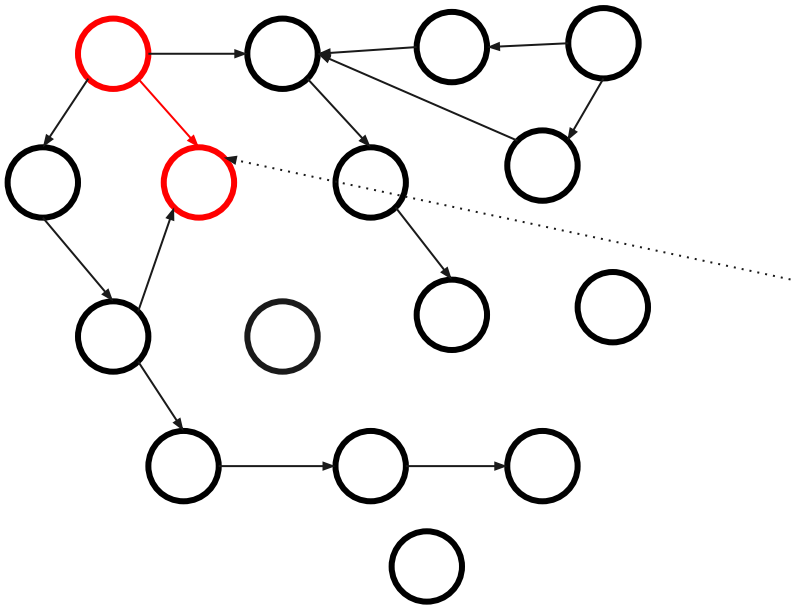
Background and Motivation



Nodes: files and programs in the servers.

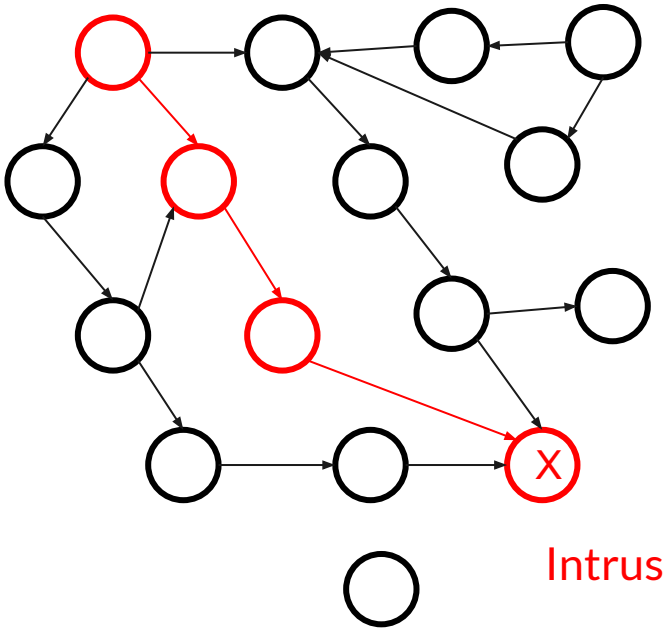
Edges (logs): read, write and execute.

Background and Motivation



Intrusion transfers to another node

Background and Motivation



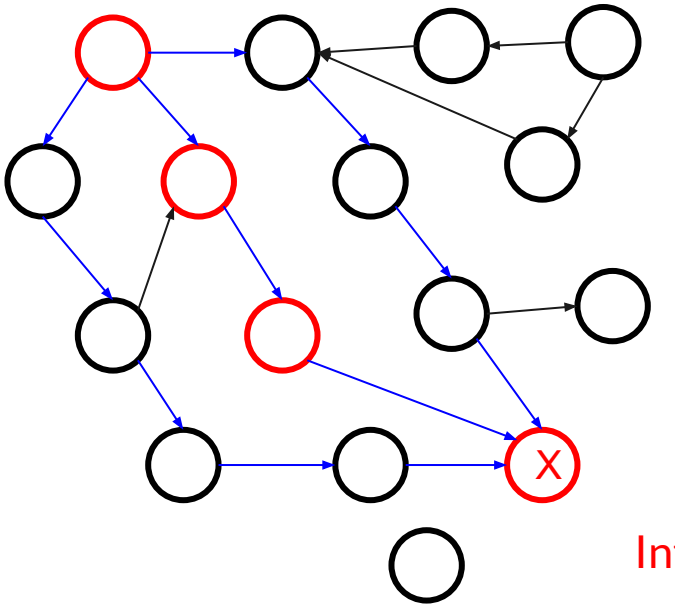
What can we do?

Use backtrack to find all the affected nodes.

(Casualty Analysis)

Intrusion detected !

Background and Motivation



Casualty Analysis:

reconstructs information flow by associating interdependent system events.

Intrusion detected !

