ORACLE

Graph-Powered Cyber-Security Intelligence

Analytics and Data Summit '20

Sungpack Hong, Research Director, Oracle Labs
Rhicheek Patra, Research Manager, Oracle Labs
Jinha Kim, Principal Member of Technical Staff, Oracle Labs

Introduction

- Lab's collaboration with various teams in Oracle
  - Many works still on-going
  - Apply graph technology to many problems in various domains

- Common Graph Benefits
  1. Graph data model: combines multiple heterogeneous data sources
  2. Graph algorithm and query
  3. Visualization and interactive exploration
  4. Combining graph and machine learning

Agenda

1 Cyber Security and Graph

2 Threat Hunting using Visual Graph Exploration

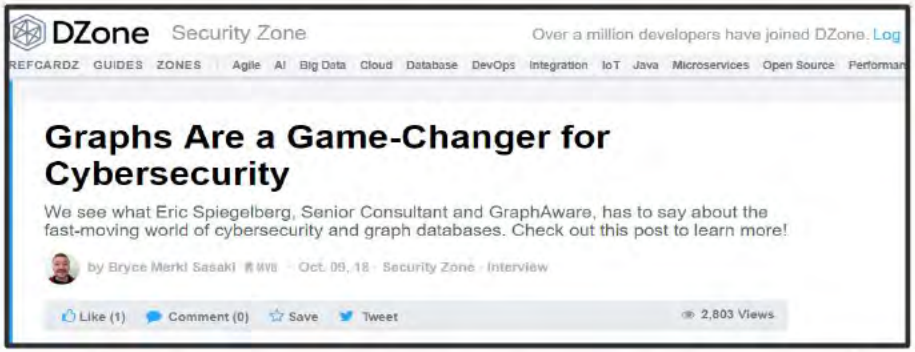
3 Threat Detection using Graph Machine Learning

(Ad-Fraud and VCN Flow)

Cyber Security and Graph

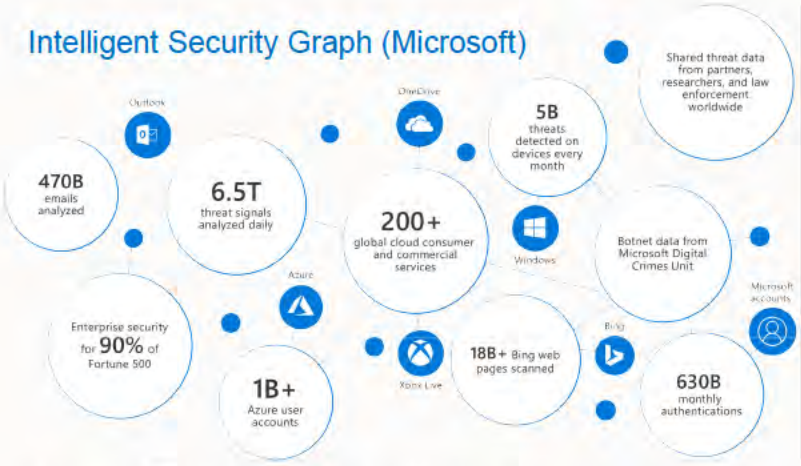
- Cyber security is a very important topic in the battle for Cloud
  - Invalid traffic detection
  - Cyber threat hunting
  - Malware detection
  - …

- Emerging approach
  - Use *graph* for enhancing cyber security



Cyber Security and Graph – How and Why?

- Use *graph* as a semi-structured *data model*
  - Combines heterogeneous data sources, i.e., security data comes from various logs, traces, events, etc.
  - Captures connections between them, e.g., same IP connects to multiple domains.



- Graph captures connections between data entities

➔ Exploit extra signals from graph for anomaly detection (e.g. via Graph ML techniques)

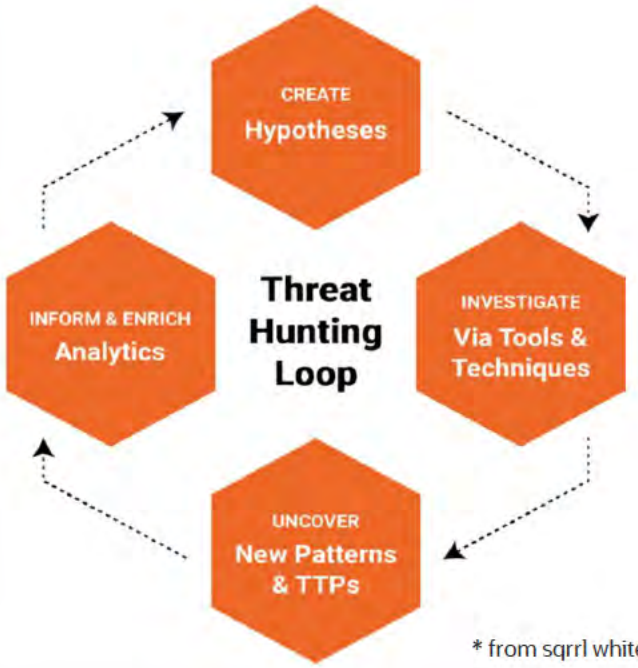
➔ Enhance cyber-threat detection

- Graph enables interactive, visual exploration of security data

➔ Environment for cyber-threat hunting

Cyber Threat Hunting

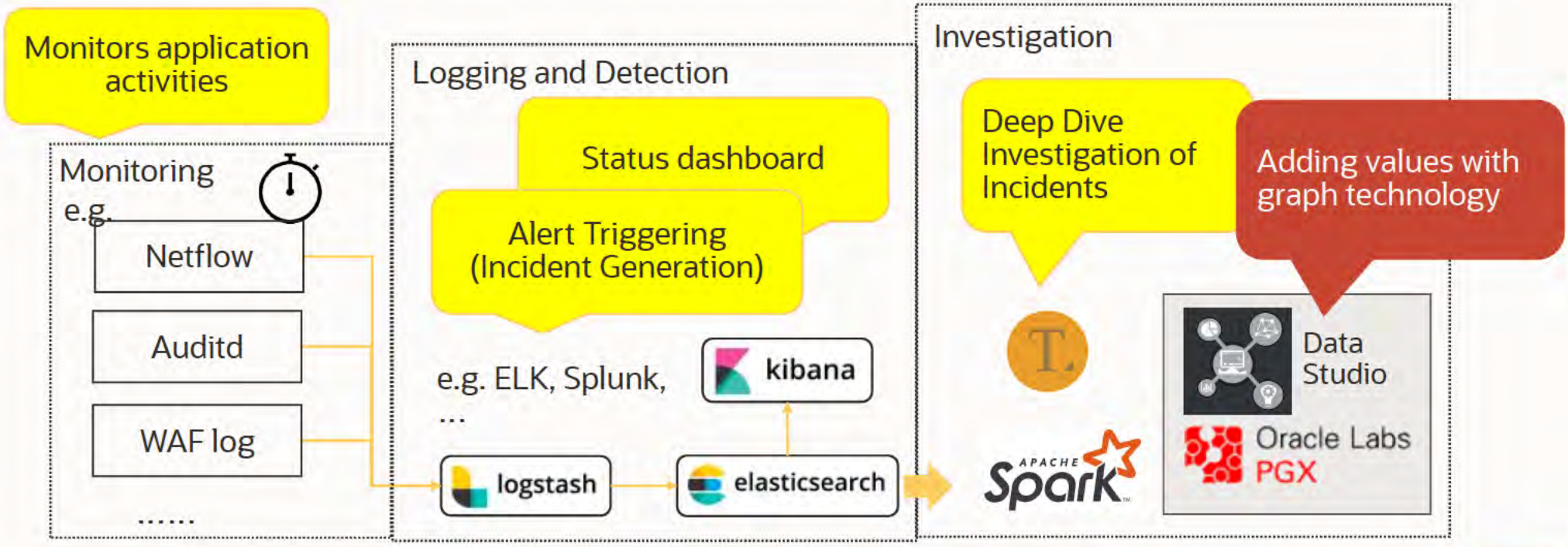
- **Cyber threat hunting** is a process of **proactively** and **iteratively** searching through data to detect and isolate advanced threats that evade existing security solutions." [1]



