Continuous Pattern Detection on Streaming Data: Present and Future

SUTANAY CHOUDHURY

JOINT WORK WITH: KHUSHBU AGARWAL, SHERMAN BEUS, DANIEL DOHNALEK, GEORGE CHIN

Workshop on “Graphs and Security”, Microsoft BlueHat, Redmond WA, 2018

Introduction and Outline

- Part 1: What we have now (StreamWorks)
  - Stream based reasoning – problem definition and algorithms
  - Cyber use cases for streaming analytics
  - Demonstration
- Part 2: The future for Autonomous Cyber systems
  - Extending stream-based reasoning towards autonomous operation
  - Conceptualizing tasks for benchmarking

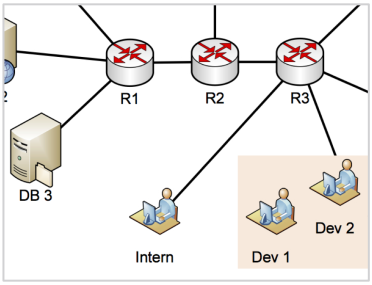
Asking a Different Question: Tell me when …

- Continuous monitoring of streaming data
  - Instead of traditional “find me all things that happened in past 24 hours” – move to a “tell me when X happens” paradigm
  - Example: Standard feature from a stock brokerage – “Tell me when MSFT goes to $150 and more than 50 million shares were traded”
  - Cyber equivalent: “Tell me when a chain of 3 logins are detected with increasing privileges?”

Malware?

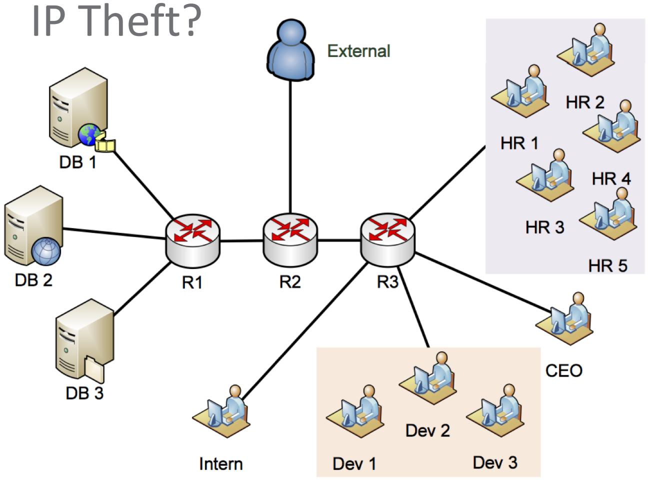
Malware installed on developer’s workstation

DB Exploit?



Project DB compromised by lateral movement

IP Theft?



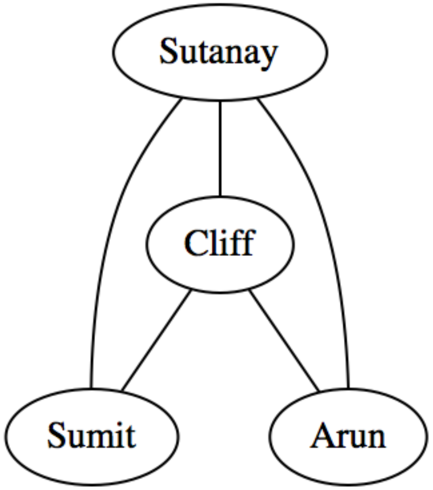
Backdoor opens to CEO’s machine when he accesses Project DB

Specific Problem Definitions

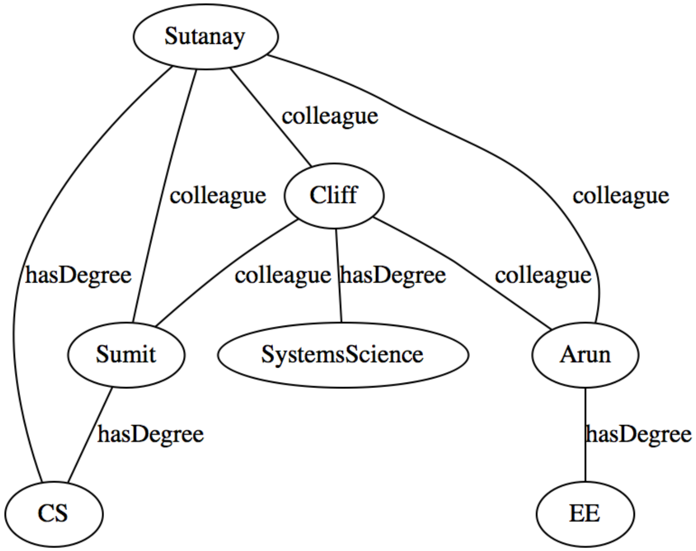
- Find the pattern from event or alert stream
  - Who is involved as source (outside) and target (inside)?
  - What type of interaction is involved?
  - What type of kill-chain behavior is represented by the pattern?
- Given (source, target and kill-chain behavior), project possible instances of the kill-chain
  - Use the Knowledge Graph to generate high probability pathways with explanations
- Interfacing with the Human Expert
  - Summarize contextual patterns to the human expert
  - Learn patterns of reviewed alerts

Property Graphs

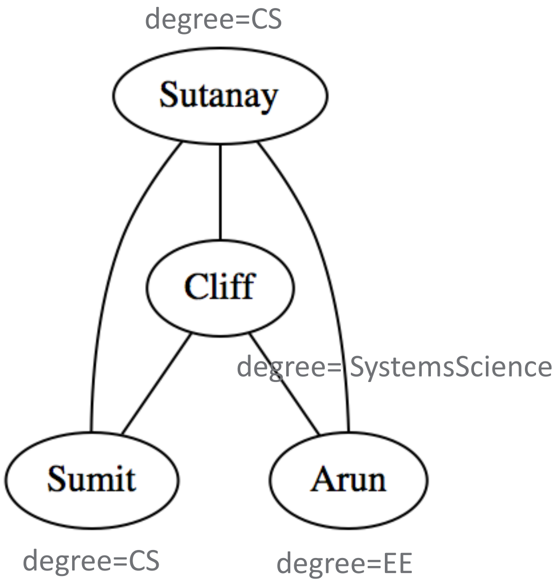
- What are property graphs, and how are they different from other representations?



