



Global Analysis with Aggregation-based Beaconing Detection across Large Campus Networks

Annual Computer Security Applications Conference (ACSAC) 2023
December 4-8, 2023, Austin, Texas, USA

Yizhe Zhang, University of Virginia

Hongying Dong, University of Virginia

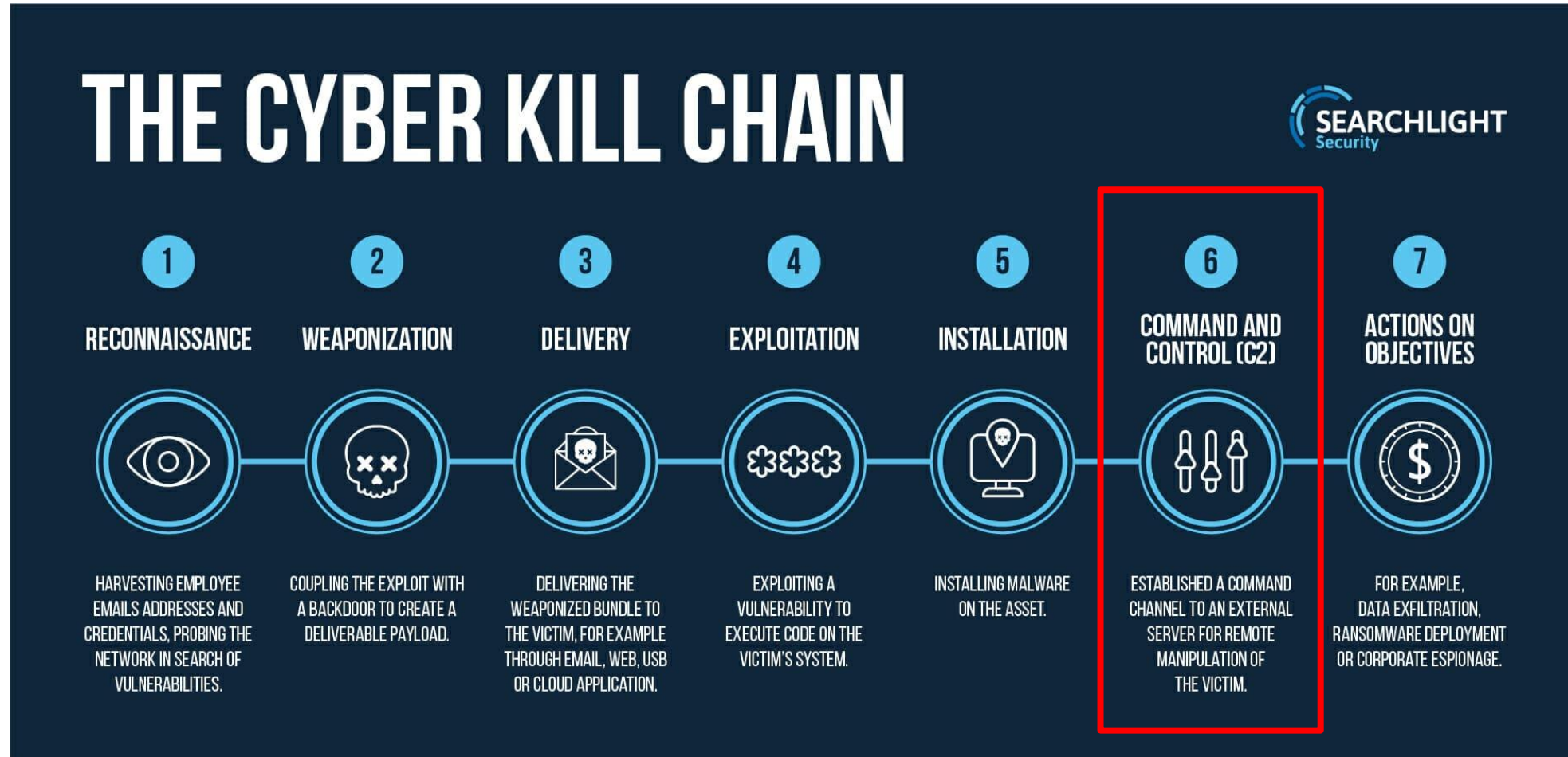
Alastair Nottingham, University of Virginia

