



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

StreamWorks: Continuous Pattern Detection on Streaming Data

SUTANAY CHOUDHURY, KHUSHBU AGARWAL, SHERMAN BEUS, DANIEL
DOHNALEK, KSHITEESH HEGDE

Pacific Northwest National Laboratory



What is StreamWorks?





The Promise of Patterns

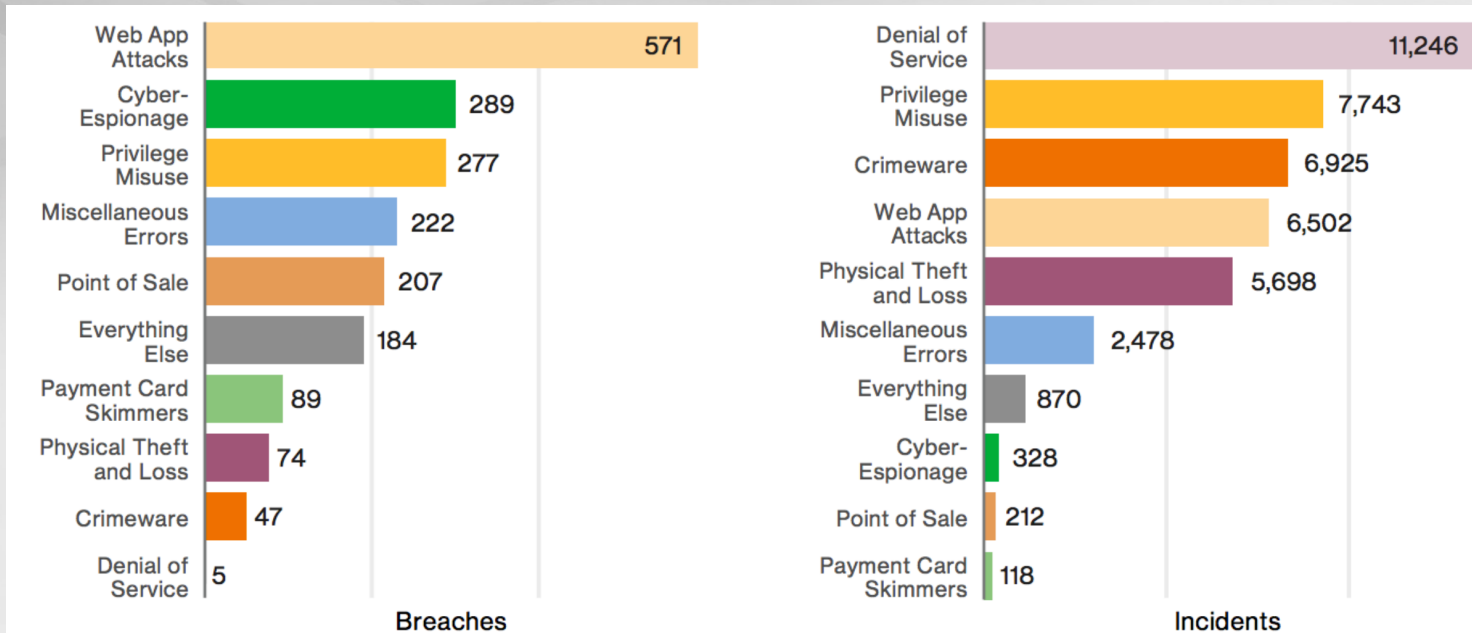


Figure 33: Percentage and count of breaches per pattern (n=1,935)

Figure 34: Percentage and count of incidents per pattern (n=42,068)

Source: Verizon 2017 Data Breach Investigations Report

- ▶ The **median number of days to detect security breaches** was **146 days** in 2015 – FireEye/Mandiant Report
- ▶ In its “Data Breach Investigations Report” in 2014, Verizon analyzed 100,000 security incidents from past decade and concluded **90% attacks fell in 10 attack patterns**

Graphs and Patterns



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** Since 1965



Red nodes are **Services**

Gray nodes are **Clients**

Users with similar role demonstrate same pattern of service usage



Tell me as soon as it happens!

▶ How do you read email?

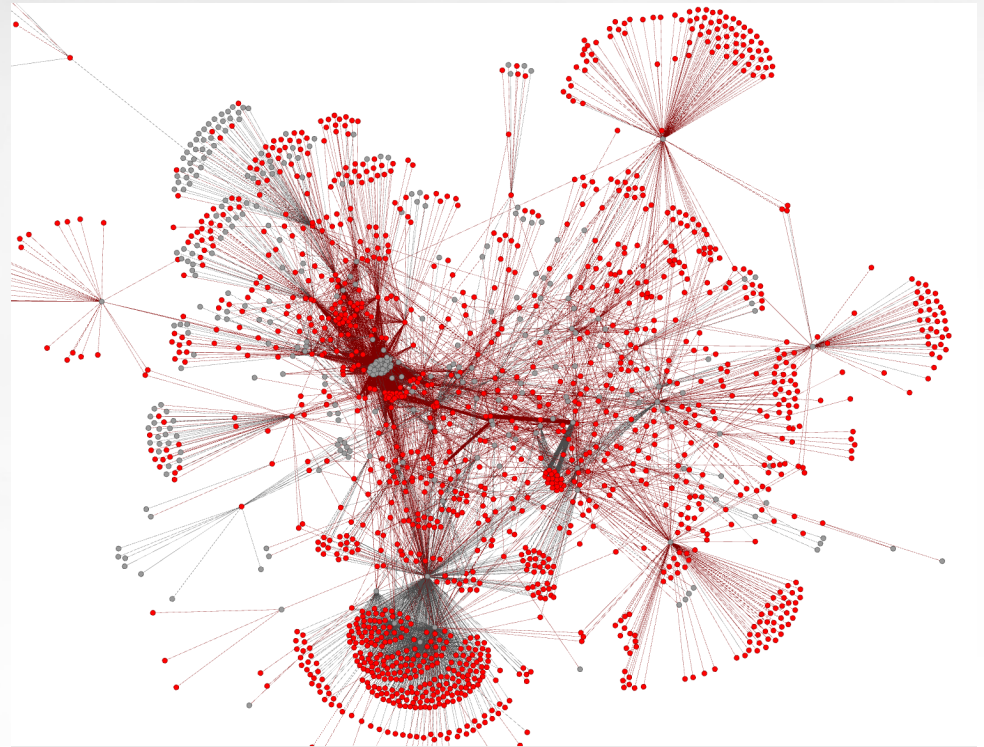
- Read every email as soon as it comes in (**Continuous Processing**)
- Read every 4 hours (**Periodic or Batched Processing**)

▶ Unfortunately, being late is not better than never in all cases

- **Cyber**: Data leaving your network or a malware spread in action
- **Finance**: Price dips intraday, your late order buys high end of the day 😊

