Effective Detection of Multimedia Protocol Tunneling using Machine Learning

USENIX Security 2018

Diogo Barradas, Nuno Santos, and Luis Rodrigues

Web Security & Privacy Lab

KAIST

Content

- Key Idea of This Paper
- Background
- Existing Metrics
  - Similarity-based Classification
- New Approach
  - Decision Tree-based Classification
- Beyond Supervised Anomaly Detection
- Discussion

Problem

- Covert channels should be unobservable.
- However, the evaluation of multimedia protocol tunneling techniques has been conducted using ad hoc methods.

Are these techniques truly unobservable?

Contribution

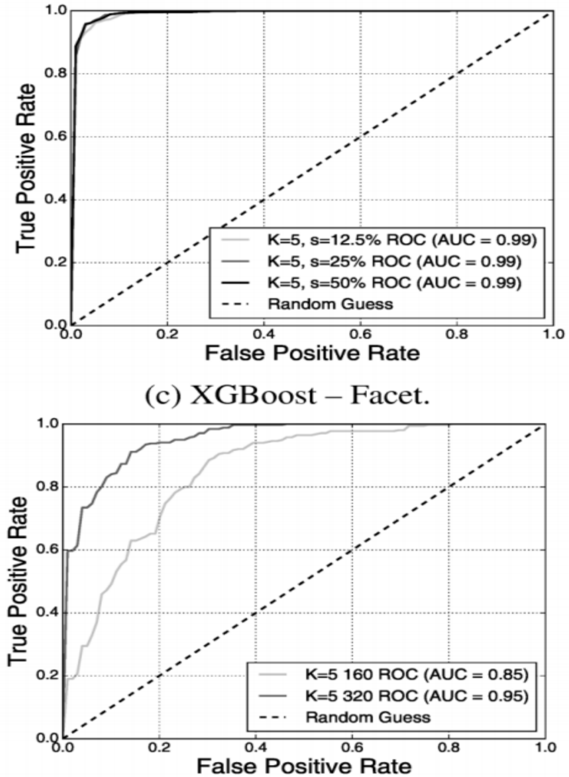
- Existing Decision-Tree based machine learning techniques can break the unobservability of the previous multimedia protocol tunneling techniques.
  - With low false positive rates
- What if datasets do not have labels?
  - Semi-supervised ML techniques
  - Unsupervised ML techniques

Web Security & Privacy Lab

KAIST

Result

- Decision-Tree based ML technique successfully detects covert traffics.
- AUC = 0.99 for Facet
- AUC = 0.85~0.95 for DeltaShaper



(c) XGBoost – Facet.

(f) XGBoost – DeltaShaper.

Meaning of the Paper

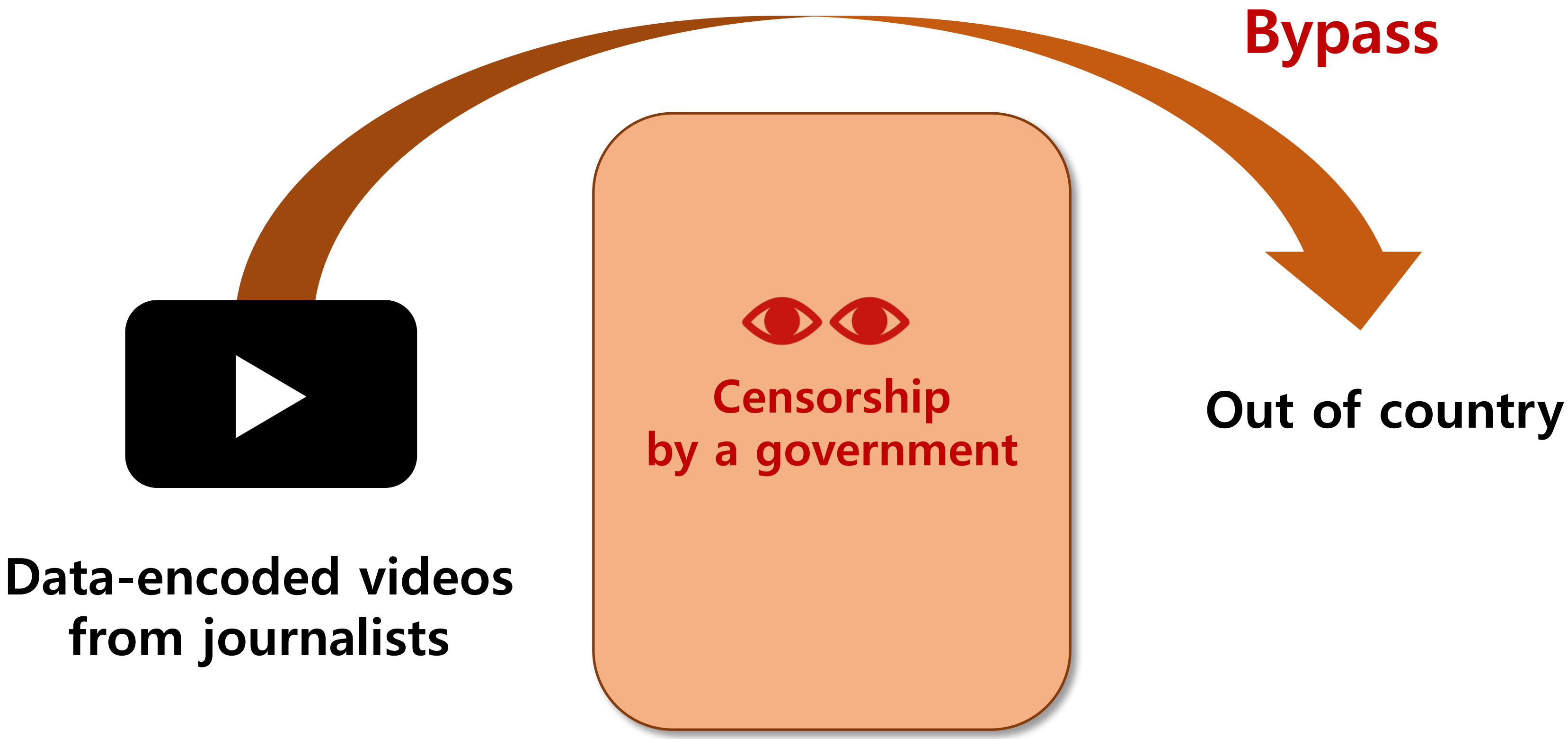
- It showed that some state-of-the-art multimedia protocol tunneling tools are flawed.
- It figured out which network features are important to detect covert channels.
- It showed that the labeled dataset is required for successful detection of covert channels

Web Security & Privacy Lab

KAIST

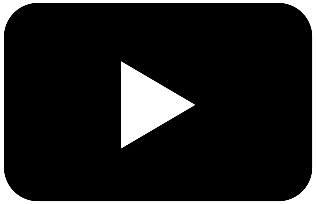
Multimedia Protocol Tunneling

- Encoding data into video channel to circumvent censorship

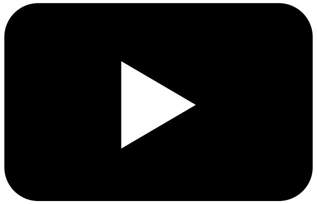


Multimedia Protocol Tunneling

- Such covert channels should be unobservable.
- Can an adversary, i.e., government, distinguish these video streams?



Legitimate video streams



Video streams that carry a covert channel

- However, evaluating the unobservability of systems providing such covert channels has been overlooked.

Web Security & Privacy Lab

KAIST

Overview

(1) Evaluations on state-of-the-art systems using existing similarity-based classifiers.

(2) Evaluations on state-of-the-art systems using other ML techniques including decision trees and their variants.

(3) Evaluations assuming the adversary is deprived of labeled data.

Target Systems

- Three state-of-the-art systems
  - Facet, CovertCast, and DeltaShaper
- Encodes data into video streams.
- Their code is publicly available.

Target System: Facet

- Allows clients to watch desired video by replacing audio/video of Skype videocalls.
- Overlays the desired video in a fraction of each frame, and fills the reaming frame with a video of a typical videocall.
  - Decreasing fraction ratio (s) means that increasing resistance against traffic analysis.
- Collected 1000 YouTube Top Liked playlist (for covert).
- Collected 1000 legitimate recorded live chat videos (Skype, for legitimate).
- **Parameter s:**
  - 50%, 25%, 12.5%

Target System: CovertCast

- Modulates web content by encoding it into colored matrix images and streams it via stream services like YouTube.
- Clients demodulate the images given through stream and get the web contents.
- Crawled 200 live-streams from YouTube.
- Generated 200 CovertCast live-streams.

Target System: DeltaShaper

- Facet + CovertCast
- Encodes data into images and transmits it.
- Encoded data (colored matrix) is overlayed in a fraction of the call screen on top of a typical chat video.
- Emulated 300 legitimate bi-directional Skype videocall.
- **Parameter** <payload frame area, cell size, number of bits, framerate>:
  - <320 X 240, 8 X 8, 6, 1>, <160 X 120, 4 X 4, 6, 1>

Web Security & Privacy Lab

KAIST

Adversary Model

- State-level adversary will attempt to detect the covert traffic.
- Providers of encrypted multimedia apps are not assumed to collude with the adversary.
  - ex: YouTube service provider will not give the raw multimedia content of arbitrary video.
- Adversary cannot control end-user’s computer.
- Domestic ISPs will cooperate with adversary so that the adversary can monitor the traffic.

Similarity-based classifiers

- Measures the similarity/dissimilarity between the distribution of legitimate video streams and video streams that carry a covert channel.

(1) Pearson's chi-squared test(χ^2)

(2) Kullback-Leibler Divergence (KL)

(3) Earth Mover's Distance (EMD)

Pearson's Chi-squared Test (χ^2)

- Is two variables differ significantly?
  - By comparing the observed & expected frequencies.
- Metric used in evaluating **Facet**.
- Used bi-gram distribution of packet lengths.
  - some extreme bi-grams are discarded.