Molly Buchanan, University of Virginia

Donald E. Brown, University of Virginia

Yixin Sun, University of Virginia

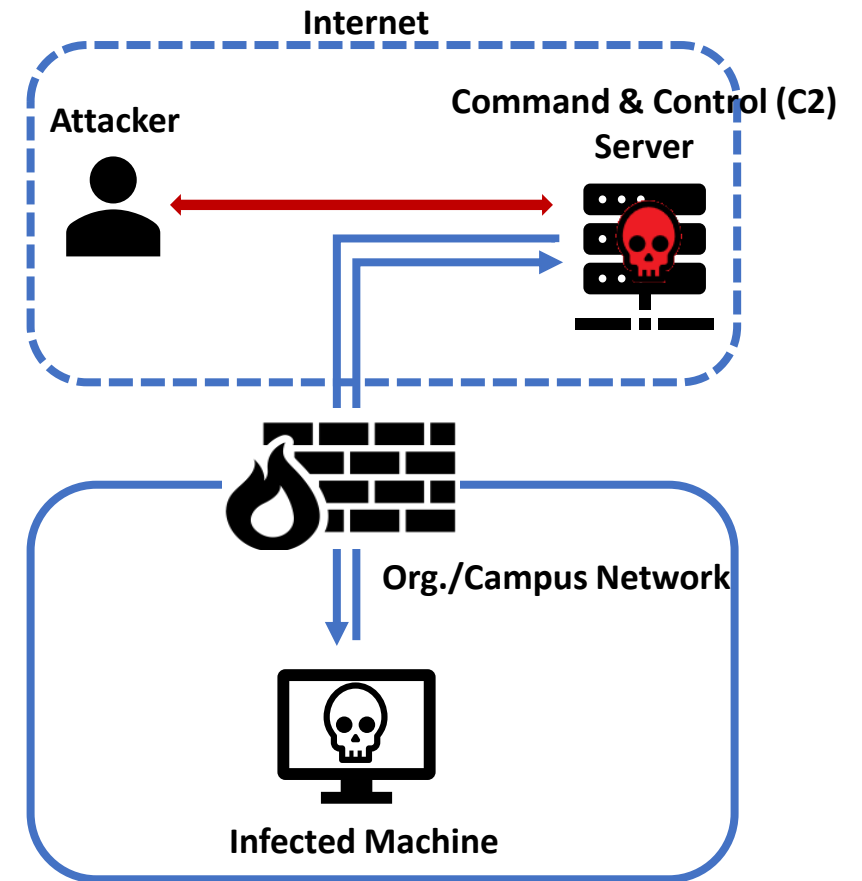
Malware Beaconsing Activity



<https://www.slcyber.io/shifting-left-in-the-cyber-kill-chain/>

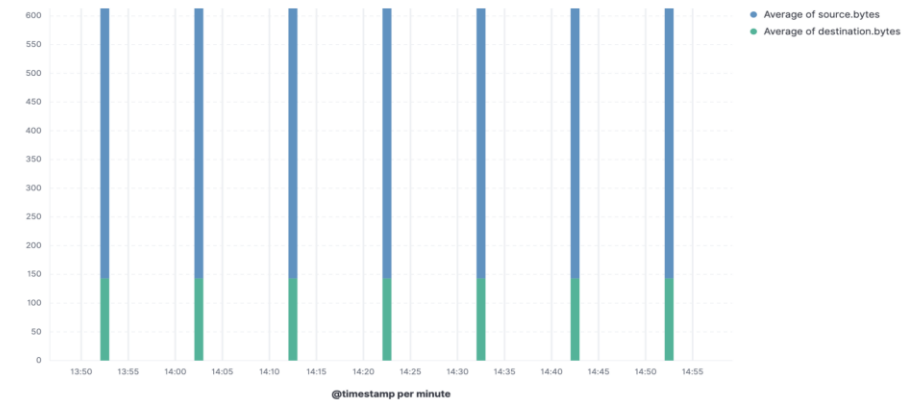
Malware Beaconing Activity

- Compromised machine/bot **regularly** announces its presence to remote C2 server.
 - Pre-programmed in malware control flow.
 - (Usually) exhibits **periodic** communication patterns.
- Effective detection tool.
 - Popular among widespread malware.
 - e.g., Zeus, Qbot, Conficker, Andromeda, njRAT.
 - No need to inspect packet payload.
 - effective when network traffic is encrypted (e.g. TLS).

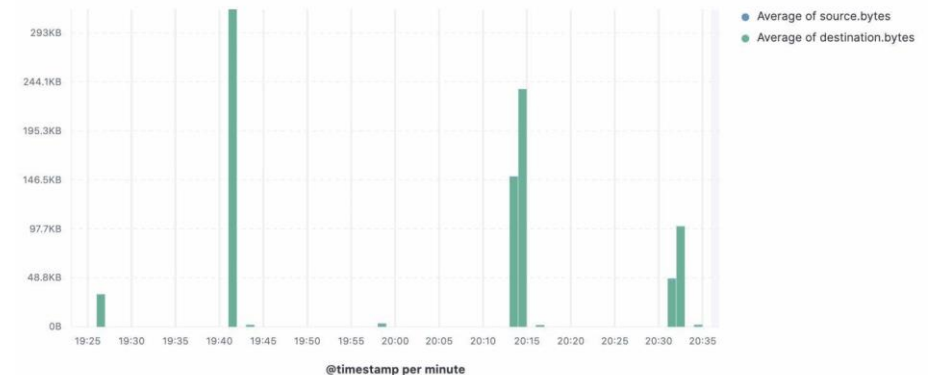


Beaconing Detection Challenges

- Disruption of the signals:
 - Network downtime, logging failure, etc.
- Adversary's countermeasures:
 - Noise/jitter.
 - DNS fast flux, etc.
- Differentiate *malicious* periodic activity from the benign ones:
 - E.g., software updates also have periodic behavior.