Approach for Continuous Pattern Detection

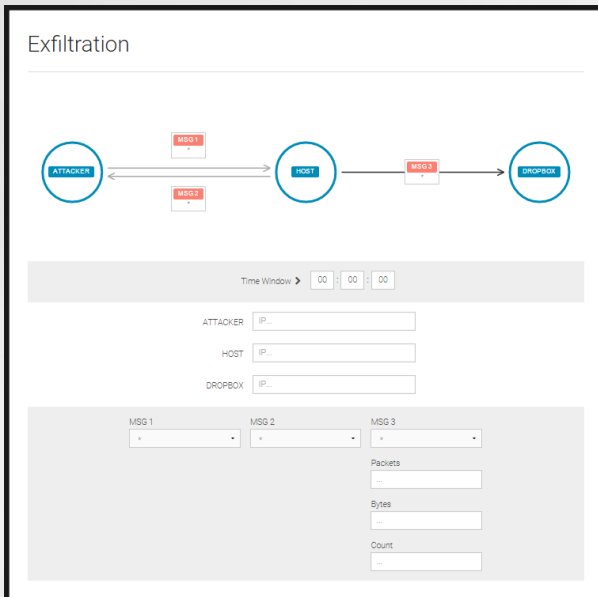
- ▶ **Incremental Querying is key to Performance**
 - We turn streaming data into a graph model
- ▶ **Guiding our insight**
 - We interviewed tens of analysts and system defenders, and asked about the top patterns they would like to detect
- ▶ **Pattern Queries in Action**
 - “Tell me when a chain of 3 logins are detected with increasing privileges?”





And last, but not the least ...

- ▶ One more “Driver”
- ▶ Visual Querying: Real users should not need to learn a new query language to use the system.



```
SELECT ?control ?target ?dropbox ?xfil WHERE {
  # Control Message from C2 to target
  ?control ?ctrlmsg ?target .
  ?ctrlmsg :FTIME ?ftime1 .
  ?ctrlmsg :STIME ?stime1 .
  ?ctrlmsg :DPKTS ?pkts1 .
  ?ctrlmsg :DOCTETS ?octets1 .
  FILTER (?pkts1 < 3 && ?octets1 < 300)

  # xFil occurs within the next hour to ?dropbox
  { SELECT ?target ?dropbox (SUM(?octets) AS ?xfil)
    WHERE {
      ?target ?flow ?dropbox .
      ?flow :DOCTETS ?octets .
      ?flow :STIME ?stime .
      FILTER (?stime > ?ftime1
        && ?stime - ?ftime1 < 3600)
    } GROUP BY ?target ?dropbox
      HAVING (SUM(?octets) > 200000)
  }

  # xFil did NOT happen from target in previous
  # hour (target usually does not send lots of
  # data to external hosts).
  { SELECT ?target
    { SELECT ?target (SUM(?octets) as ?outRate)
      WHERE {
        ?target ?flow ?dst .
        ?flow :DOCTETS ?octets .
        ?flow :STIME ?stime .
        FILTER (?stime < ?stime1
          && ?stime1 - ?stime < 3600)
      } GROUP BY ?target ?dst
    } GROUP BY ?target
      HAVING (MAX(?outRate) < 100000)
  }
}
```