* from sqrrl whitepaper

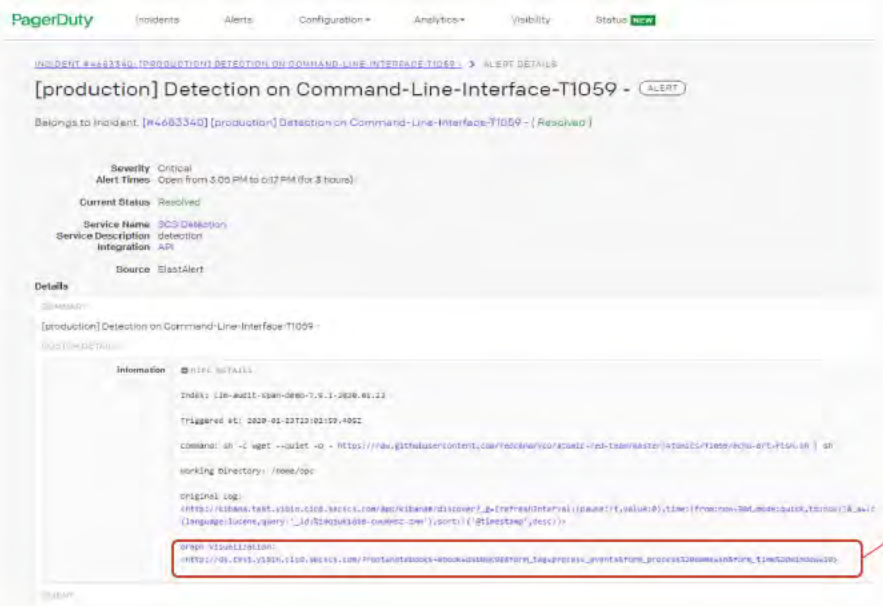
[1] https://www.techrepublic.com/article/cyber-threat-hunting-why-this-active-strategy-gives-analysts-an-edge/

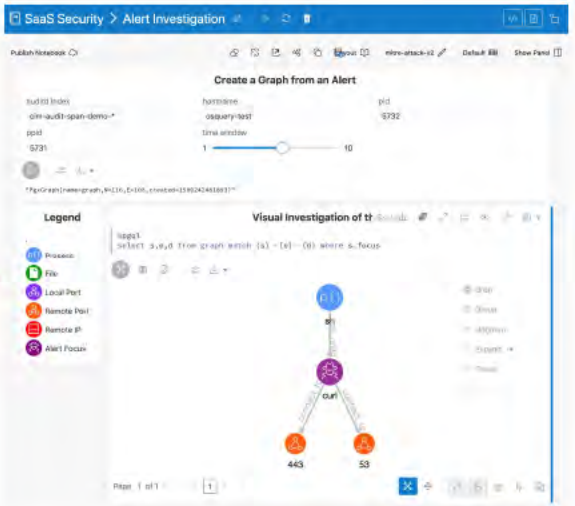
Typical Cyber Intelligence System



Collaboration with SaaS Security Team

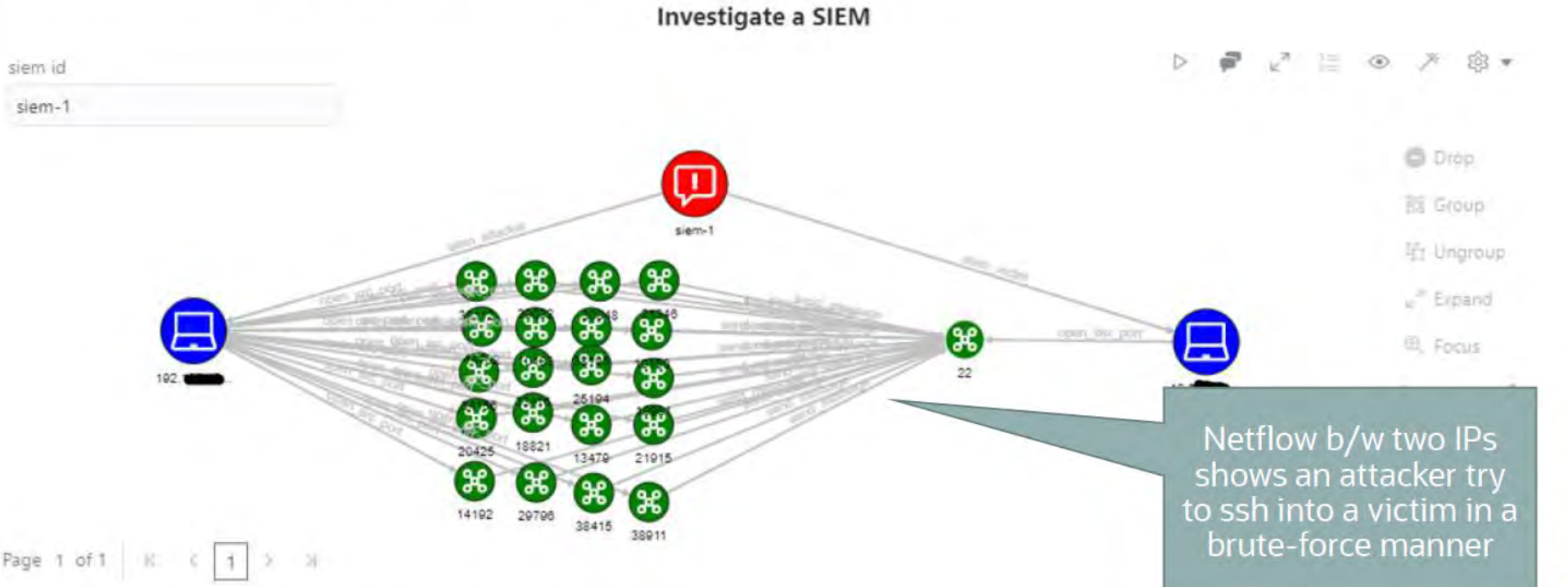
- Collaboration with SaaS Security Team
- Adding *graph-based* visual threat investigation





Example 1 (Synthetic Data)