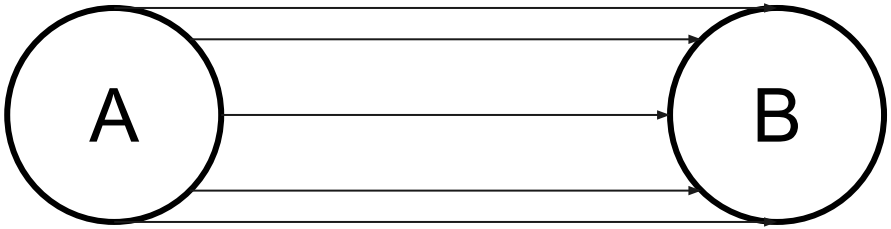
Background and Motivation

- High pressure on storage:
  - e.g. NEC Laboratories America only sustains 3 month log data with more than 5TB from only 100 hosts.

- Compelling need for a solution to scale storage and processing capacity to meet the enterprise-level requirement.

Background and Motivation

Let's do some reductions for casualty analysis.



Background and Motivation

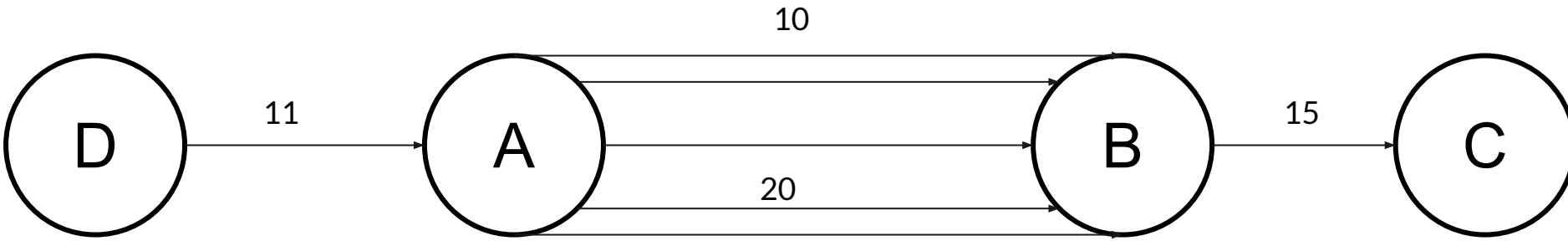
Let's do some reductions

5 times reduction rate!

Background and Motivation

However…

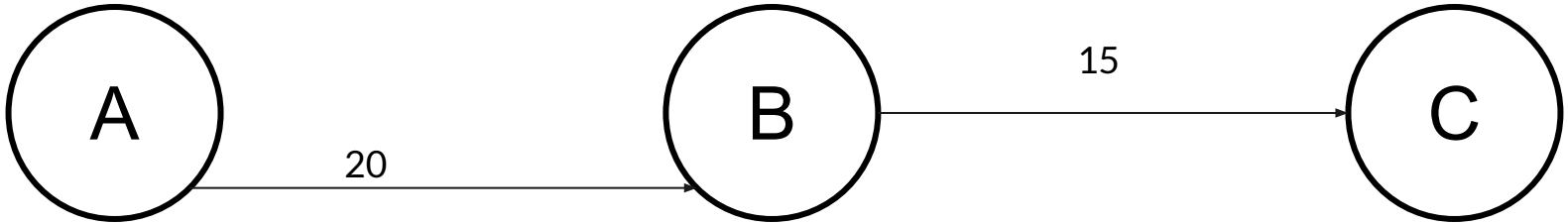
- Time constraint



Background and Motivation

However…

- Time constraint

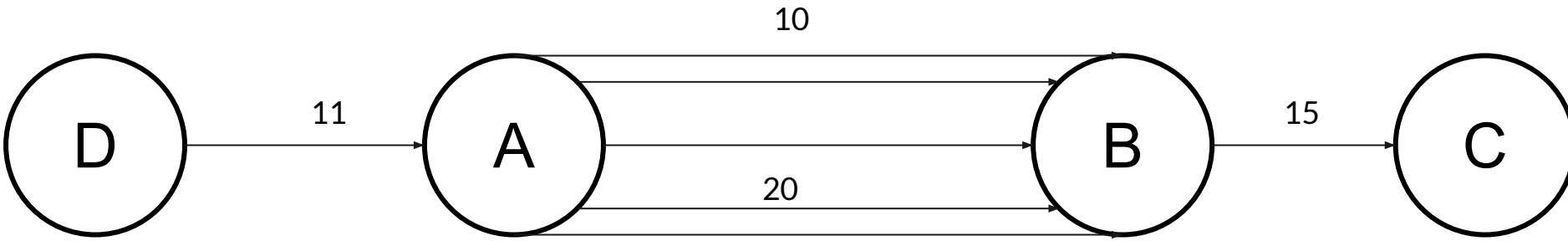


C can no longer backtrack to A!