Querying for Chains of Activity



Pacific Northwest
NATIONAL LABORATORY

Originally Operated by Battelle Since 1965

Path Query



Time Window >

00 : 00 : 00

Message Count >



HOST 1

MSG 1

HOST 2

MSG 2

HOST 3

MSG 3

HOST 4

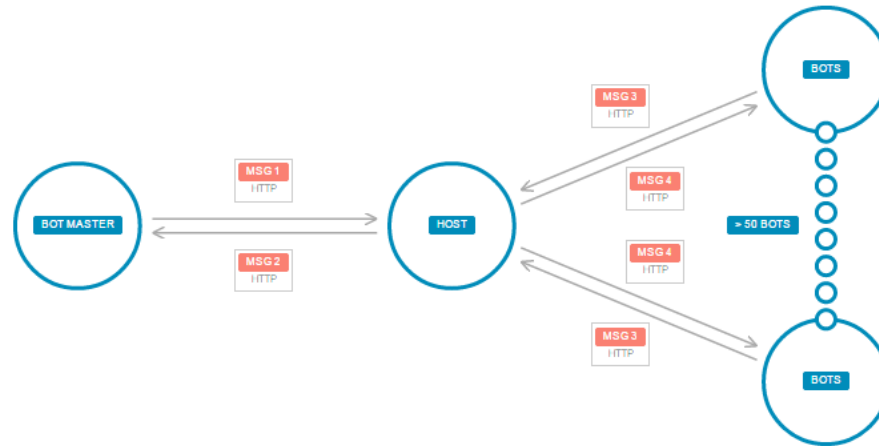
Botnet Command and Control



Pacific Northwest
NATIONAL LABORATORY

Originally Operated by Battelle Since 1965

Botnet Command and Control



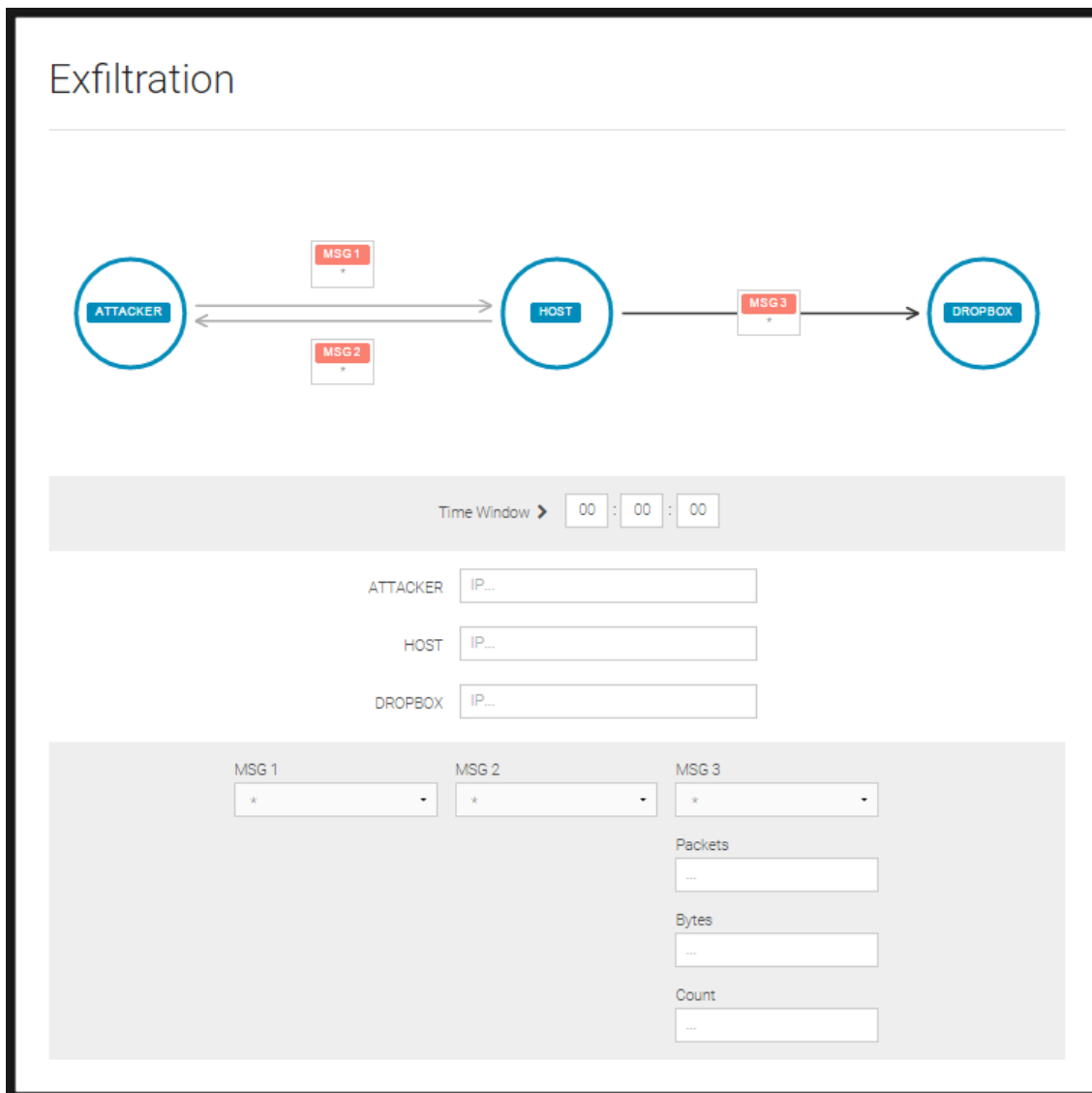
Time Window > 00 : 00 : 00

BOT MASTER

HOST

BOTS >50

MSG 1	MSG 2	MSG 3	MSG 4
<input type="text" value="HTTP"/>	<input type="text" value="HTTP"/>	<input type="text" value="HTTP"/>	<input type="text" value="HTTP"/>
Packets		Packets	
<input type="text" value="..."/>		<input type="text" value="..."/>	
Bytes		Bytes	
<input type="text" value="..."/>		<input type="text" value="..."/>	