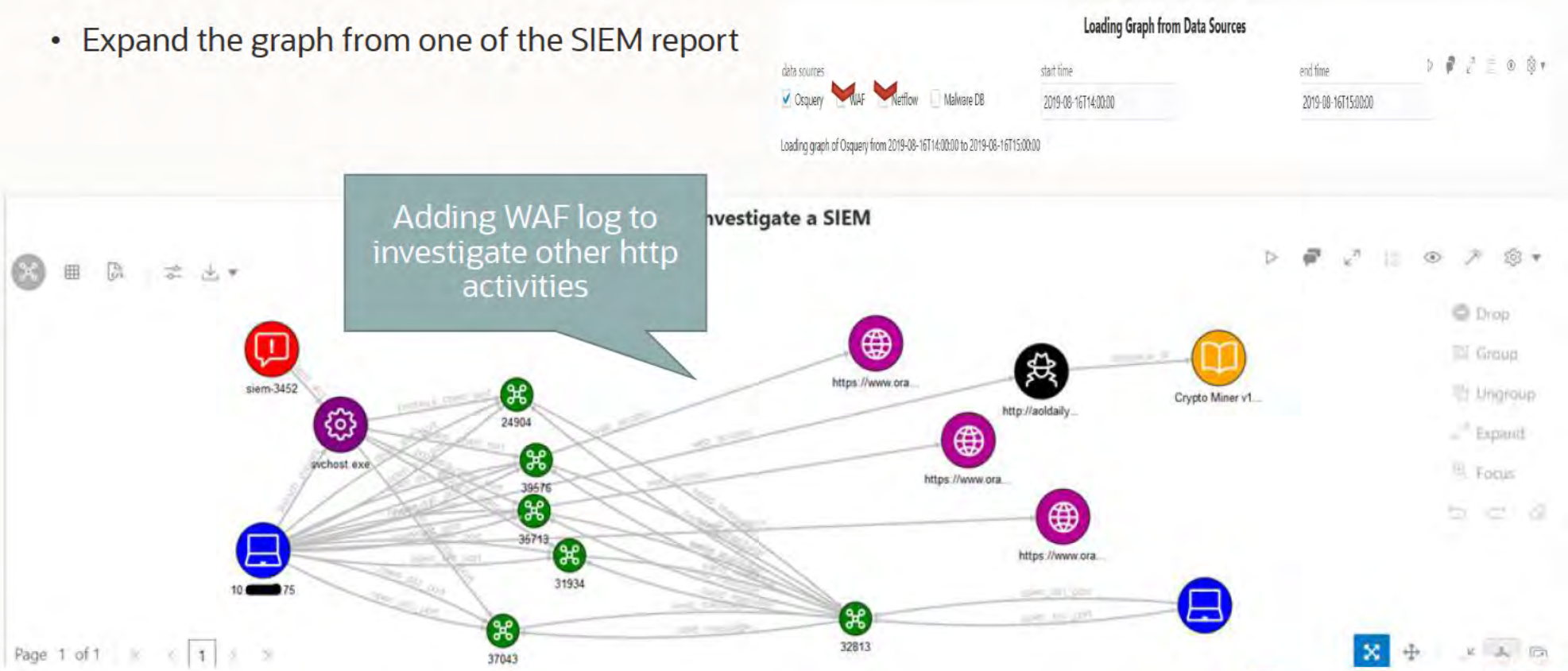
- Expand the graph from one of the SIEM report



Netflow b/w two IPs shows an attacker try to ssh into a victim in a brute-force manner

Example 2 (Synthetic Data)

- Expand the graph from one of the SIEM report



Agenda

1 Cyber Security and Graph

2 Threat Hunting using Visual Graph Exploration

3 Threat Detection using Graph Machine Learning (Ad-Fraud and VCN Flow)

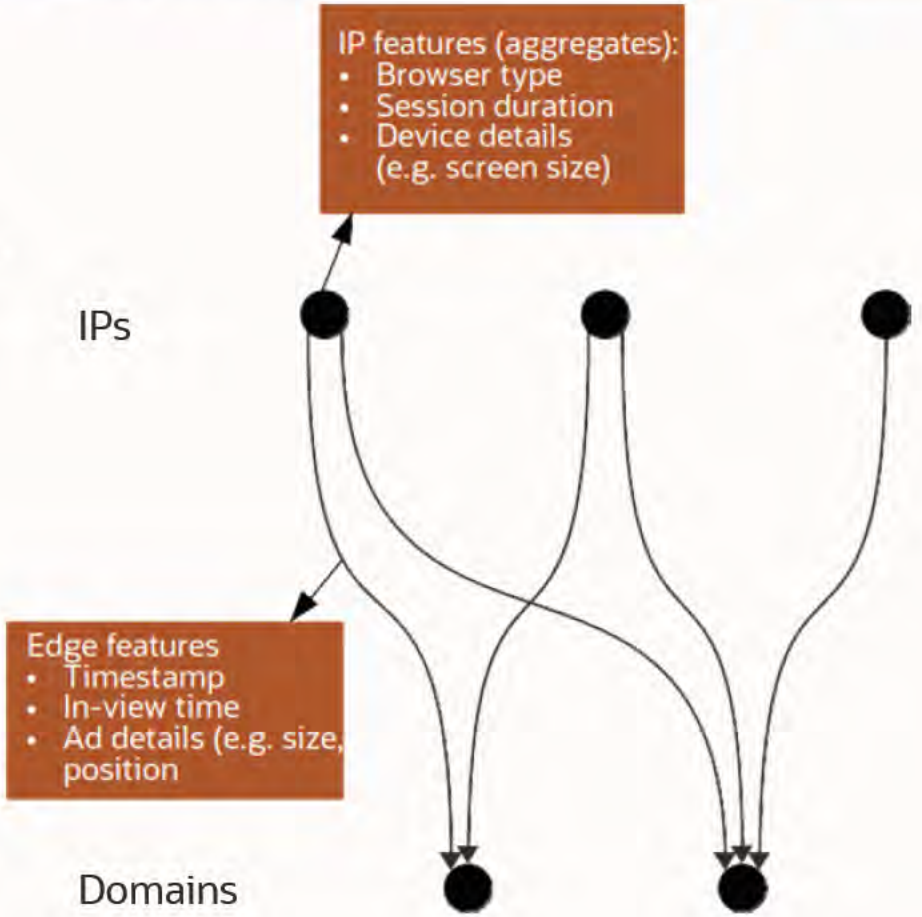
Ad-Fraud Detection

- Collaboration w/ Moat (Oracle Data Cloud)
- Non-human actors click online ads and create impressions
  - Advertisers gain no value from these → they should not be charged
- Sources of *Invalid Traffic* (IVT) include:
  - **Ad-fraud:** fraudulent actors create botnets that click on ads to generate revenue
  - Crawlers, scrapers, indexers and automated browsers

- Overall, IVT creates large costs to advertisers → estimated **$23 billion in 2019**
- Challenging to detect
  - Dynamic Adversaries (bot-masters try to avoid detection)
  - Labeled data is scarce and unreliable (most advanced bots will not be labeled)

Graph based IVT detection

- The problem can naturally be framed as a *node classification problem* on a **graph**
  - Graphs express both small and large scale patterns in user behavior
  - *Small scale*: how often and in what patterns do users access domains?
  - *Large scale*: what types of websites does a user like to visit, what websites have related user bases?
- We are interested in **false positives**, i.e., IPs that Moat labeled benign, and we label as bots
  - Moat has rule-based systems that detect simple bots with high precision
  - They may have been missing many advanced bots



Graph Signals: Modeling your Data into Graph

(1) Graph captures additional signals from your data

How data entities are inter-connected to each other

VIDEO_SALES_ORDERS

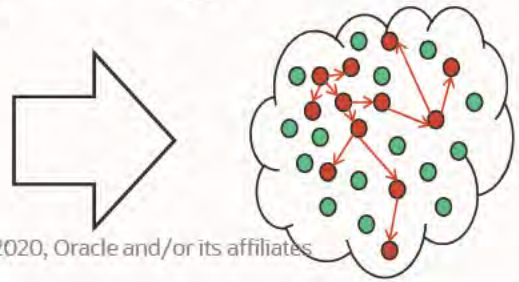
SALES_ID	CUST_NAME
10	SMITH
20	JONES
30	TURNER
40	ADAMS