Background and Motivation

However…

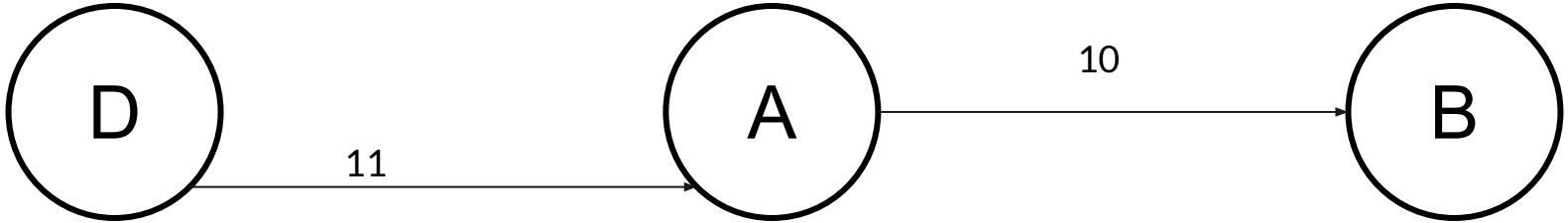
- Time constraint



Background and Motivation

However…

- Time constraint



B can no longer backtrack to D!

Background and Motivation

Current Compaction Techniques:

- Lossy compression: removes logs matching predefined patterns in the dependency graph. [Lee, K. H.,et all. 2013][Tang, Y., et all, 2018][Xu, Z., 2016][Hossain, M. 2018]
    - Remove repeated edges under full dependency preservation (FD). [Hossain, M. 2018]
    - Remove nodes that only contains read-only file. [Lee, K. H.,et all. 2013]

- No 100% guarantee for the same casualty analysis outcome.

Background and Motivation

How about Lossless Compression?

- Standard tools[Deutsch, Peter. 1996][Bartík, et all, 2015]
  - (Gzip, LZ4, etc)
- DataBase Compression[Roth, et all 1993]

	compression rate	decompression overhead	support time constraint
Standard tools	high	high	Yes
DataBase Compression	high	low	No

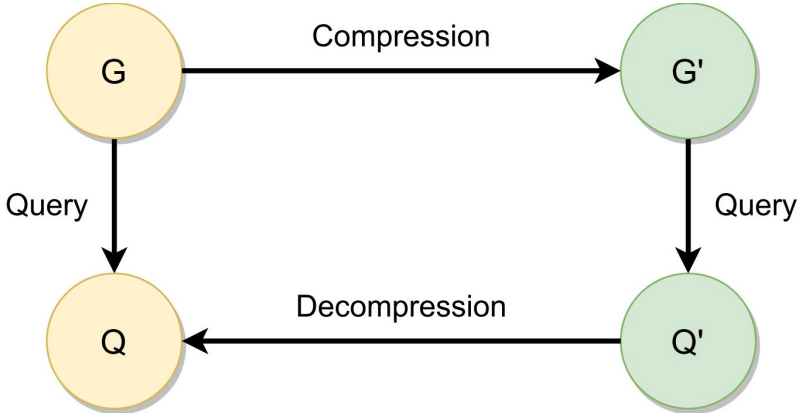
- **Our Solution: Query-friendly Lossless compression.**

Background and Motivation

Query-friendly Lossless compression. (Inspired by [Fan, W, et al. 2012])

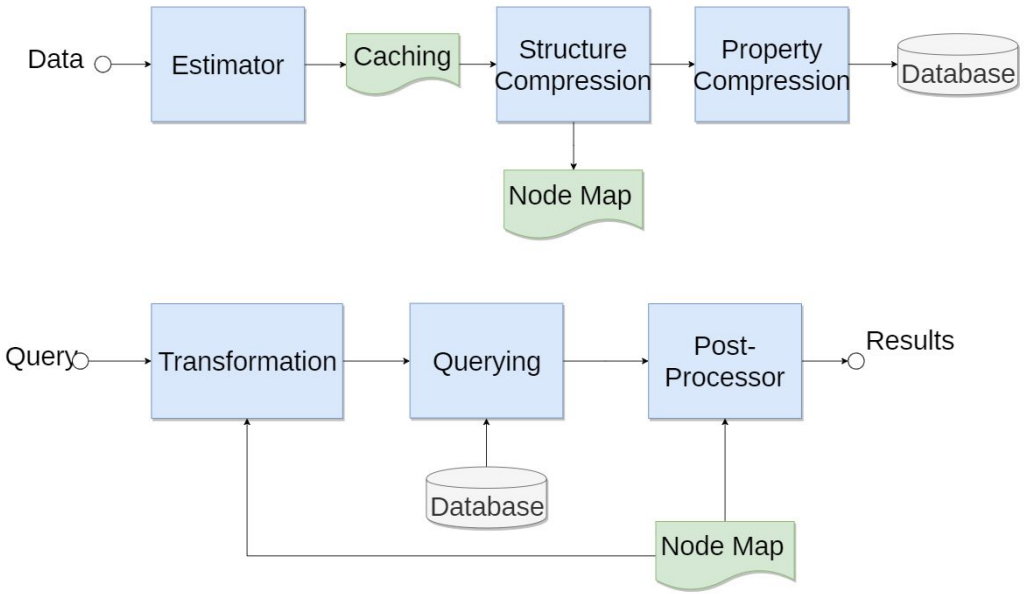
G' : compressed graph.