- Compute two models: Legitimate, Covert
  - For a given distribution T, compute the minimum distance between (T, Legitimate) and (T, Covert)
  - Pick one with the minimum distance. (Naïve version)

Web Security & Privacy Lab

KAIST

Kullback-Leibler Divergence (KL)

- Measuring relative entropy between two targets by computing the information lost when trying to approximate one distribution with the other.
- Two target distributions
  - YouTube videos carrying modulated data.
  - YouTube videos which are legitimate.
- Compares the quantized frequency distribution of packet lengths.
- A metric used for building a classifier for CovertCast.

Earth Movers' Distance (EMD)

- Measures the dissimilarity between two distributions, where the distance between single features can be defined in a distance matrix.
- The dissimilarity represents the necessary amount of work to convert one into another.
- A metric used for building classifier for DeltaShaper.
- Compute two groups: Legitimate, Covert
  - For a given distribution T, compute a EMD distance pair for each (member_Legitimate, T) and (member_Covert, T)
  - Pick one with the minimum average distance.

Results & Findings

Multimedia Protocol Tunneling System	χ^2 Classifier ACC	χ^2 Classifier TPR	χ^2 Classifier TNR	KL Classifier ACC	KL Classifier TPR	KL Classifier TNR	EMD Classifier ACC	EMD Classifier TPR	EMD Classifier TNR
Facet (s=50%)	0.743	0.797	0.689	0.575	0.675	0.476	0.575	0.578	0.572
Facet (s=25%)	0.713	0.795	0.630	0.558	0.615	0.500	0.535	0.827	0.242
Facet (s=12.5%)	0.772	0.793	0.750	0.551	0.596	0.506	0.530	0.793	0.267
DeltaShaper $\langle 320 \times 240, 8 \times 8, 6, 1 \rangle$	0.690	0.716	0.663	0.546	0.628	0.464	0.567	0.500	0.633
DeltaShaper $\langle 160 \times 120, 4 \times 4, 6, 1 \rangle$	0.540	0.437	0.650	0.515	0.531	0.500	0.528	0.223	0.833
CovertCast	0.990	1.000	0.980	0.923	0.999	0.846	0.830	0.965	0.695

Table 1: Accuracy, true positive, and true negative rates when detecting covert channels on different multimedia protocol tunneling systems. For the EMD classifier, the threshold value was chosen to be the one providing the highest accuracy, irrespective of the trade-off between the true positive and true negative rates of the classifier.

Unobservability guaranteed

Web Security & Privacy Lab

KAIST

Results & Findings

Multimedia Protocol Tunneling System	χ^2 Classifier ACC	χ^2 Classifier TPR	χ^2 Classifier TNR	KL Classifier ACC	KL Classifier TPR	KL Classifier TNR	EMD Classifier ACC	EMD Classifier TPR	EMD Classifier TNR
Facet (s=50%)	0.743	0.797	0.689	0.575	0.675	0.476	0.575	0.578	0.572
Facet (s=25%)	0.713	0.795	0.630	0.558	0.615	0.500	0.535	0.827	0.242
Facet (s=12.5%)	0.772	0.793	0.750	0.551	0.596	0.506	0.530	0.793	0.267
DeltaShaper $\langle 320 \times 240, 8 \times 8, 6, 1 \rangle$	0.690	0.716	0.663	0.546	0.628	0.464	0.567	0.500	0.633
DeltaShaper $\langle 160 \times 120, 4 \times 4, 6, 1 \rangle$	0.540	0.437	0.650	0.515	0.531	0.500	0.528	0.223	0.833
CovertCast	0.990	1.000	0.980	0.923	0.999	0.846	0.830	0.965	0.695

Table 1: Accuracy, true positive, and true negative rates when detecting covert channels on different multimedia protocol tunneling systems. For the EMD classifier, the threshold value was chosen to be the one providing the highest accuracy, irrespective of the trade-off between the true positive and true negative rates of the classifier.

χ^2 outperforms other classifiers

Results & Findings

Multimedia Protocol Tunneling System	χ^2 Classifier ACC	χ^2 Classifier TPR	χ^2 Classifier TNR	KL Classifier ACC	KL Classifier TPR	KL Classifier TNR	EMD Classifier ACC	EMD Classifier TPR	EMD Classifier TNR
Facet (s=50%)	0.743	0.797	0.689	0.575	0.675	0.476	0.575	0.578	0.572
Facet (s=25%)	0.713	0.795	0.630	0.558	0.615	0.500	0.535	0.827	0.242
Facet (s=12.5%)	0.772	0.793	0.750	0.551	0.596	0.506	0.530	0.793	0.267
DeltaShaper $\langle 320 \times 240, 8 \times 8, 6, 1 \rangle$	0.690	0.716	0.663	0.546	0.628	0.464	0.567	0.500	0.633
DeltaShaper $\langle 160 \times 120, 4 \times 4, 6, 1 \rangle$	0.540	0.437	0.650	0.515	0.531	0.500	0.528	0.223	0.833
CovertCast	0.990	1.000	0.980	0.923	0.999	0.846	0.830	0.965	0.695

Table 1: Accuracy, true positive, and true negative rates when detecting covert channels on different multimedia protocol tunneling systems. For the EMD classifier, the threshold value was chosen to be the one providing the highest accuracy, irrespective of the trade-off between the true positive and true negative rates of the classifier.

CovertCast failed to guarantee unobservability

Web Security & Privacy Lab

KAIST

Results & Findings

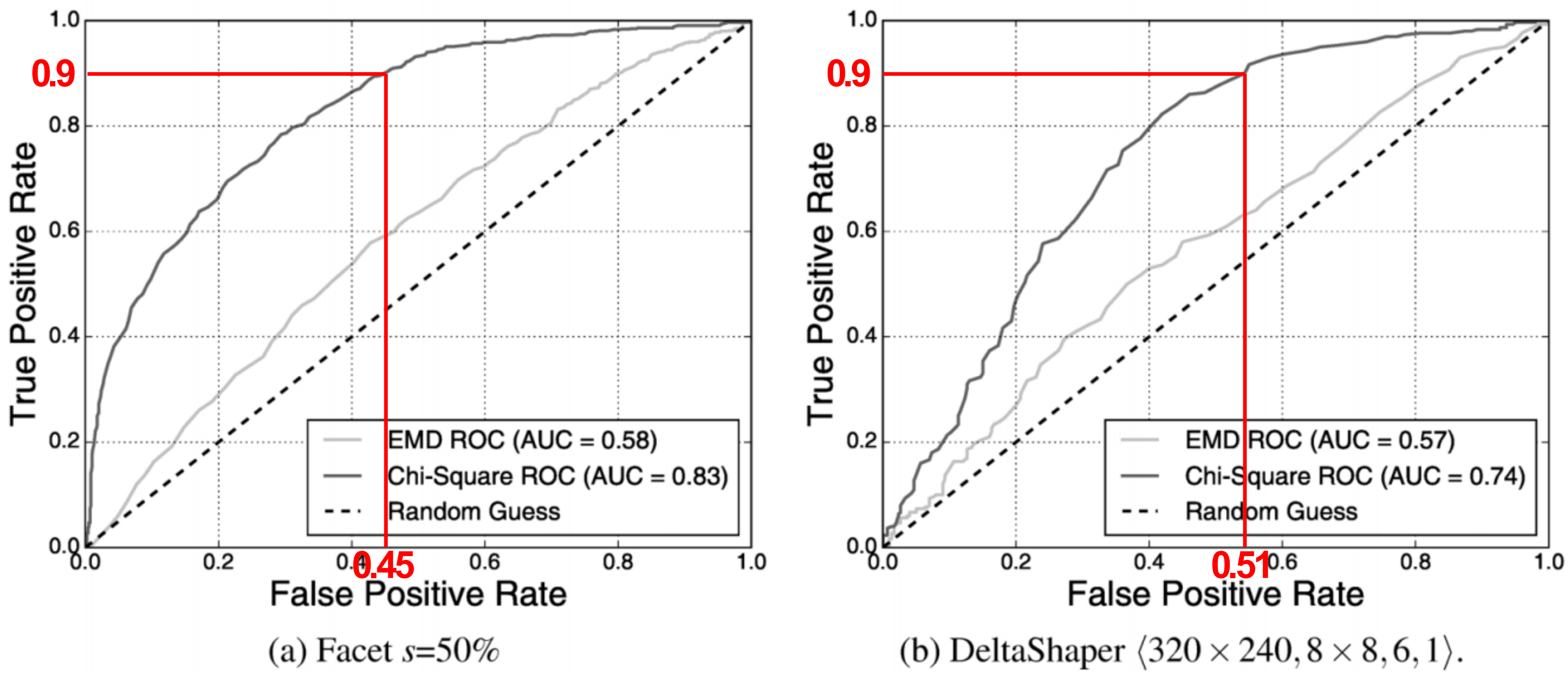


Figure 1: ROC curve for the χ^2 and EMD classifiers when identifying Facet and DeltaShaper traffic.

χ^2 produces large false positive rates (Facet, DeltaShpaer)

Web Security & Privacy Lab

KAIST

Decision-Tree Based Classification

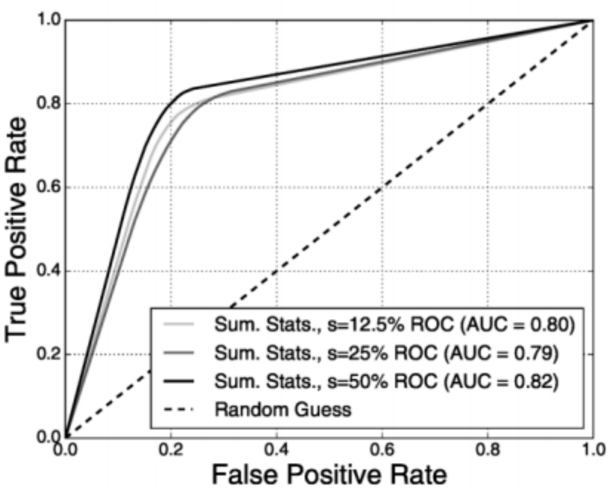
- Decision Trees (DT)
  - Each tree node is either a decision or leaf node.
  - Decision node split the current branch by an attribute.
- Random Forests (RF)
  - An ensemble learning method.
  - Selects result from a majority vote of multiple DTs.
- eXtreme Gradient Boosting (XGBoost)
  - Creates a new tree which optimizes the predictions.
  - Has a benefit to control overfitting.