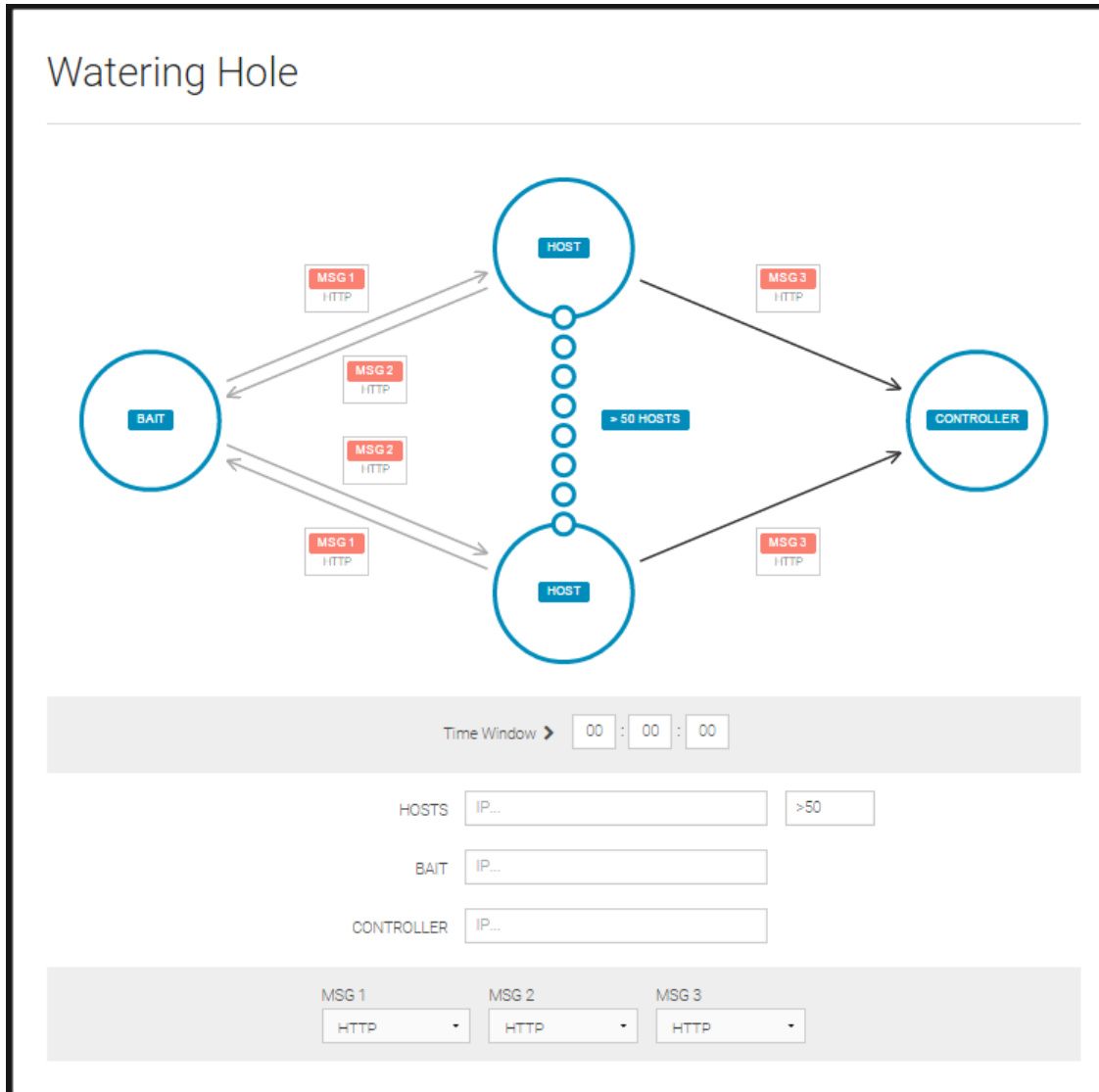


Watering Hole



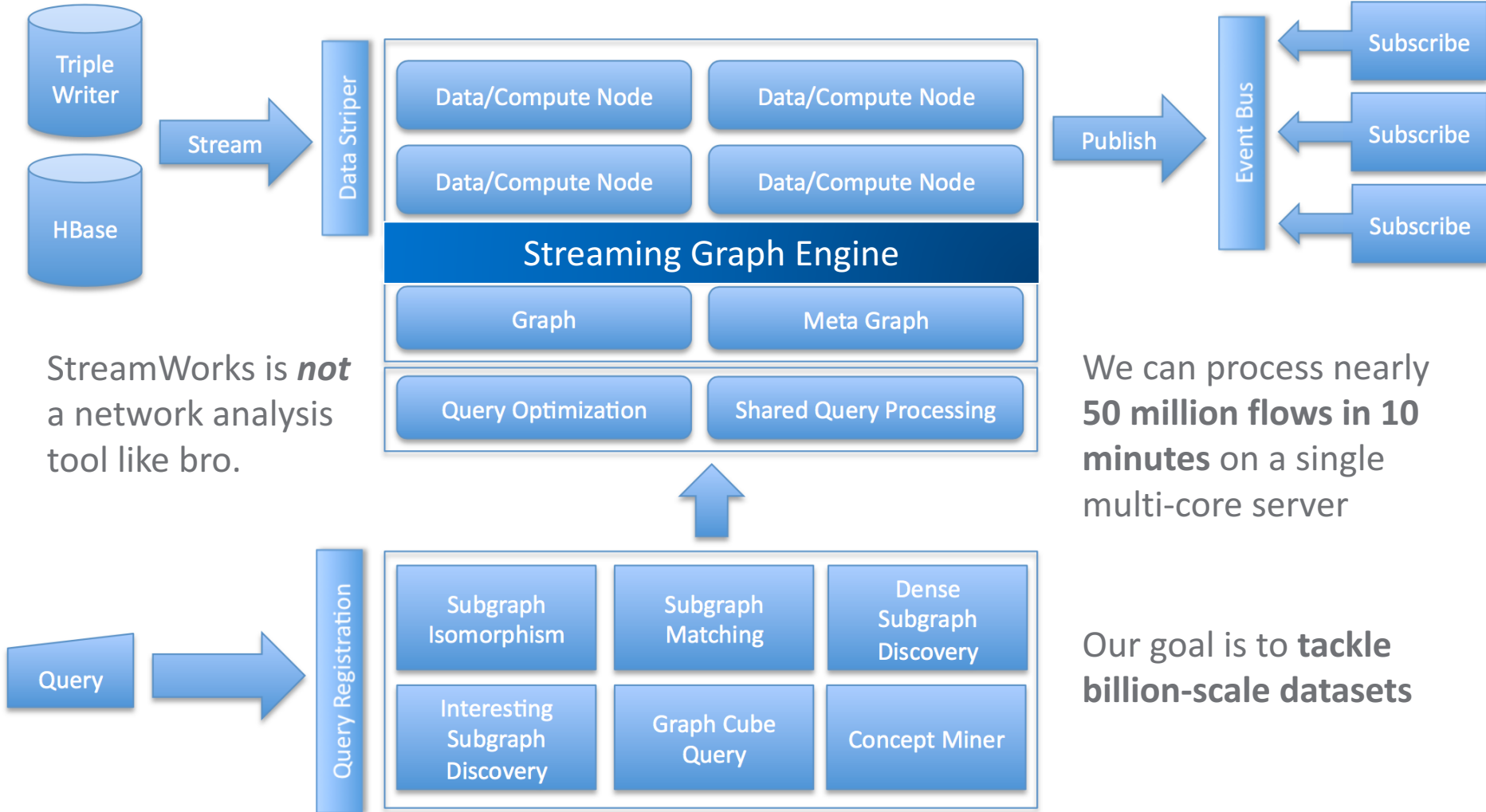
Pacific Northwest
NATIONAL LABORATORY

Originally Operated by Battelle Since 1965





The StreamWorks Architecture



StreamWorks is *not* a network analysis tool like bro.

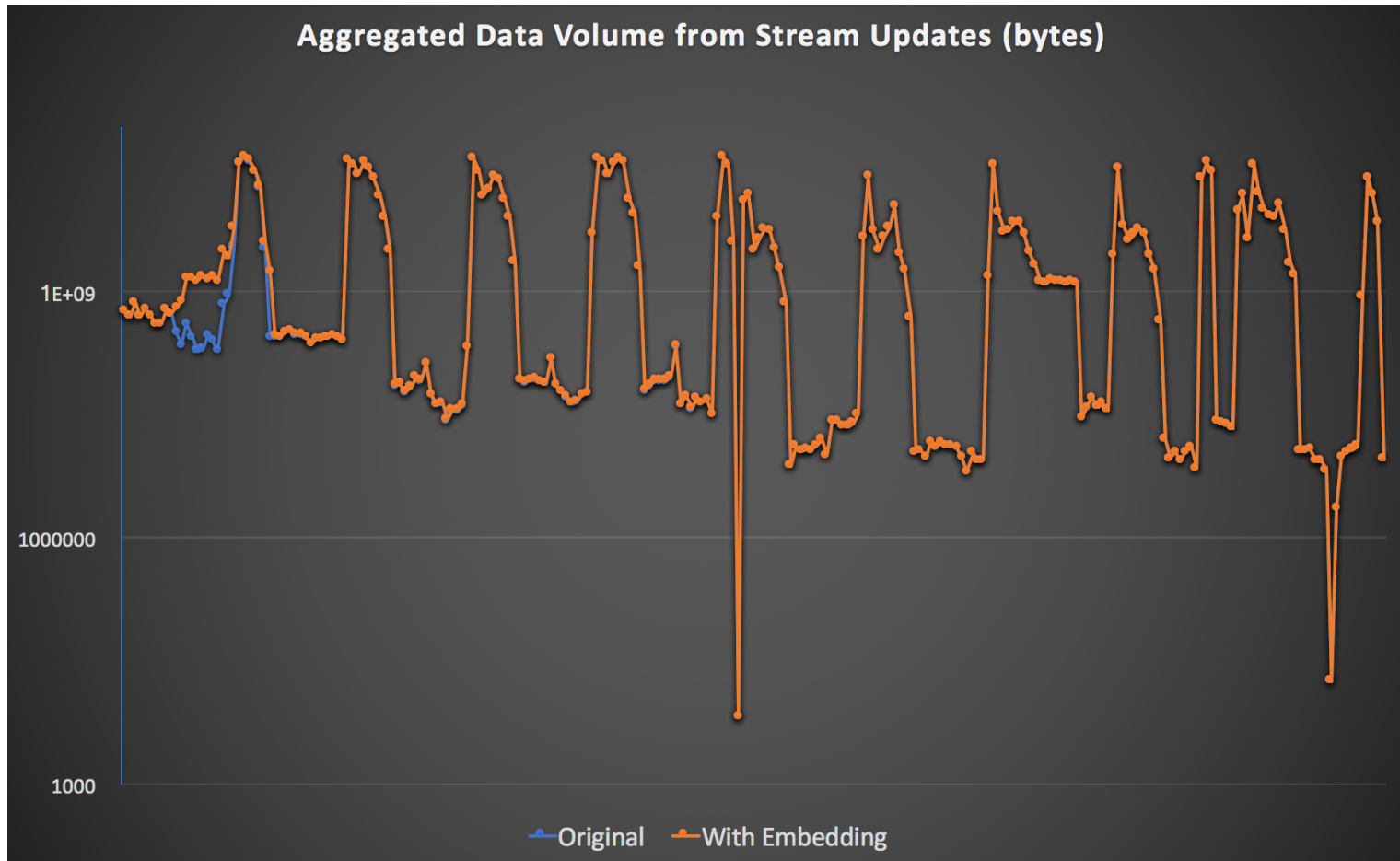
We can process nearly **50 million flows in 10 minutes** on a single multi-core server