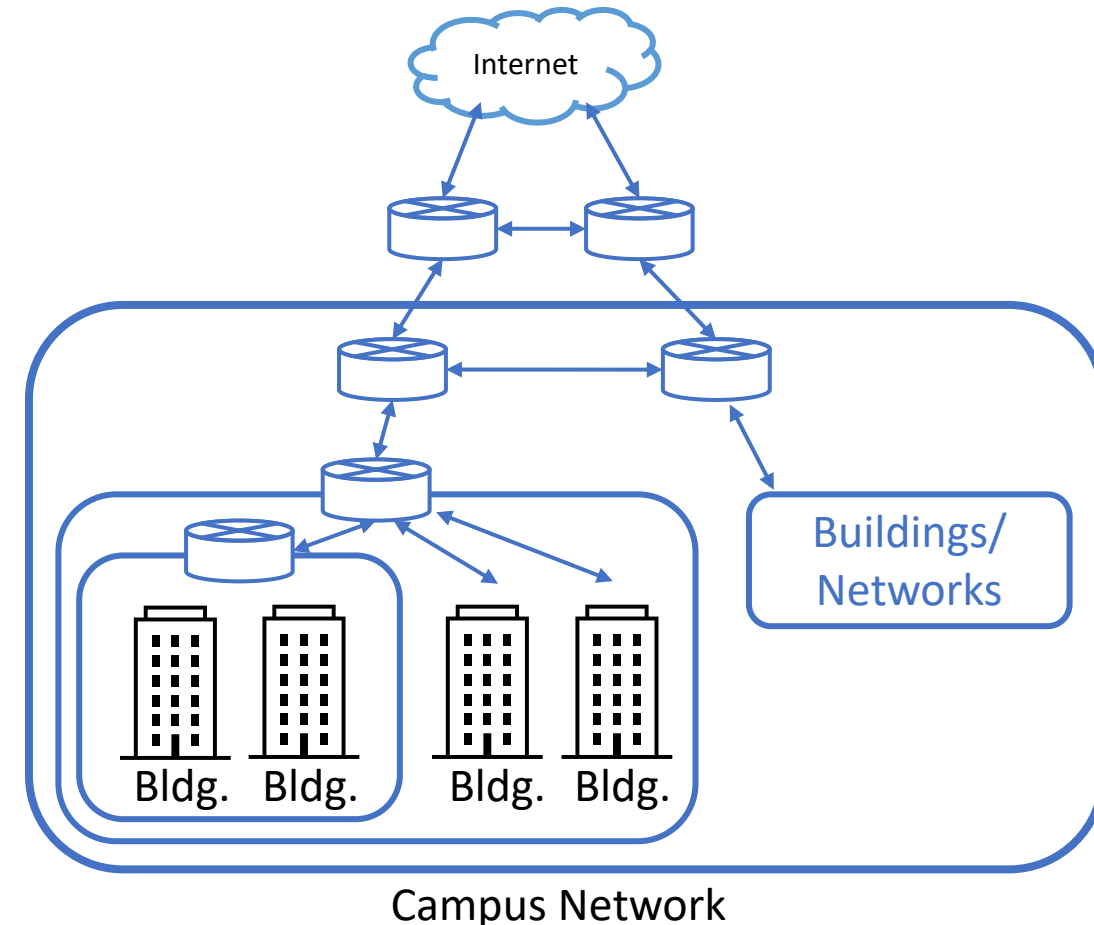
Koadic C2 beacons*



Beaconing activity with noise (jitter) *

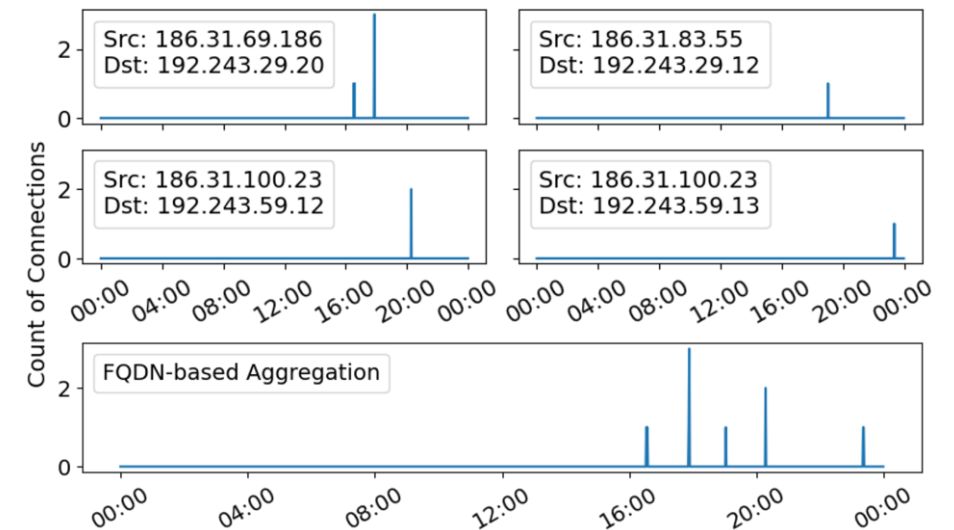
Beaconing Detection Challenges in Campus Network

- Large traffic volume.
- Ad-hoc devices.
- Difficulties in campus network host tracking:
 - Limited visibility into uninstrumented subnets.
 - Record loss due to logging infrastructure under heavy load.



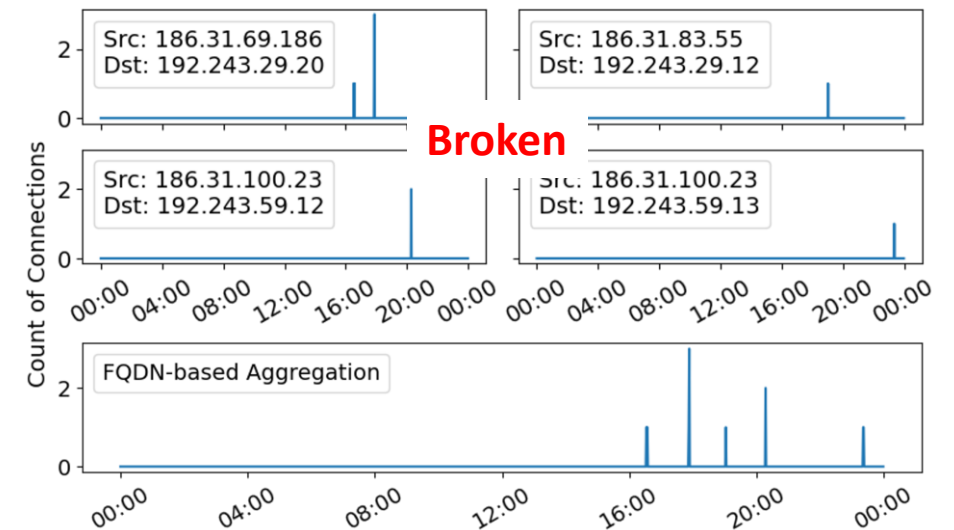
Fine-grained Detector (Prior Works)