Social Networks

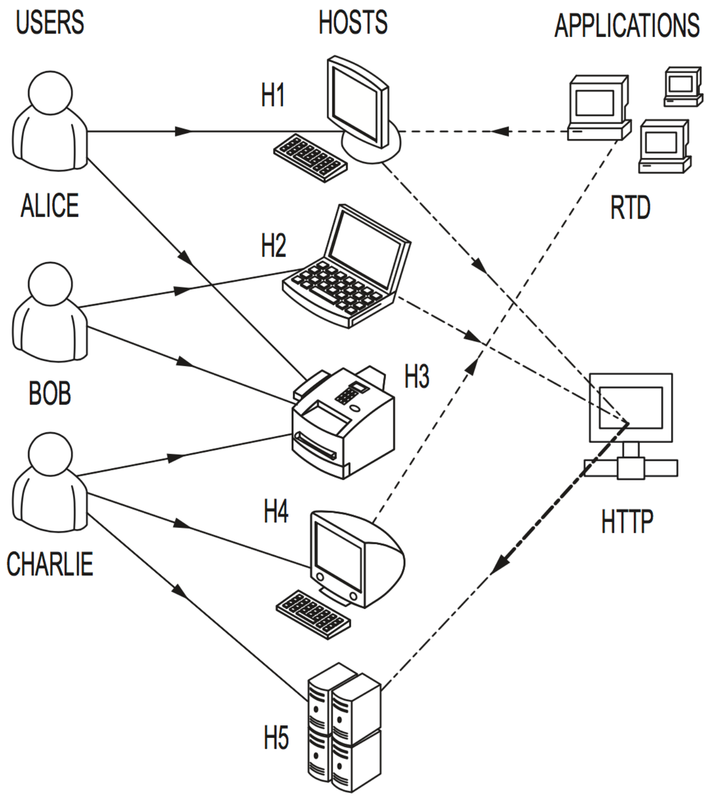


Multi-Relational RDF Graphs



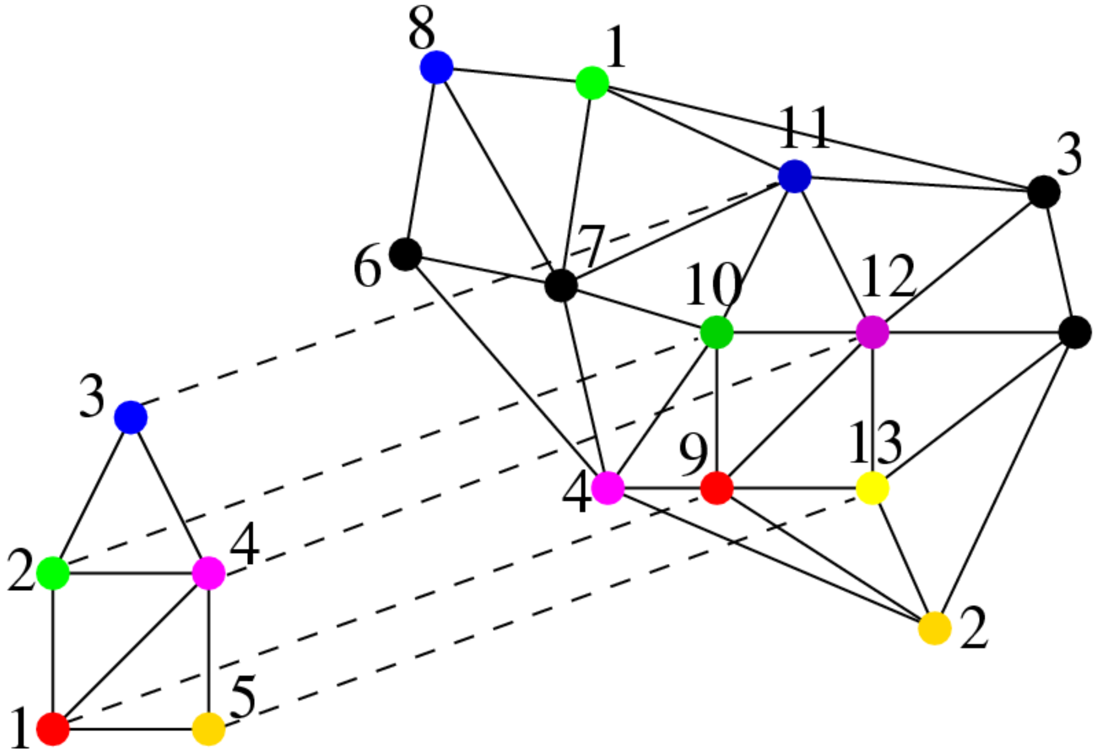
Property Graphs

Example Graph Models for Cyber Data

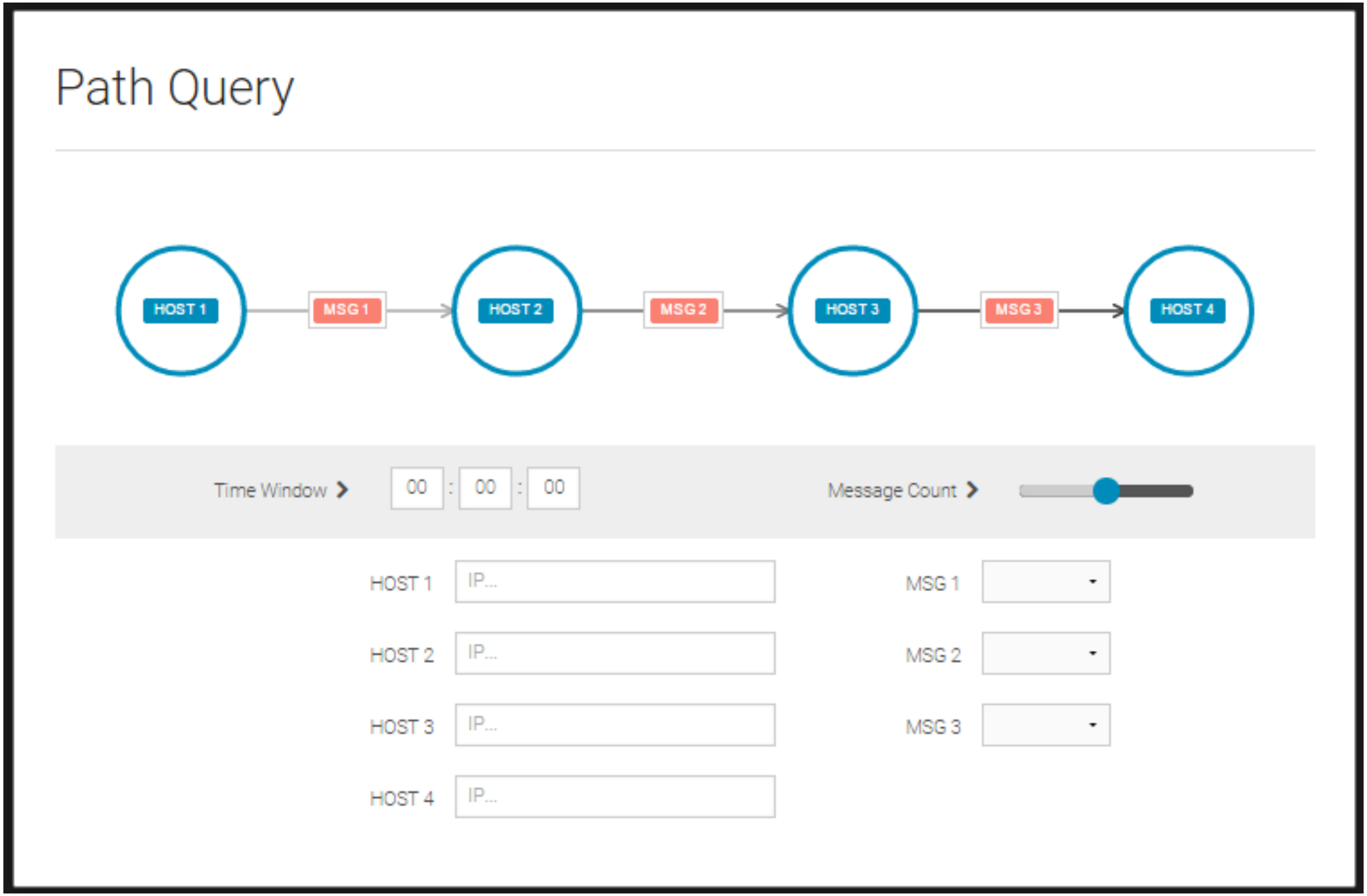


Subgraph Pattern Matching

- Query and database are both represented as graphs
- Report embedding of query pattern instances in a dynamic graph
- **Pattern Queries in Action**
  - "Tell me when a chain of 3 logins are detected with increasing privileges?"



StreamWorks Demonstration – Path Query



Path Query

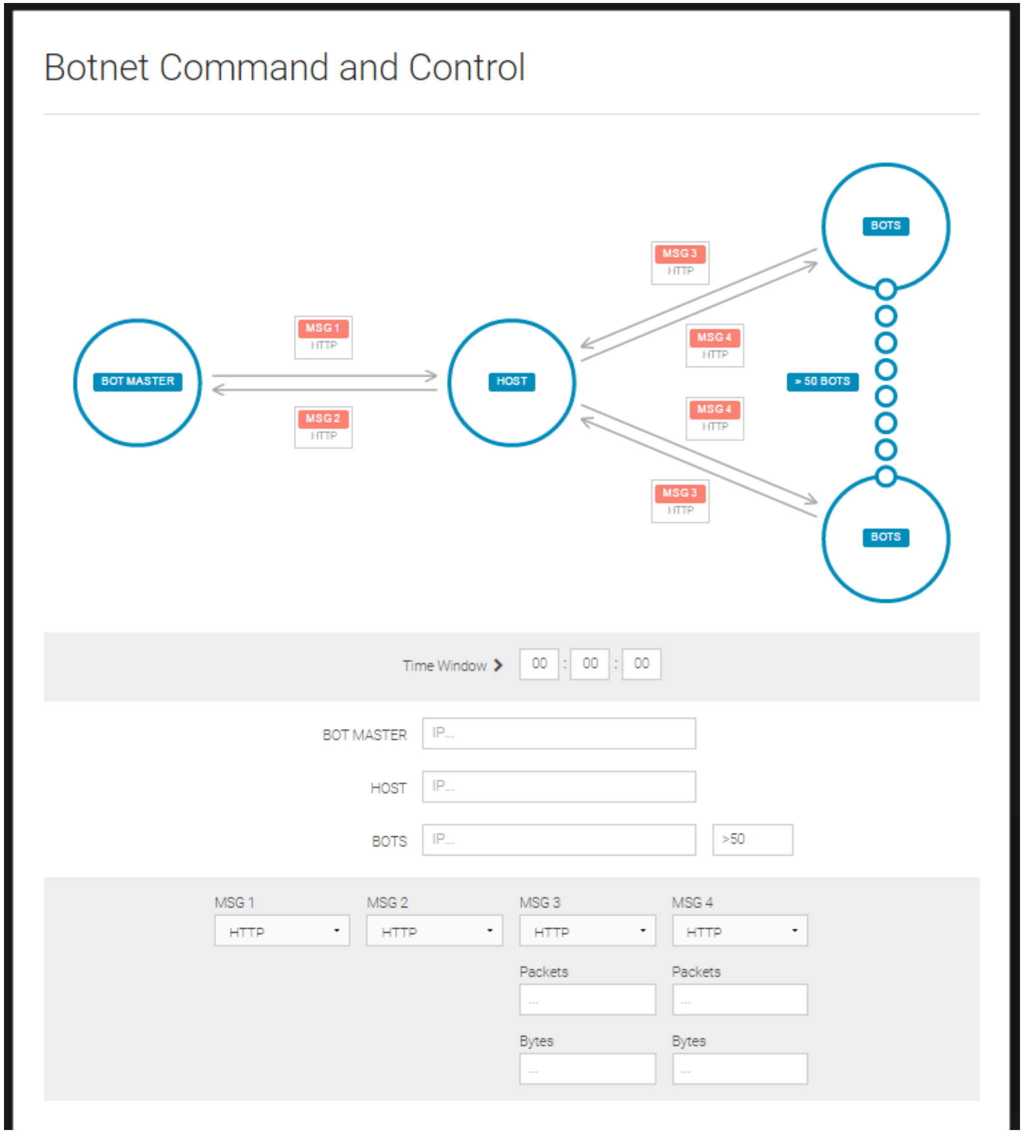
Time Window › 00 : 00 : 00

Message Count ›

HOST 1	IP...	MSG 1	
HOST 2	IP...	MSG 2	
HOST 3	IP...	MSG 3	
HOST 4	IP...		

Joslyn, C., Choudhury, S., Haglin, D., Howe, B., Nickless, B., & Olsen, B. (2013, June). Massive scale cyber traffic analysis: a driver for graph database research. In *First International Workshop on Graph Data Management Experiences and Systems* (p. 3). ACM.

StreamWorks Demonstration – Botnet Control and Command



Botnet Command and Control

Time Window ❯ 00 : 00 : 00

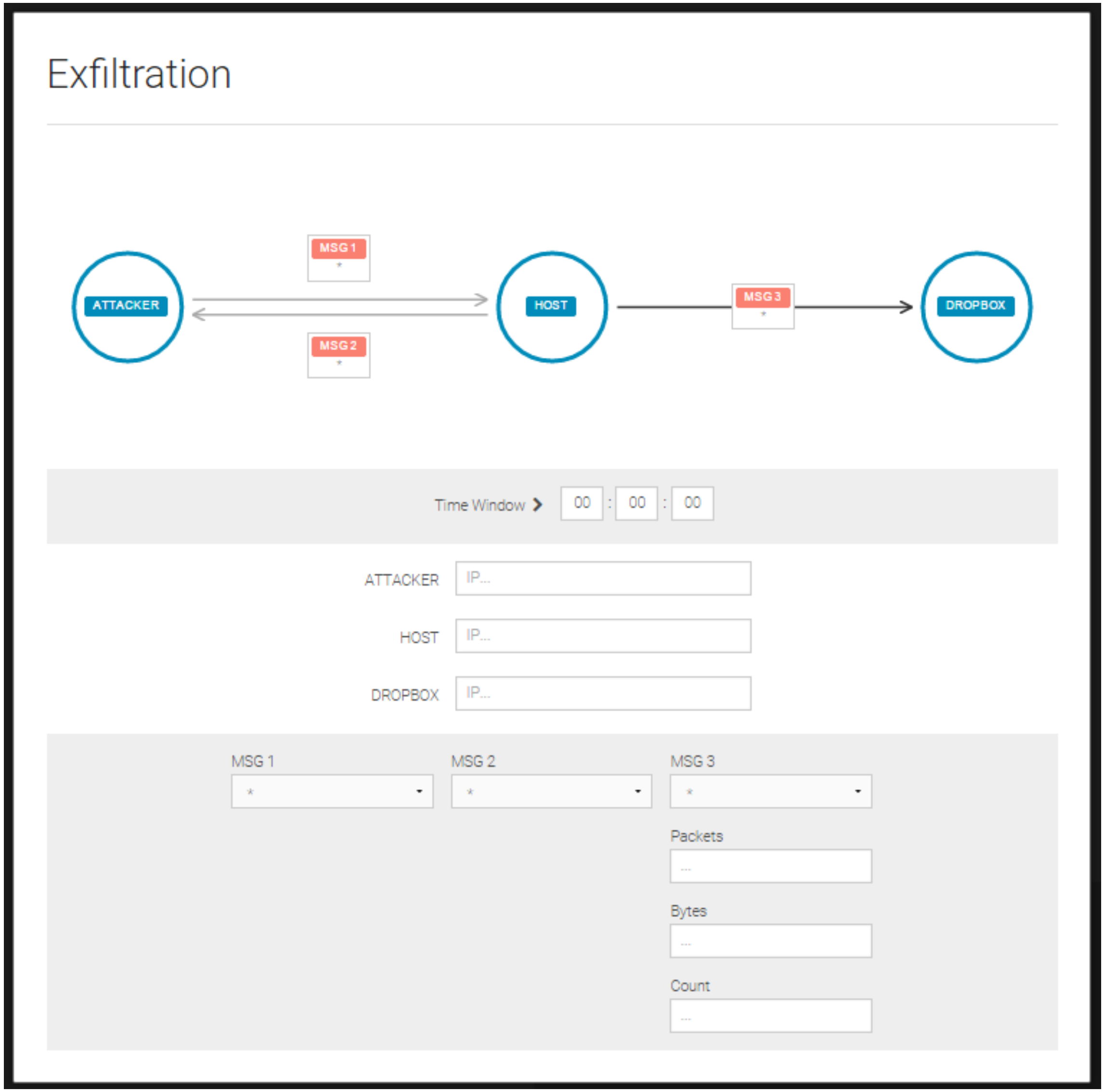
BOT MASTER IP...

HOST IP...

BOTS IP... >50

MSG 1	MSG 2	MSG 3	MSG 4
HTTP ▾	HTTP ▾	HTTP ▾	HTTP ▾
		Packets ...	Packets ...
		Bytes ...	Bytes ...

StreamWorks Demonstration – Exfiltration



Exfiltration

Time Window > 00 : 00 : 00

ATTACKER IP...

HOST IP...

DROPBOX IP...