Feature Sets

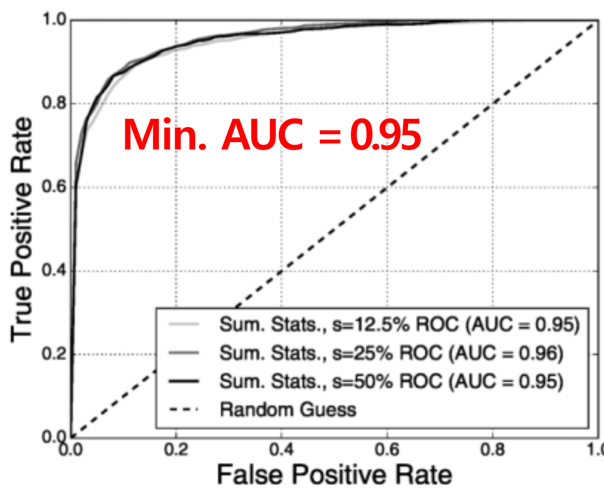
- Summary Statistics (ST)
  - A timeseries of packet lengths.
  - A timeseries of packet inter-arrival times.
  - Burst behavior.
- Quantized Packet Lengths (PL)
  - Quantized frequency distribution of packet lengths.

Results & Findings - Facet

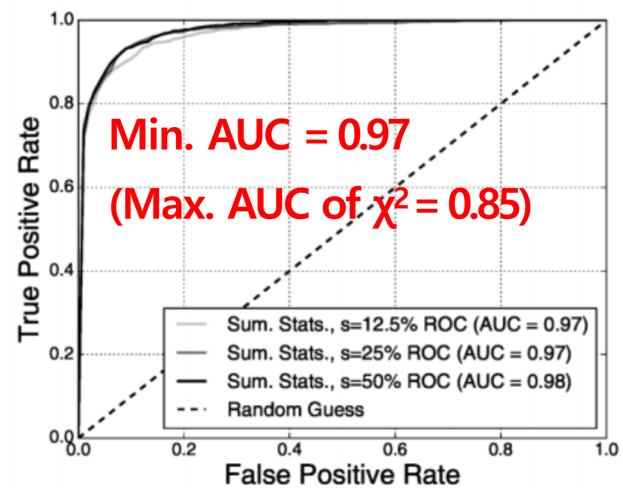
- ROC curves w/ Feature Set 1: Summary Statistics



(a) Decision Tree – Facet.



(b) Random Forest – Facet.



(c) XGBoost – Facet.

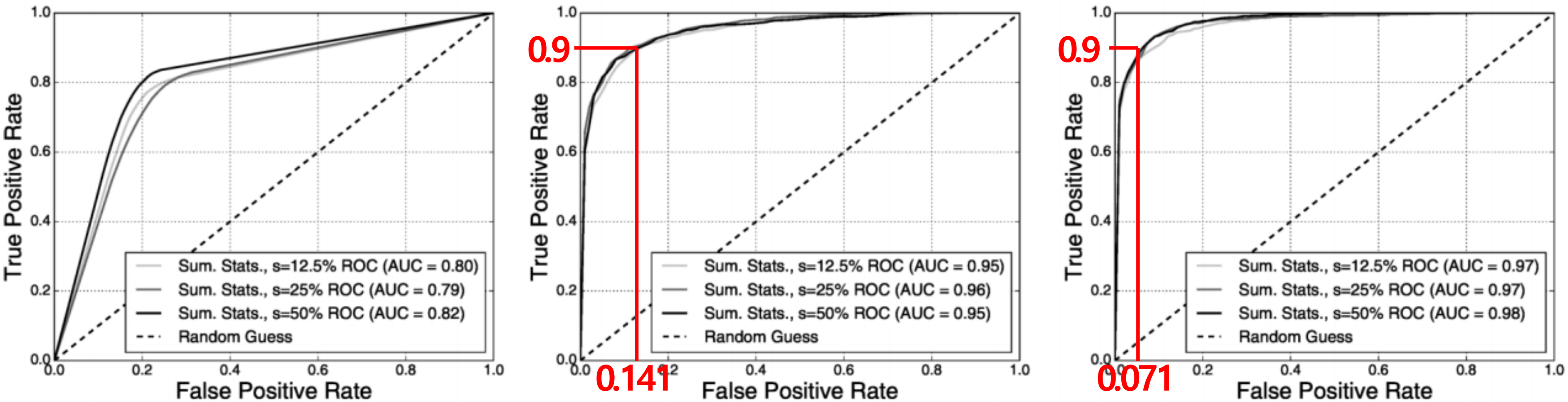
Random Forest/XGBoost breaks unobservability.

Web Security & Privacy Lab

KAIST

Results & Findings - Facet

- ROC curves w/ Feature Set 1: Summary Statistics



(a) Decision Tree – Facet. (b) Random Forest – Facet. (c) XGBoost – Facet.

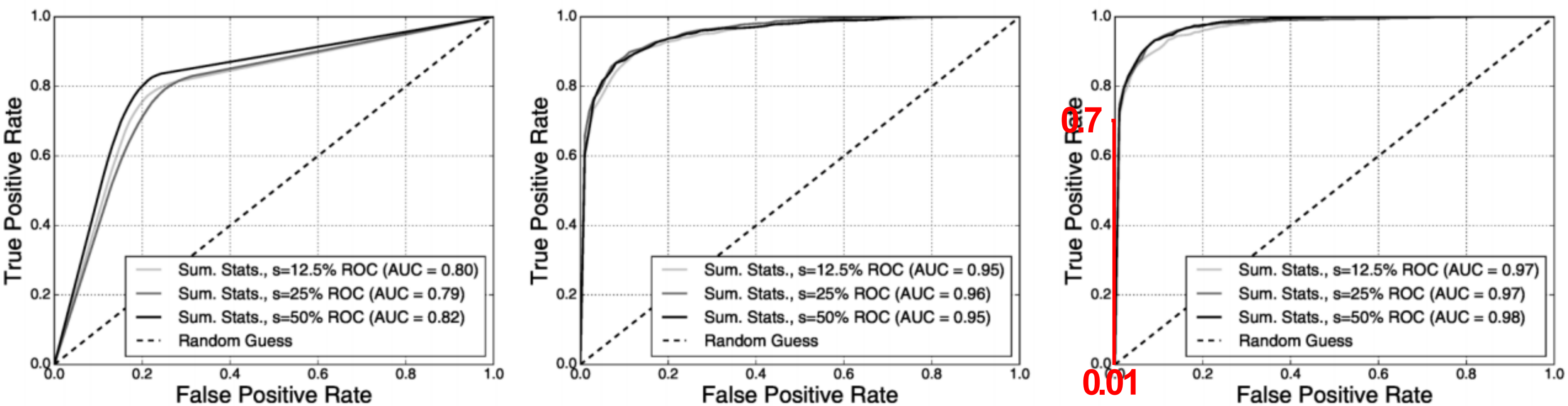
Low false positive rates

Web Security & Privacy Lab

KAIST

Results & Findings - Facet

- ROC curves w/ Feature Set 1: Summary Statistics

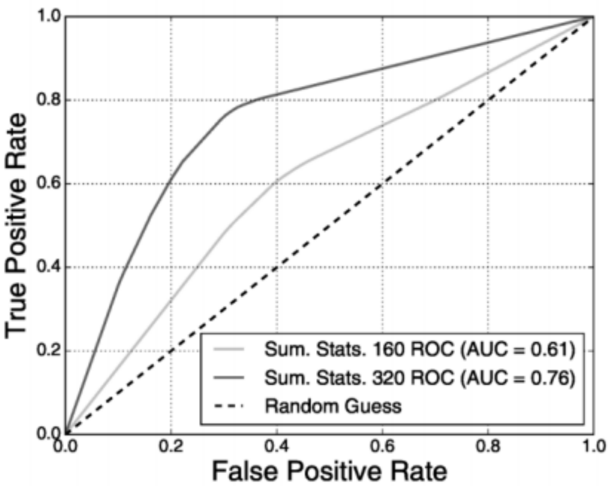


(a) Decision Tree – Facet. (b) Random Forest – Facet. (c) XGBoost – Facet.

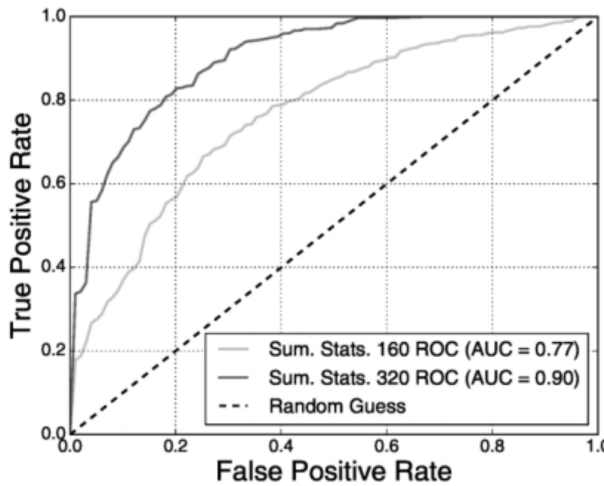
Low false positive rates

Results & Findings - DeltaShaper

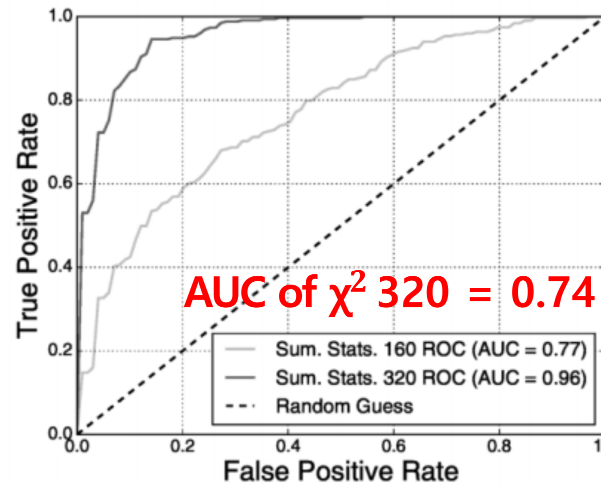
- ROC curves w/ Feature Set 1: Summary Statistics



(d) Decision Tree – DeltaShaper.