- Fine-grained Detector:
 - Reconstruct time-series of network activity based on {source, destination} pair.
 - Source: {IP, port, MAC, user agent, etc.}.
 - Destination: {IP, port, FQDN, AS, URL, etc.}.
 - Build time-series as **precise** as possible.
- Limitations:
 - Not applicable in campus networks.
 - Vulnerable to common attacker evasion techniques.



Fine-grained Detector (Prior Works)

- Fine-grained Detector:
 - Reconstruct time-series of network activity based on {source, destination} pair.
 - Source: {IP, port, MAC, user agent, etc.}.
 - Destination: {IP, port, FQDN, AS, URL, etc.}.
 - Build time-series as **precise** as possible.
- Limitations:
 - Not applicable in campus networks.
 - Vulnerable to common attacker evasion techniques.



Broken

Periodic pattern is not evident in *fine-grained* analysis.

Aggregation-based Detector

- Aggregation-based Detector:
 - Focus on the time-series of **server-side** network activity.
 - Applicable in real-world campus network (or any large network that is heavily NATted or highly dynamic).
- Challenges:
 - **Noisier** signals as compared to fine-grained time series .



The aggregation of multiple periodic signals is still periodic.

Global Analysis with Aggregation-based Detector

- Global Analysis:
 - **Intuition:** Events that are undetectable in a single network become more pronounced/obvious when viewed across many heterogeneous networks?
- Aggregation-based Detector:
 - **Across protocols.**
 - **Across multiple organizations.**

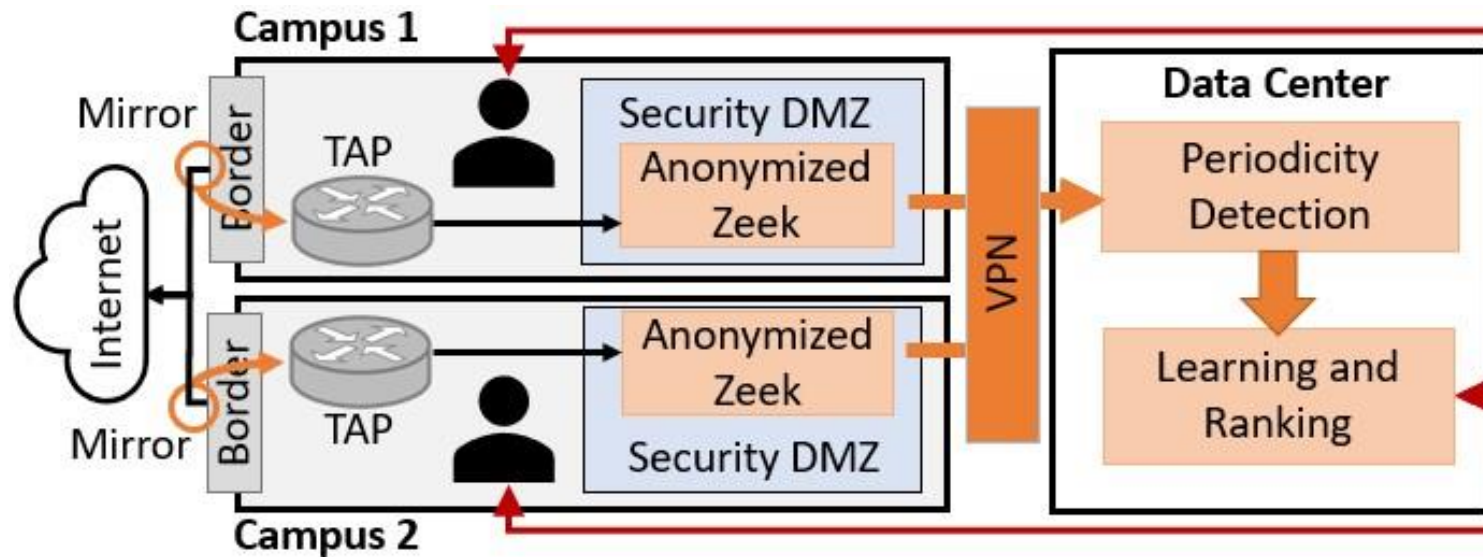
Global Analysis with Aggregation-based Detector

- Campus beaconing detection challenges:
 - Disruption of the signals.
 - Attacker's countermeasures (noise/jitter).
 - Difficulties to identify malicious beaconing activity from the benign ones.
 - **Campus network environment:**
 - Ad-hoc devices.
 - Record loss.
 - Huge amount of network traffic.
 - Lack of ground truth labels.
 - privacy concerns (anonymized data).
- Key features of our solution:
 - Aggregation-based detector focusing on server-side beaconing activity across protocols and universities.
 - New periodicity detection algorithm to handle noisy data.
 - Self-learning and active-learning pipeline to prioritize suspicious activities with limited labels.