MSG 1 *

MSG 2 *

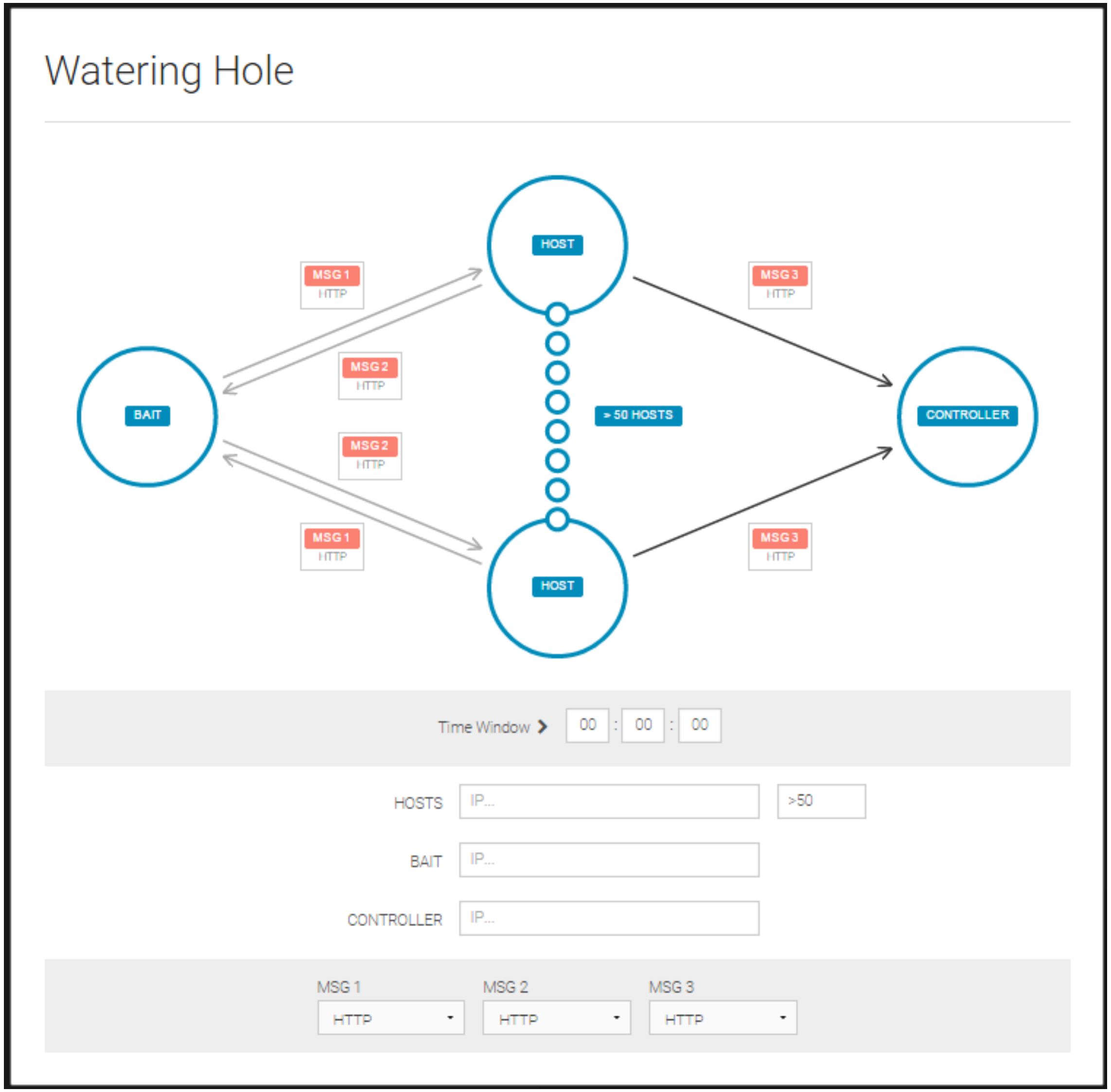
MSG 3 *

Packets ---

Bytes ---

Count ---

StreamWorks Demonstration – Watering Hole



Watering Hole

Time Window › 00 : 00 : 00

HOSTS IP... >50

BAIT IP...

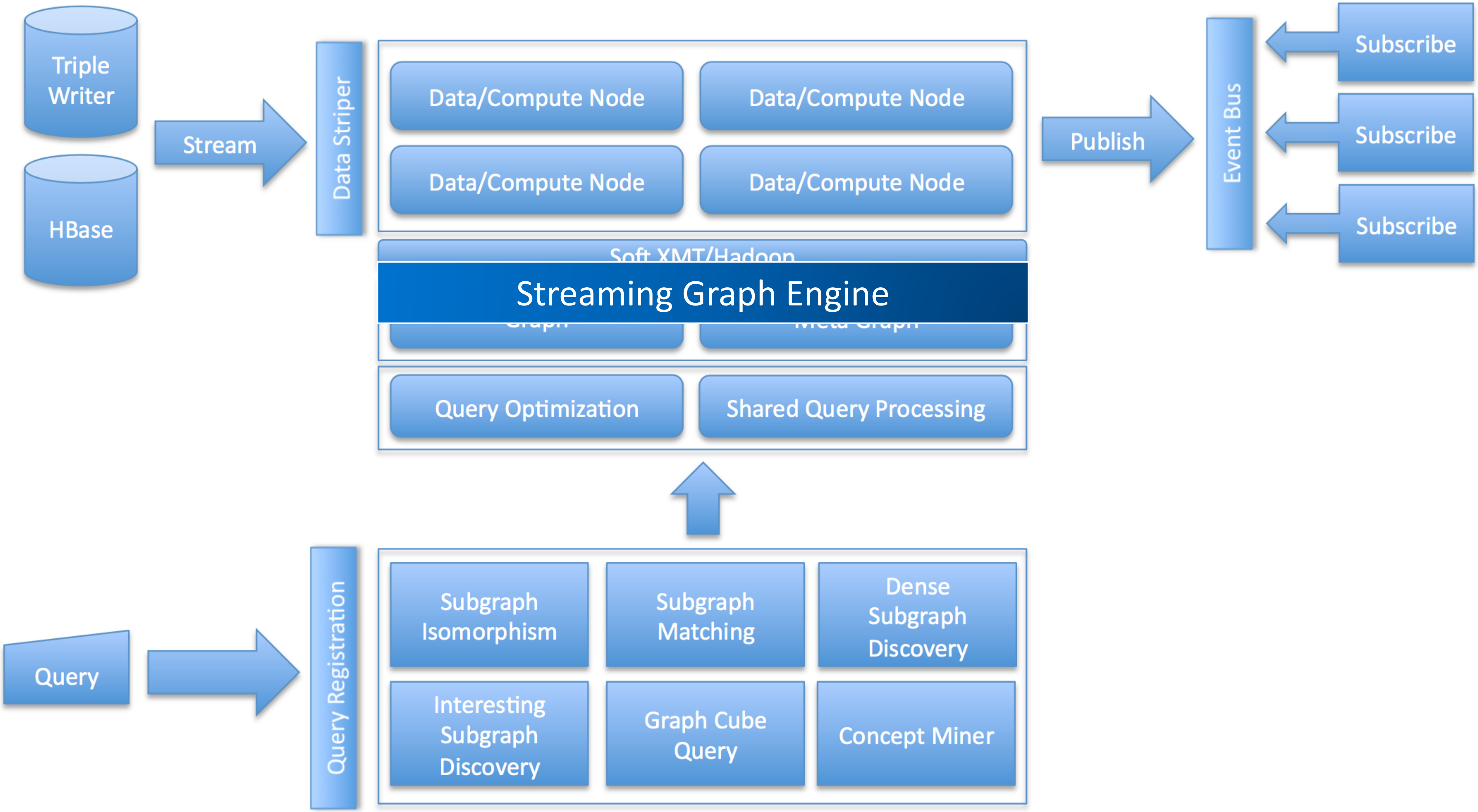
CONTROLLER IP...

MSG 1 HTTP

MSG 2 HTTP

MSG 3 HTTP

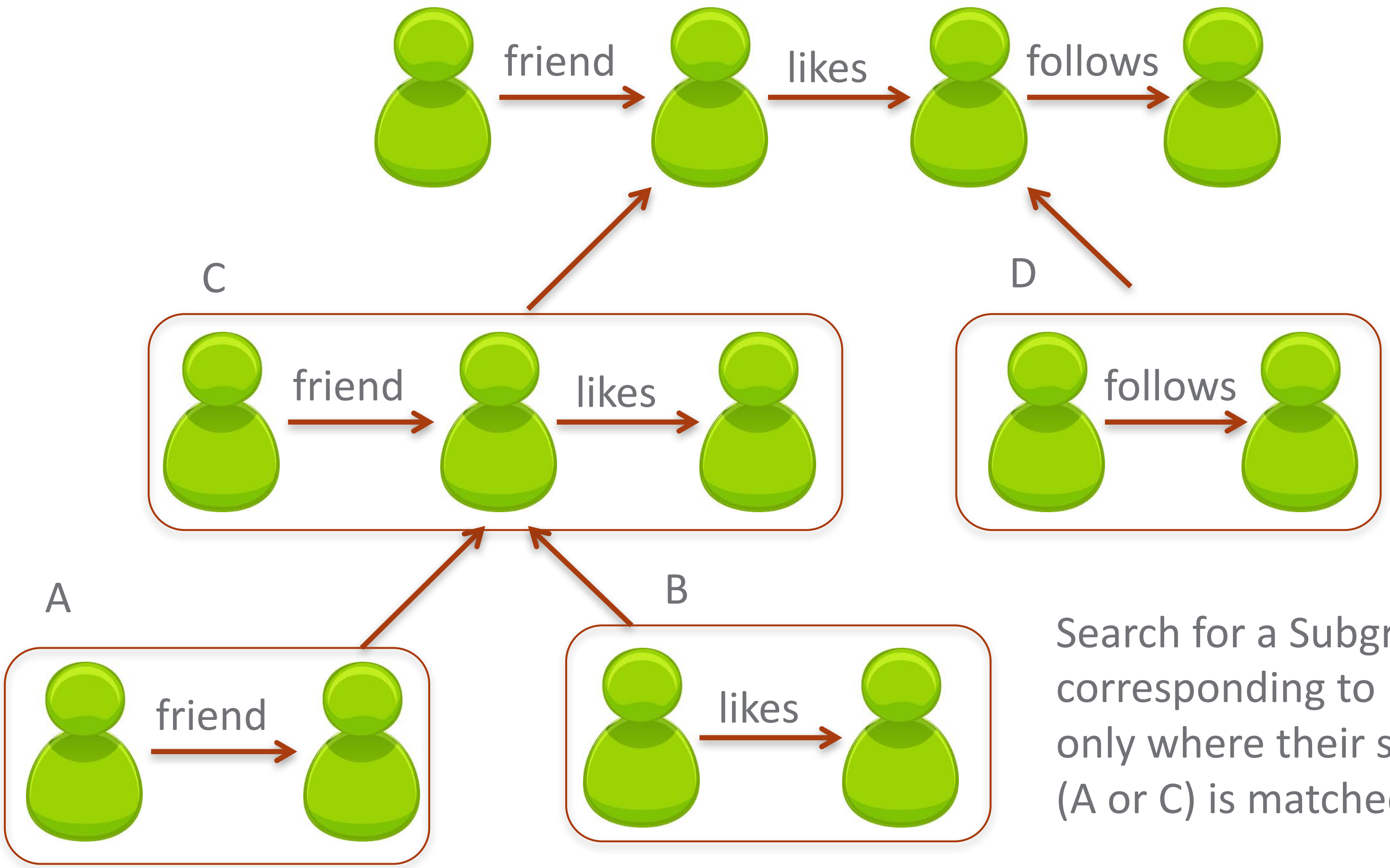
The StreamWorks Architecture



Choudhury, S., Holder, L., Chin, G., Ray, A., Beus, S., & Feo, J. (2013, June). Streamworks: a system for dynamic graph search. SIGMOD.