Q' : query result on the compressed graph.



Background and Motivation

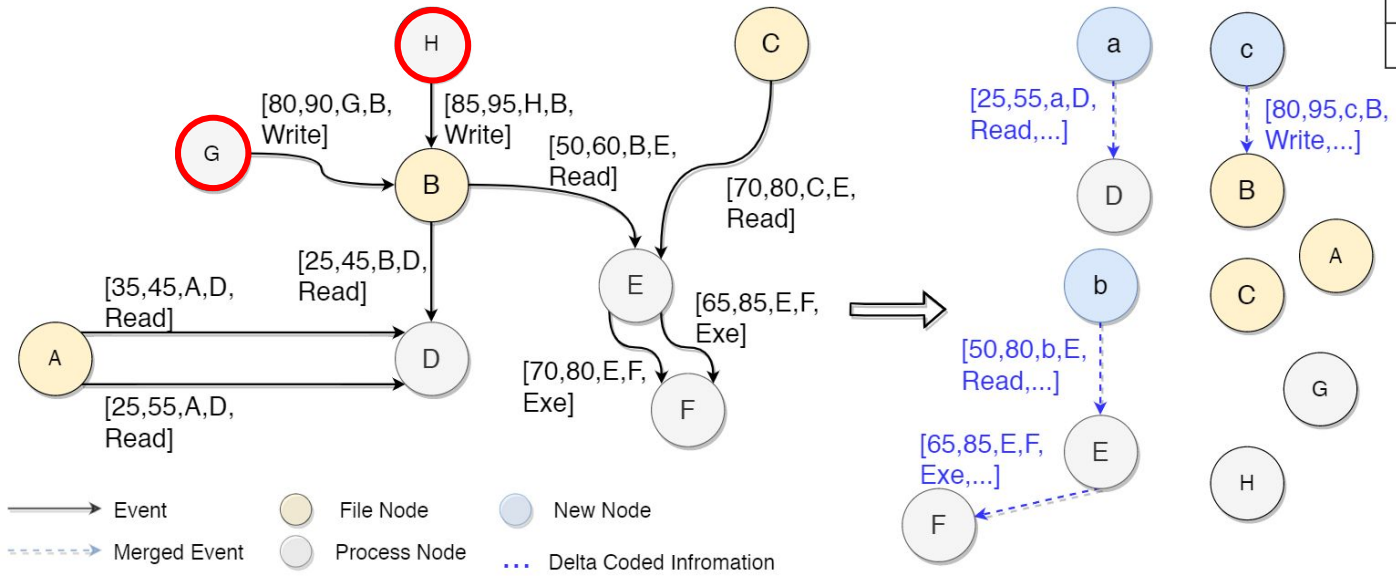
Design Architecture



Compression on Logs dependency graph

Compression on Logs dependency graph

Compression on Graph Structure: Nodes merge

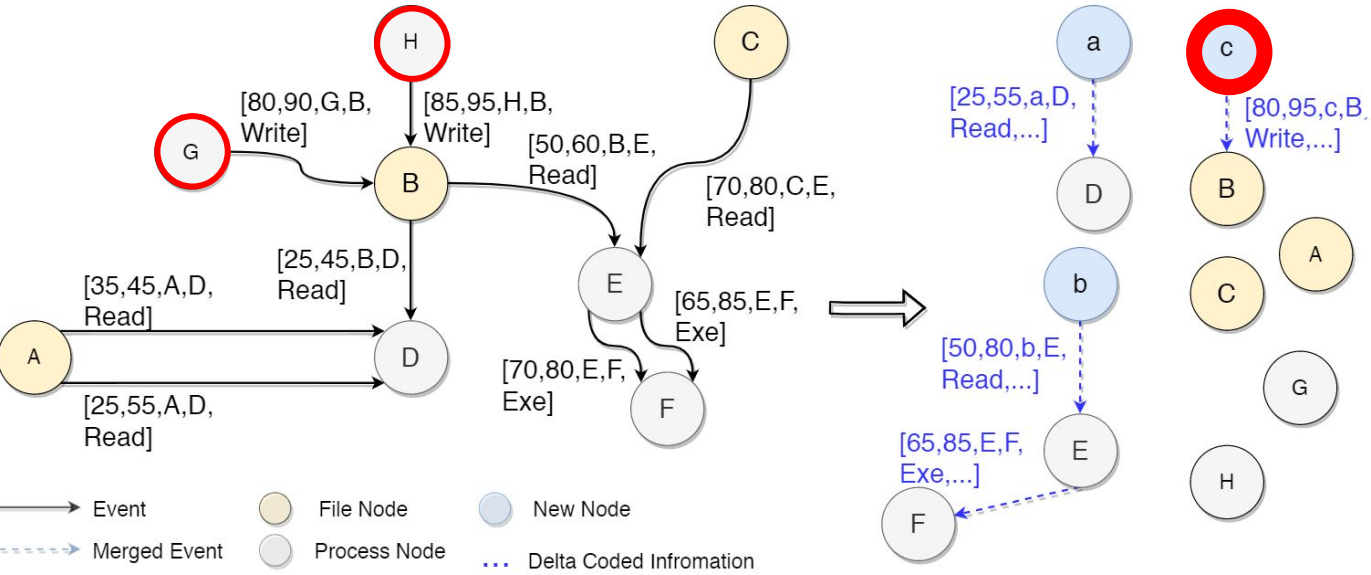


New Node	Represented Nodes
a	A, B
b	B, C
c	G, H

Node Map

Compression on Logs dependency graph

Compression on Graph Structure: Nodes merge

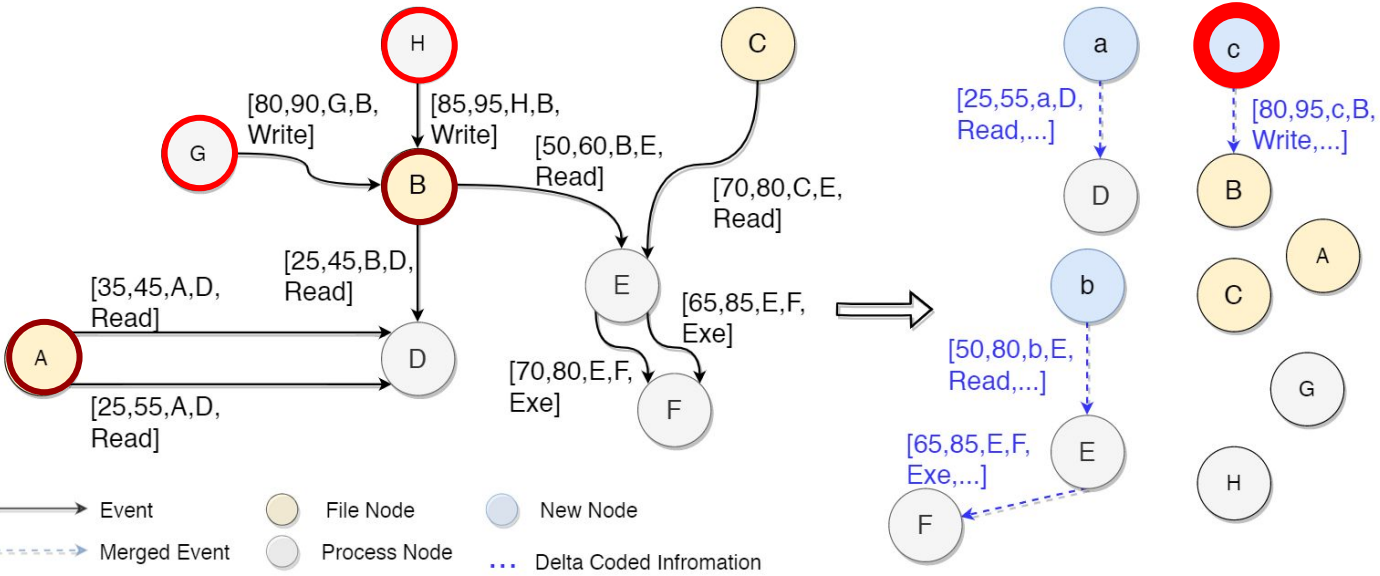


New Node	Represented Nodes
a	A, B
b	B, C
c	G, H

Node Map

Compression on Logs dependency graph