System Overview

	Log	Log Size (gzipped)	Connections	Data Coverage
Campus1	HTTP	0.8 TB	8.92 B	96.97%
	SSL	3.8 TB	34.05 B	97.56%
Campus2	HTTP	0.8 TB	7.82 B	97.22%
	SSL	2.2 TB	24.44 B	97.22%
Total	-	7.6 TB	75.23 B	-

Zeek is a Network Security Monitoring Tool that used by many Security Operational Centers (SOCs).

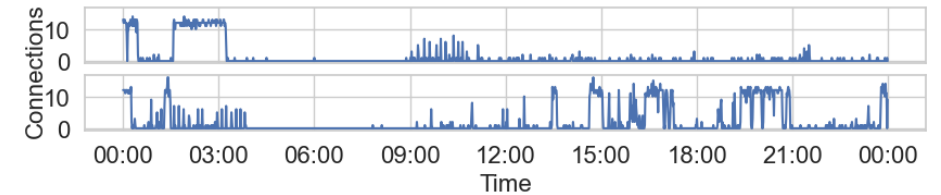


- **Aggregation-based periodicity detection** to reconstruct server-side time-series and identify periodic activities.
- **Learning and ranking pipeline** to prioritize malicious beaconing activity with limited human involvement.
- Ethical concerns: we follow IRB process and other regulations. Anonymization details can be found in our prior work*.

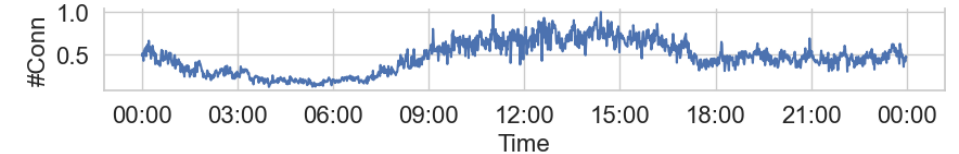
*Alastair Nottingham, Molly Buchanan, Mark Gardner, Jason Hiser, and Jack Davidson. 2022. Sentinel: A Multi-institution Enterprise Scale Platform for Data-driven Cybersecurity Research. In 2022 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW). IEEE, 252–257.

Periodicity Detection

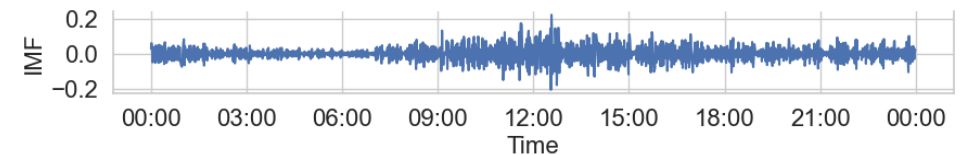
- Time series reconstruction:
 - Build time-series based on server's Fully Qualified Domain Name (FQDN).
- Signal decomposition:
 - Use Empirical Mode Decomposition (EMD) to decompose the time-series signal.
 - Use the first extracted intrinsic mode functions (IMF) to represent server's communication pattern.
- Periodicity detection:
 - Random permutation.
 - Fourier analysis.
 - Auto-correlation function (ACF).



Example of fine-grained time series.

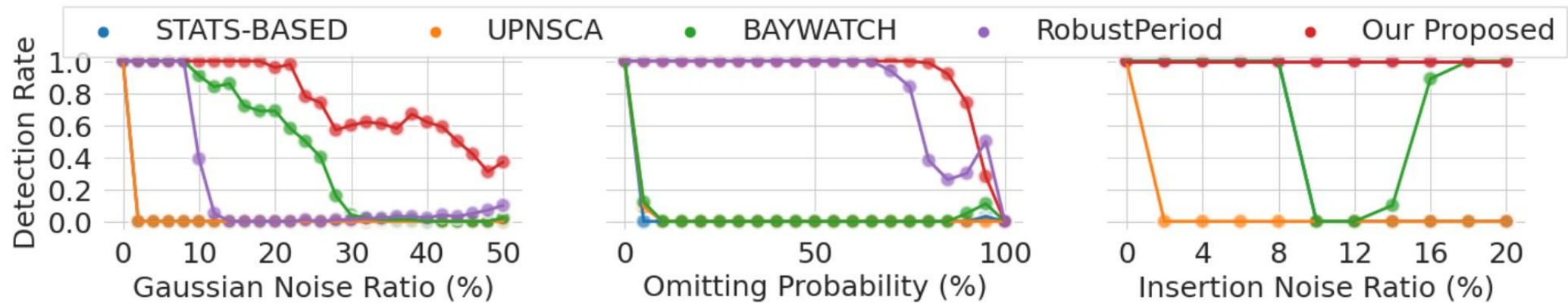


Aggregated time series.



Extract first IMF to represent server-side activity.

