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

U.S. DEPARTMENT OF ENERGY

What is StreamWorks?

The Promise of Patterns

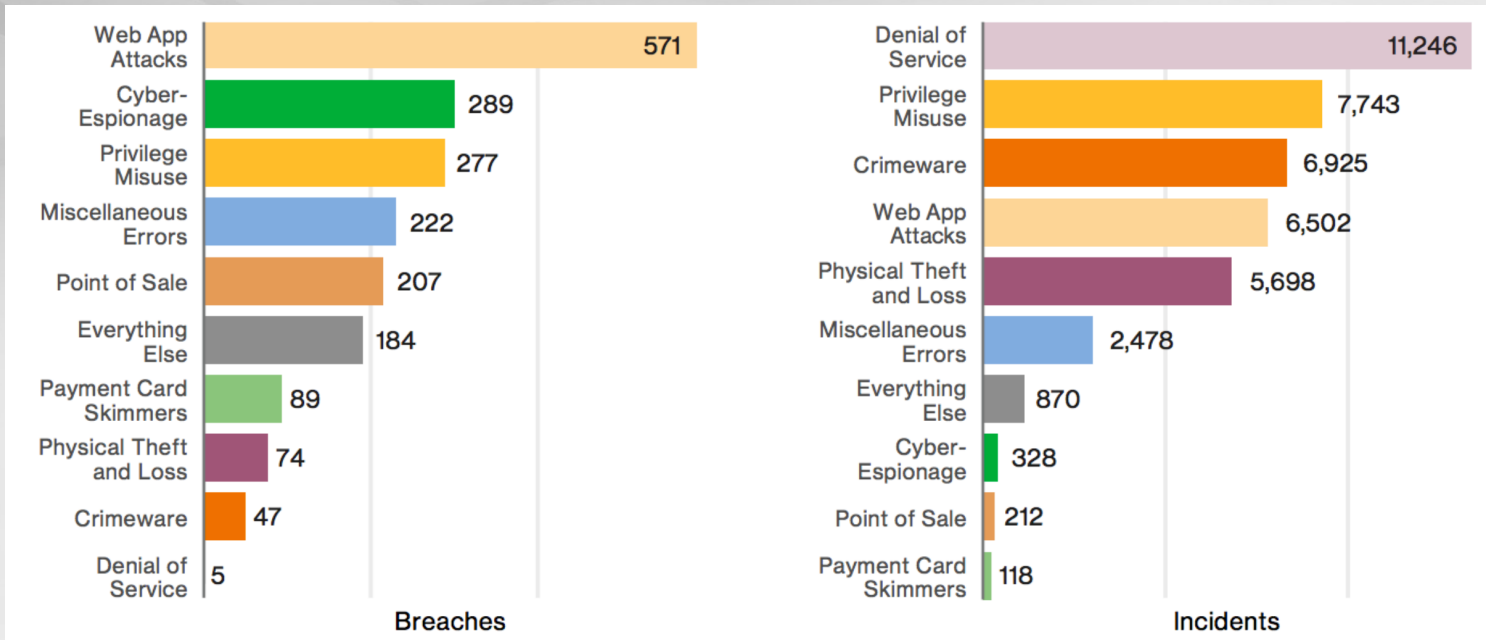


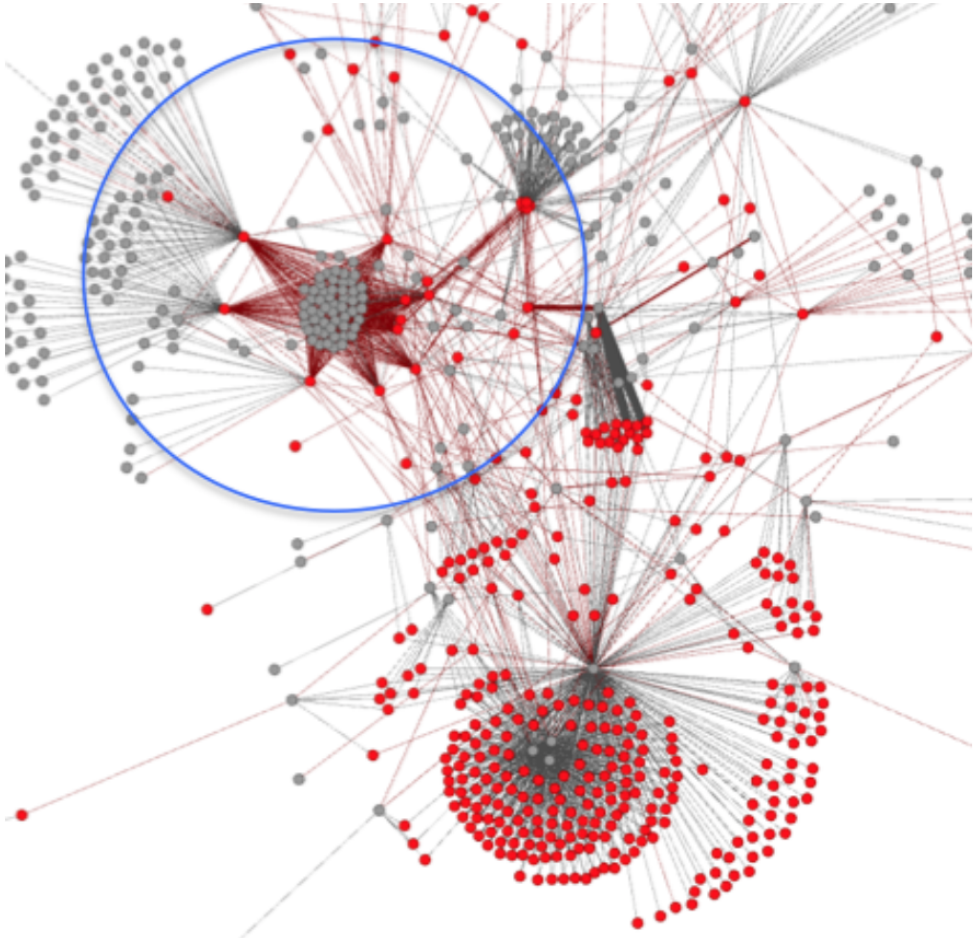
Figure 33: Percentage and count of breaches per pattern (n=1,935)

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



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!

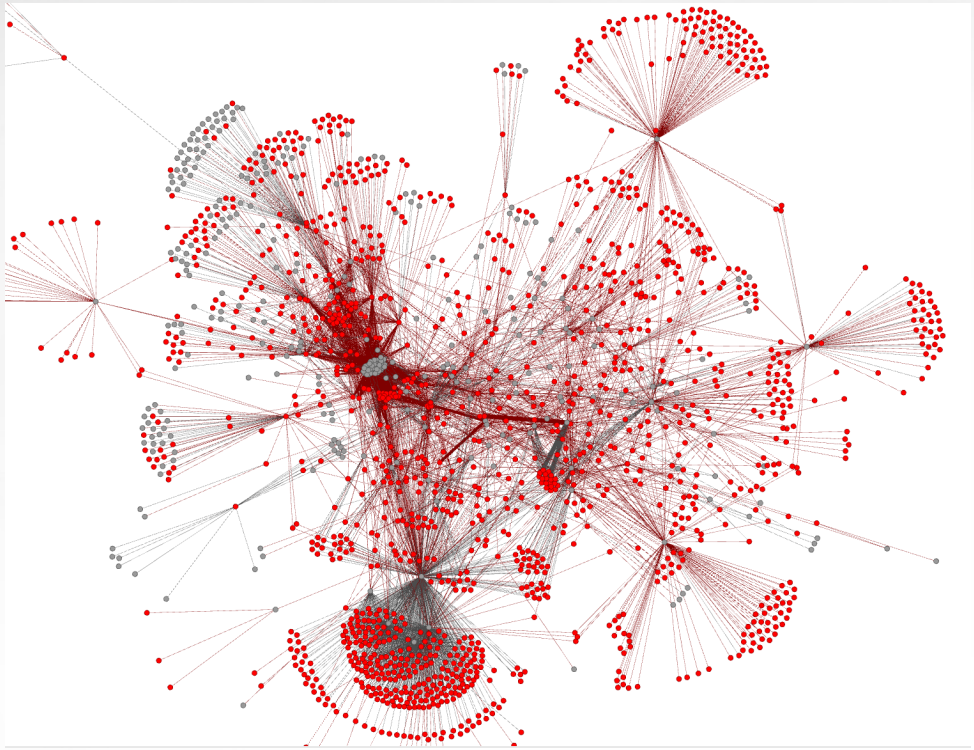
- **How do you read email**?
  - Read every email as soon as it comes in (**Continuous Processing**)
  - Read every 4 hours (**Periodic or Batched Processing**)