Compression on Graph Structure: Nodes merge

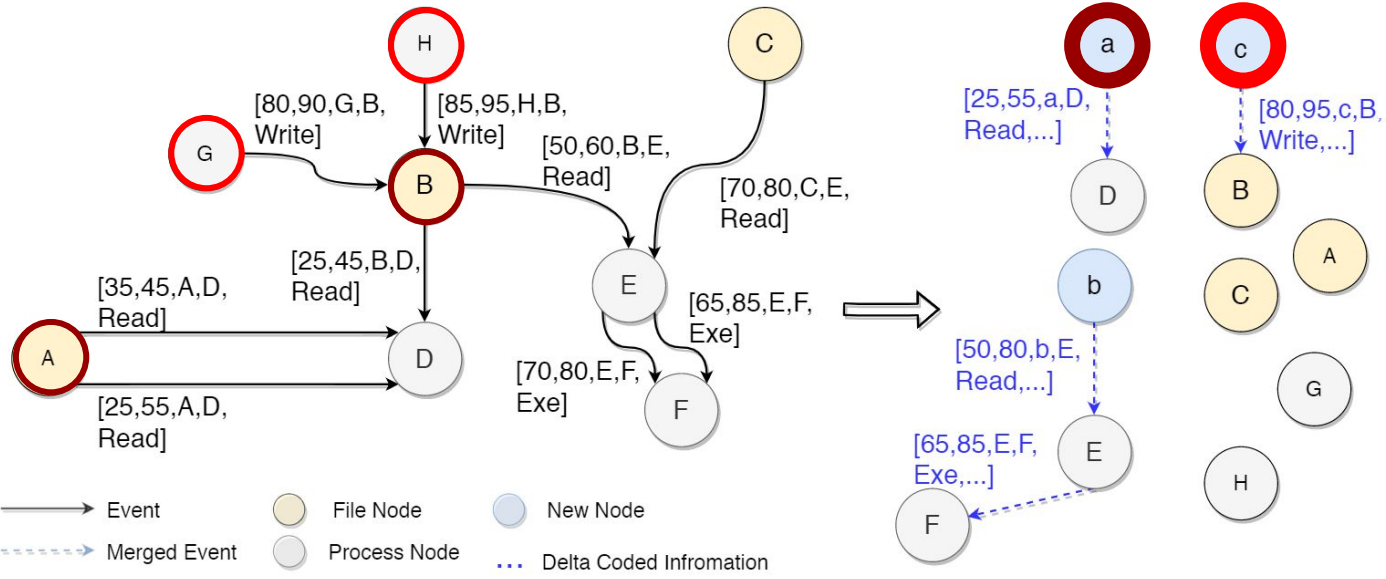


New Node	Represented Nodes
a	A, B
b	B, C
c	G, H

Node Map

Compression on Logs dependency graph

Compression on Graph Structure: Nodes merge



New Node	Represented Nodes
a	A, B
b	B, C
c	G, H

Node Map

Logs dependency graph compression

Compression on Edge properties:

Sorting: ascending for the starttime, descending for the endtime

Long integer sequences that share a long prefix.

starttime	endtime	srcid	dstid	agentid	accessright
[1562723488971, 1562723488991]	[1562734588984, 1562734588995]	15	27	-582777938	Execute

Table: Example entry of FileEvent

Logs dependency graph compression

Compression on edge properties:

- Delta Coding: Represents a sequence of values with the differences(or delta).

Long integer sequences that share a long prefix.

starttime	endtime	srcid	dstid	agentid	accessright
[1562723488971, 1562723488991]	[1562734588984, 1562734588995]	15	27	-582777938	Execute

Table: Example entry of FileEvent

Logs dependency graph compression

Compression on Edge properties:

- Delta Coding: Represents a sequence of values with the differences(or delta).