SALES_ORDER_LINE_ITEMS

SALES_ID	LINE_ID	PROD_ID
10	1	1000
10	2	3000
20	1	4000
20	2	3000
20	3	2000
30	1	1000
30	2	1000
40	1	4000

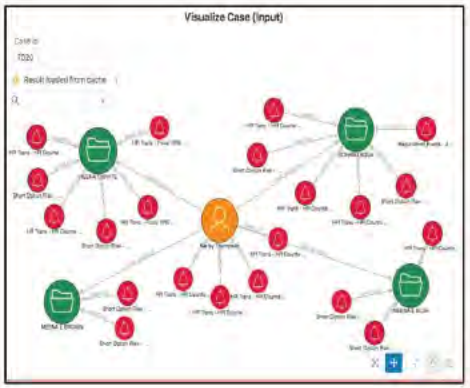
VIDEO_PRODUCTS

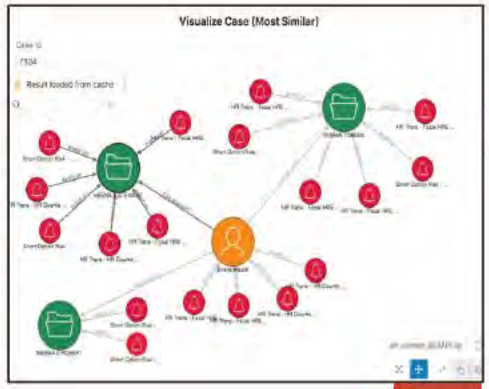
PROD_ID	PROD_DESC
1000	TOY STORY
2000	TRUE LIES
3000	POPCORN
4000	STARGATE



(2) Graph gives a way to capture unstructured or ad-hoc sets of data

e.g. Two cases in financial crime investigation:
➔ How similar are they?





Machine Learning and Graphs

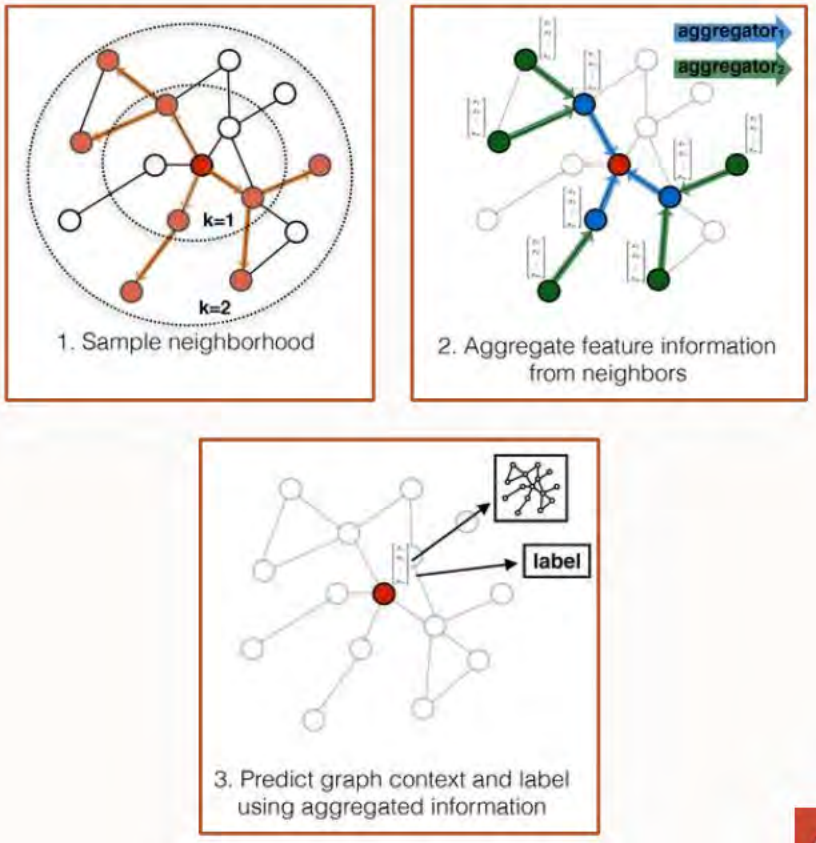
Need a systematic methodology that turns graph information into n-dimensional numeric representation, i.e. embedding

➔ We discuss two embedding techniques: vertex embedding and graph embedding

GraphSAGE

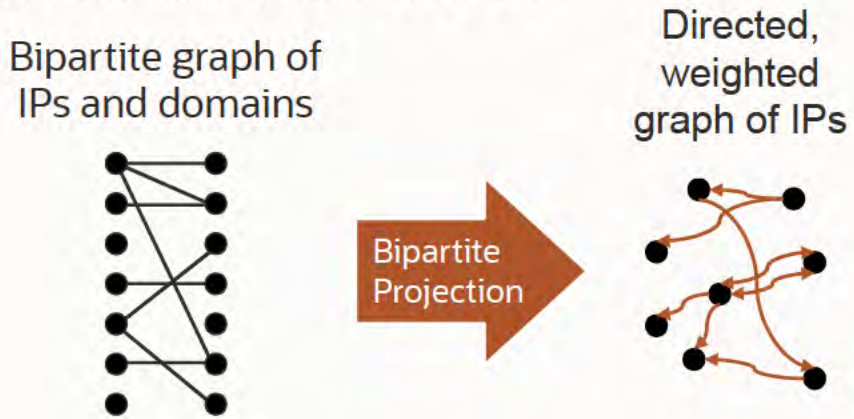
Inductive Representation Learning on Large Graphs, Hamilton et al. NIPS'17

- GraphSAGE is an **inductive**, **sampling** based Graph Neural Network (GNN) algorithm
- Unlike DeepWalk, it does not learn embeddings (slow, global optimization), but rather a **function that generates embeddings** (efficient, local optimization)
- Many Advantages:
  - **Inductive** → once trained on one day, can predict on any day (i.e., unseen nodes or unseen graph)
  - **Sampling based** → high efficiency makes it viable for very large graphs
  - **Feature aware** → makes predictions based on graph topology and graph features



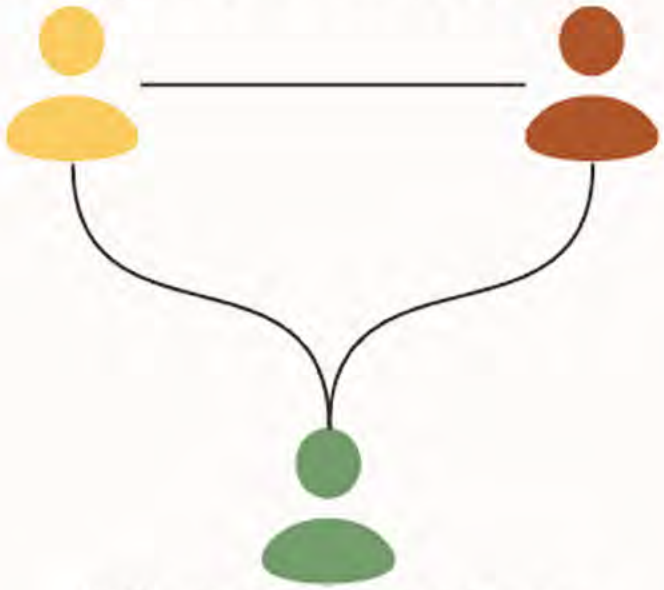
Homogeneous Graph Projection

- GraphSAGE is designed to work on homogeneous graphs (single type of nodes)
  - Our IP-Domain graph is heterogeneous!
- **Idea**: project the bipartite graph onto a homogeneous IP-IP graph
- Doing this allows us to run GraphSAGE on our data
- Homogeneous GraphSage is available from PGX 19.4.0!