Periodicity Detection Evaluation



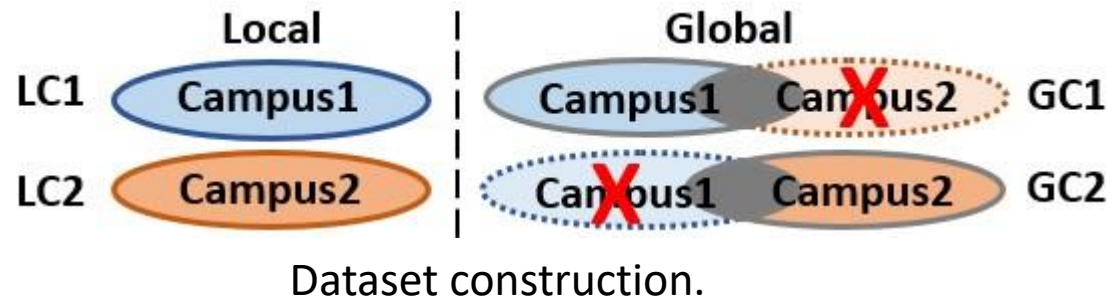
Comparison with existing algorithms using synthetic signals under various noise.

	STAT-based	UPNSCA	BAYWATCH	Our Proposed
Count of unique FQDNs	0	0	10,841	13,837

Comparison with existing algorithms using one-month real-world campus traffic.

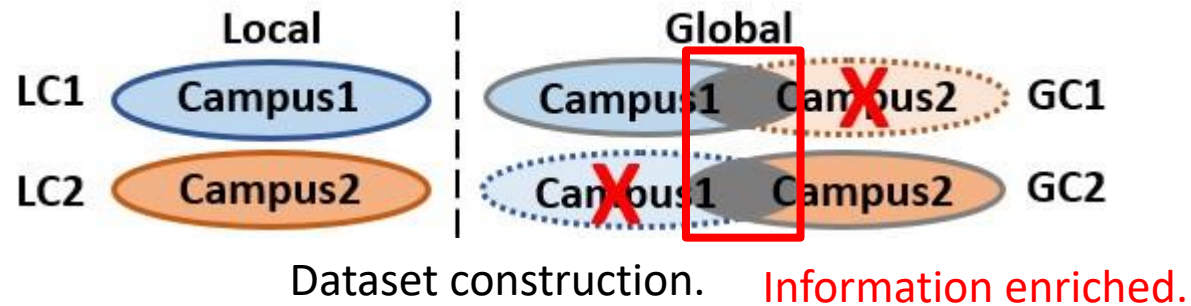
The unofficial implementation of RobustPeriod is too slow (one-minute to process one time-series) to be evaluated on campus traffic.

First Glance of the Efficacy of Global Analysis



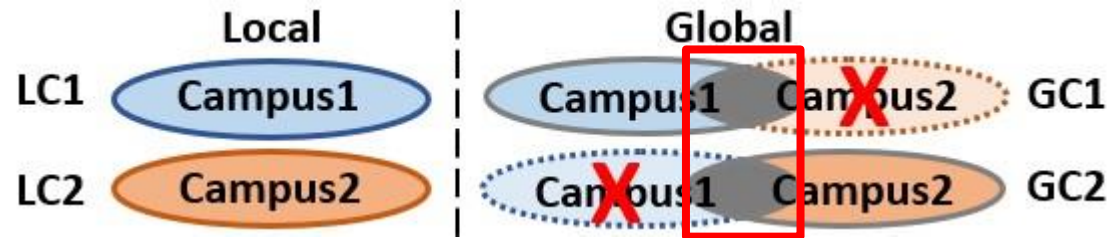
1. GC1 and LC1 have the same number of FQDNs.
2. GC2 and LC2 have the same number of FQDNs.
3. When FQDN_A is visited by both campuses (gray area), its information is enriched during the aggregation.

First Glance of the Efficacy of Global Analysis



1. GC1 and LC1 have the same number of FQDNs.
2. GC2 and LC2 have the same number of FQDNs.
3. When FQDN_A is visited by both campuses (gray area), its information is enriched during the aggregation.

First Glance of the Efficacy of Global Analysis

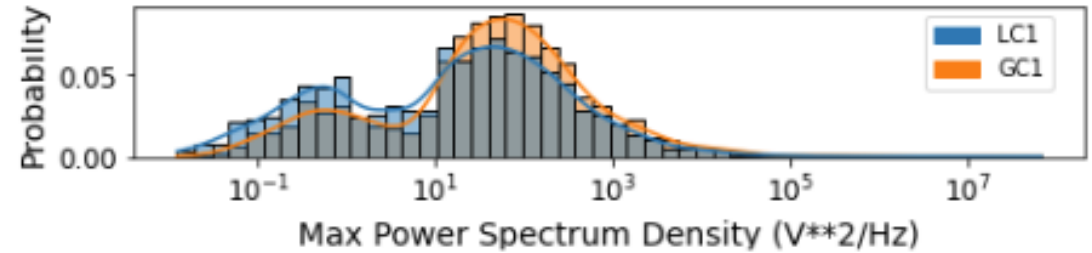


Dataset construction. Information enriched.

Average count of distinct FQDNs per day.

	LC1	LC2	GC1	GC2
Periodic	12,246	9,190	17,528	15,310
Total	514,777	357,644	514,777	357,644

1. GC1 and LC1 have the same number of FQDNs.
2. GC2 and LC2 have the same number of FQDNs.
3. When FQDN_A is visited by both campuses (gray area), its information is enriched during the aggregation.



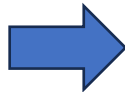
Learning and Ranking Pipeline

Prioritize malicious beaconing activity for SOC analysts.