(e) Random Forest – DeltaShaper.



(f) XGBoost – DeltaShaper.

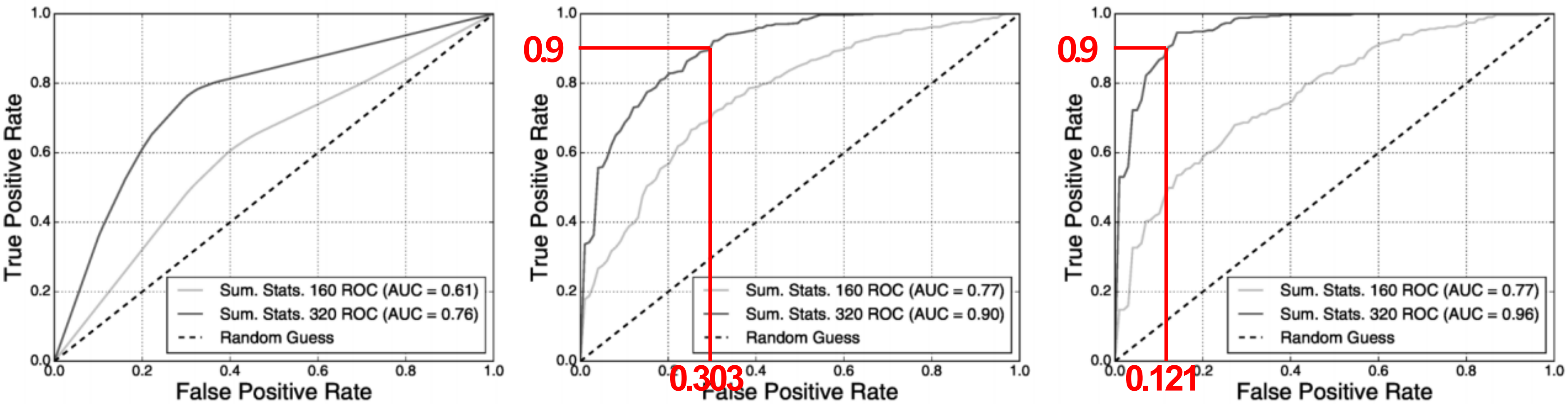
Random Forest/XGBoost breaks unobservability.

Web Security & Privacy Lab

KAIST

Results & Findings - DeltaShaper

- ROC curves w/ Feature Set 1: Summary Statistics

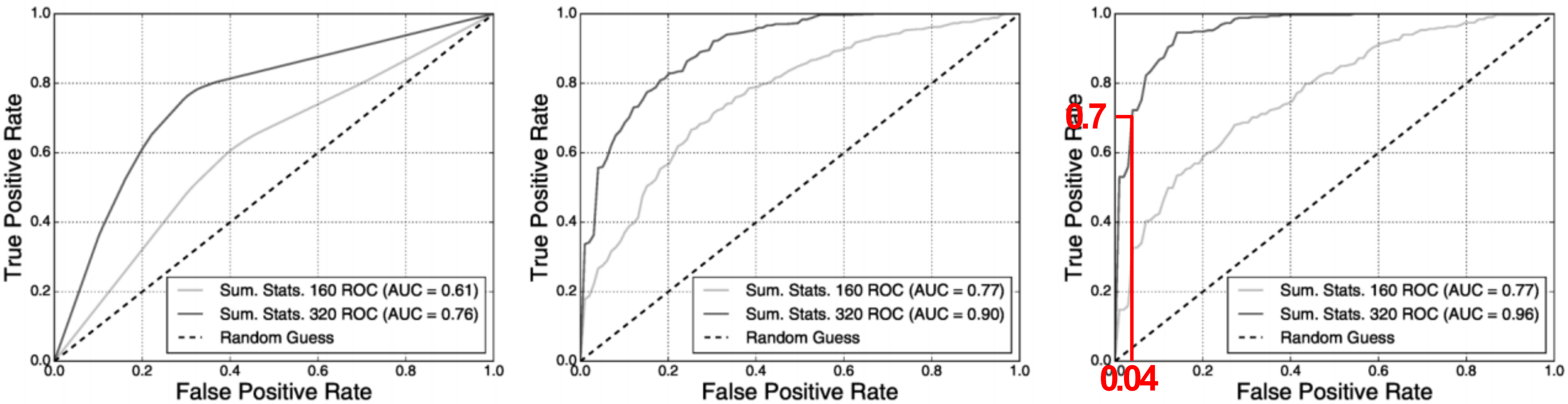


(d) Decision Tree – DeltaShaper. (e) Random Forest – DeltaShaper. (f) XGBoost – DeltaShaper.

Low false positive rates

Results & Findings - DeltaShaper

- ROC curves w/ Feature Set 1: Summary Statistics

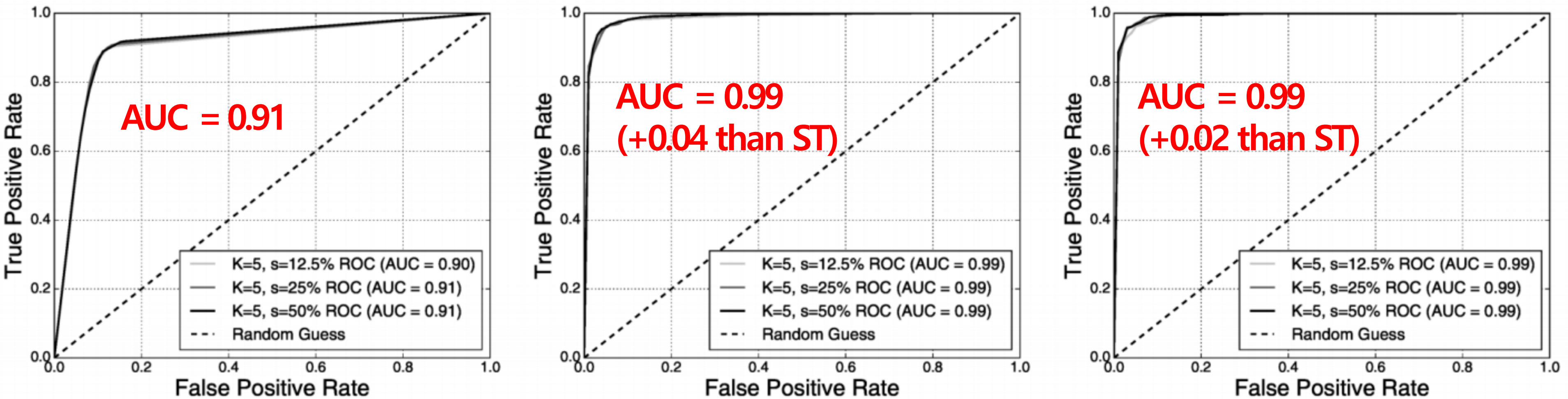


(d) Decision Tree – DeltaShaper. (e) Random Forest – DeltaShaper. (f) XGBoost – DeltaShaper.

Low false positive rates

Results & Findings - Facet

- ROC curves w/ Feature Set 2: Quantized PLs

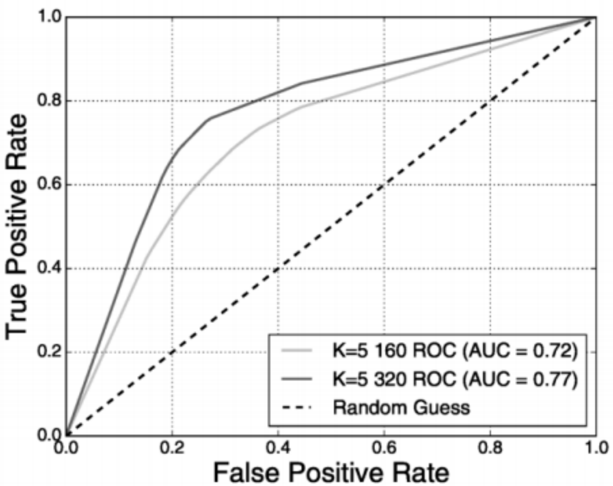


(a) Decision Tree – Facet. (b) Random Forest – Facet. (c) XGBoost – Facet.

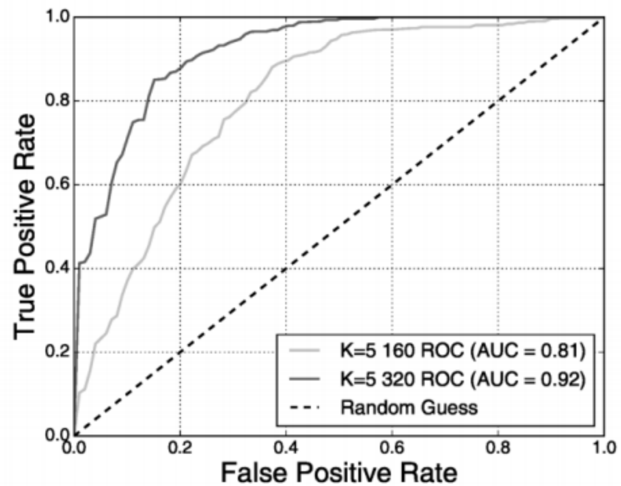
Quantized PLs outperform the use of summary statistics.