- **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

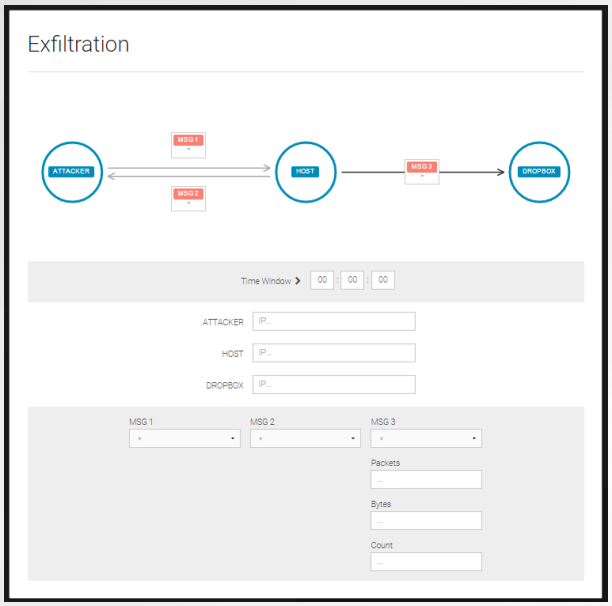
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- **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?”



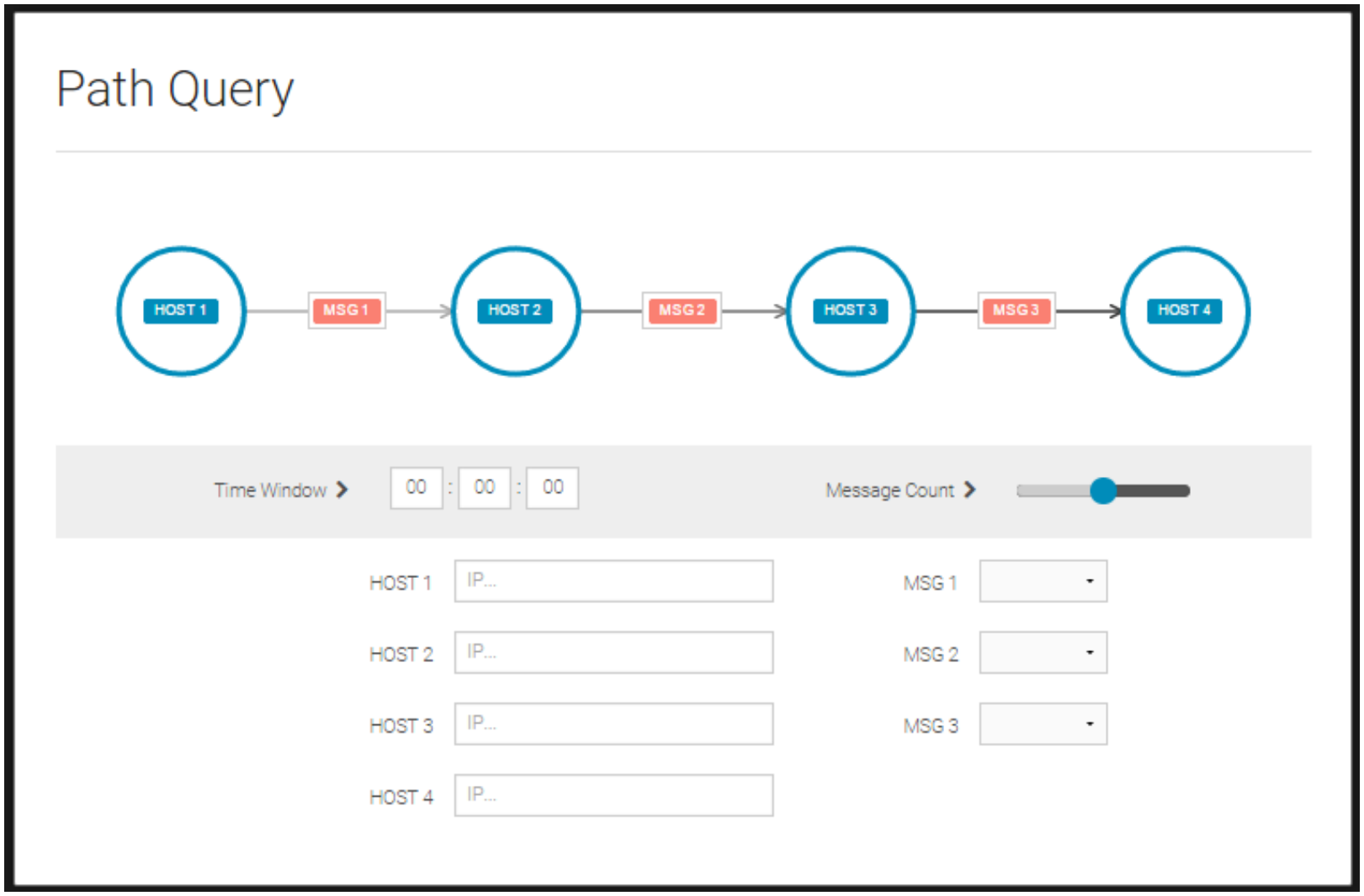
And last, but not the least …

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

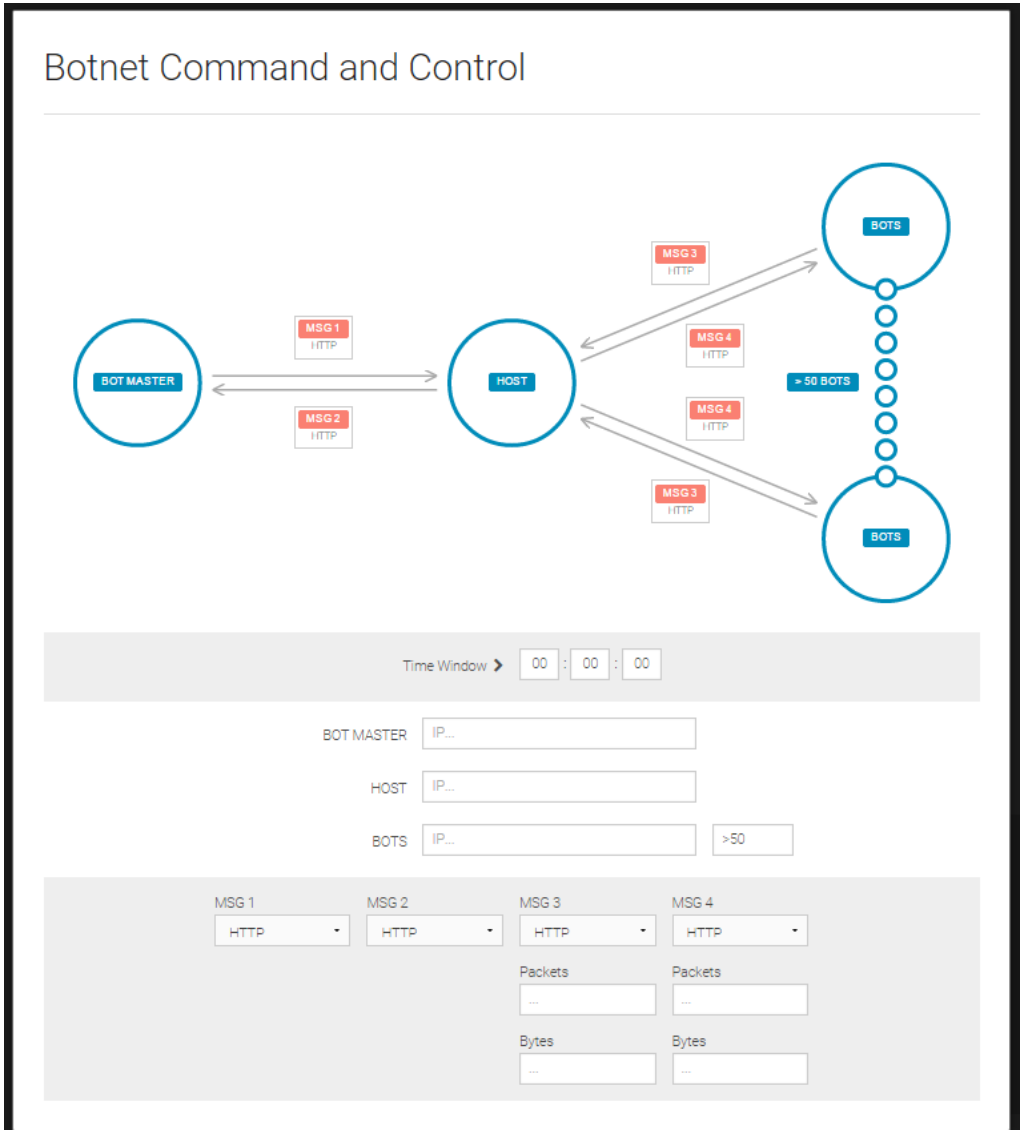


```
SELECT ?control ?target ?dropbox ?xfil WHERE {
 # Control Message from C2 to target
  ?control ?ctrlmsg ?target .
  ?ctrlmsg :FTIME ?ftime1 .
  ?ctrlmsg :STIME ?stime1 .
  ?ctrlmsg :DPKTS ?pkts1 .
  ?ctrlmsg :DOCTETS ?octets1 .
  FILTER (?pkts1 < 3 && ?octets1 < 300)

  # 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)
  }
}
```