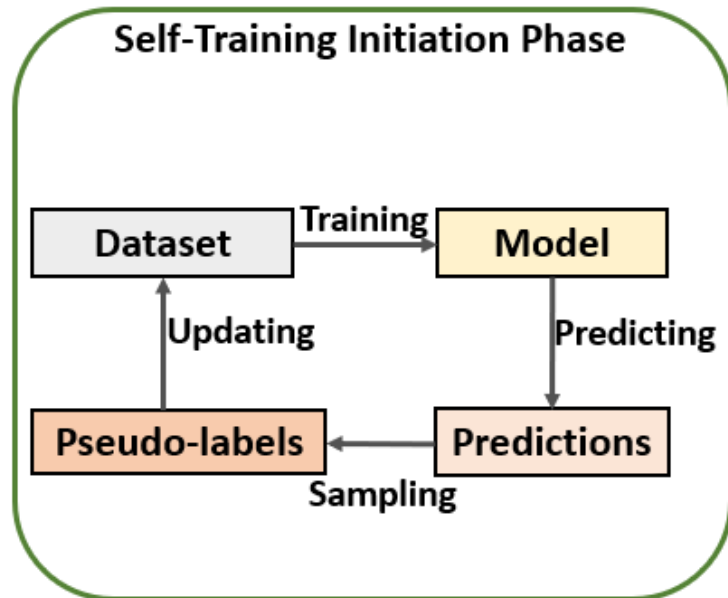
- Huge amount of data:
 - More than 500K FQDNs per day.
 - More than 15K periodic FQDNs per day.
- Limited ground truth:
 - General issue when dealing with real-world network traffic data.
 - Given >500K FQDNs per day, it's impossible to query all FQDNs.
 - Lagging issues with threat intelligence platform.
- Highly-imbalanced datasets:
 - Huge amount of benign traffic.
 - Small amount of malicious activity.

Learning and Ranking Pipeline

Prioritize malicious beaconing activity for SOC analysts.

- Huge amount of data:
 - More than 500K FQDNs per day.
 - More than 15K periodic FQDNs per day.
 - Limited ground truth:
 - General issue when dealing with real-world network traffic data.
 - Given >500K FQDNs per day, it's impossible to query all FQDNs.
 - Lagging issues with threat intelligence platform.
 - Highly-imbalanced datasets:
 - Huge amount of benign traffic.
 - Small amount of malicious activity.
- 
- Randomly sample and partially label the dataset in the starting phase using VirusTotal.
 - Use self-training to re-balance label distribution.
 - Use active-learning to continuously learn from experts and improve model performance throughout the time.

Learning and Ranking Pipeline

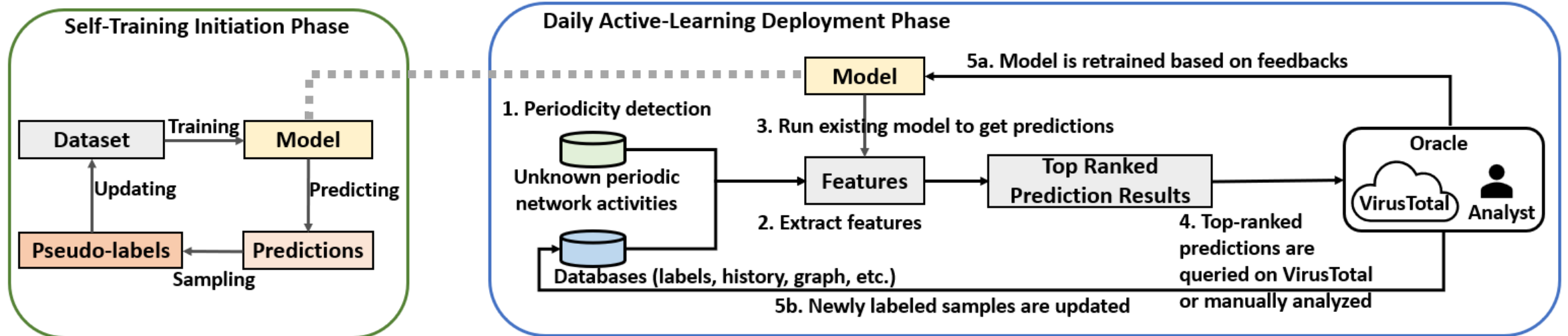


4-class RF model.

Each class corresponds to the number of the number of VirusTotal engines ($\#MalEng$) that detects a specific FQDN as malicious.

- Features:
 - Periodicity-based features.
 - Graph-based features.
 - Historical-based features.
 - Other features.
- Employ CReST self-training processing.
- Our sampling strategy:
 - $\alpha = 1$ when $\#MalEng \geq 2$
 - $\alpha = 0.05$ when $\#MalEng == 1$
 - $\alpha = 0$ when $\#MalEng == 0$
 - For highly-imbalanced dataset, the minority class is observed to have a very high precision.