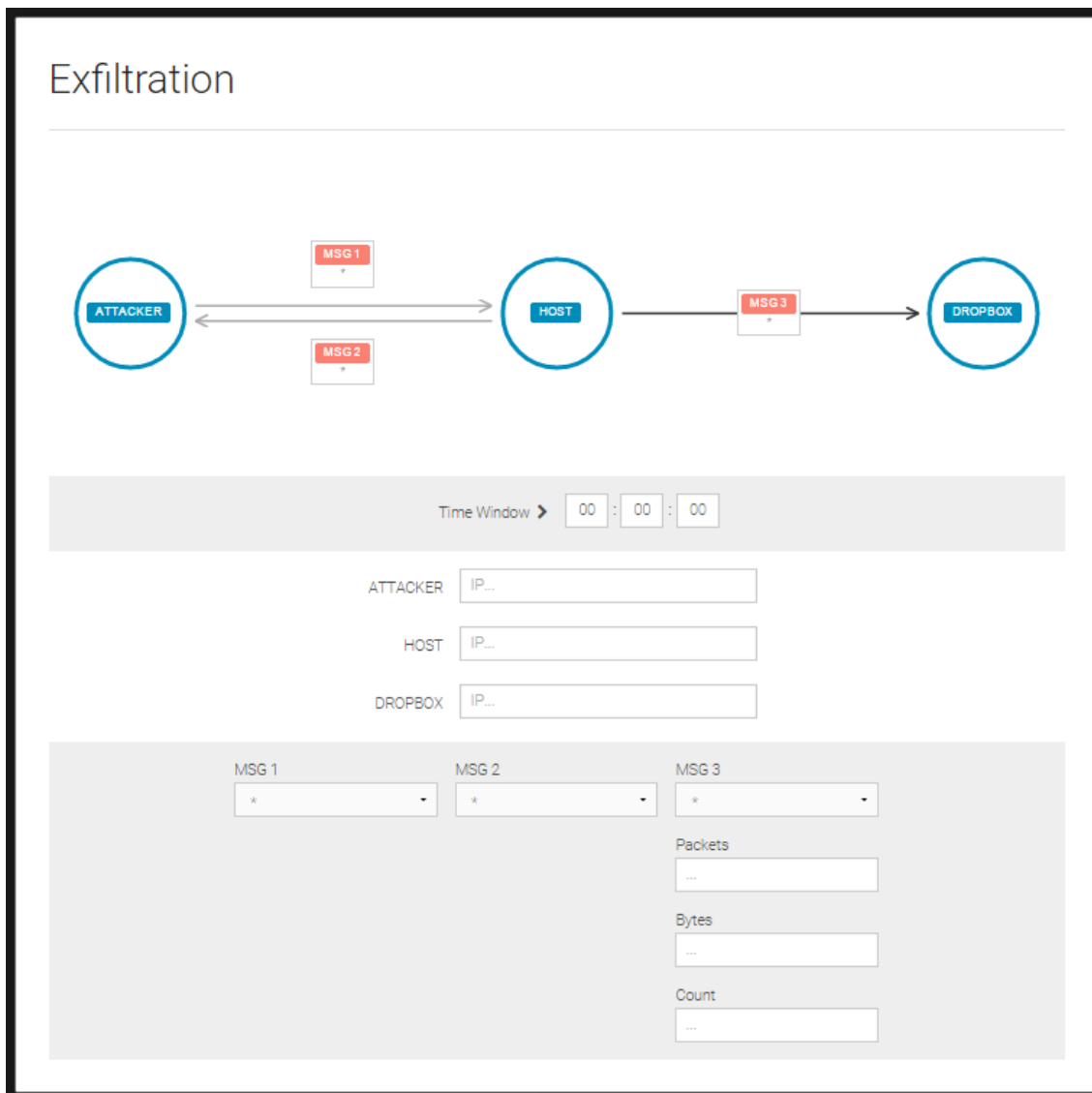
Our goal is to **tackle billion-scale datasets**