Dynamic Graph Query Optimization

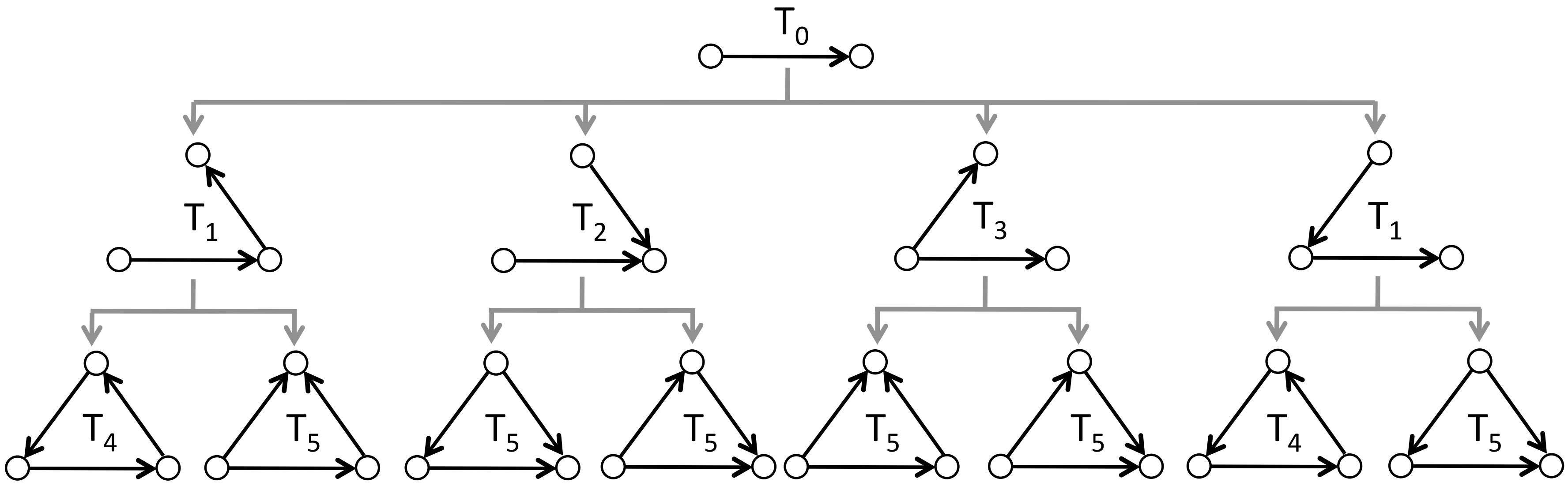
Assume selectivity order
sel(A) > sel(B) > sel(C) > sel (D)



Search for a Subgraph corresponding to B or D only where their sibling (A or C) is matched.

Selectivity Estimation

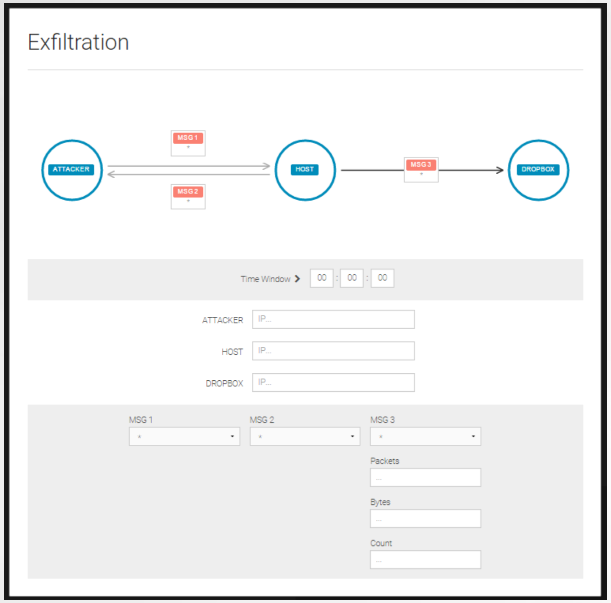
- Selectivity. We compute the selectivity of all *primitives* by counting their frequencies
- Frequency counting is expensive beyond 2-edge subgraphs

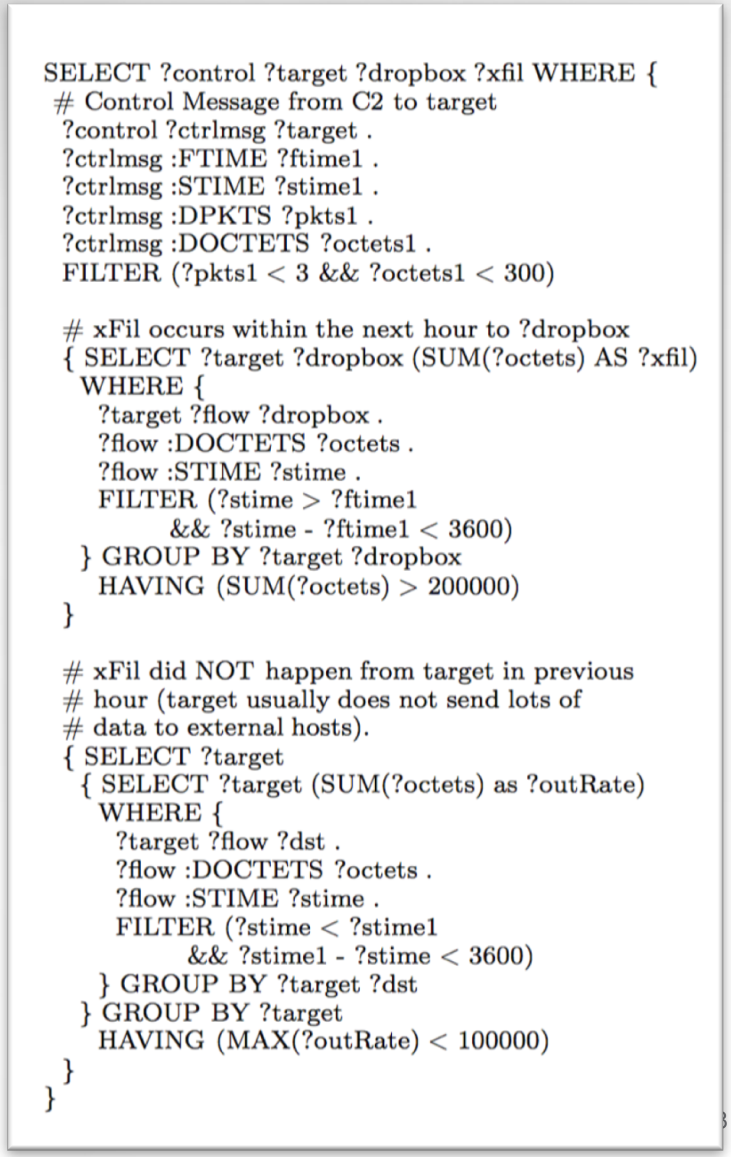


Choudhury, S., Holder, L., Chin, G., Agarwal, K., & Feo, J. (2015). A selectivity based approach to continuous pattern detection in streaming graphs. *EDBT*.

Demo: StreamWorks

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





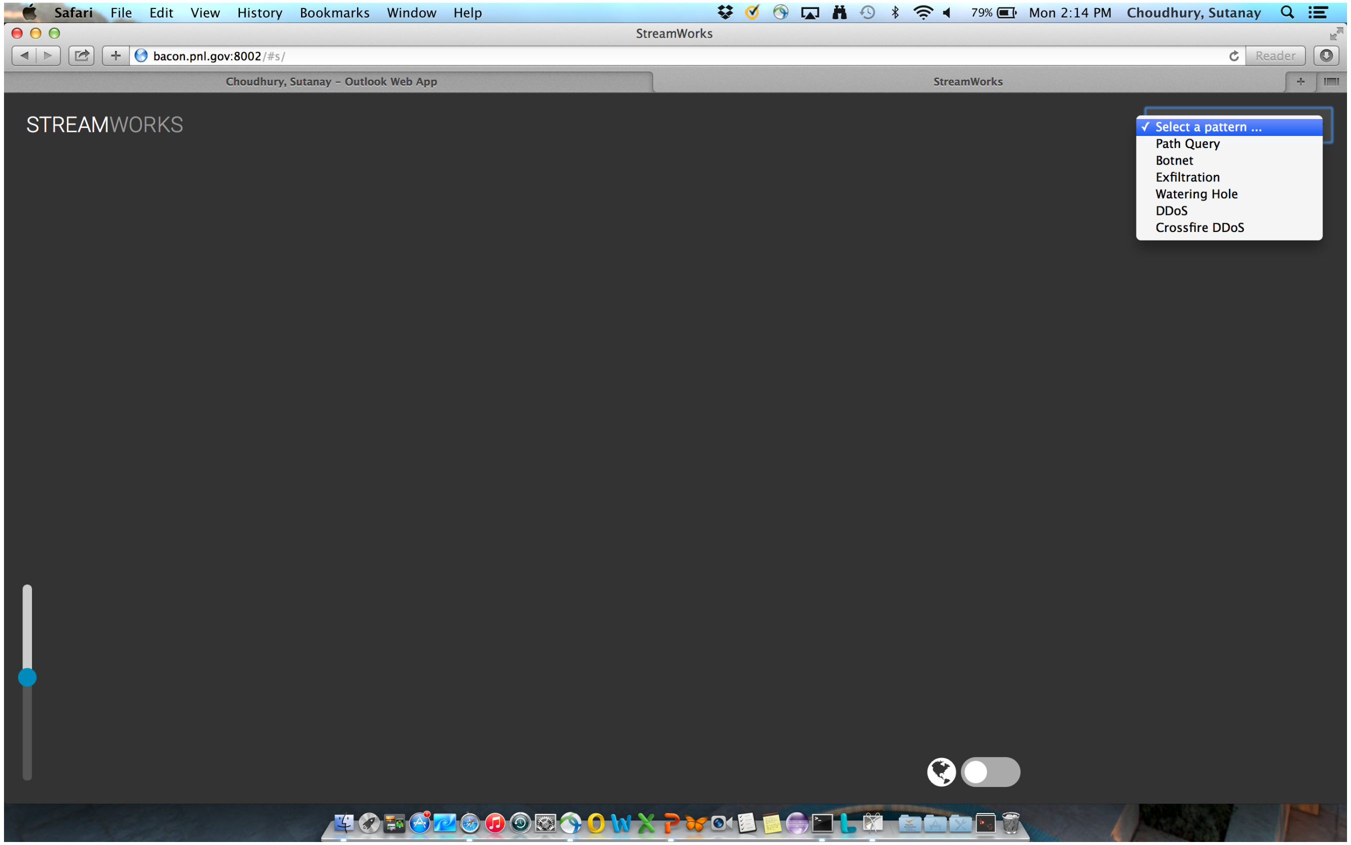
```
SELECT ?control ?target ?dropbox ?xfil WHERE {
 # Control Message from C2 to target
  ?control ?ctrlmsg ?target .
  ?ctrlmsg :FTIME ?ftime1 .
  ?ctrlmsg :STIME ?stime1 .
  ?ctrlmsg :DPKTS ?pkts1 .
  ?ctrlmsg :DOCTETS ?octets1 .
  FILTER (?pkts1 < 3 && ?octets1 < 300)

  # 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)
  }
}
```

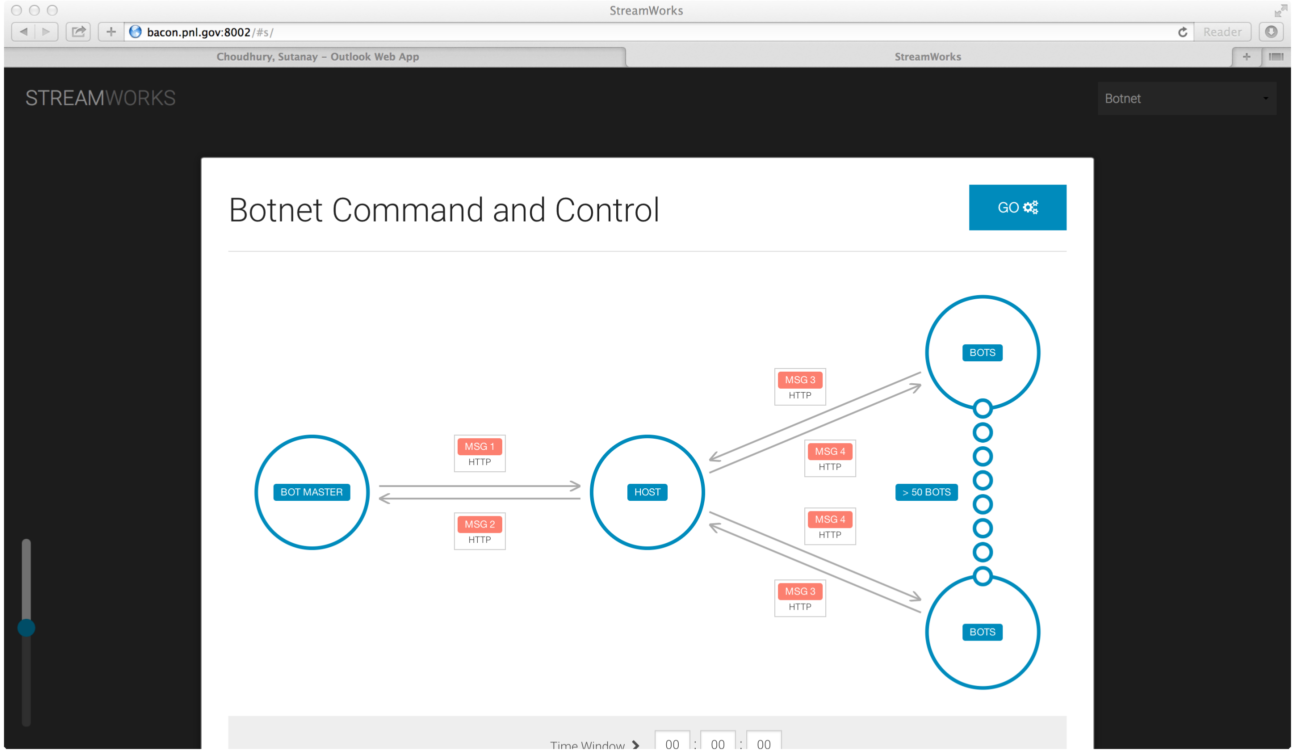
Use in 3 simple steps – No need to learn SQL/SPARQL