Results & Findings - DeltaShaper

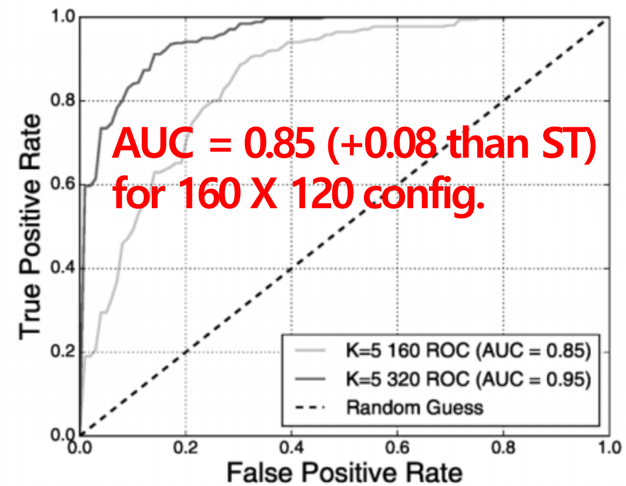
- ROC curves w/ Feature Set 2: Quantized PLs



(d) Decision Tree – DeltaShaper.



(e) Random Forest – DeltaShaper.

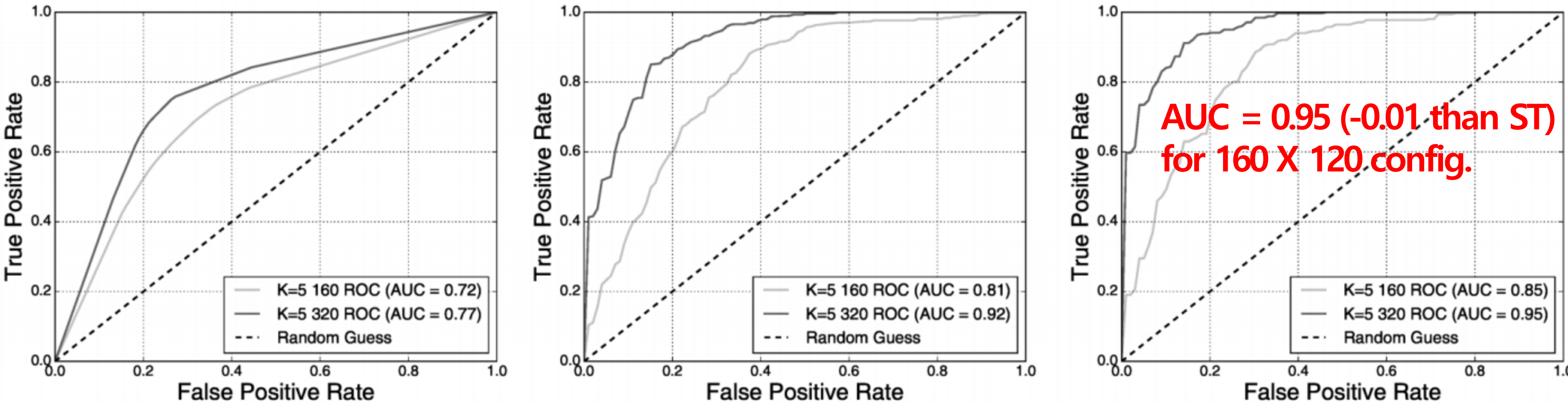


(f) XGBoost – DeltaShaper.

Quantized PLs outperform the use of summary statistics.

Results & Findings - DeltaShaper

- ROC curves w/ Feature Set 2: Quantized PLs



(d) Decision Tree – DeltaShaper. (e) Random Forest – DeltaShaper. (f) XGBoost – DeltaShaper.

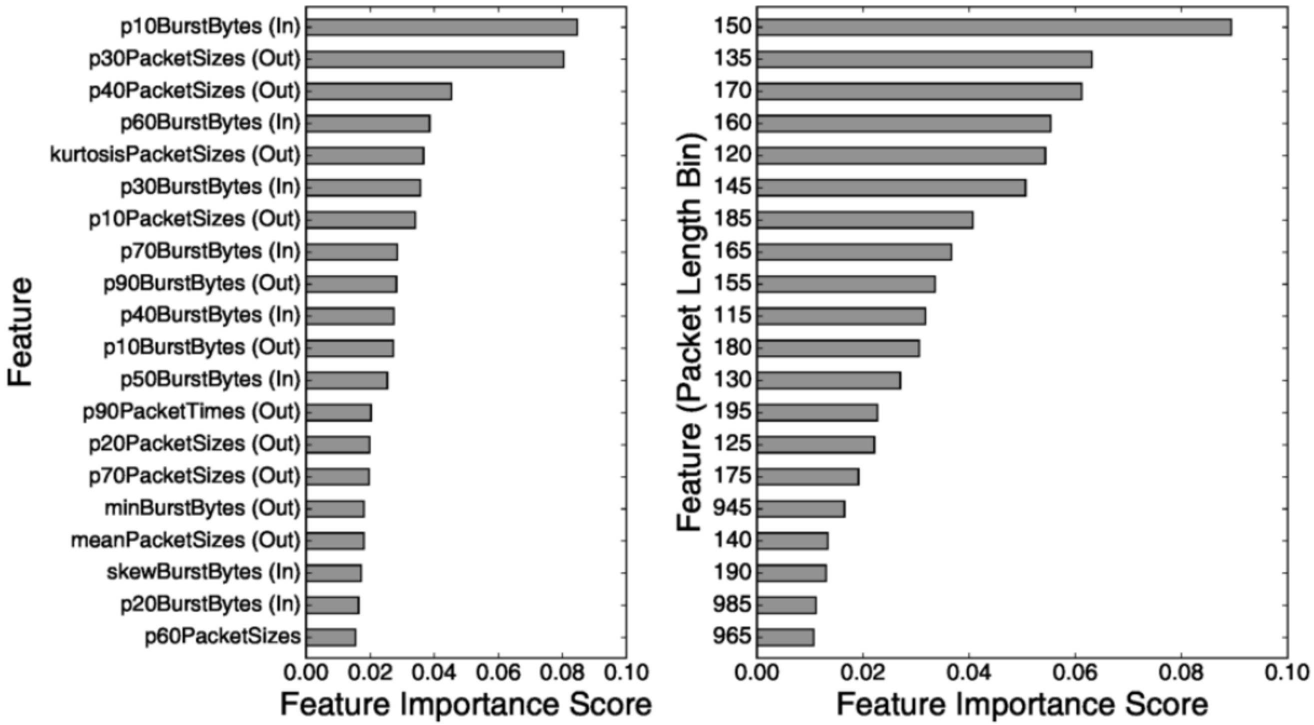
Quantized PLs underperform on XGBoost with 320 X 240 config.

Web Security & Privacy Lab

KAIST

Feature Importance - Facet

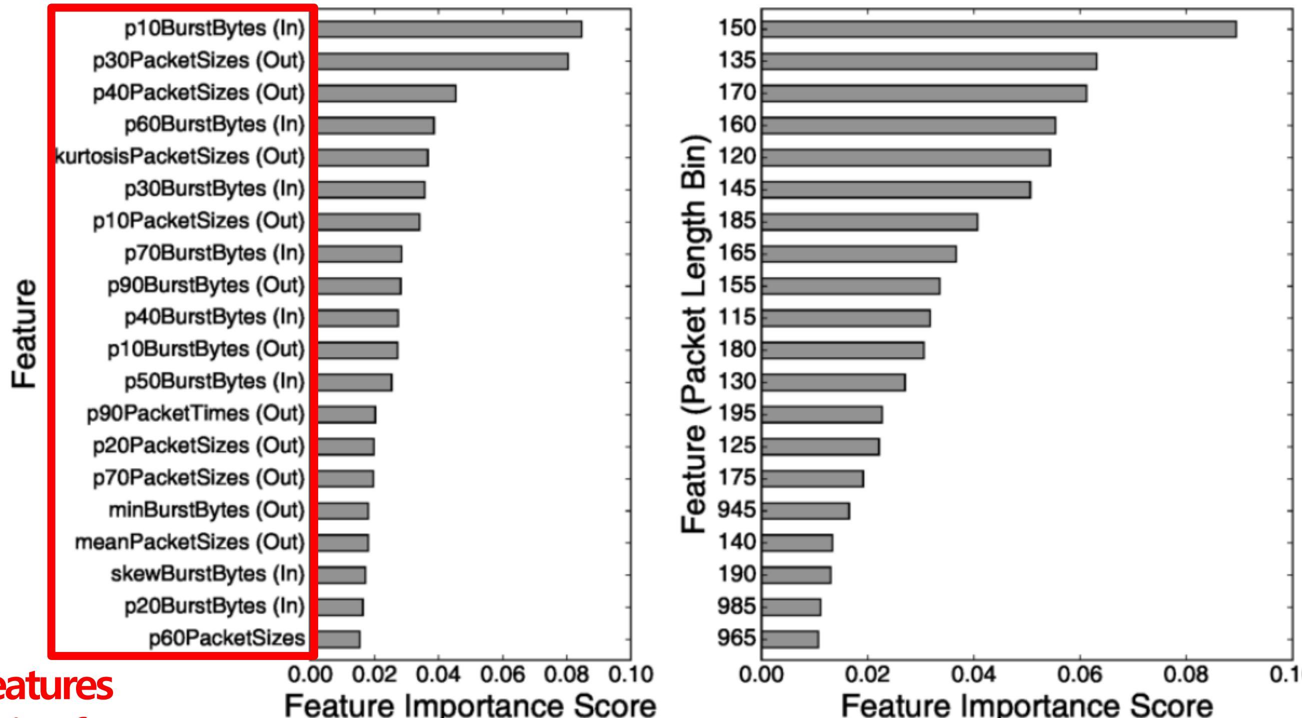
- **TOP 20 Features for ST/PL by XGBoost algorithm, s=50%.**



(a) ST - Facet. (b) PL - Facet.