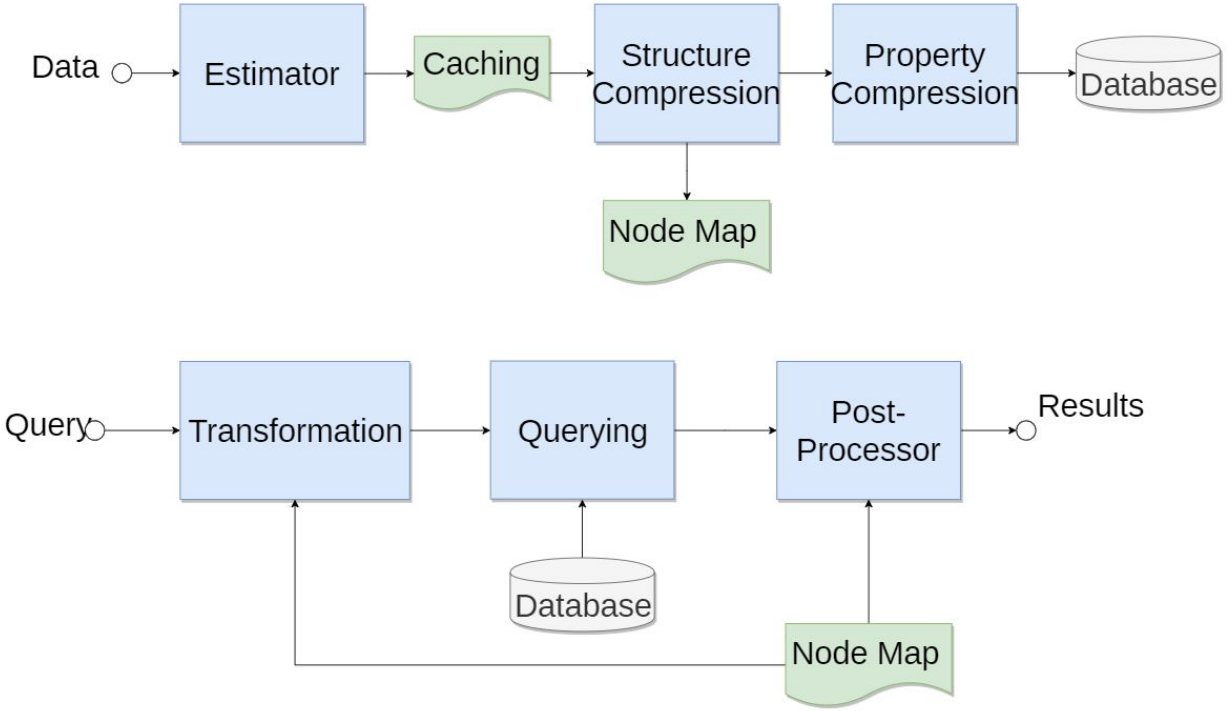
Long integer sequences that share a long prefix.

starttime	endtime	srcid	dstid	agentid	accessright
[1562723488971, 20]	[1562734588984, 11]	15	27	-582777938	Execute

Table: Example entry of FileEvent

Query and decompression

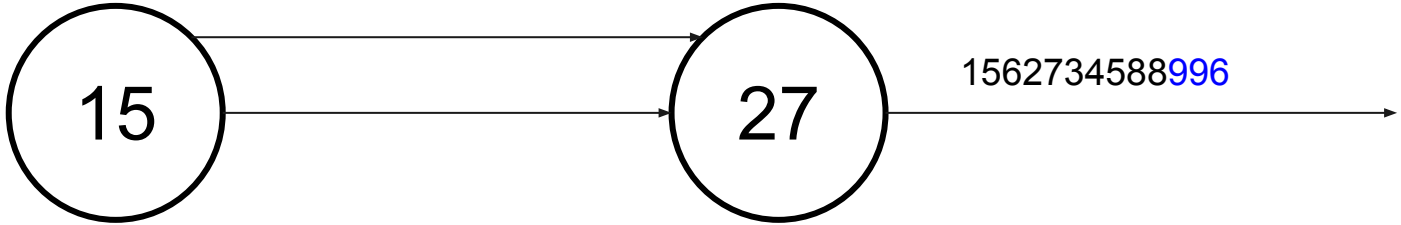
Query and decompression



Query and decompression

select * from id == 27 and endtime < 1562734588996

We can backtrack to node 15 without any decompression.

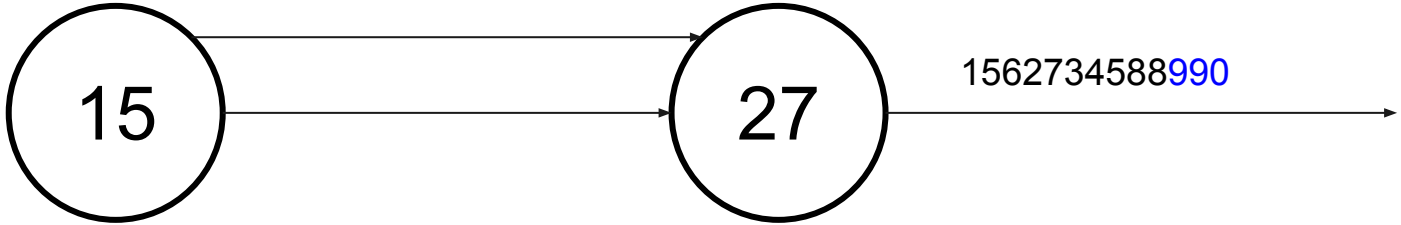


starttime	endtime	srcid	dstid	agentid	accessright
[1562723488971, 20]	[1562734588995, -11]	15	27	-582777938	Execute

Query and decompression

select * from id == 2 and endtime < 1562734588990

After decompress [1562734588995, -11] in to [1562734588995, 1562734588984].
Return node 15.



starttime	endtime	srcid	dstid	agentid	accessright
[1562723488971, 20]	[1562734588995, -11]	15	27	-582777938	Execute

Evaluation

Evaluation

Compression ratio:

- NEC Lab: 7.6GB/20GB (5M events)

- DARPA Transparent Computing program [DARPA/I20, 2020]: 18GB/233GB (1,183M events)