- We provide templates to users to specify questions and auto-generate queries
- **Step 1:** Choose a query



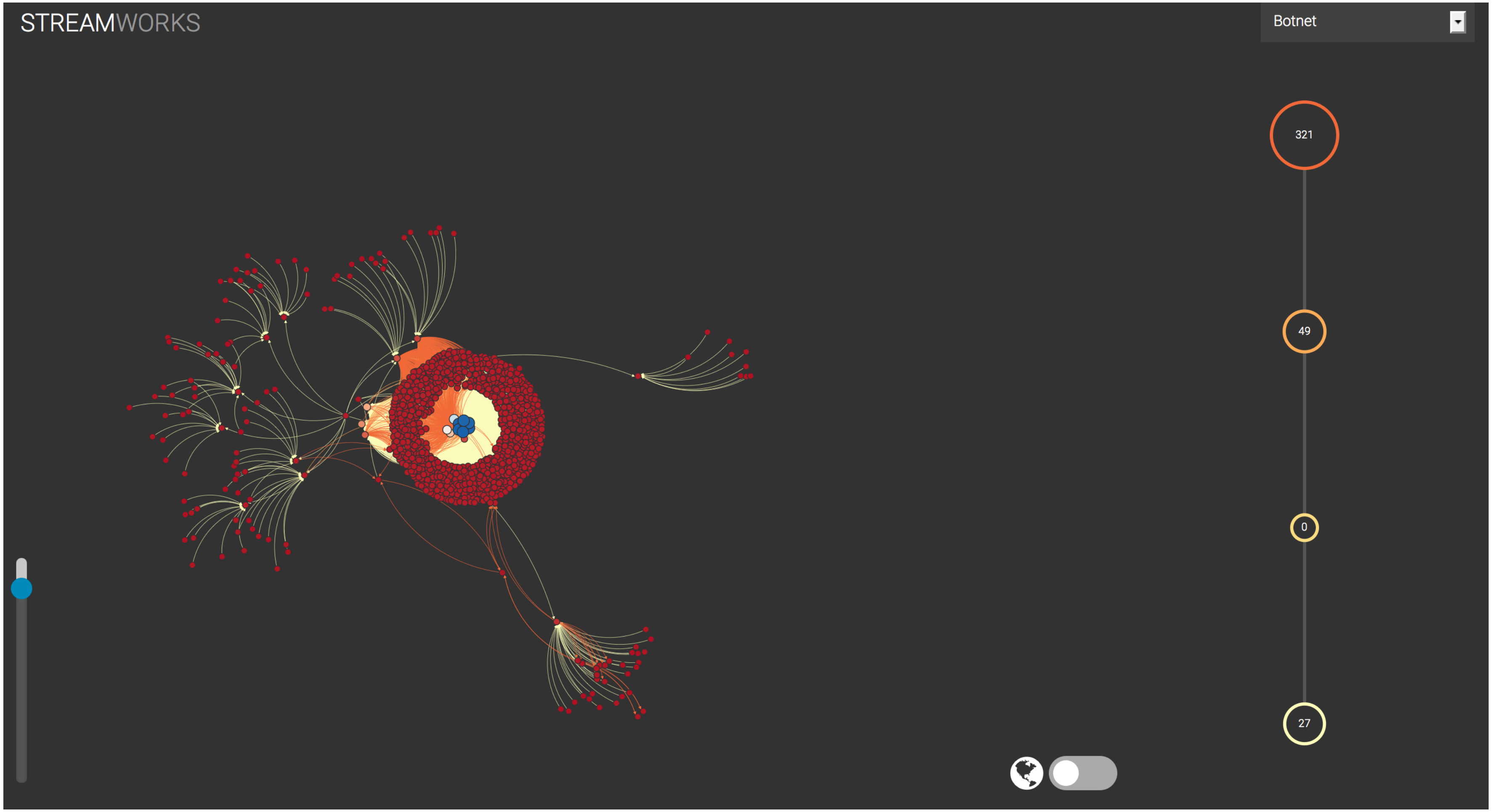
Build your query visually

- **Step 2:** Specify constraints, choose properties to specialize the queries and hit Go!
- This notifies the Apache Spark cluster to register the query and start pattern matching on the stream



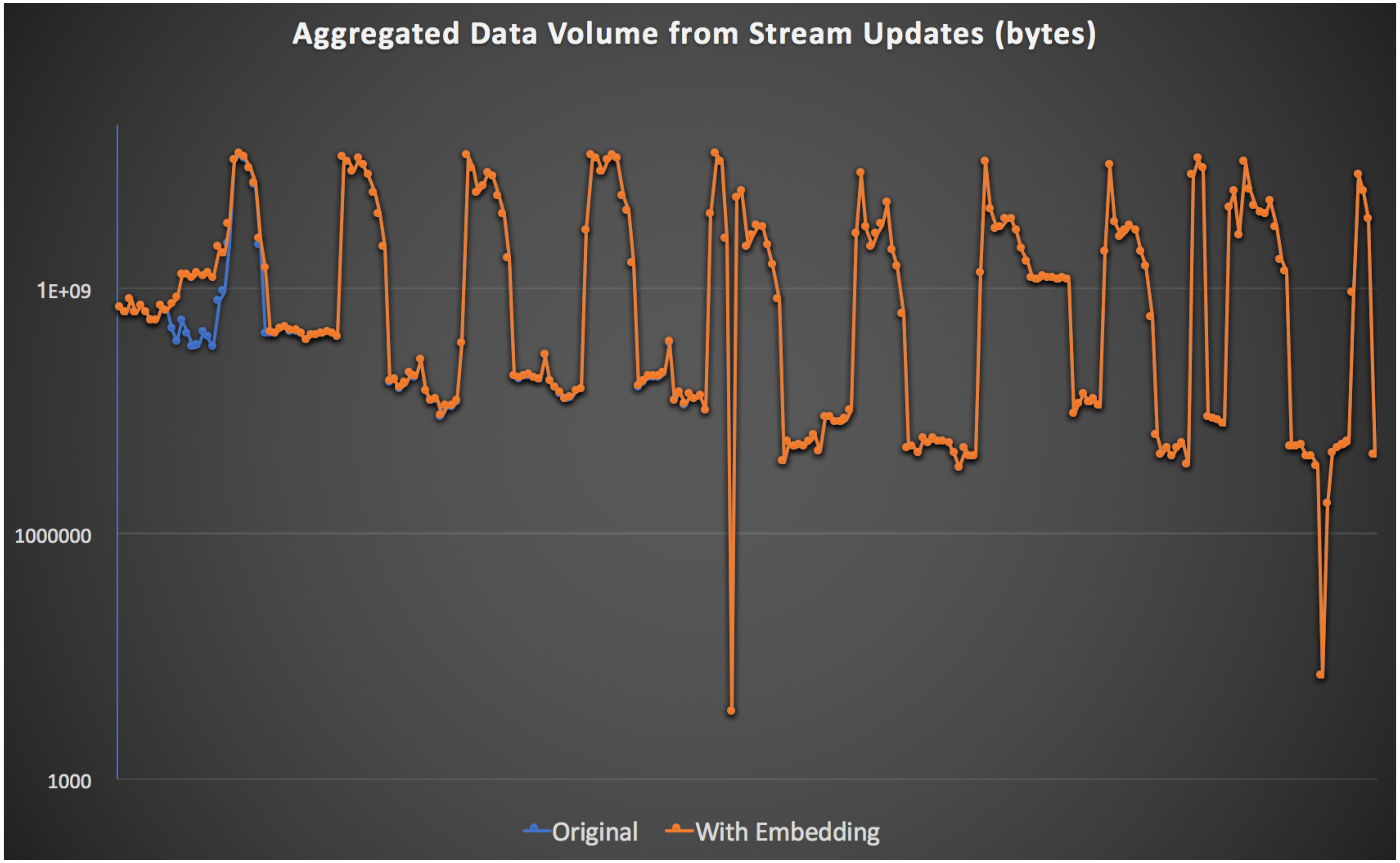
Visualize the results

- **Step 3**: When matches are found in the stream, results are send back to the web server for visualization
- Showing a botnet match. Color gradient used to indicate confidence.

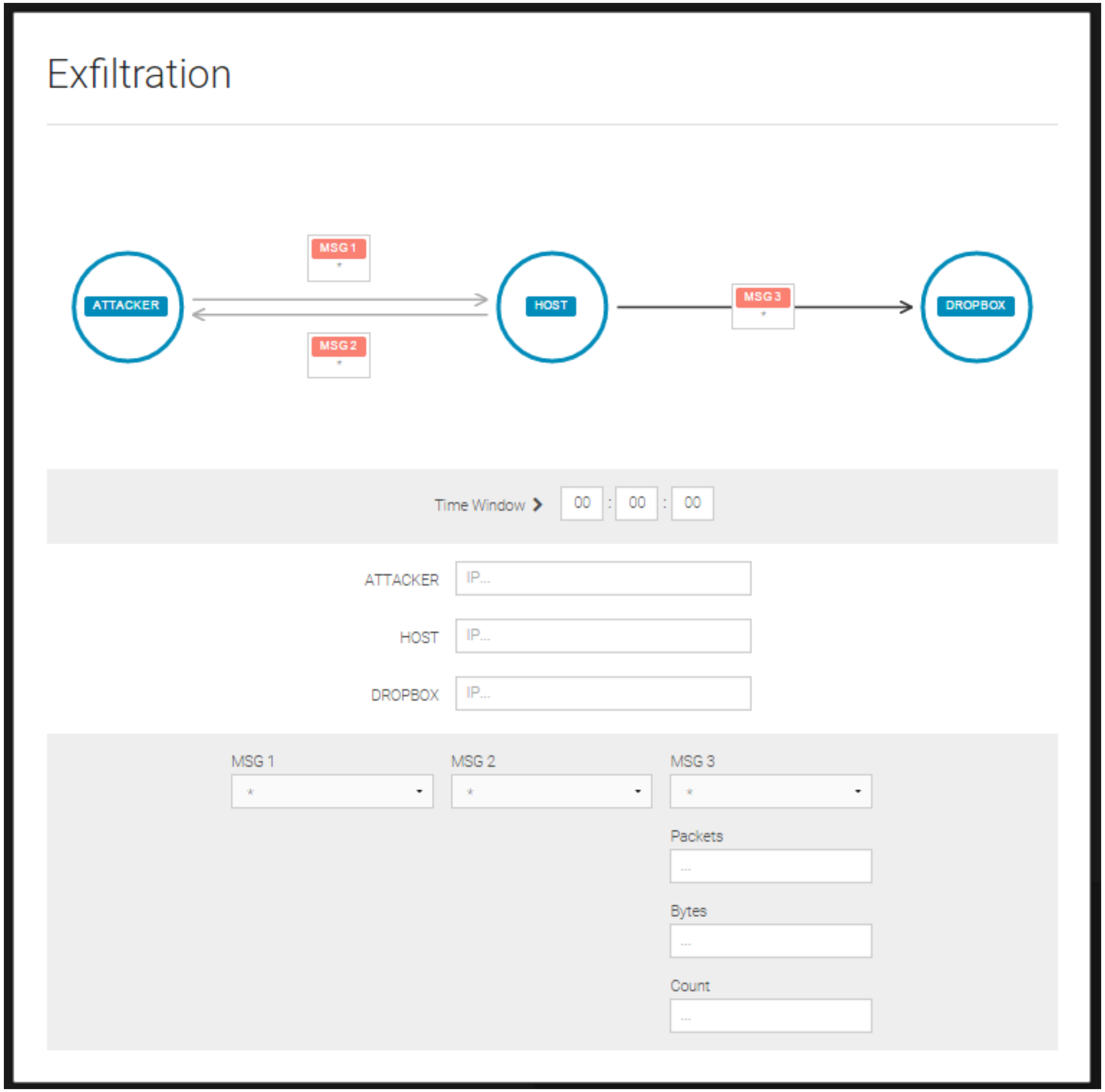


Finding the Needle in a Haystack

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



Exfiltration



Exfiltration

Time Window > 00 : 00 : 00

ATTACKER	IP...
HOST	IP...
DROPBOX	IP...