Feature Importance - Facet

- **Facet is more vulnerable to analysis based on PL & Burst.**



8 packet size features
11 burst behavior features
(Total 19 out of TOP 20)

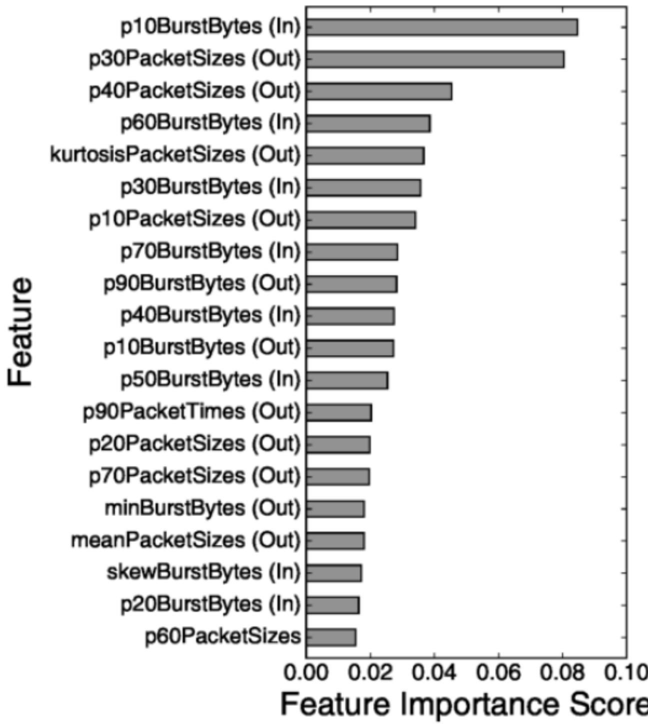
(a) ST - Facet. (b) PL - Facet.

Web Security & Privacy Lab

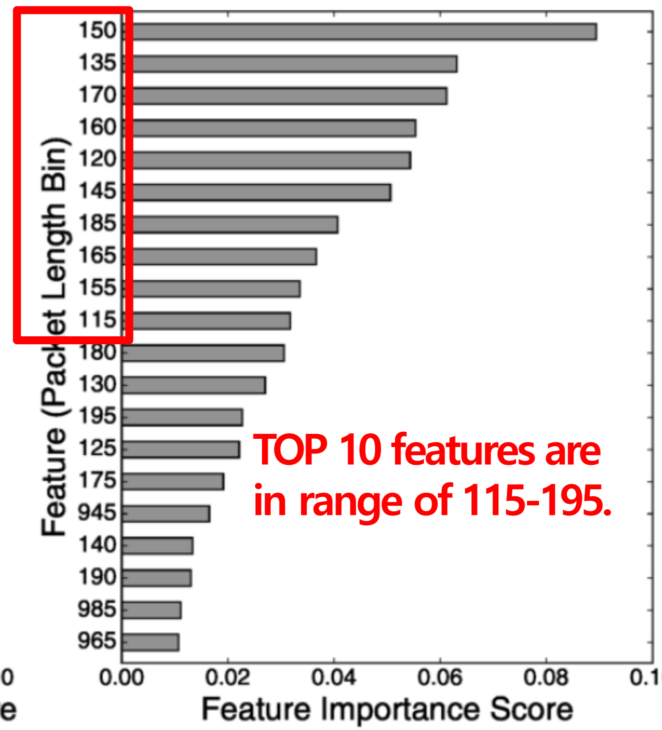
KAIST

Feature Importance - Facet

- **Facet covert channels are spotted by PL b/w 115-195 bytes.**



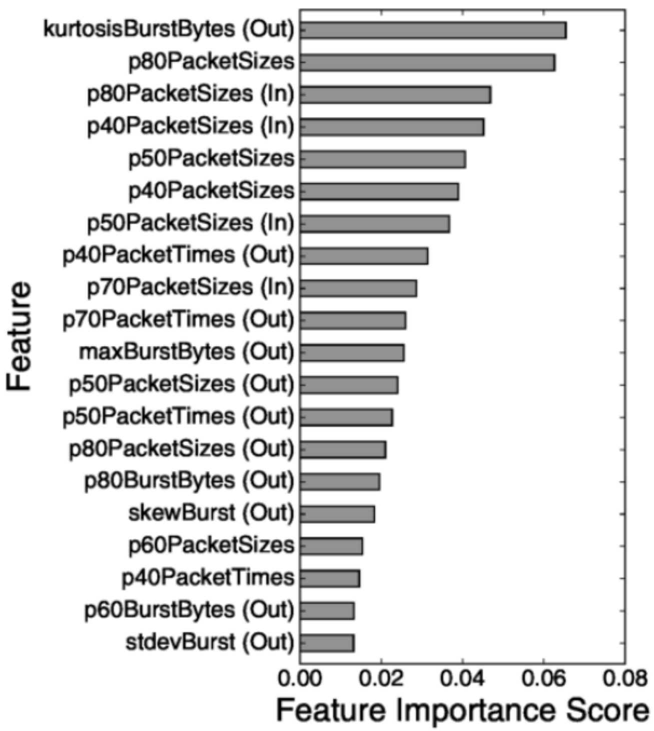
(a) ST - Facet.



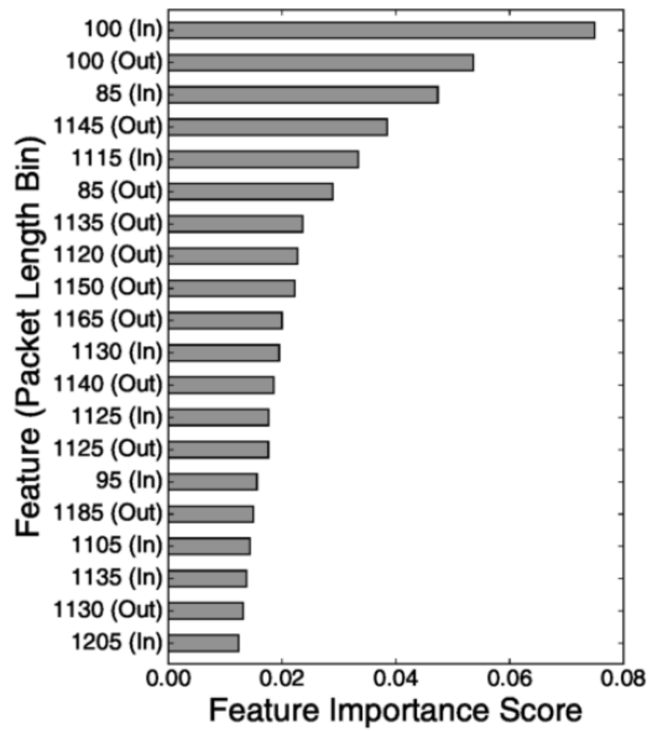
(b) PL - Facet.

Feature Importance - DeltaShpaer

- TOP 20 Features by XGBoost, <320 X 240, 8 X 8, 6, 1>



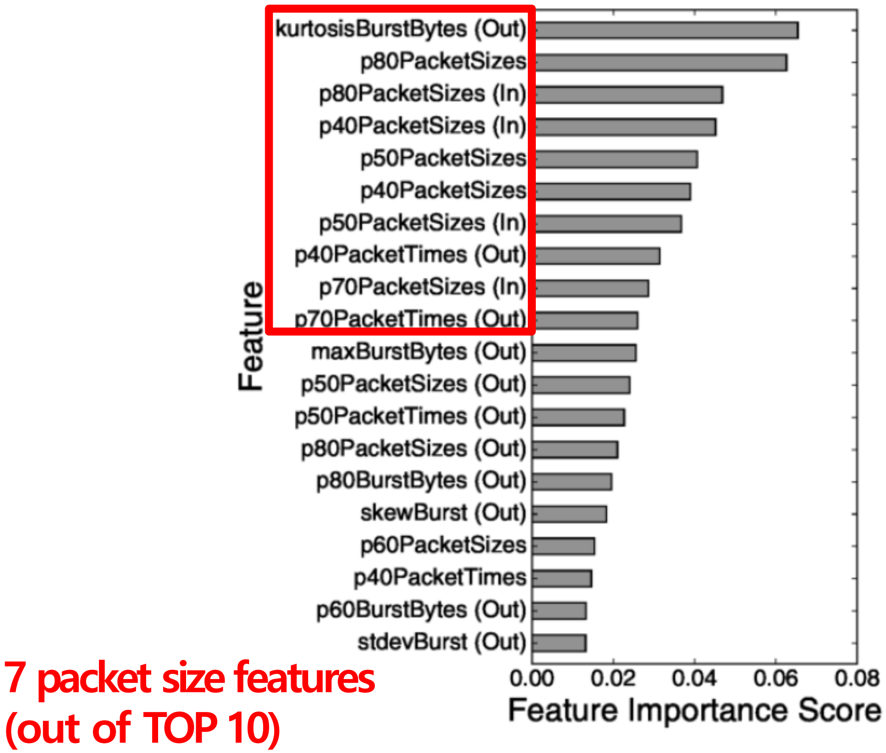
(c) ST - DeltaShaper.



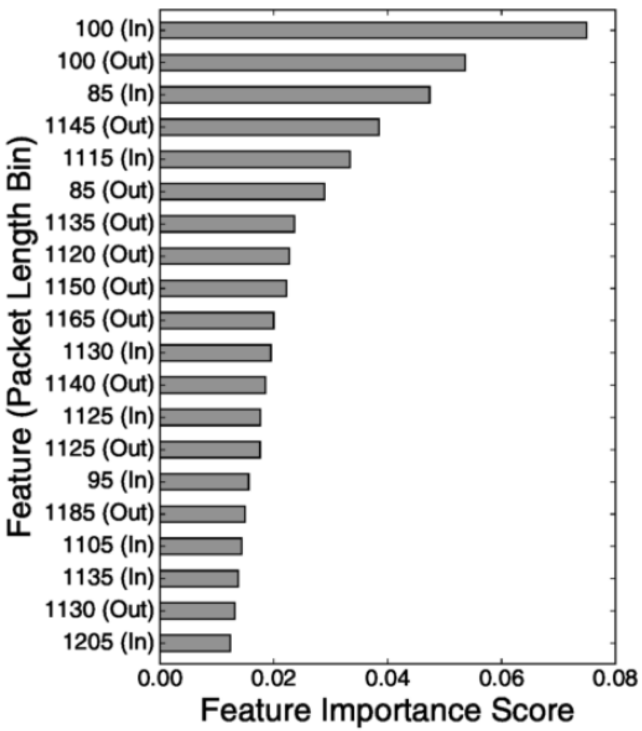
(d) PL - DeltaShaper.

Feature Importance - DeltaShpaer

- DeltaShpaer is more vulnerable to analysis based on PL.