Learning and Ranking Pipeline

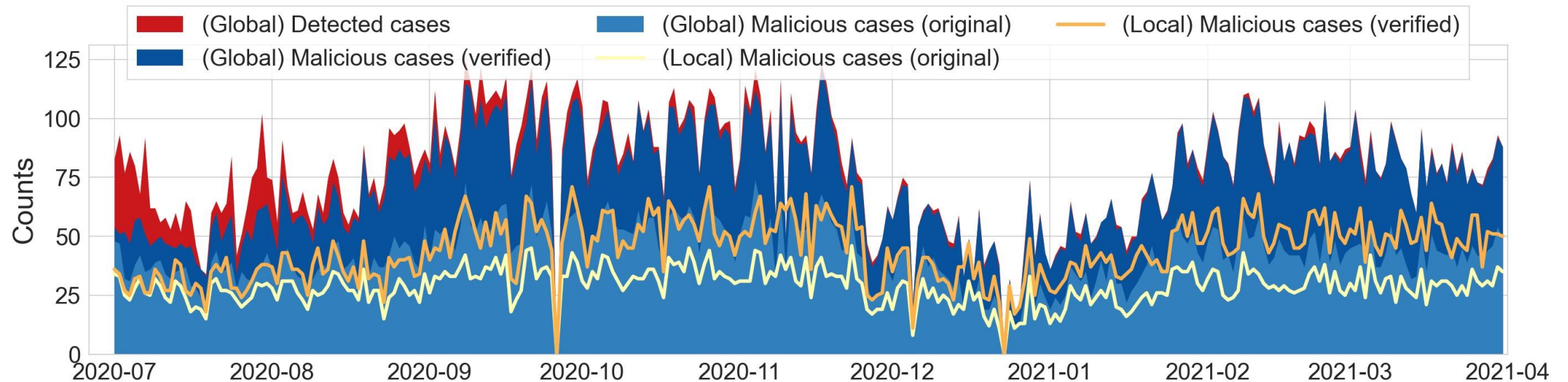


Two phase learning and ranking pipeline.

The final model in the self-training phase is the initial deployment model in active-learning pipeline.

Learning and Ranking Pipeline Evaluation

- User-centric performance:
 - For real-world SOC operation, the primary goal is to minimize False Positives and maintain a reasonable cases for manual verification.



Case for manual verification:

10 cases on average are reported to analysts for further investigation per day.

	Local Pipeline	Global Pipeline	Diff.
Detected	46.93	77.15	30.22
Malicious (Original)	28.79	42.44	13.65
Malicious (Verified)	14.84	29.69	10.22
Malicious (Total)	43.63	72.13	28.50
Unknown	3.30	5.02	1.72
Accuracy	92.97%	93.49%	-

Average daily detection.

Learning and Ranking Pipeline Evaluation

- Overall model performance:

	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>
Original Analysis	0.9957	0.9464	0.9969
Retrospective Analysis	0.9701	0.9675	0.9701

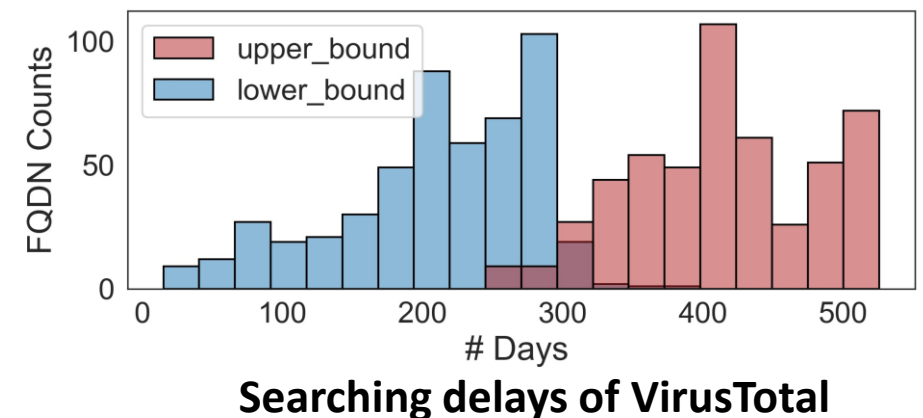
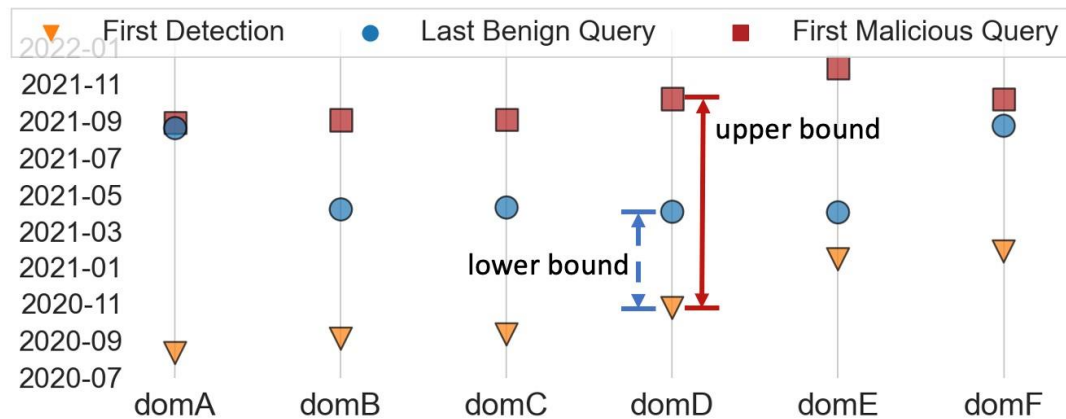
Randomly sample 10% of all domains from last 3-month (Jan-Apr 2021) and query them on VirusTotal.

Original analysis: scores by the time of detection (Jan-Apr 2021)

Retrospective analysis: scores computed by re-query all the domains in December 2022

Assessing VirusTotal's Searching Delay

- VirusTotal mechanism:
 - Searching: query database.
 - Scanning: request to scan the submitted request.
- Searching is a widely used mechanism, however:



Conclusion

- Global analysis:
 - Leverage data across multiple organizations.
- Aggregation-based periodicity detection algorithm:
 - Detect periodic activity with presence of large noise.
- Self-training and active-learning pipeline:
 - Perform detection on large volume of traffic with limited labels and human involvement.
- Evaluate and deploy the system across large-scale real-world campus networks.

Thank you.
Questions?