MSG 1: *

MSG 2: *

MSG 3: *

Packets: ---

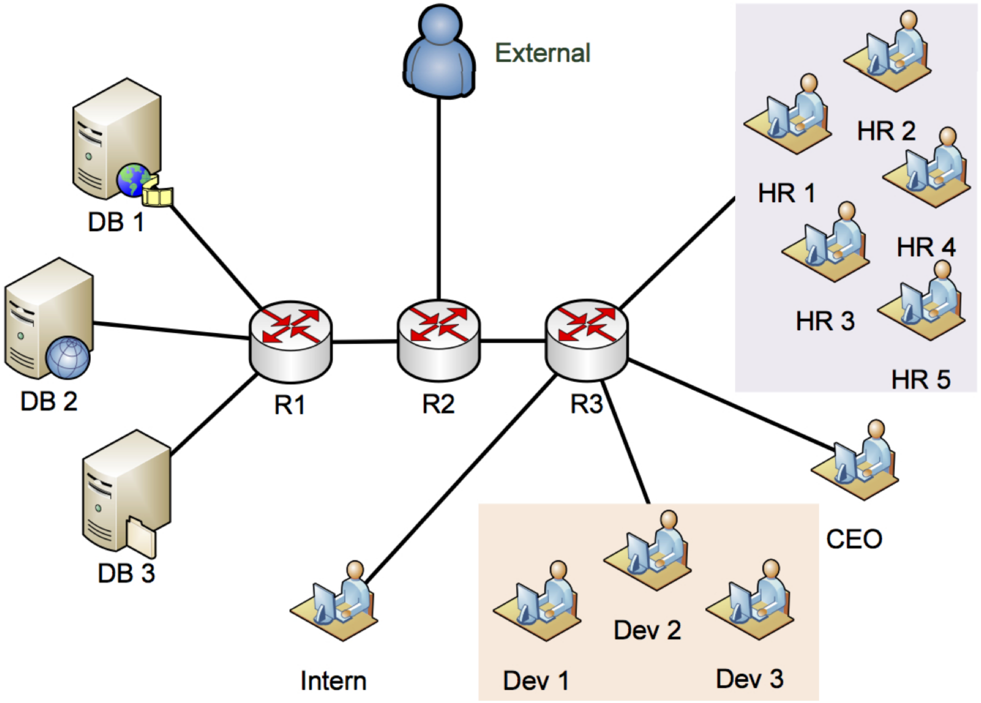
Bytes: ---

Count: ---

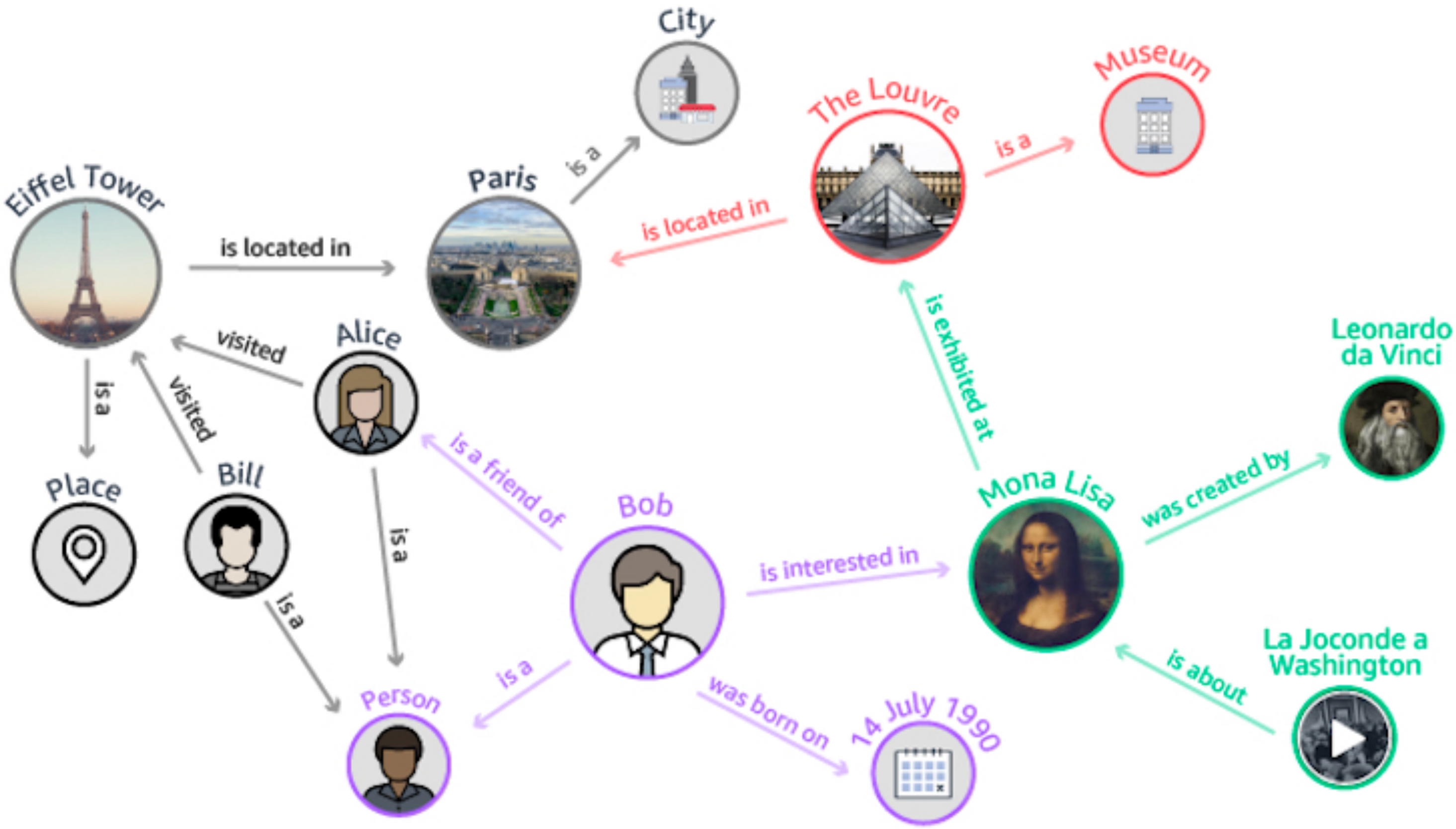
- Demo

Explaining Top Matches

- Finding a pattern is the first step. The context is more important.
  - Given an alert on host Dev1, provide explanations such as “**Dev1 frequent connects to service DB2. CE0’s machine frequently connects to DB2**”

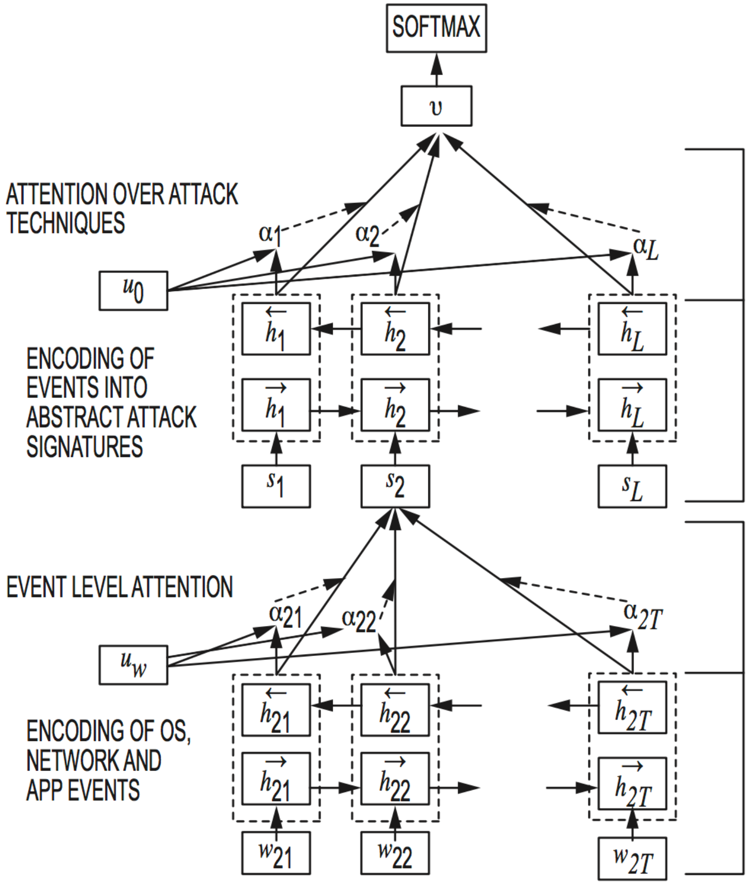


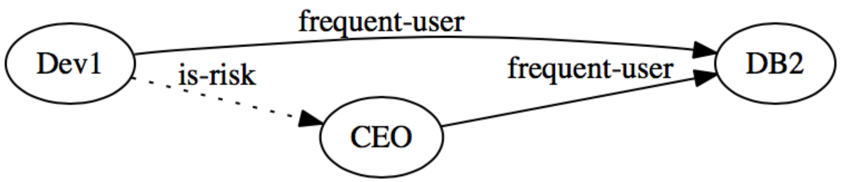
Introducing the Knowledge Graph



Building Knowledge Graph for Cyber

- A Knowledge Graph that learns and stores behavioral summary about entities of interest
  - **Role mining**: grouping systems based on their functional roles
  - **Rules**: Relational constraint using properties of entities
  - **Event prediction models**: recurrent neural network based approaches to score an event