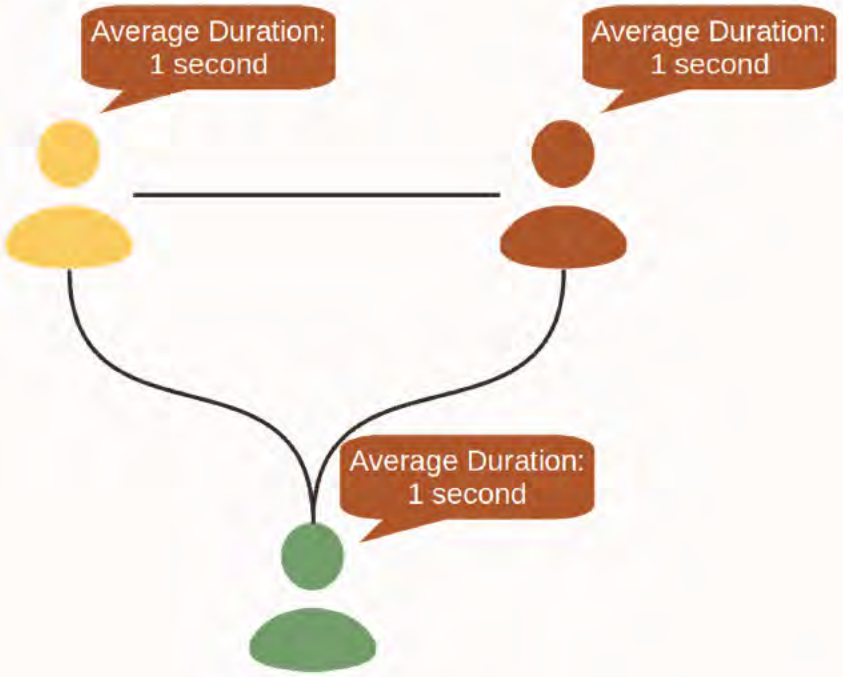


What do we gain from the IP features?

Botnets often have unusual IP features, and cluster together with similar bots



Difficult to make conclusions

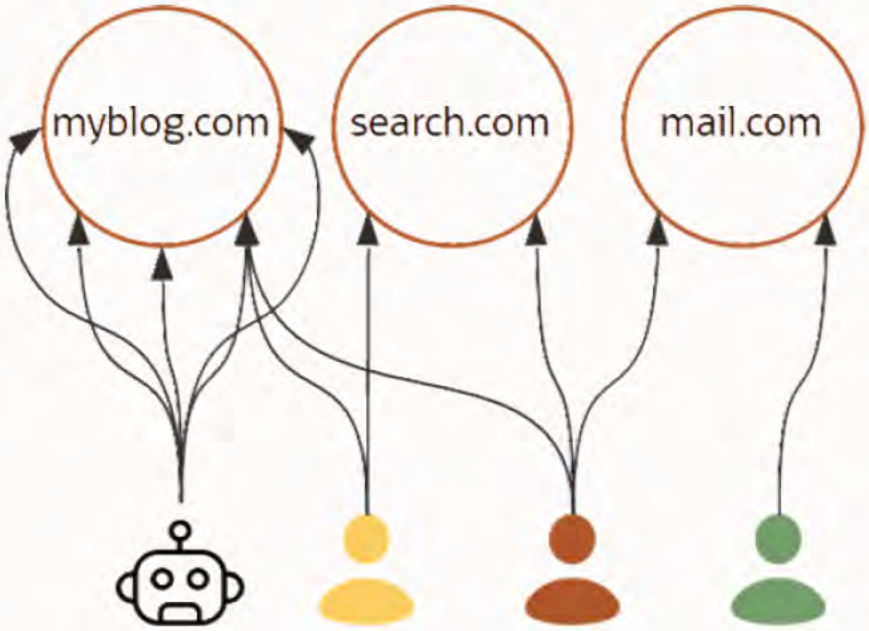


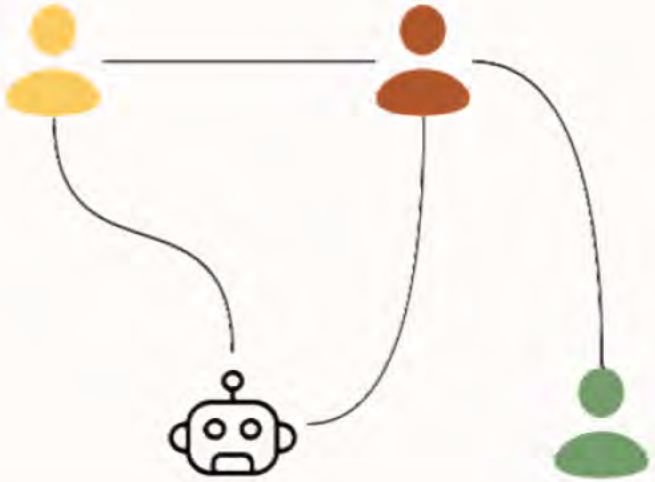
Clearly suspicious

Homogeneous GraphSAGE pipeline

- Classification accuracy: ~80% → **10% improvement over DeepWalk**
- We gain all the advantages of GraphSAGE:
  - Making use of both **features** and **graph topology**
  - Inductive
  - No longer have to re-train every day
  - End-to-end training

What is lost in Projection?