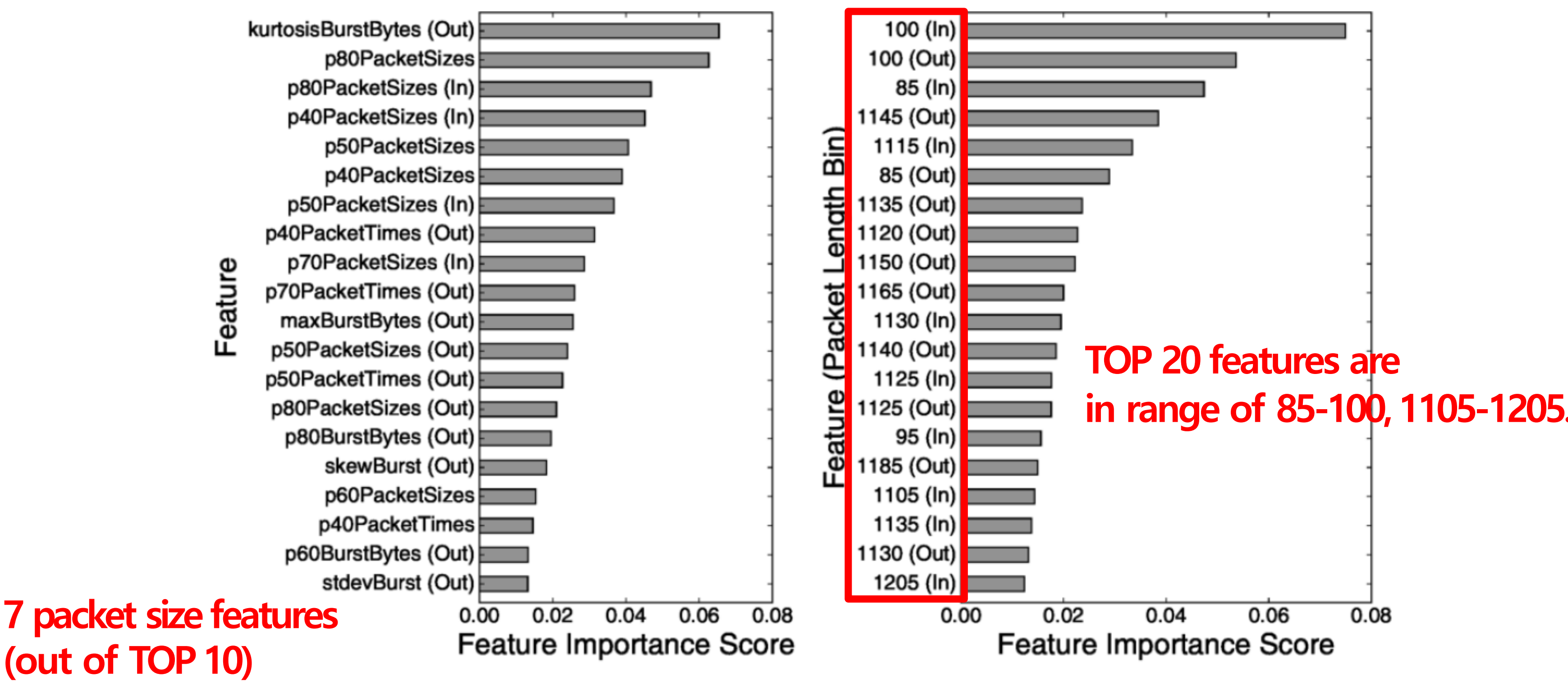
(c) ST - DeltaShaper.



(d) PL - DeltaShaper.

Feature Importance - DeltaShpaer

- DeltaShpaer: 85-100, 1105-1205 bytes are important PL.



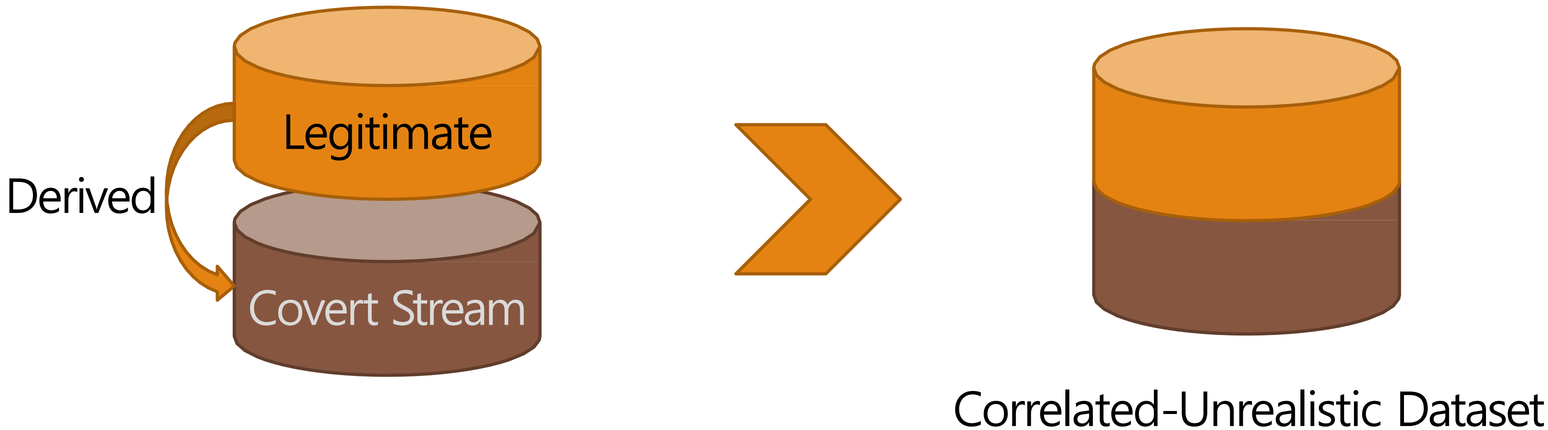
(c) ST - DeltaShaper. (d) PL - DeltaShaper.

Web Security & Privacy Lab

KAIST

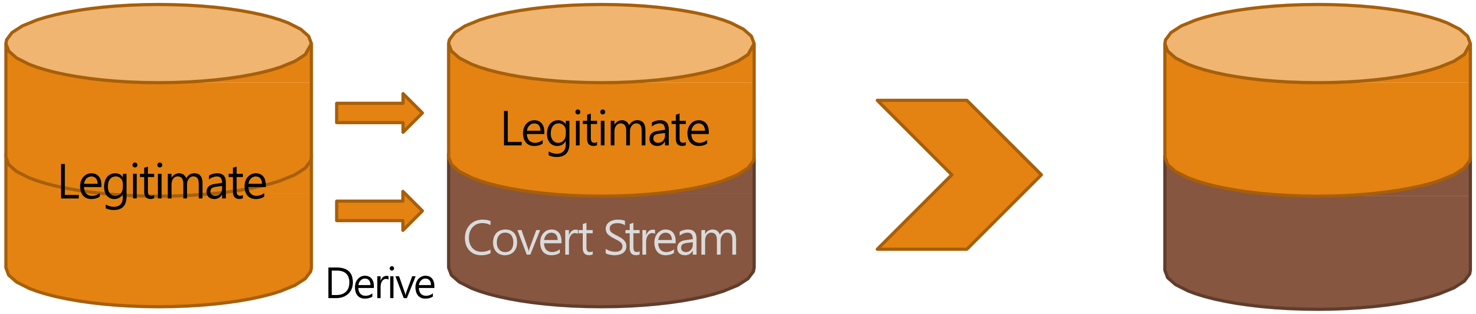
Alternative Dataset Evaluation

- Current dataset is not realistic
  - **Currently**: Covert streams are produced using legitimate videos
  - **Problem 1**: Introduce correlation among classes
  - **Problem 2**: Positive class : Negative class = 1 : 1 Unrealistic!



Alternative Dataset Evaluation

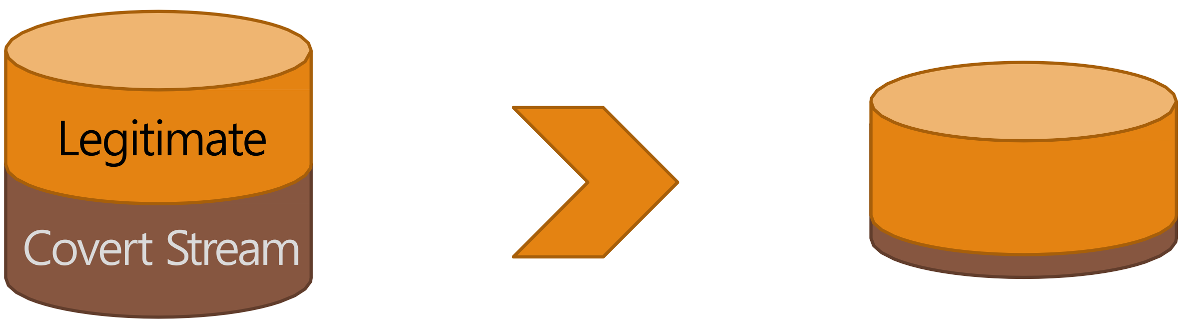
- Experiment 1
  - **Solution of Problem 1**
    - Produce covert streams using the half of legitimate videos
  - 10-fold cross-validation, 10 times repeated to prevent the overfitting



Not Correlated Dataset

Alternative Dataset Evaluation

- Experiment 2
  - **Solution**: Keep the low pos-to-neg ratio during testing
  - Training set : Test set = 7 : 3
  - Pos-to-neg ratio: 1 : 1 (Training set), 1 : 100 (Testing)



1:1 Ratio Unrealistic Dataset

1:100 Ratio Realistic Dataset

Alternative Dataset Evaluation

- Experiment 1
  - Used XGBoost
  - AUC=0.95 for DeltaShpaer <320 X 240, 8 X 8, 6, 1>
    - 0.01 less than the original result
  - AUC=0.99 for Facet s=50%
- Experiment 2
  - Identify 90% of Facet s=50% traffic with FPR of 2%
  - Identify 90% of DeltaShpaer <320> traffic with FPR of 18%
    - Only 4% larger than original result
- **Possible Correlation among classes & sampling ratio do not limit the accuracy of this work**

Beyond Supervised Anomaly Detection

- Assumes an adversary who does not have an access to labeled anomalies.