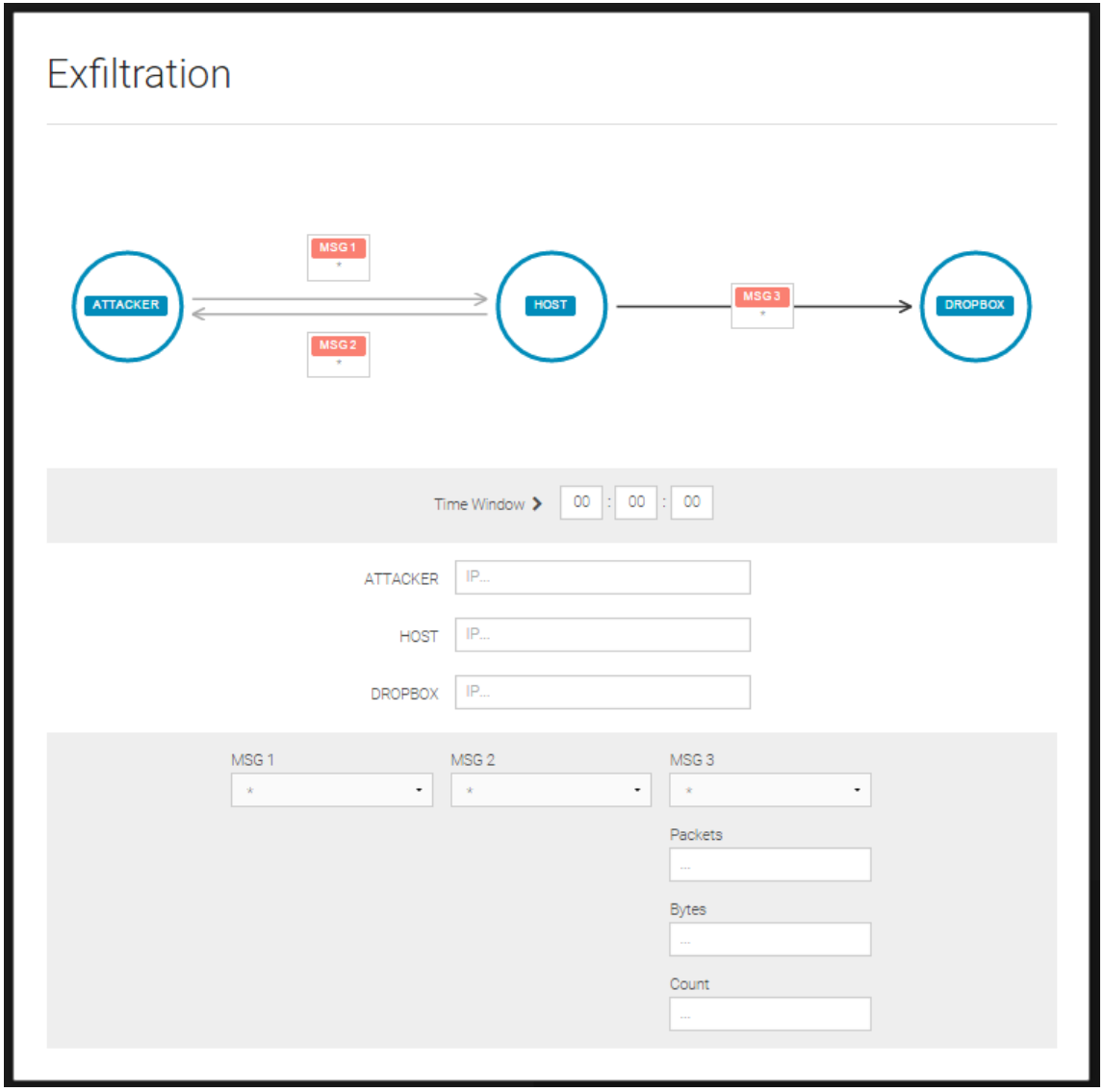
Querying for Chains of Activity



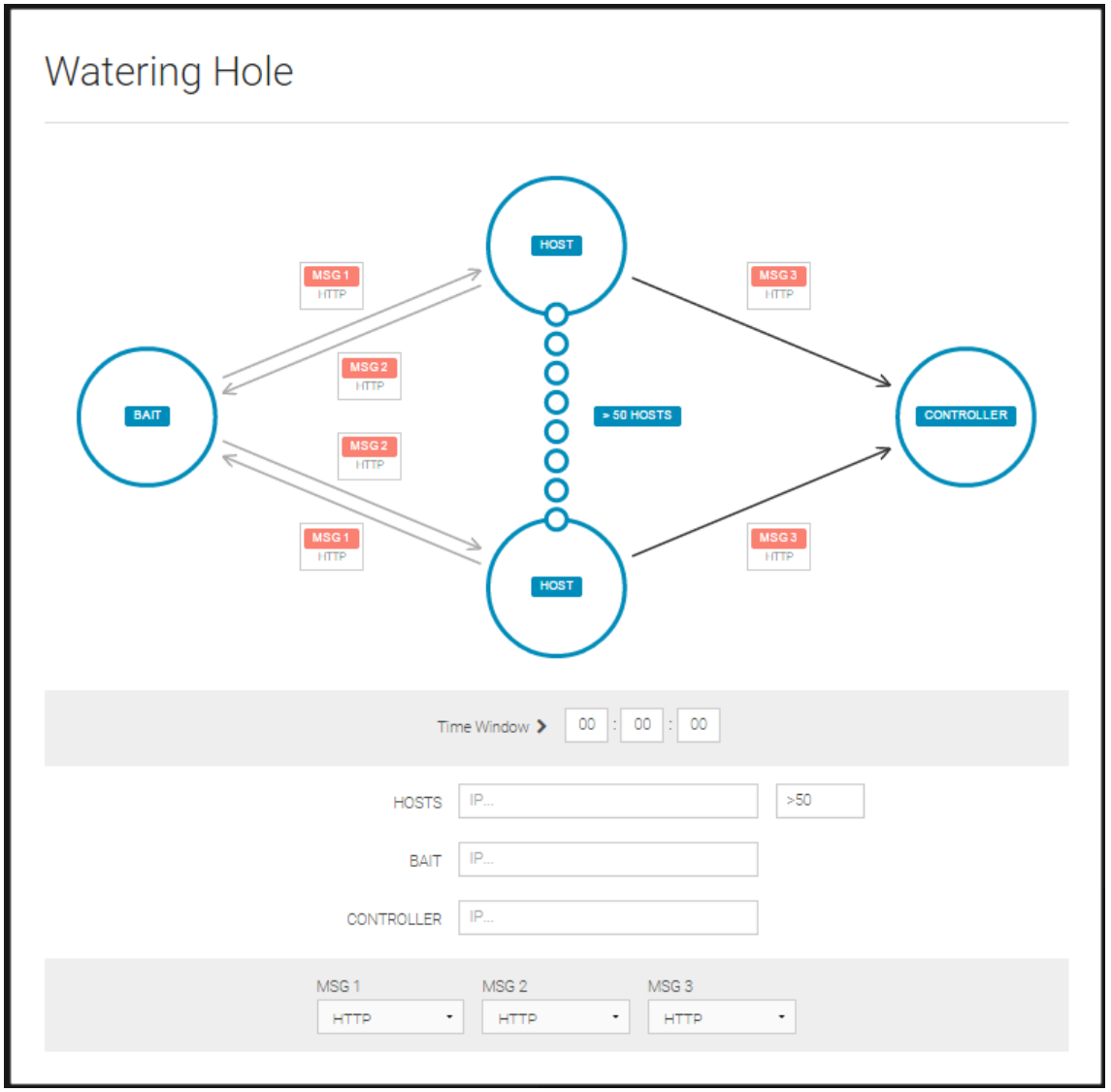
Botnet Command and Control



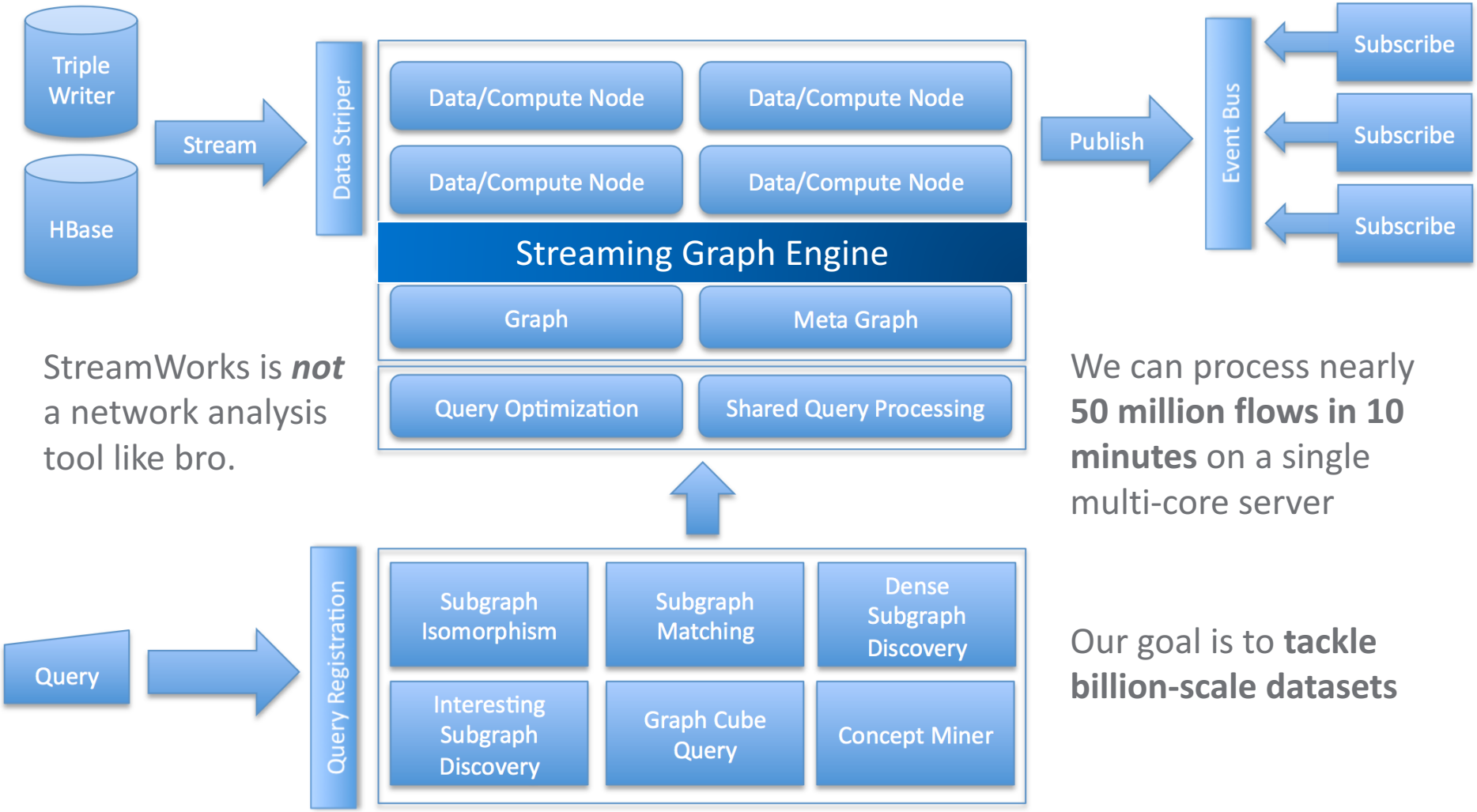