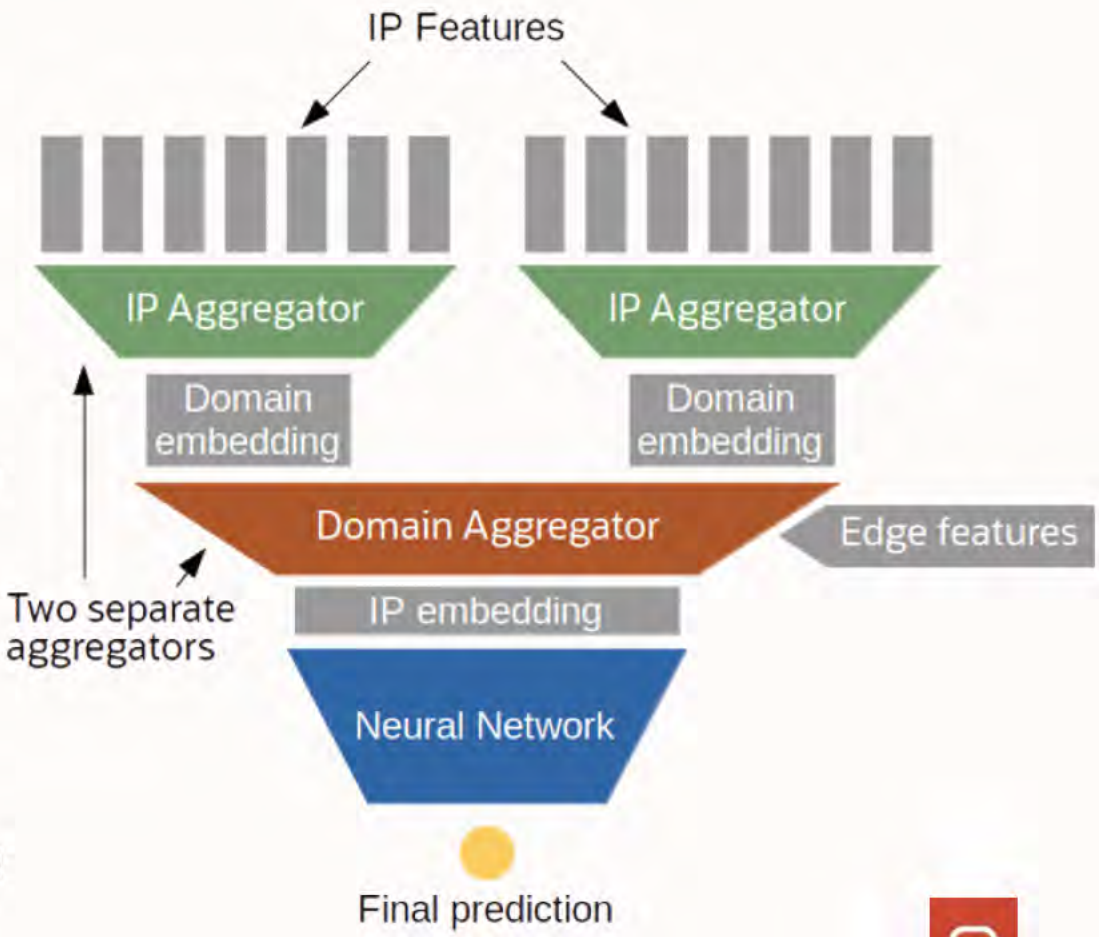




The bot's behavior is entirely hidden by the projection!

Heterogeneous Architecture

- With some changes in architecture, we design a GraphSAGE variant that can run directly on the heterogeneous graph
- Unlike the homogeneous approach, we retain all multi-edges, and can make use of edge features
  - Supporting edge features is not mentioned in the original paper, but they are a relatively simple extension
- Technical challenge: the graph becomes considerably larger
  - Need to add nodes for all the domains
  - Many more edges (including reverse edges which need to be stored in our data-structure)



Heterogeneous Architecture

- Classification accuracy: ~89% → **9% improvement over Homogeneous GraphSage**
- Inherits/retains all advantages of Homogeneous GraphSAGE
- First model that is able to model **temporality**
  - Temporality is very crucial in this domain and captured via edge features as well as the RNN-based Domain aggregator
  - Some prior art even focuses on detecting bots using *only temporal features*
- The model was strong enough to conclude our investigation stage and move ahead with deployment

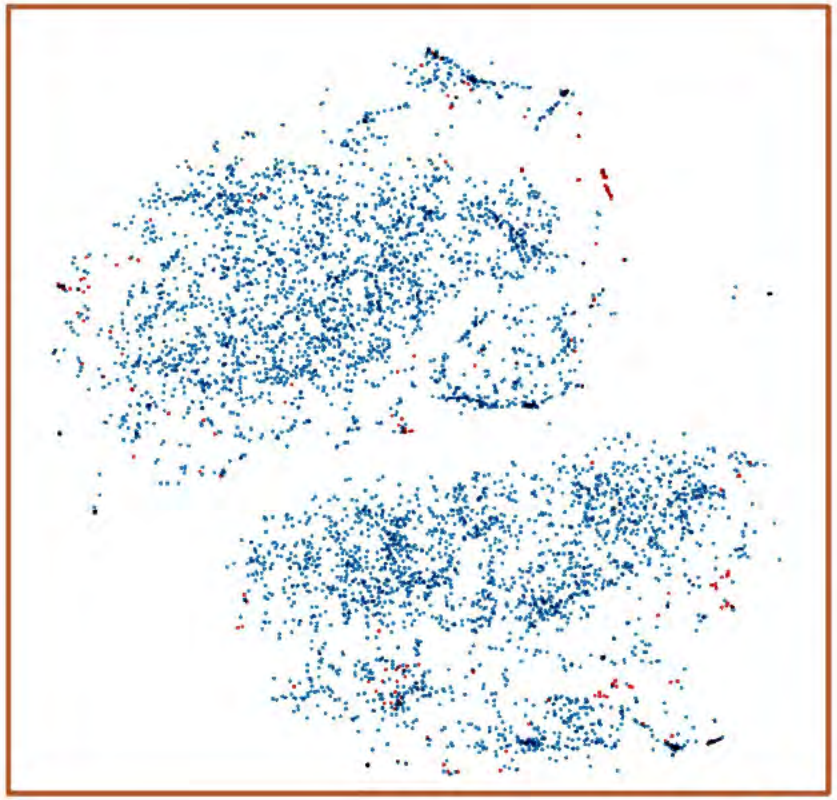
(e.g. DeBot: Twitter Bot Detection via Warped Correlation – Chavoshi et al. ICDM'2016)

Outlook – Unsupervised Bot Detection

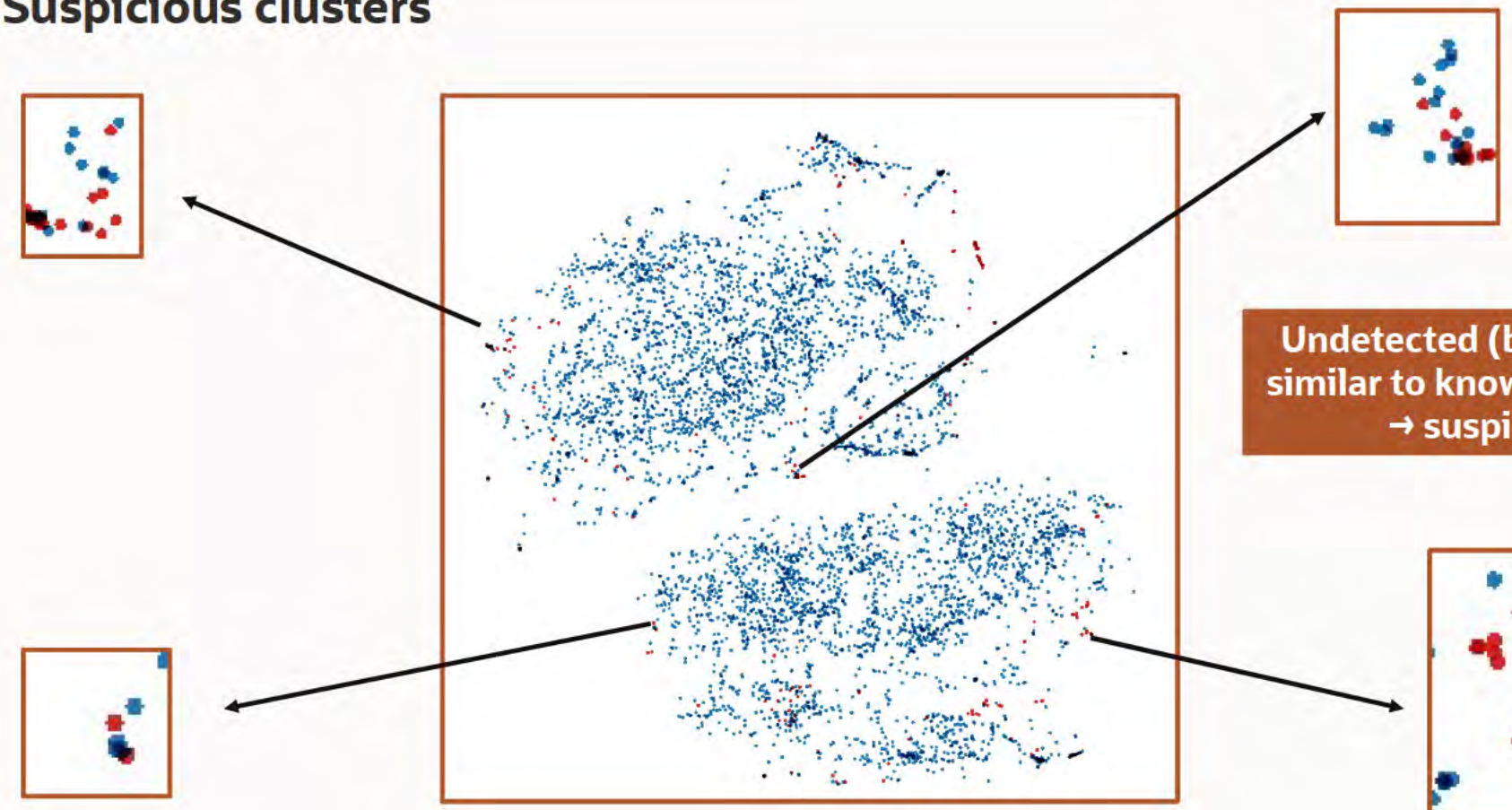
- **Supervised**: train using labels
- **Unsupervised**: train without labels
- Moat labels encode already known bots
  - Moat labels capture only a subset of all bots (skewed towards simpler bots captured via pre-defined rules)
  - Learning using the labels implies we will always find similar bots to those that are known
  - We want to find very **different types of bots** as well as capture **outliers**
- Promising direction and current research: Unsupervised Heterogeneous GraphSAGE
  - Application of our heterogeneous architecture
  - Produces embeddings **without employing the labels**