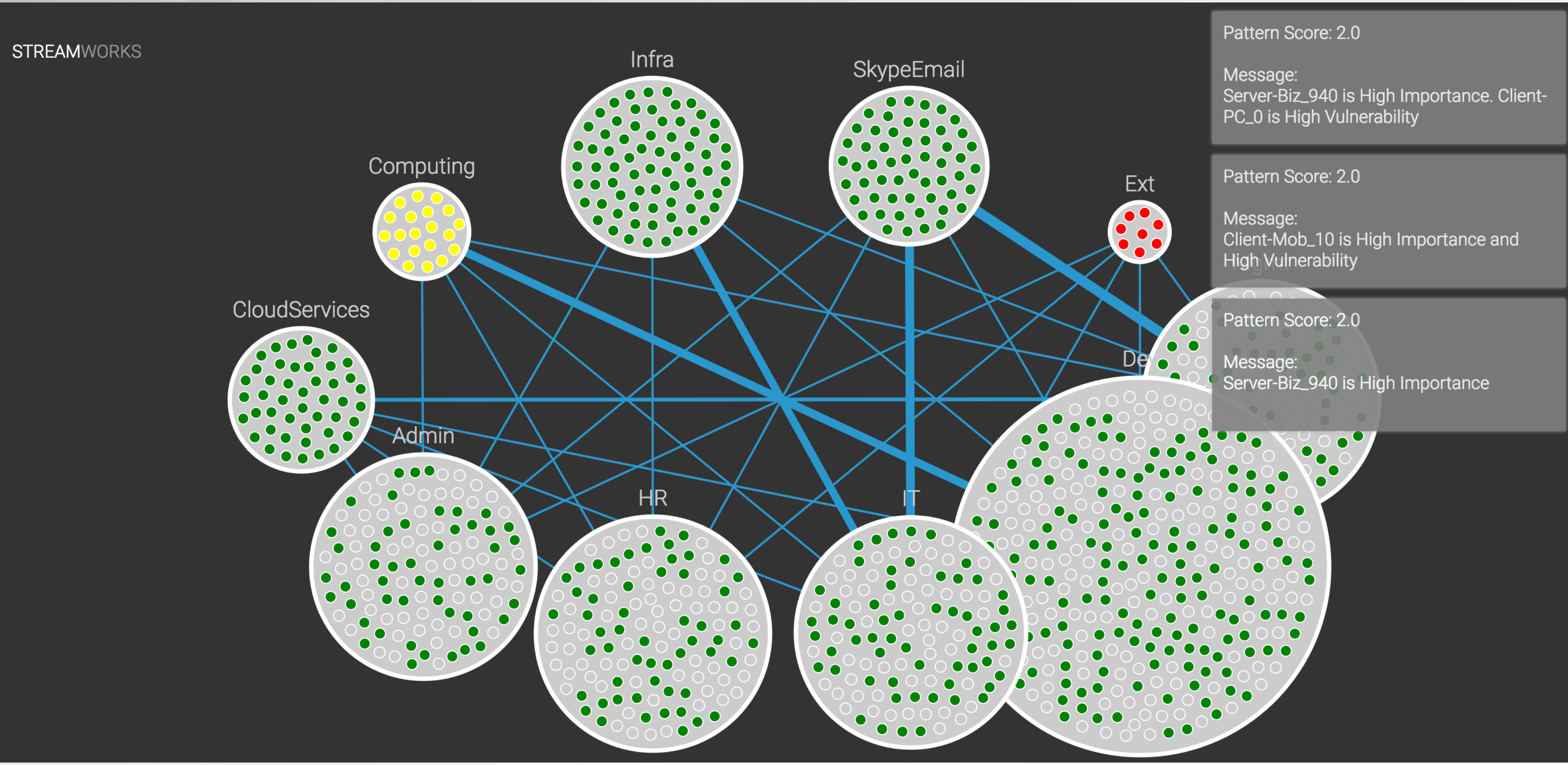


Choudhury, S., Agarwal, K., Purohit, S., Zhang, B., Pirrung, M., Smith, W., & Thomas, M. (2017, April). Nous: Construction and querying of dynamic knowledge graphs. In *Data Engineering (ICDE), 2017 IEEE 33rd International Conference on* (pp. 1563-1565). IEEE.

Demo: Explaining Top Matches

Pacific Northwest
NATIONAL LABORATORY

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



Pattern Score: 2.0

Message:
Server-Biz_940 is High Importance. Client-PC_0 is High Vulnerability

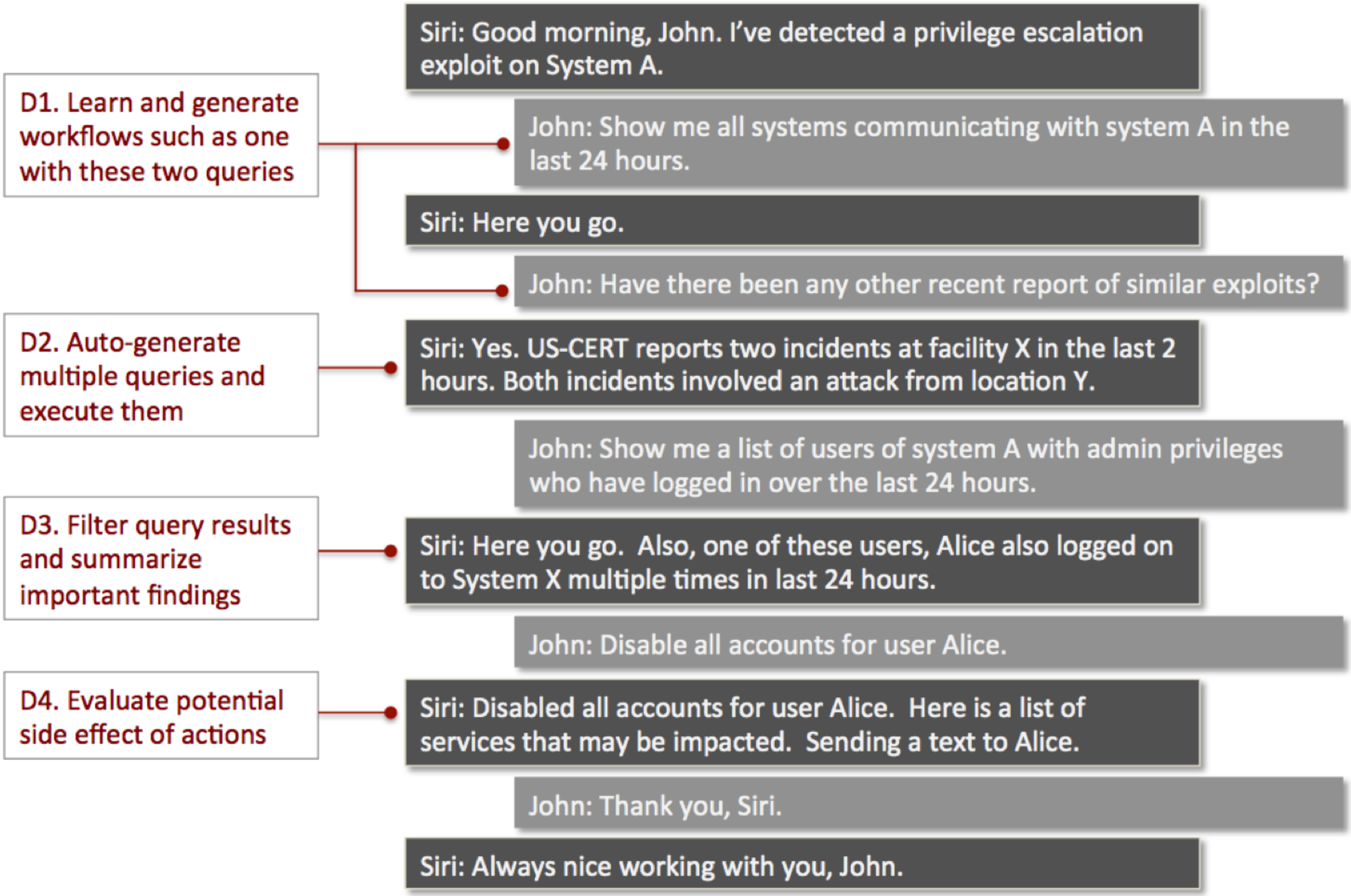
Pattern Score: 2.0

Message:
Client-Mob_10 is High Importance and High Vulnerability

Pattern Score: 2.0

Message:
Server-Biz_940 is High Importance

The Future: Towards a Passive AI for Cyber



Siri: Good morning, John. I've detected a privilege escalation exploit on System A.

D1. Learn and generate workflows such as one with these two queries

John: Show me all systems communicating with system A in the last 24 hours.

Siri: Here you go.

John: Have there been any other recent report of similar exploits?

D2. Auto-generate multiple queries and execute them

Siri: Yes. US-CERT reports two incidents at facility X in the last 2 hours. Both incidents involved an attack from location Y.

John: Show me a list of users of system A with admin privileges who have logged in over the last 24 hours.

D3. Filter query results and summarize important findings

Siri: Here you go. Also, one of these users, Alice also logged on to System X multiple times in last 24 hours.

John: Disable all accounts for user Alice.

D4. Evaluate potential side effect of actions

Siri: Disabled all accounts for user Alice. Here is a list of services that may be impacted. Sending a text to Alice.

John: Thank you, Siri.

Siri: Always nice working with you, John.

How do we evaluate AI systems operating in a streaming environment?

- Develop a benchmark of tasks and dataset
- **Proposing 5 “Cyber IQ” levels representing progressive levels of complexity**

- IQ level 1: Find Anomalous Events of Interest
  - Can perform basic tasks such as detecting unusual events (VPN logins from multiple locations for same user)
- IQ level 2: Learn from human experts and build its Knowledge Graph
  - Observe data and generate questions for the human expert
    - Example: Which systems/services have more strategic importance?
    - What does action X accomplish? Which systems does this affect?
    - Don’t ask, but observe. Try to observe the impact/success of action X in steering the system back to normalcy.

Passive AI (level 3)

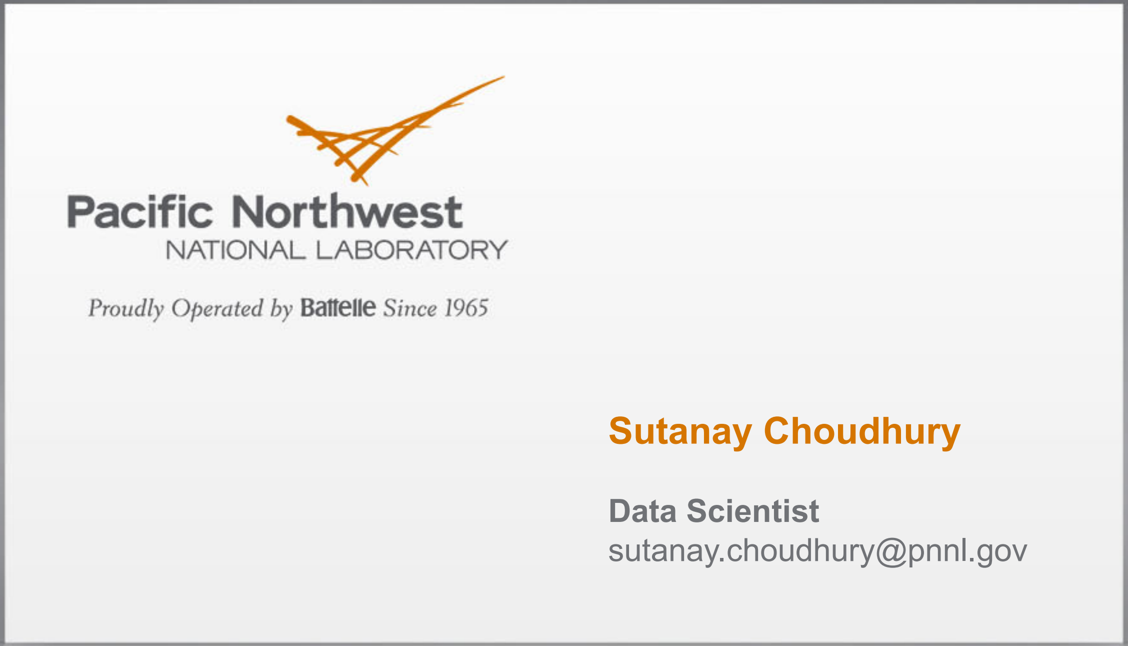
- IQ level 3: Work with human experts
  - Start sending recommendations “In past, we saw case X on days A, B and C, and you took actions P and Q. Q had higher success rate.”
  - Explanations: “You took actions P and Q in situations like X. However, this time we are also seeing strong presence of features M and N that were not present in X before.”