Exfiltration



Watering Hole



The StreamWorks Architecture



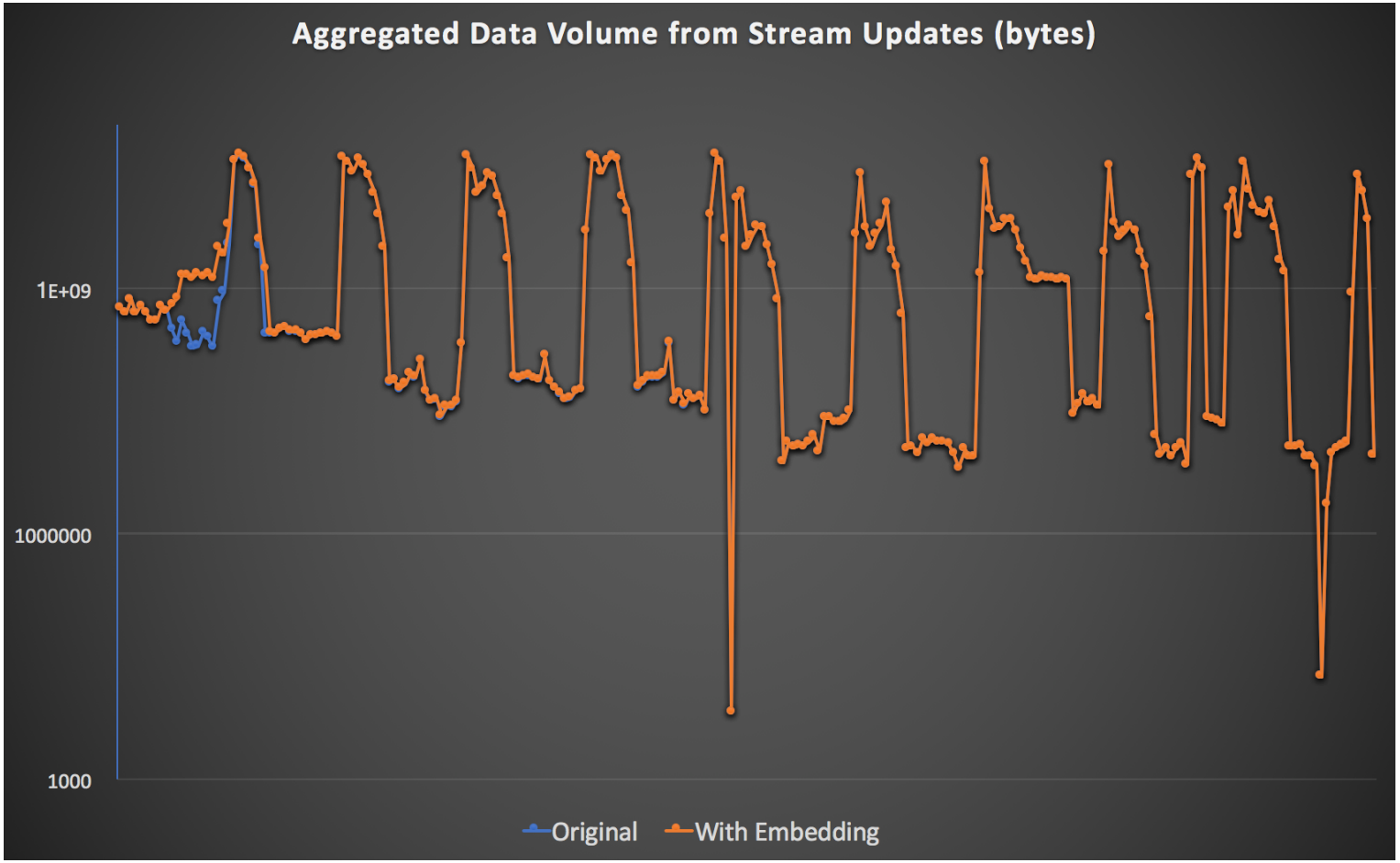
StreamWorks is ***not*** a network analysis tool like bro.