Unsupervised Embeddings

- The model is trained to produce high dimensional embeddings (visualized here with TSNE) **without knowing labels**
- Embedding space encodes **IP behavior**
  - **IPs close in embedding space behave similarly**
  - Visit similar websites, have similar temporal access patterns, etc.
- We can use this structured space to find new bots

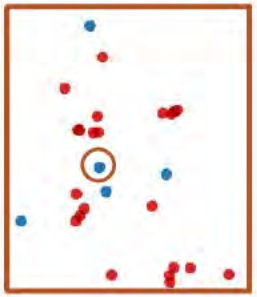


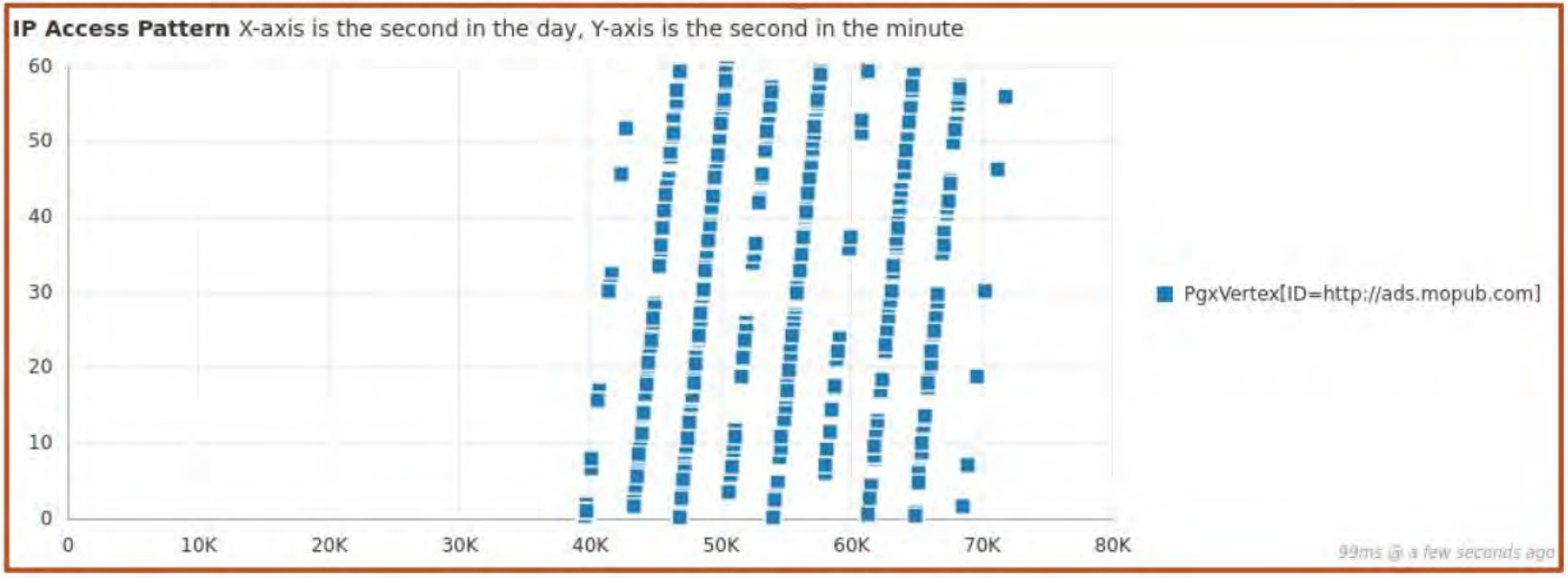
Suspicious clusters



Undetected (blue) IPs act similar to known bots (red) → suspicious

Ad Fraud Detection: Exploring a specific Cluster





Only accesses the same domain for 9 hours → rather suspicious behavior!

10800 seconds = 3 Hours

VCN Flow Analysis: Goal

- Collaboration with Data Scientists in OCI
- Analyze Virtual Cloud Network (VCN) logs
- Identify suspicious activities, and IPs related to it
- Machine-Learning Approach
  - Apply outlier detection technique

Overview

VCN Flow Logs keeps detailed records of every flow that passes through your VCN and presents this data for analysis in the Oracle Cloud Infrastructure Logging service. The data includes information about the source and destination of the traffic, along with the quantity of traffic and the "permit" or "deny" action taken, based on your network security rules. You can use this information for network monitoring, troubleshooting, and compliance. Through integration with the Logging service, you can view, search, and retrieve log files.

	version	source_ip	destination_ip	source_port	destination_port	protocol	packets	bytes	start_time	end_time	action	capture_status
0	2	-	-	-	-	-	-	-	1563510404	1563510464	-	NODATA
1	2	147.[redacted].185	10.[redacted].10	6443	49262	6	44	22516	1563514696	1563514722	ACCEPT	OK
2	2	10.[redacted].10	147.[redacted].185	49262	6443	6	44	2498	1563514696	1563514722	ACCEPT	OK
3	2	147.[redacted].185	10.[redacted].10	6443	36808	6	46	22752	1563514696	1563514722	ACCEPT	OK
4	2	10.[redacted].10	147.[redacted].185	36808	6443	6	46	2654	1563514696	1563514722	ACCEPT	OK

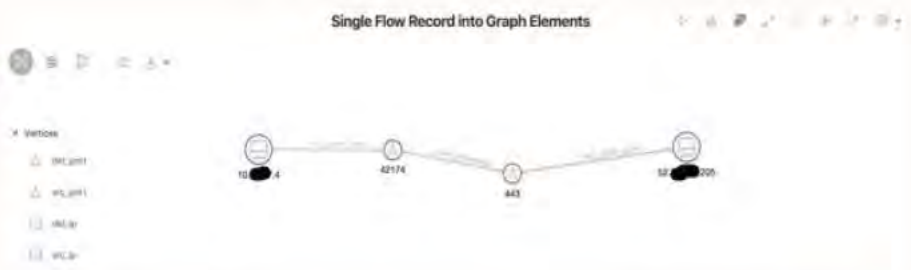
Examples of VCN flow logs

VCN Flow Analysis: Approaches

- Approach 1 - Handcrafted features
  - Aggregate original features per IP
    - 18 features

```
Index(['to_external_num_of_flows', 'to_external_unique_source_ports',
       'to_external_unique_internal_ips', 'to_external_avg_packets',
       'to_external_sum_packets', 'to_external_avg_bytes',
       'to_external_sum_bytes', 'to_external_avg_duration',
       'to_external_sum_duration', 'to_internal_num_of_flows',
       'to_internal_unique_destination_ports',
       'to_internal_unique_internal_ips', 'to_internal_avg_packets',
       'to_internal_sum_packets', 'to_internal_avg_bytes',
       'to_internal_sum_bytes', 'to_internal_avg_duration',
       'to_internal_sum_duration'],
      dtype='object')
```