Autopilot (Level 4 and 5)

- Autopilot (Level 4)
  - Develop a conservative "autopilot mode”
  - Generate reports providing reasoning for every action chosen
  - It has to be better than shutting down everything in the event of an anomaly, but smartly using tools such as throttling traffic to certain VMs or isolating them
- Meta-System: A system to audit the primary
  - Understand when human and machines underperform
  - Compare votes of the human expert and that of the machine
  - Inject purposeful disruptions to test robustness against unseen problems

Summary

- A tale of two graphs
  - **A fast moving graph** – Dynamic graph updated in near real-time and maintained as sliding window in time
  - **A slowly changing graph** – The Knowledge graph reflects what we learn about behavioral patterns in the data
- **Applications**: Build complex systems with the graphs as their memory
  - **Today**: Continuously maintain the memory, search and serve top-k queries with summarization
  - **Future**: Build complex systems using graphs algorithms and machine-learning techniques as tools that evolve the memory and operate on it



Pacific Northwest
NATIONAL LABORATORY

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

Sutanay Choudhury

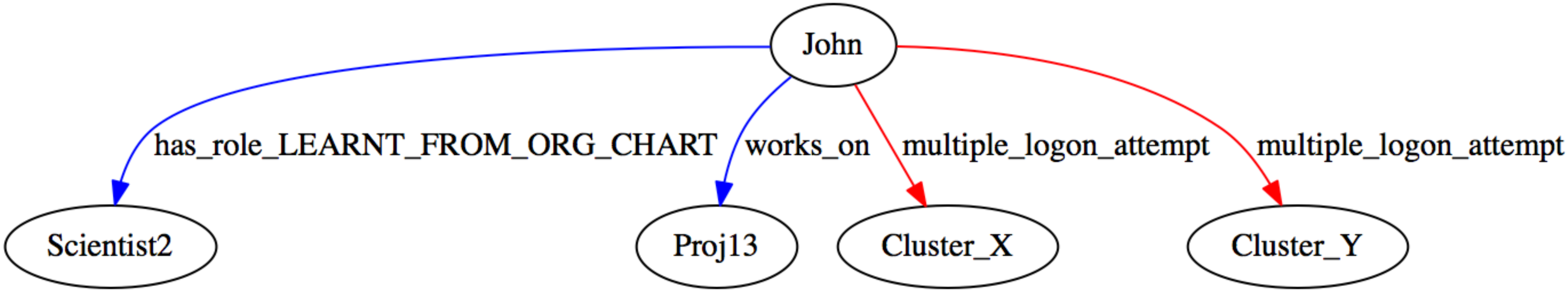
Data Scientist
sutanay.choudhury@pnnl.gov

Backup

Motivating Example

- Let’s take a very simple action (**John, an employee working on a DOE project launches a job on a cluster**) and try to categorize that as normal or unusual
- Classifying someone as an Insider Threat builds on multiple indicators. Categorization such as above is a basic building block in that process

What do we know about John?

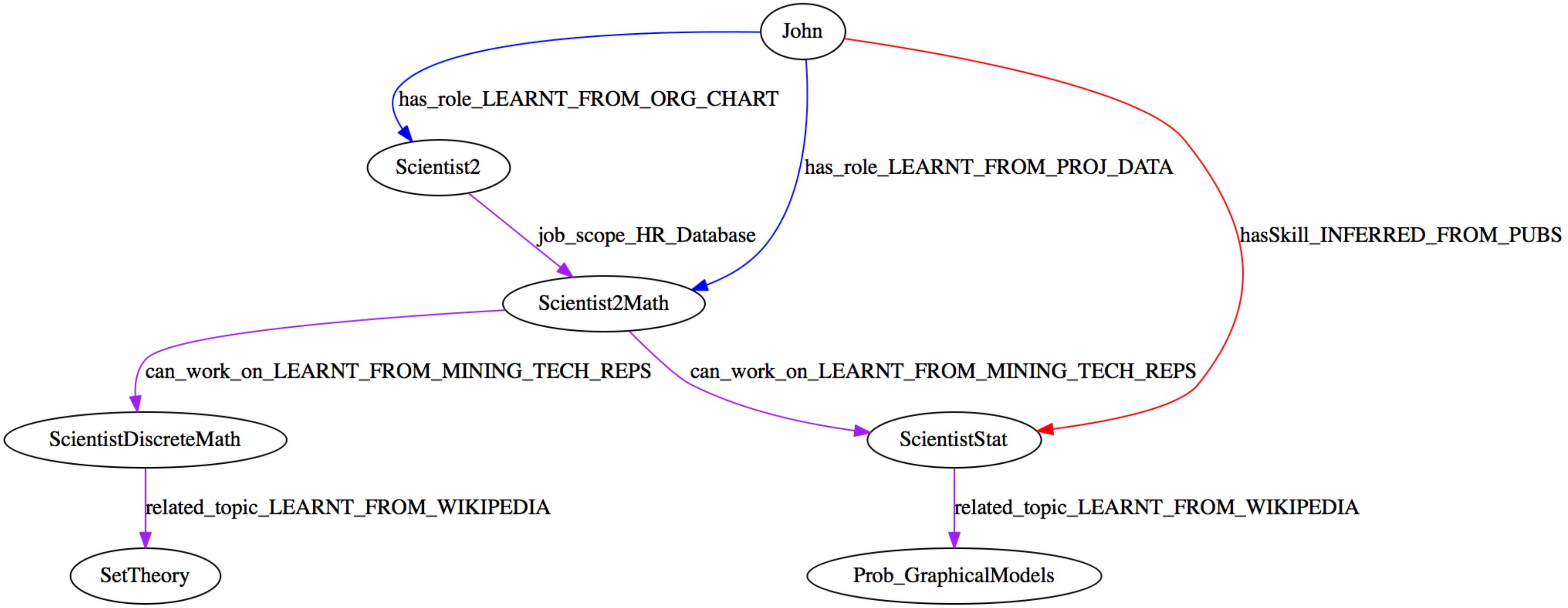


Red facts are summaries obtained from streaming data

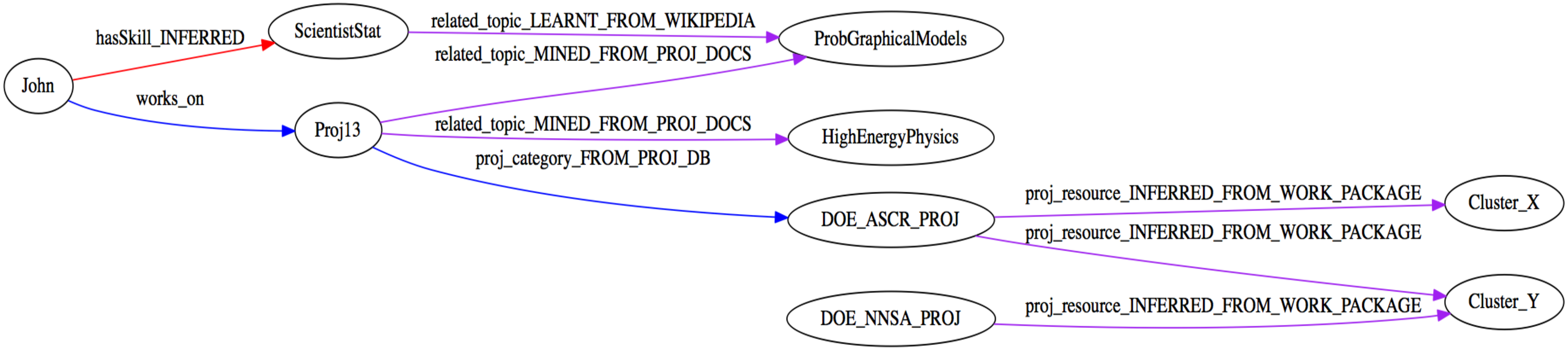
Blue facts are static information

What does John do in PNNL?

- Let's expand by adding data from HR database and project publications

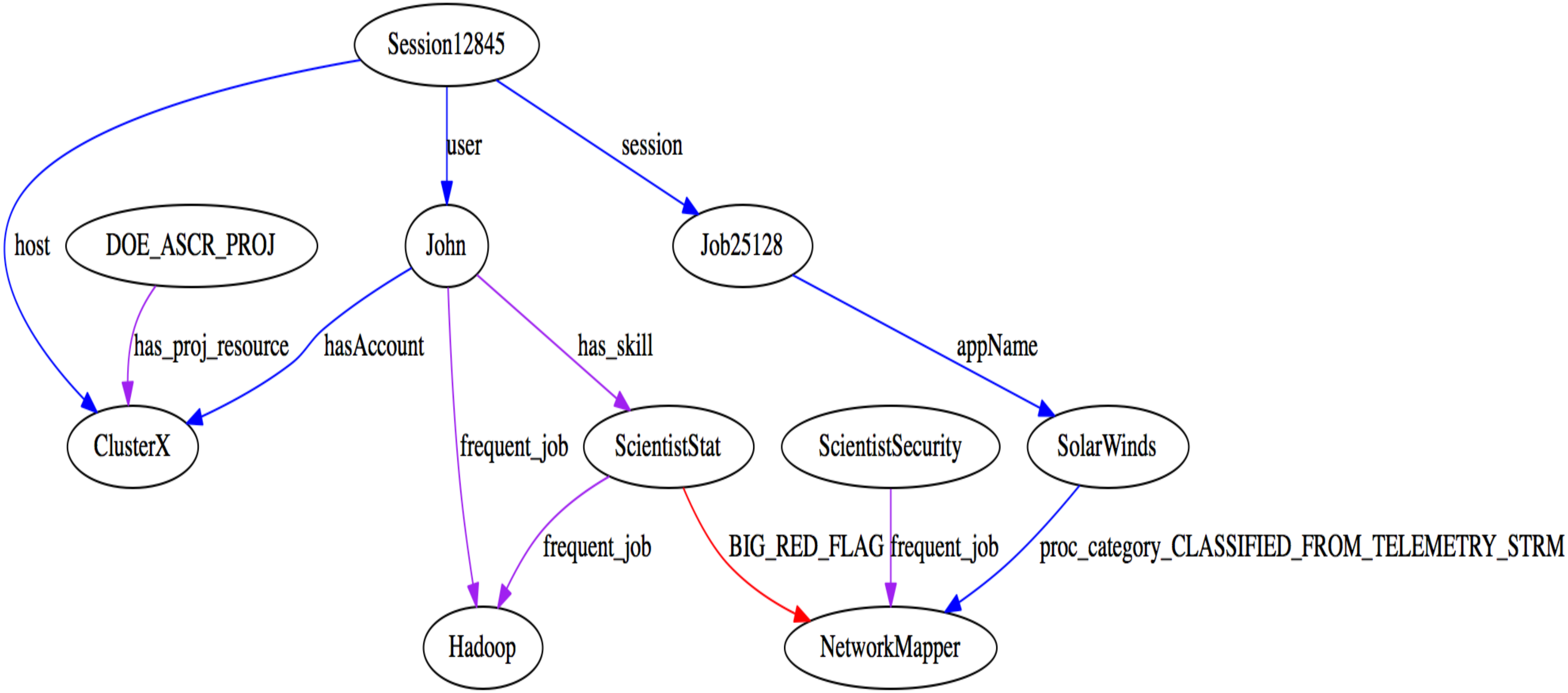

John
has_role_LEARNT_FROM_ORG_CHART
Scientist2
has_role_LEARNT_FROM_PROJ_DATA
job_scope_HR_Database
hasSkill_INFERRED_FROM_PUBS
Scientist2Math
can_work_on_LEARNT_FROM_MINING_TECH_REPS
can_work_on_LEARNT_FROM_MINING_TECH_REPS
ScientistDiscreteMath
ScientistStat
related_topic_LEARNT_FROM_WIKIPEDIA
related_topic_LEARNT_FROM_WIKIPEDIA
SetTheory
Prob_GraphicalModels

Tell me more about his Projects


John
hasSkill_INFERRED
ScientistStat
works_on
Proj13
related_topic_LEARNT_FROM_WIKIPEDIA
related_topic_MINED_FROM_PROJ_DOCS
ProbGraphicalModels
related_topic_MINED_FROM_PROJ_DOCS
HighEnergyPhysics
proj_category_FROM_PROJ_DB
DOE_ASCR_PROJ
DOE_NNSA_PROJ
proj_resource_INFERRED_FROM_WORK_PACKAGE
proj_resource_INFERRED_FROM_WORK_PACKAGE
proj_resource_INFERRED_FROM_WORK_PACKAGE
Cluster_X
Cluster_Y

What do we know about the specific network activity?

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*


Session12845
user
session
host
DOE_ASCR_PROJ
John
Job25128
has_proj_resource
hasAccount
has_skill
appName
ClusterX
frequent_job
ScientistStat
ScientistSecurity
SolarWinds
frequent_job
BIG_RED_FLAG
frequent_job
proc_category_CLASSIFIED_FROM_TELEMETRY_STRM
Hadoop
NetworkMapper