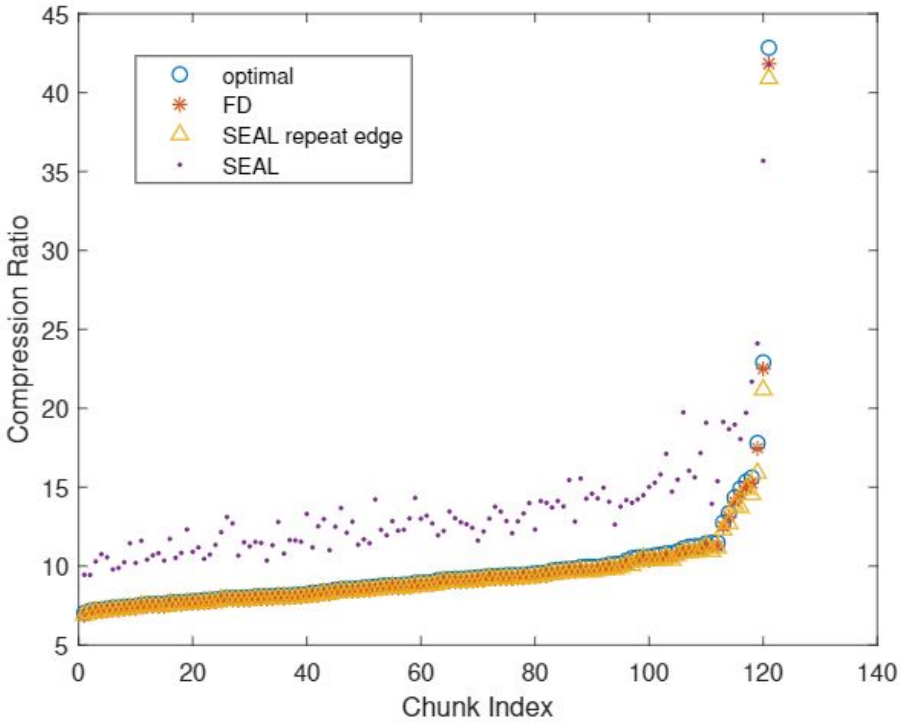
Evaluation

Compression Ratio Comparison :

- Optimal ratio for edges compression. Only leave one edge between one pair of nodes.

  Full dependency(FD) and our edge only method both gain similar results.

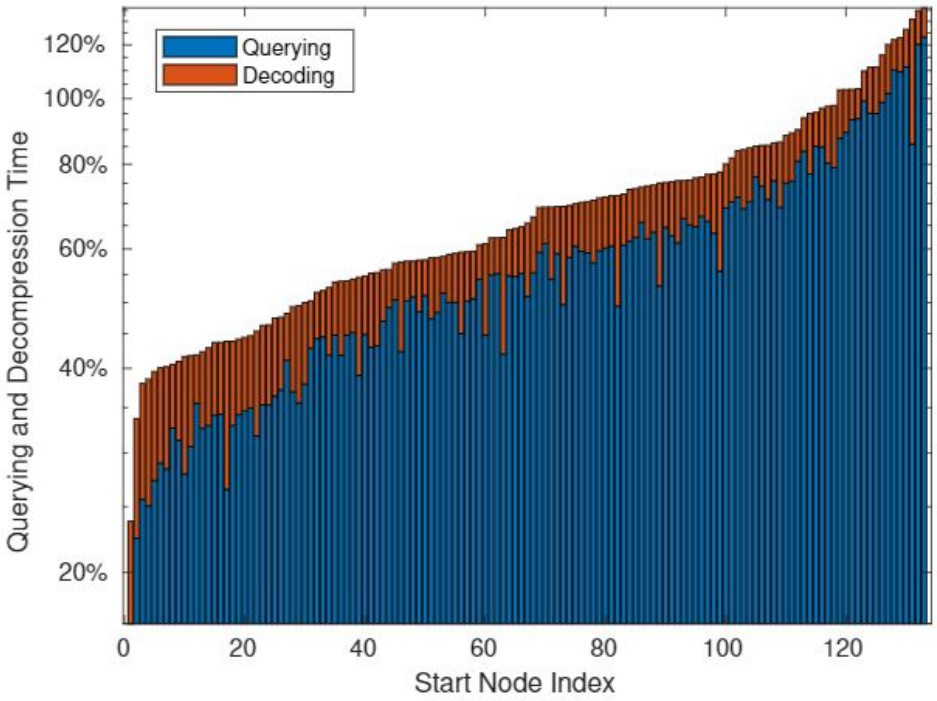
- Better Ratio if node merged is added.



Evaluation

Query efficiency:

- 89% cases have less total time than the uncompressed data. (< 100%).
- On average decompression only takes 18.66% of the overall time.



Thank you!