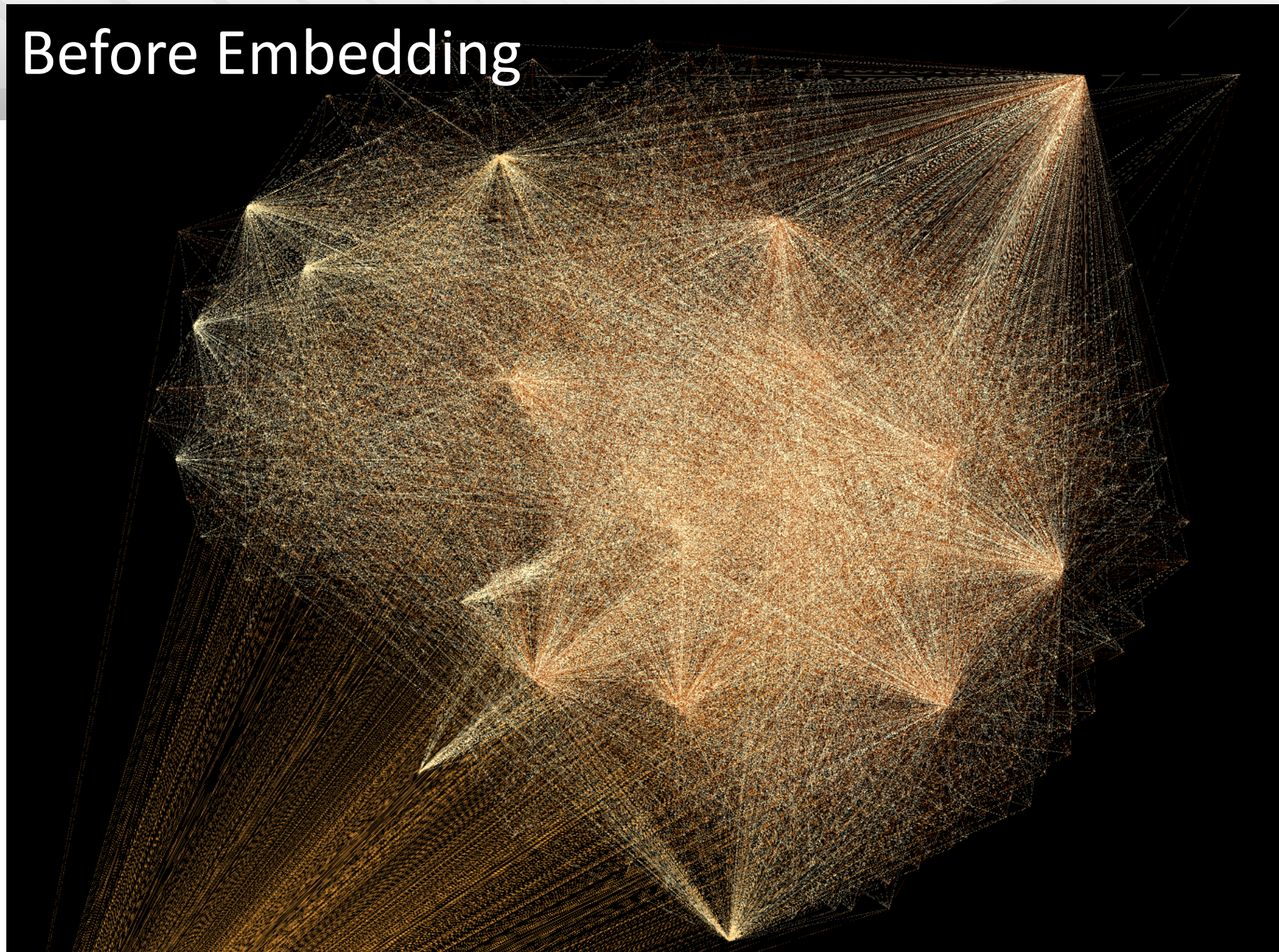
Finding the Needle in a Haystack

- ▶ Embedded multiple embeddings of exfiltration into a large-scale dataset

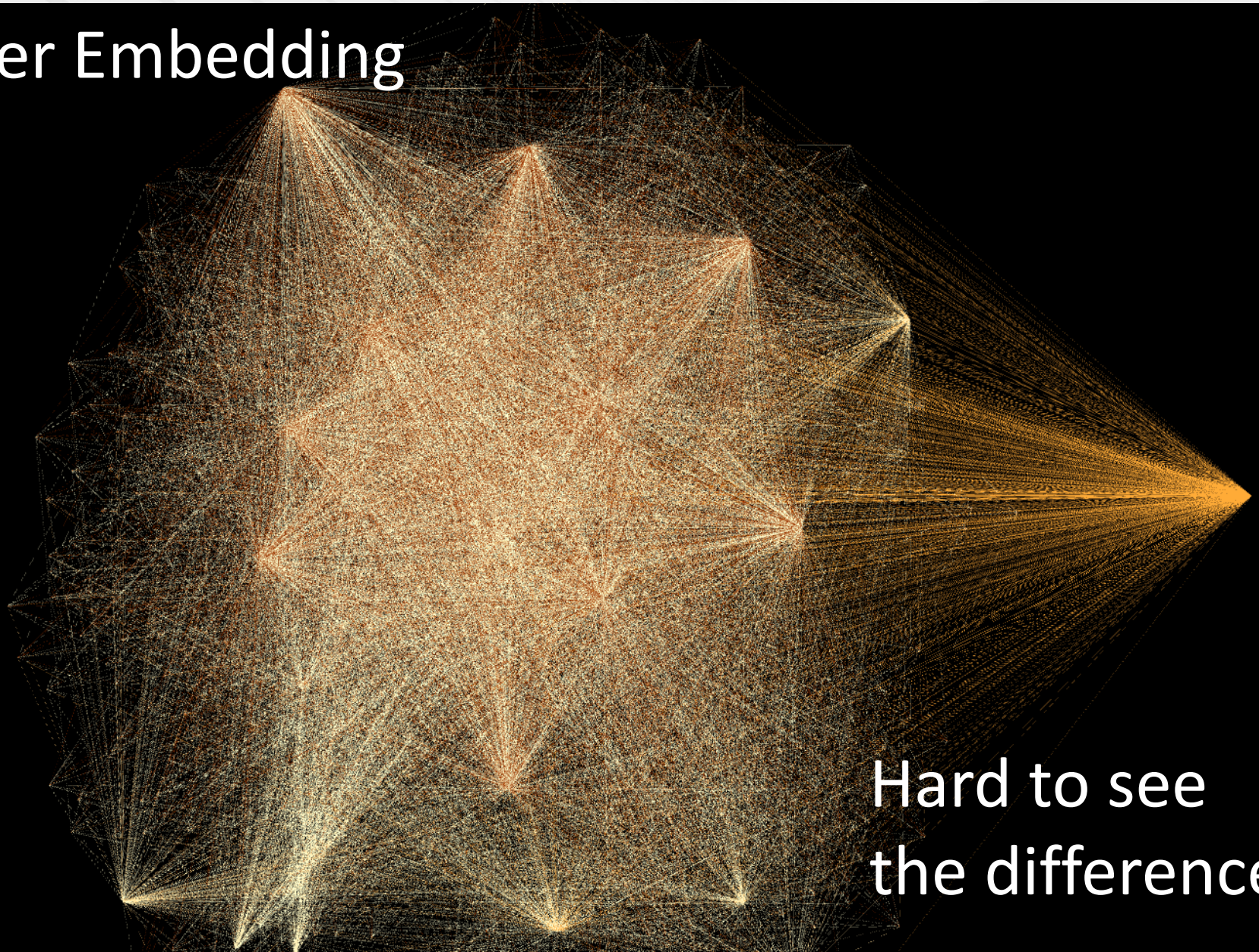




Before Embedding



After Embedding



Hard to see
the difference!

Visualization of Matching Patterns



Pacific Northwest
NATIONAL LABORATORY

Developed, Operated by Battelle Since 1965

The screenshot displays the StreamWorks application interface. On the left, a large network graph is visualized with nodes in red, blue, and white, connected by orange arrows. The graph is titled "STREAMWORKS" and "Exfiltration". A video player overlay is visible at the bottom, showing a progress bar at 01:43 and a duration of -01:57. On the right, a terminal window shows the following output:

```
[d3x771@constance01 streamworks]$ ./runBroker.sh
Starting!
Submitted batch job 1397630
Sending message (424 bytes) to /topic/cass.demo-result
Sending message (424 bytes) to /topic/cass.demo-result
Sending message (660 bytes) to /topic/cass.demo-result
Sending message (660 bytes) to /topic/cass.demo-result
Sending message (931 bytes) to /topic/cass.demo-result
Sending message (931 bytes) to /topic/cass.demo-result
Sending message (1434 bytes) to /topic/cass.demo-result
Sending message (1434 bytes) to /topic/cass.demo-result
Sending message (1319 bytes) to /topic/cass.demo-result
Sending message (1319 bytes) to /topic/cass.demo-result
Sending message (1595 bytes) to /topic/cass.demo-result

```

Below the terminal, a log window shows the following output:

```
17/05/12 17:09:32 INFO StreamingSearchWithJoinTree$: Processin
/pic/projects/streaming_graph/vast/multi_exfil_data/vast_3791
.r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,0,0,0,0)
17/05/12 17:09:35 INFO StreamingSearchWithJoinTree$: Processin
/pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,0,0,11,11)
17/05/12 17:09:37 INFO StreamingSearchWithJoinTree$: Processin
/pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,11,11,17,17)
17/05/12 17:09:40 INFO StreamingSearchWithJoinTree$: Processin
/pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,28,28,24,24)
17/05/12 17:09:43 INFO StreamingSearchWithJoinTree$: Processin
/pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,52,52,37,37)
17/05/12 17:09:46 INFO StreamingSearchWithJoinTree$: Processin
/pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,89,89,34,34)
17/05/12 17:09:49 INFO StreamingSearchWithJoinTree$: Processin
/pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,123,123,41,41)

```

Visualization of matching patterns



Pacific Northwest
NATIONAL LABORATORY

Developed, Operated by Battelle Since 1965

The screenshot displays the StreamWorks application interface. The main window, titled "STREAMWORKS", shows a complex network graph with numerous nodes and edges. A specific node is highlighted with a red circle and labeled "12". Below the graph, there is a video player interface with a progress bar showing 02:00 and -01:40, and two circular buttons labeled "360".