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

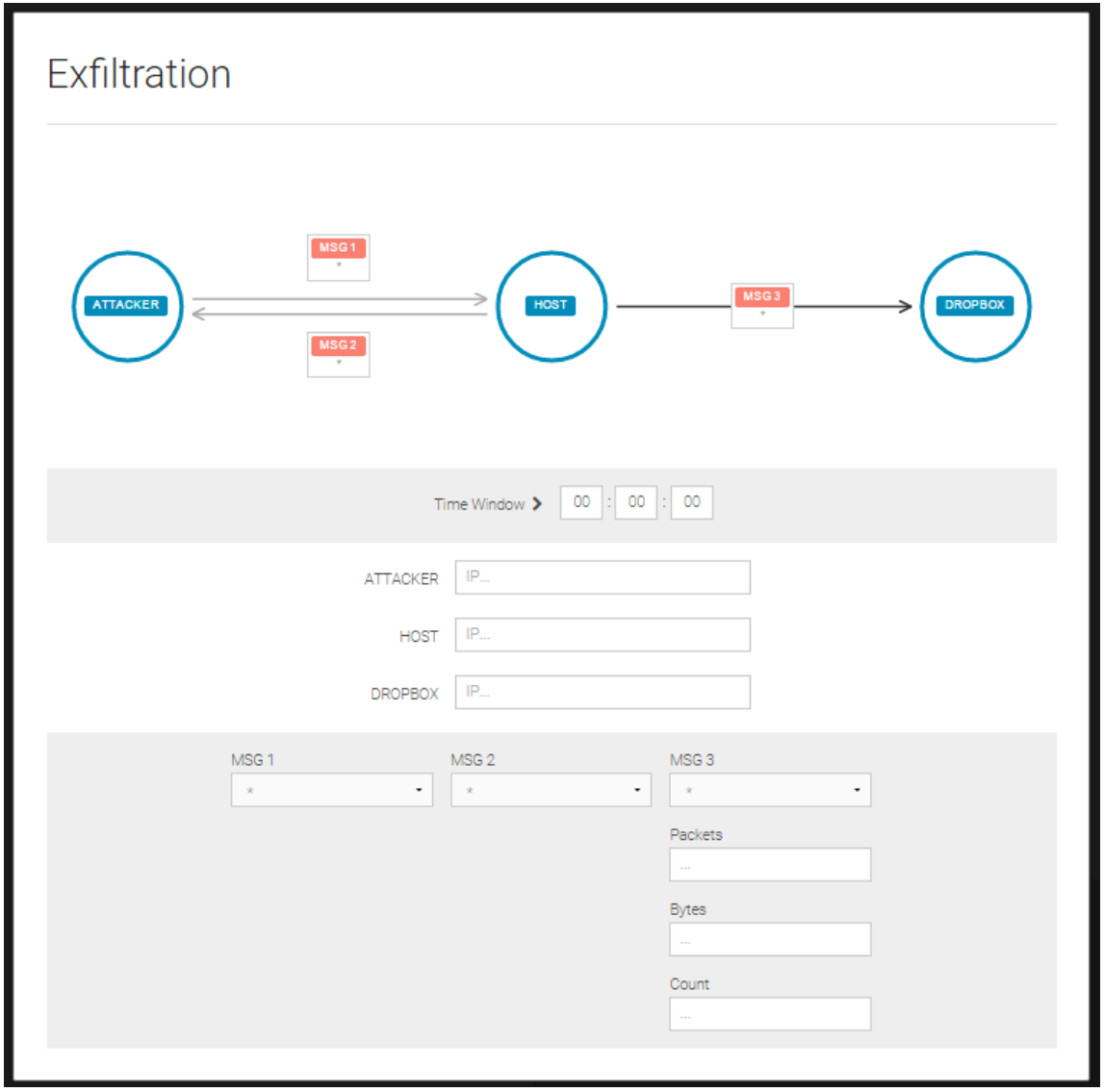
Our goal is to **tackle billion-scale datasets**

Finding the Needle in a Haystack

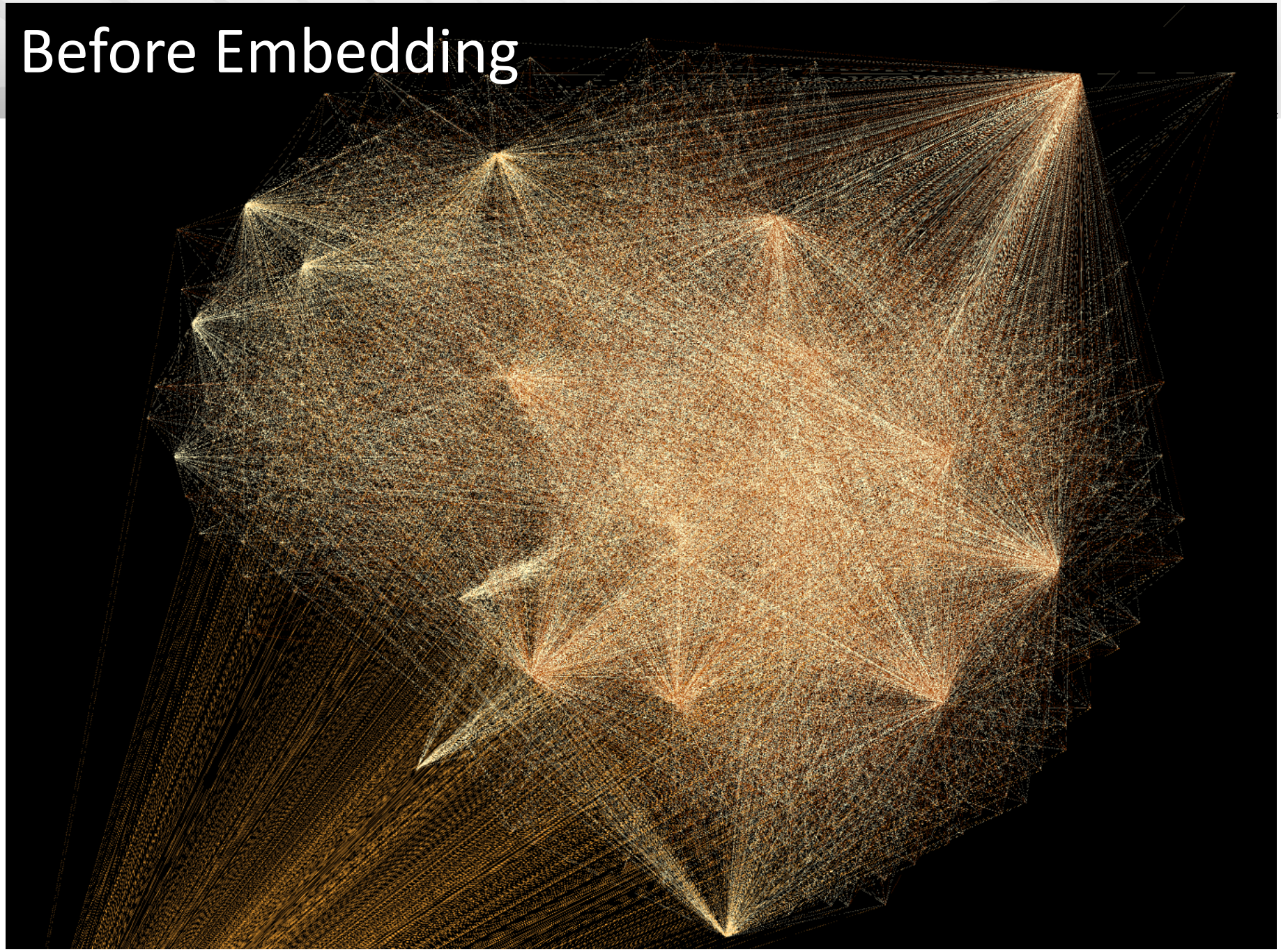
- Embedded multiple embeddings of exfiltration into a large-scale dataset