- Approach 2 – Adding Graph Features
  - Construct a graph from the VCN flows
  - Extract Graph features
  - Using Graph ML Technique
  - Feed to outlier detection model



Approaches

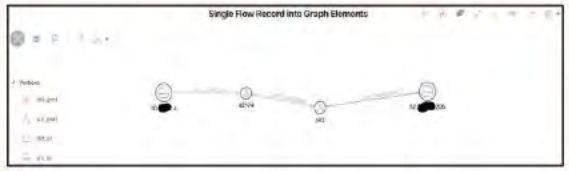
Original Dataset

version	source ip	destination ip	source port	destination port	protocol	packets	bytes	start_time	end_time	action	capture status
0	2	-	-	-	-	-	-	1565210434	1565210464	-	NODATA
1	2	147.[redacted]185	9[redacted]10	6443	49282	6	44	22516	1565614696	1565614722	ACCEPT	OK
2	2	10[redacted]10	147.[redacted]85	49282	6443	6	44	3493	1565614696	1565614722	ACCEPT	OK
3	2	147.[redacted]185	10[redacted]10	6443	38808	6	48	22753	1565614696	1565614722	ACCEPT	OK
4	2	10[redacted]10	147.[redacted]85	38808	6443	6	48	3954	1565614696	1565614722	ACCEPT	OK

Original Features

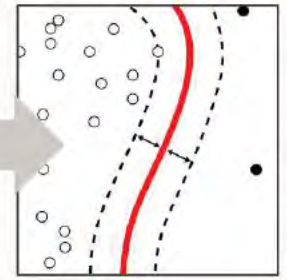
Graph Rep

+ Graph Features



ML Model

Outlier Detection



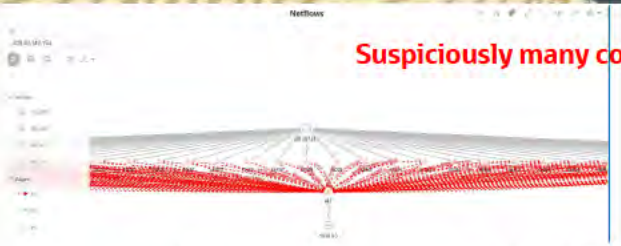
Machine Learning Model (Outlier Detection)

- Isolation Forest
  - Learn multiple decision trees
  - Outliers would be located in lower level in the trained decision trees
- Autoencoder
  - Learn an autoencoder
  - Outliers would have higher reconstruction loss

VCN Flow Analysis: Evaluation

	ip	asname	outlier_score_isoforest	report_count
11004	147 [redacted] 185	ORACLE-BMC- [redacted] - Oracle Corporation, US	1.000000	0
9728	129 [redacted] 93	STORTEK-INT - Oracle Corporation, US	0.987607	0
9727	129 [redacted] 189	STORTEK-INT - Oracle Corporation, US	0.987607	0
9963	134 [redacted] 2	ORACLE-BMC- [redacted] - Oracle Corporation, US	0.984114	0
9964	134 [redacted] 2	ORACLE-BMC- [redacted] - Oracle Corporation, US	0.983498	0
9968	134 [redacted] 2	ORACLE-BMC- [redacted] - Oracle Corporation, US	0.983088	0
11005	147 [redacted] 192	ORACLE-BMC- [redacted] - Oracle Corporation, US	0.976131	0
38922	54 [redacted] 8	AMAZON-02 - Amazon.com, Inc., US	0.972461	0
9715	129 [redacted] 39	ORACLE-BMC- [redacted] - Oracle Corporation, US	0.967784	0
9719	129 [redacted] 134	ORACLE-BMC- [redacted] - Oracle Corporation, US	0.967581	0
9711	129 [redacted] 149	ORACLE-BMC- [redacted] - Oracle Corporation, US	0.967175	0
9718	129 [redacted] 236	ORACLE-BMC- [redacted] - Oracle Corporation, US	0.960290	0
38919	54 [redacted] 168	AMAZON-02 - Amazon.com, Inc., US	0.960088	0
9722	129 [redacted] 243	ORACLE-BMC- [redacted] - Oracle Corporation, US	0.959886	0
9712	129 [redacted] 201	ORACLE-BMC- [redacted] - Oracle Corporation, US	0.959886	0
22225	216 [redacted] 20	PSFT-INC - Oracle Corporation, US	0.958473	0
14890	18 [redacted] 225	AMAZON-02 - Amazon.com, Inc., US	0.958272	0
10707	140 [redacted] 6	GITHUB - GitHub, Inc., US	0.958272	0
19238	192 [redacted] 117	GITHUB - GitHub, Inc., US	0.958070	0
19237	192 [redacted] 116	GITHUB - GitHub, Inc., US	0.957868	0

Handcrafted Feature



Reports in AbuseIP.com

	ip	asname	outlier_score_isoforest	eport_count
24148	165 [redacted] 32	DIGITALOCEAN-ASN - DigitalOcean, LLC, US	1.000000	34
25947	167 [redacted] 212	DIGITALOCEAN-ASN - DigitalOcean, LLC, US	0.961955	36
19846	165 [redacted] 121	DIGITALOCEAN-ASN - DigitalOcean, LLC, US	0.961702	0
31487	185 [redacted] 16	FLOKINET, SC	0.959724	306
23202	185 [redacted] 52	NOVOGARA-AS, NL	0.926304	857
39731	37 [redacted] 232	ESTROWEB, NL	0.924728	0
8542	82 [redacted] 6	From ipremoval.sms.symantec.com/	920884	77
22713	209 [redacted] 154	ATT-INTERNET4 - AT&T Services, Inc., US	0.919210	0
14940	89 [redacted] 199	INT-NETWORK, SC	0.914807	0
42240	104 [redacted] 210	DIGITALOCEAN-ASN - DigitalOcean, LLC, US	0.911335	35
41746	37 [redacted] 67	ESTROWEB, NL	0.910755	302
18727	185 [redacted] 7	Datashield, Inc., SC	0.903205	237
242	37 [redacted] 216	ESTROWEB, NL	0.901549	397
9223	165 [redacted] 129	DIGITALOCEAN-ASN - DigitalOcean, LLC, US	0.895824	22
25155	212 [redacted] 177	Online SAS, FR	0.892718	54
10511	160 [redacted] 120	SUN-DSEO - Oracle Corporation, US	0.889860	0
26629	37 [redacted] 202	ESTROWEB, NL	0.889400	3847
30874	160 [redacted] 40	SUN-DSEO - Oracle Corporation, US	0.886877	0
30319	129 [redacted] 193	STORTEK-INT - Oracle Corporation, US	0.884799	0
11407	185 [redacted] 13	MYHOSTING-AS, RO	0.884309	159

spammer

malware distributer

+ Graph Feature

Summary

- **Threat Detection using Graph Machine Learning**
  - We demonstrate the applicability of Graph Machine Learning for advanced IVT detection
- **Threat Hunting using Graph Visualization**
  - We employ Graph Query and Visualization to investigate Suspected Alerts and hunt for possible Cyber threats
- **Applications of Graphs to other Security domains**

Thank you