The terminal window on the right shows the following log output:

```
Sending message (931 bytes) to /topic/cass.demo-result
Sending message (1434 bytes) to /topic/cass.demo-result
Sending message (1434 bytes) to /topic/cass.demo-result
Sending message (1319 bytes) to /topic/cass.demo-result
Sending message (1319 bytes) to /topic/cass.demo-result
Sending message (1595 bytes) to /topic/cass.demo-result
Sending message (1595 bytes) to /topic/cass.demo-result
Sending message (1757 bytes) to /topic/cass.demo-result
Sending message (1757 bytes) to /topic/cass.demo-result
Sending message (1715 bytes) to /topic/cass.demo-result
Sending message (1715 bytes) to /topic/cass.demo-result
Sending message (2028 bytes) to /topic/cass.demo-result
Sending message (2136 bytes) to /topic/cass.demo-result
Sending message (2136 bytes) to /topic/cass.demo-result
Sending message (150 bytes) to /topic/cass.demo-result
Sending message (222 bytes) to /topic/cass.demo-result
Sending message (144 bytes) to /topic/cass.demo-result
Sending message (149 bytes) to /topic/cass.demo-result
Sending message (72 bytes) to /topic/cass.demo-result
Sending message (148 bytes) to /topic/cass.demo-result

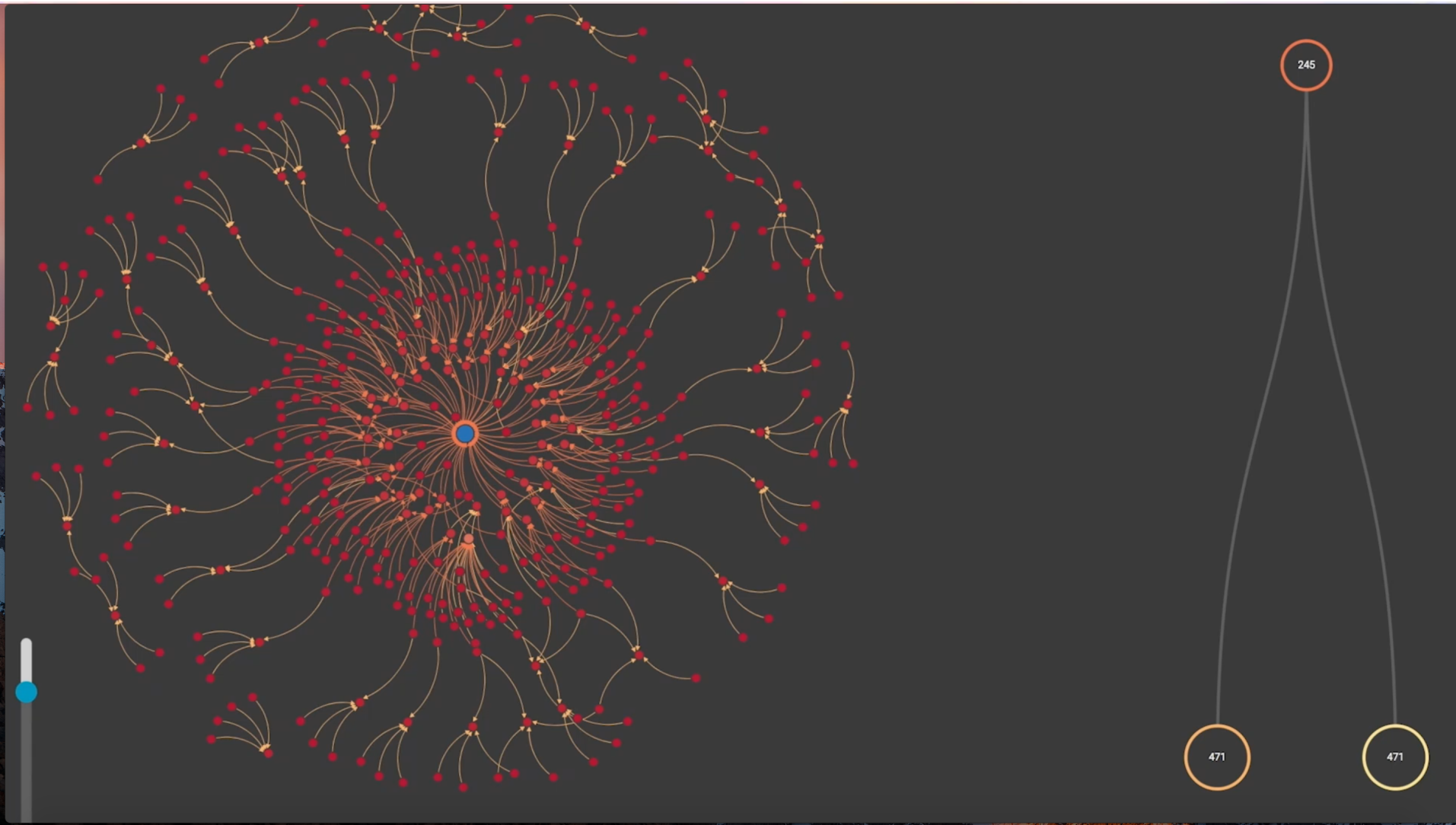
17/05/12 17:09:49 INFO StreamingSearchWithJoinTree$: Processin
l /pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,123,123,41,41)
17/05/12 17:09:52 INFO StreamingSearchWithJoinTree$: Processin
l /pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,164,164,45,45)
17/05/12 17:09:56 INFO StreamingSearchWithJoinTree$: Processin
l /pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,209,209,44,44)
17/05/12 17:10:03 INFO StreamingSearchWithJoinTree$: Processin
l /pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,253,253,52,52)
17/05/12 17:10:07 INFO StreamingSearchWithJoinTree$: Processin
l /pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,305,305,55,55)
17/05/12 17:10:12 INFO StreamingSearchWithJoinTree$: Processin
l /pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,360,360,40,40)
17/05/12 17:10:16 INFO StreamingSearchWithJoinTree$: Processin
l /pic/projects/streaming_graph/vast/multi_exfil_data/vast_379
r.tsv
(LHS(prevBatch), RHS(prevBatch), LHS, RHS,400,400,29,29)
```