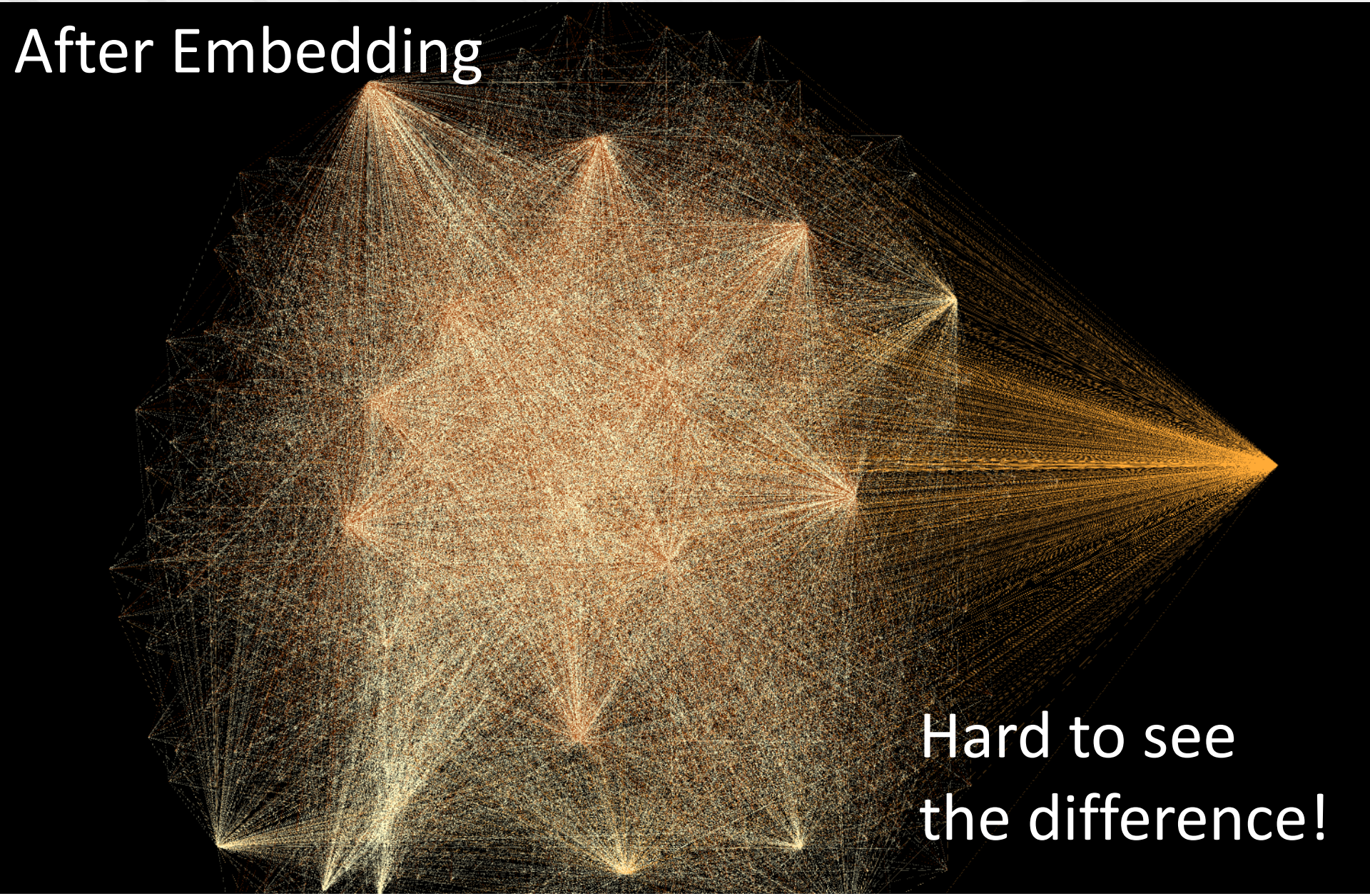
Exfiltration



Before Embedding

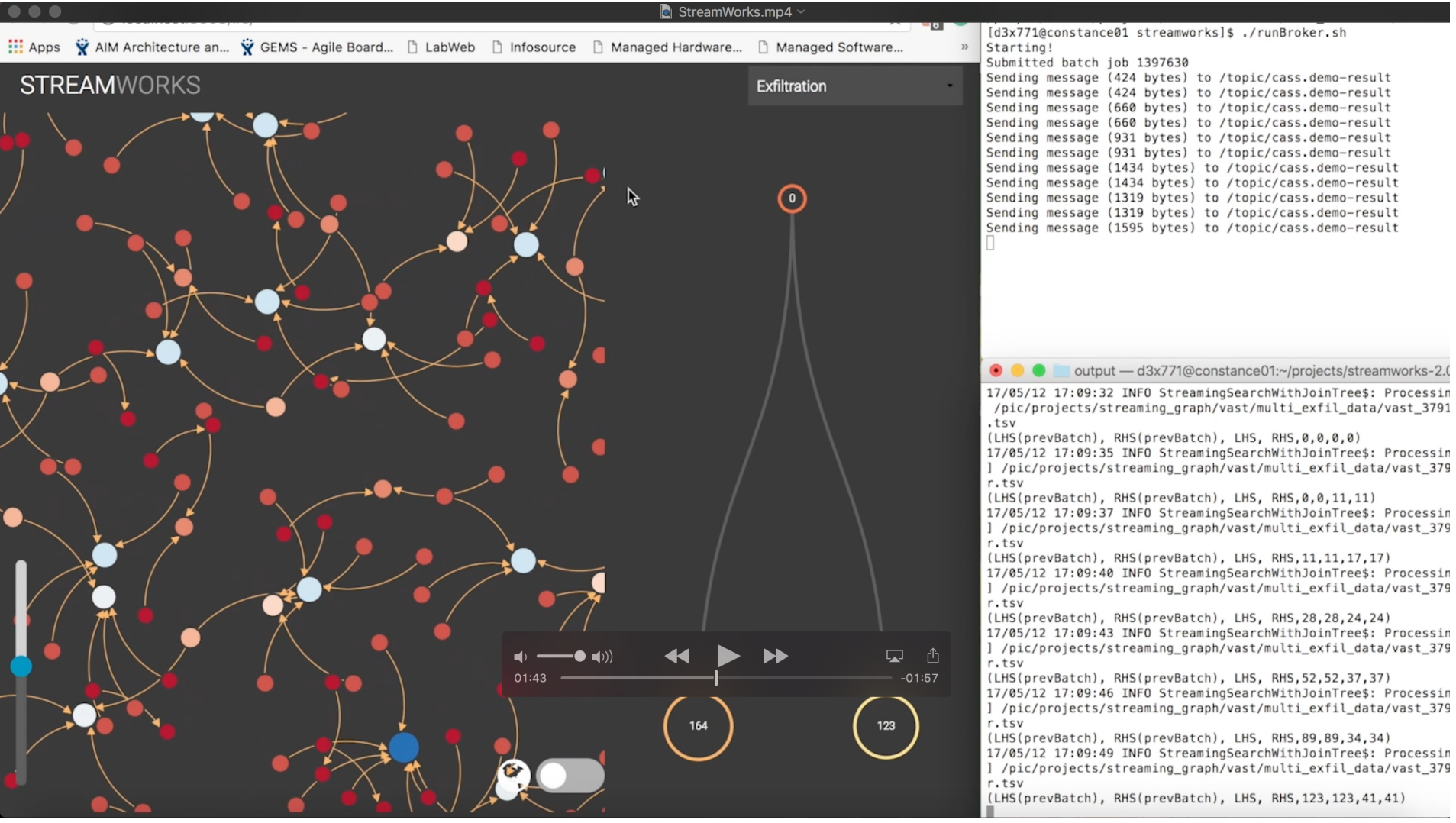


After Embedding



Visualization of Matching Patterns

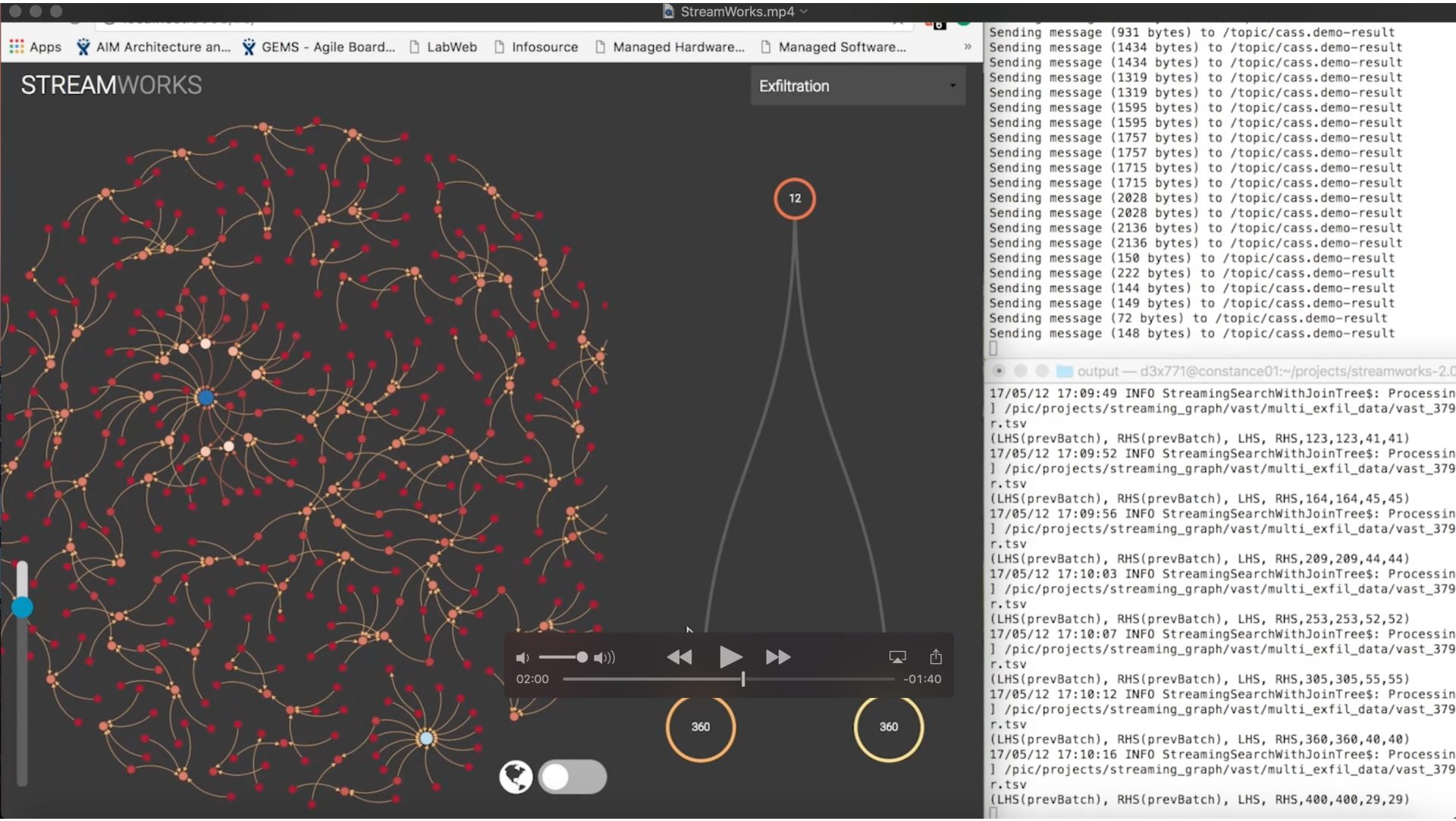
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965