Visualization of Matching Patterns



Pacific Northwest
NATIONAL LABORATORY

Originally Operated by Battelle Since 1965



Visualization of Matching Patterns

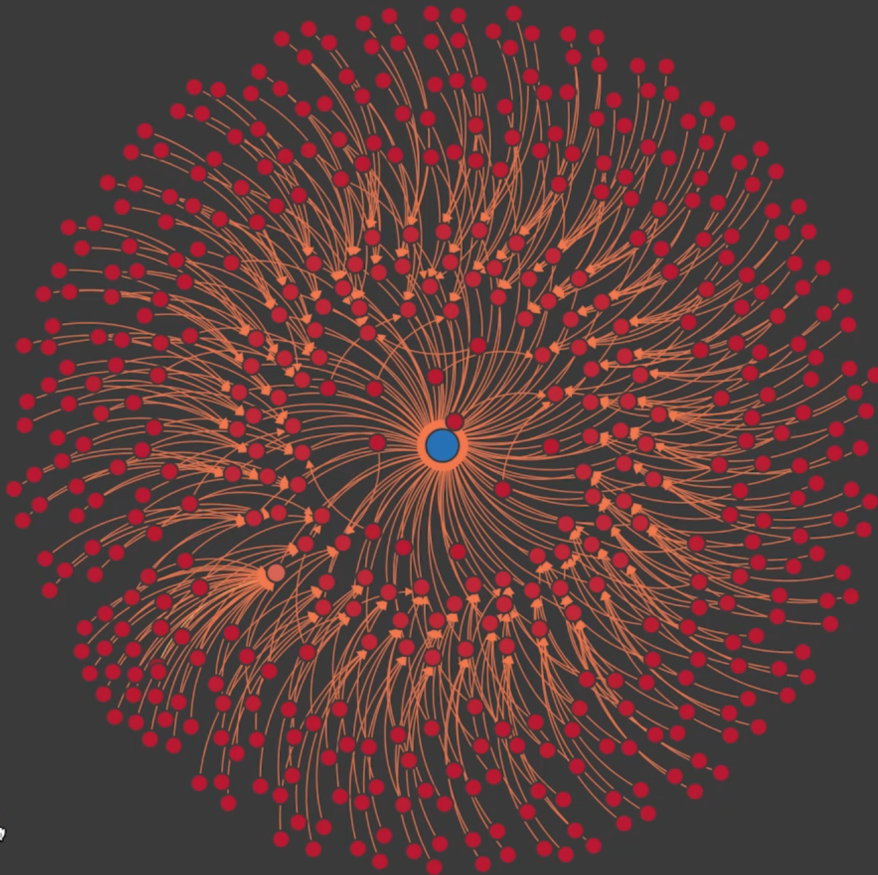


Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

STREAMWORKS

Exfiltration



Providing a geographical perspective



Another example of Geo-View



Pacific Northwest
NATIONAL LABORATORY

Originally Operated by Battelle Since 1965



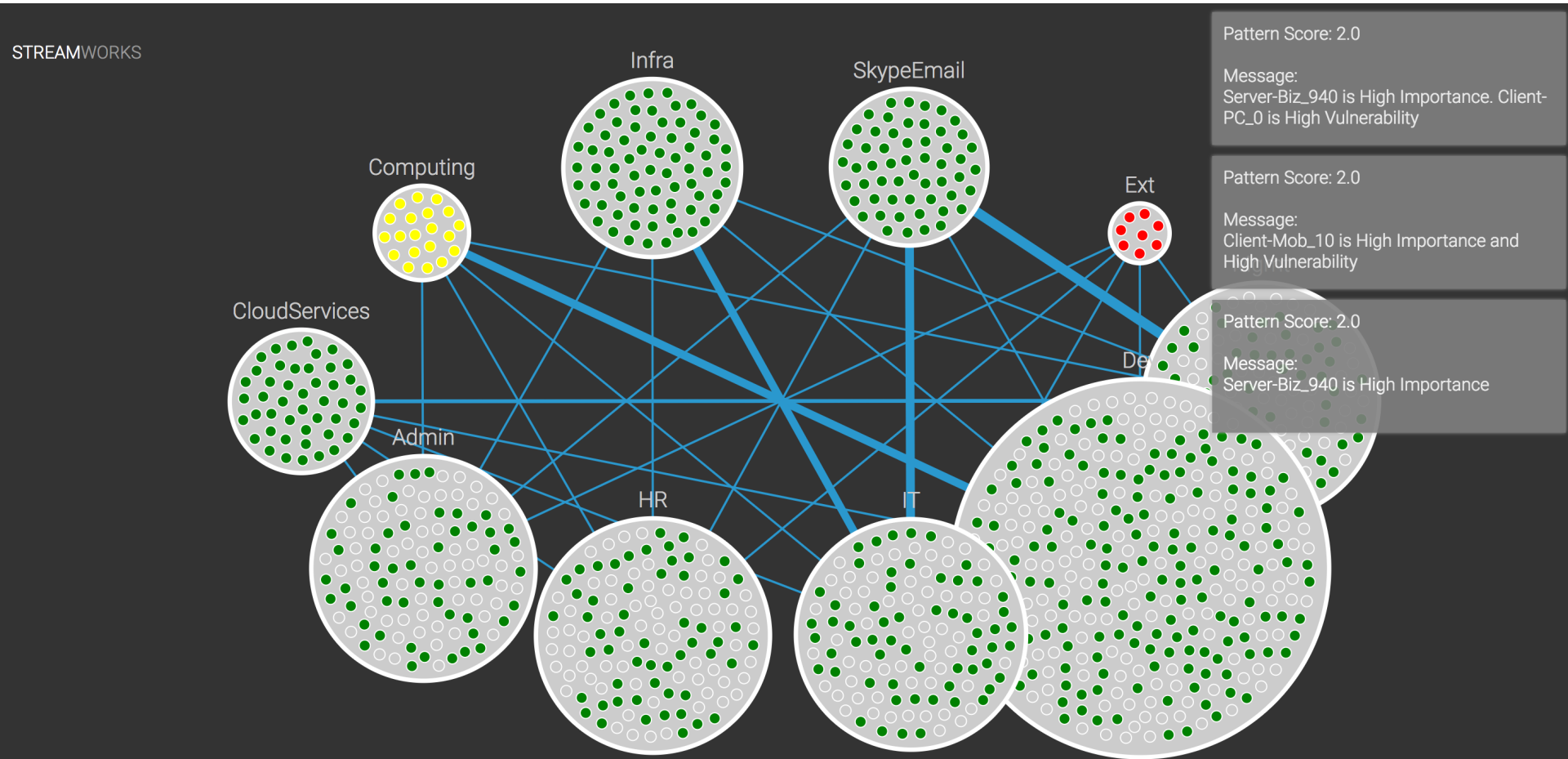
Tell Me Why!



Pacific Northwest
NATIONAL LABORATORY

Originally Operated by Battelle Since 1965

- ▶ Too many matches is a problem
- ▶ Rank and Explain through background knowledge and behavioral patterns learnt from data

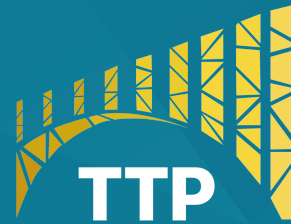




Competition

Product	Streaming	Graph Search	Visual Analytics
StreamWorks	✓	✓	✓
SQRRL Enterprise	✗	✓	✓
Apache Spark	✓	✗	✗
Neo4J	✗	✓	✗

- We obtained 10-100x improvement in runtime on an internet backbone traffic flow dataset.
- Filed US Patent on graph based pattern matching technology



TRANSITION TO PRACTICE

THANK YOU!

StreamWorks
sutanay.choudhury@pnnl.gov