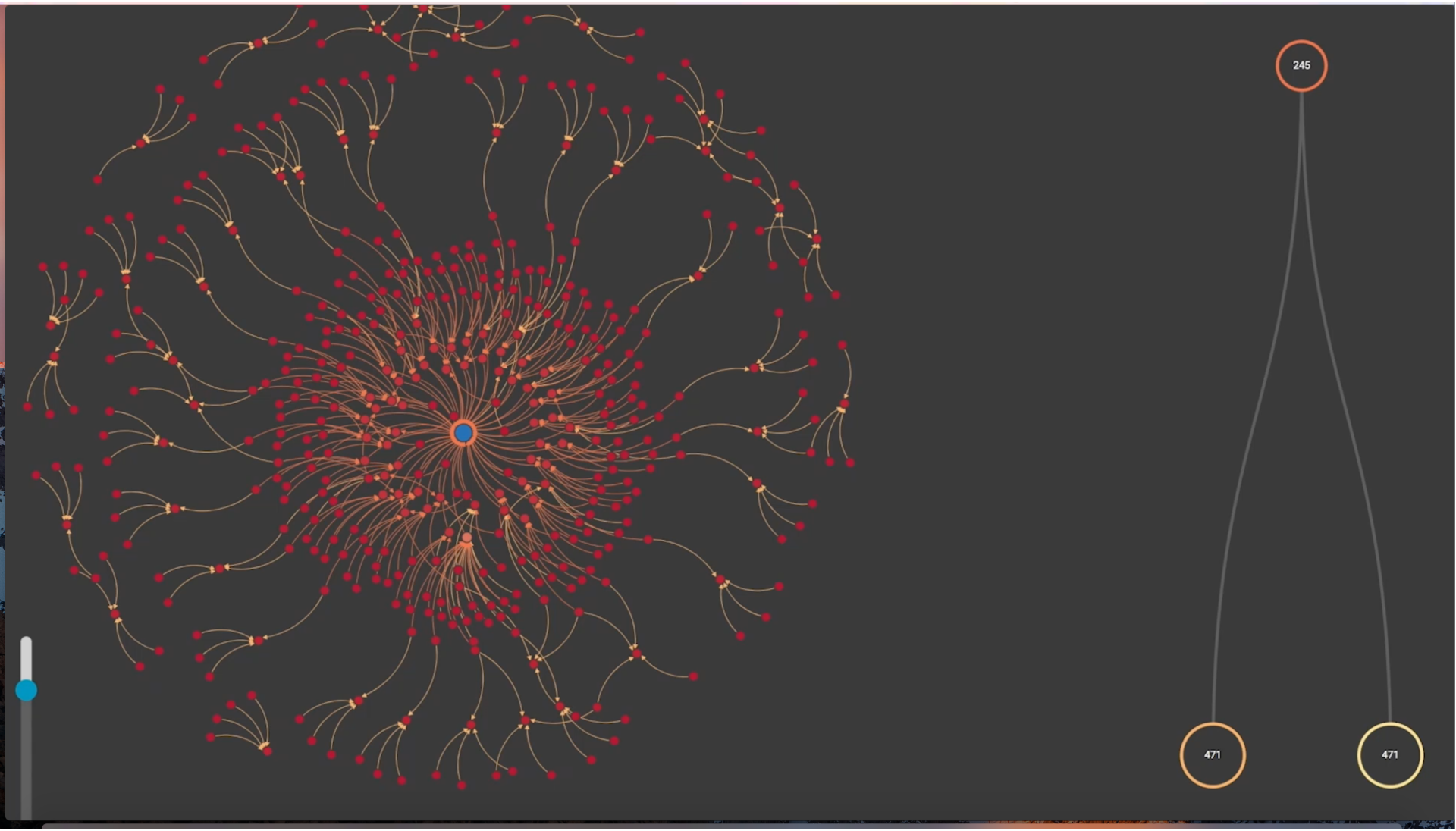
Visualization of matching patterns

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

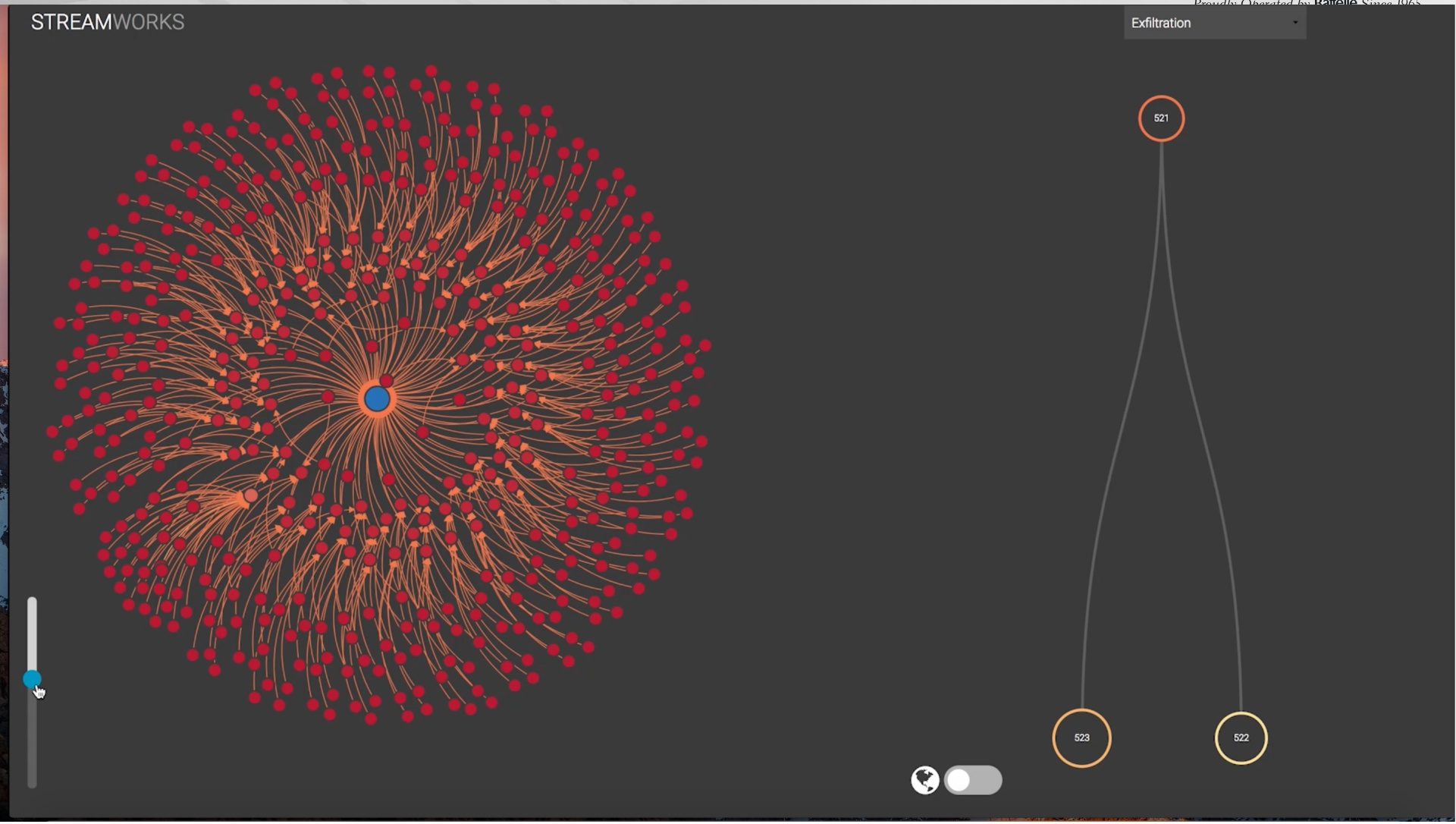


Visualization of Matching Patterns



Visualization of Matching Patterns

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965



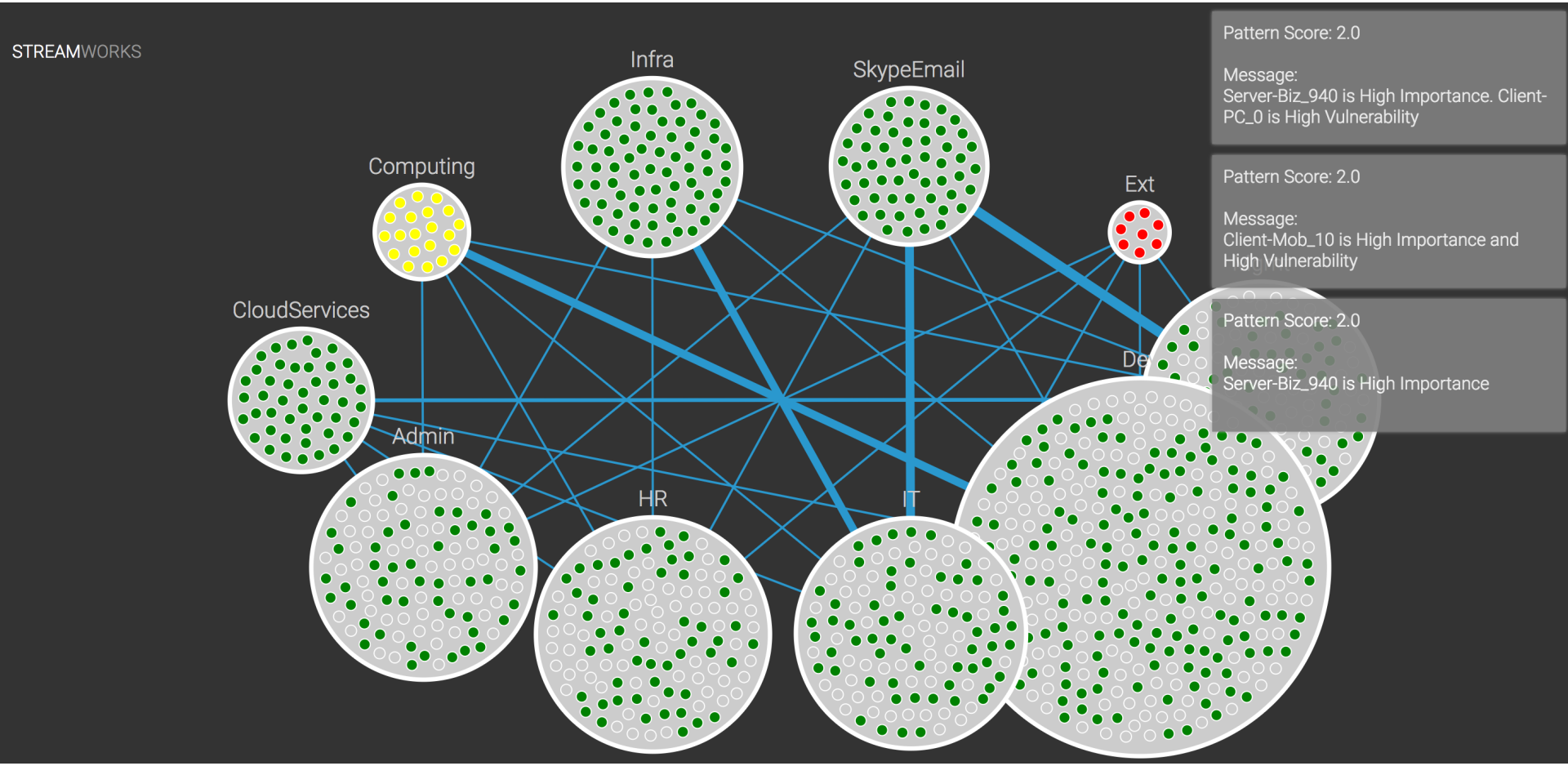
Providing a geographical perspective

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Another example of Geo-View

Tell Me Why!

- 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**

Pacific Northwest
NATIONAL LABORATORY

TTP
TRANSITION TO PRACTICE

THANK YOU!

StreamWorks

sutanay.choudhury@pnnl.gov

Homeland Security
Science and Technology

This technology has been brought to you by the DHS S&T Cyber Security Division Transition to Practice (TTP) Program. For more information, contact ST.TTP@hq.dhs.gov