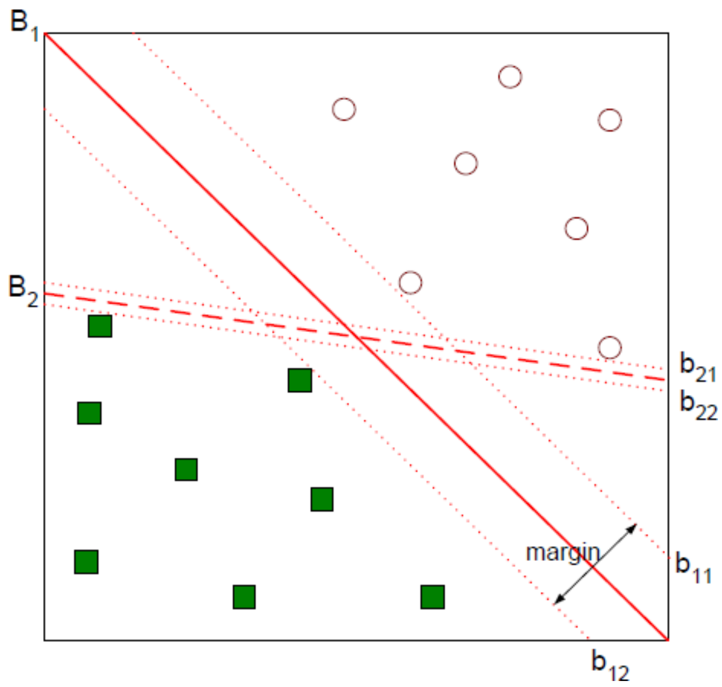
(1) One-class SVM (OCSVM)

(2) Autoencoder

(3) Isolation Forest

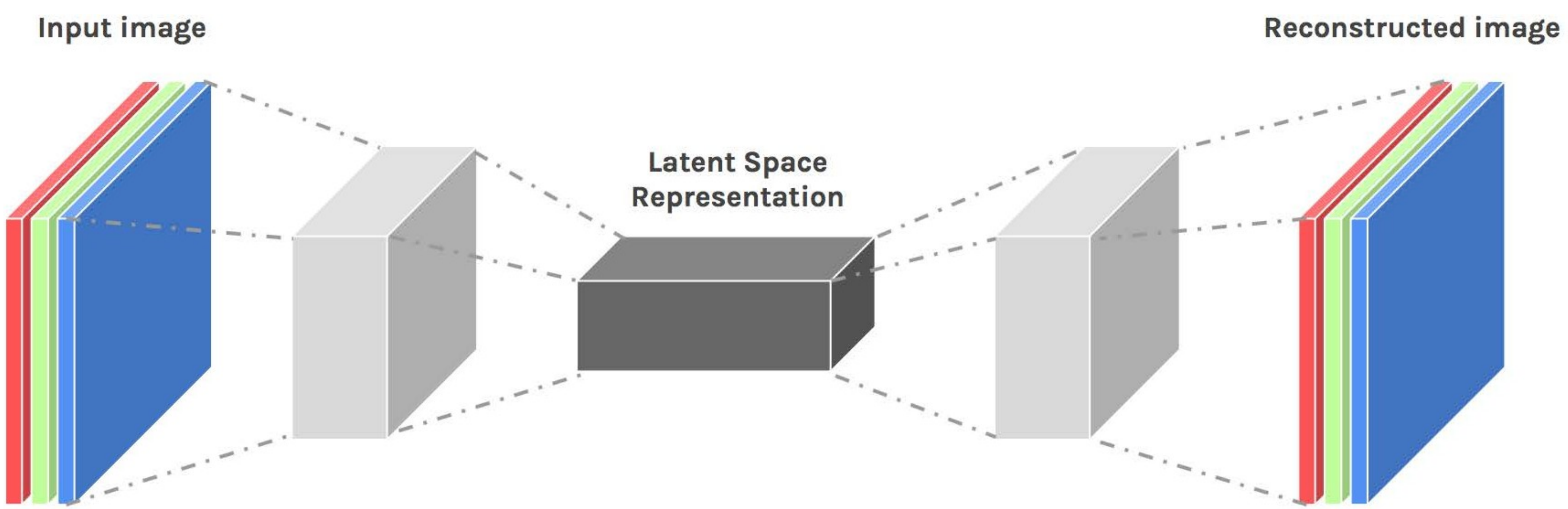
One-class SVM

- Defines a boundary between normal/anomaly.
- Finds maximal margin hyperplane.



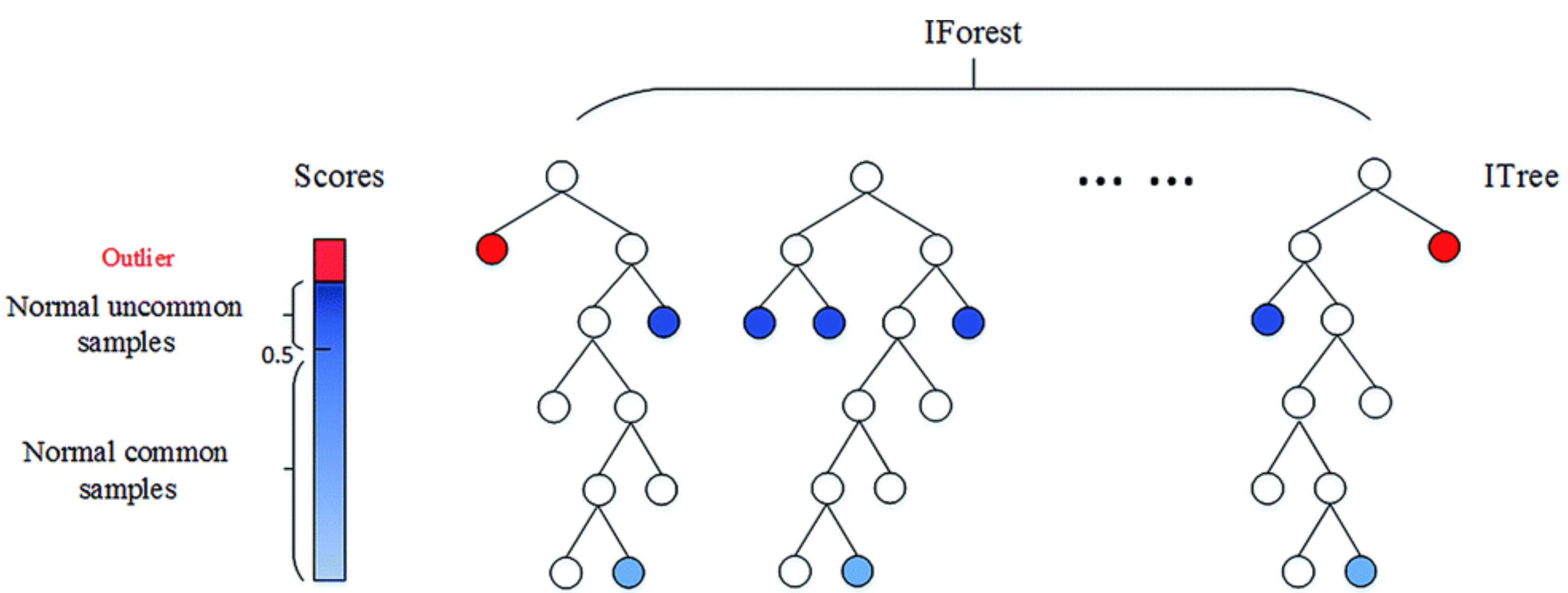
Autoencoder

- Approximates the identity function through a compressed representation of its inputs.
- Anomaly detection using reconstruction error



Isolation Forest

- Detects outliers by isolating anomalous samples
- Selects a split between its min and max values



Findings

Multimedia Protocol Tunneling System	OCSVM		Autoencoder		Isolation Forest	
	Max AUC	Avg AUC	Max AUC	Avg AUC	Max AUC	Avg AUC
Facet (s=50%)	0.631	0.576	0.702	0.638	0.561	0.551
Facet (s=25%)	0.629	0.580	0.700	0.650	0.528	0.519
Facet (s=12.5%)	0.639	0.584	0.706	0.647	0.536	0.520
DeltaShaper $\langle 320 \times 240, 8 \times 8, 6, 1 \rangle$	0.567	0.531	0.662	0.574	0.580	0.557
DeltaShaper $\langle 160 \times 120, 4 \times 4, 6, 1 \rangle$	0.548	0.518	0.576	0.544	0.553	0.532

Table 5: Maximum and average AUC of OCSVM, Autoencoder and Isolation Forest when classifying Facet and DeltaShaper traffic. Search (min, max, step): OCSVM (ν(0.1, 1, +0.1), γ(0.01, 1, +0.01)); Autoencoder (hidden_layers(4,512,*2), compressed_representation(4,512,*2), learning_rate[0.001,0.01], epochs[1000]); Isolation Forest (n_trees(50,200,*2), n_samples(64,512,*2))

OCSVMs cannot identify covert traffics properly.

Findings

Multimedia Protocol Tunneling System	OCSVM		Autoencoder		Isolation Forest	
	Max AUC	Avg AUC	Max AUC	Avg AUC	Max AUC	Avg AUC
Facet (s=50%)	0.631	0.576	0.702	0.638	0.561	0.551
Facet (s=25%)	0.629	0.580	0.700	0.650	0.528	0.519
Facet (s=12.5%)	0.639	0.584	0.706	0.647	0.536	0.520
DeltaShaper $\langle 320 \times 240, 8 \times 8, 6, 1 \rangle$	0.567	0.531	0.662	0.574	0.580	0.557
DeltaShaper $\langle 160 \times 120, 4 \times 4, 6, 1 \rangle$	0.548	0.518	0.576	0.544	0.553	0.532

Table 5: Maximum and average AUC of OCSVM, Autoencoder and Isolation Forest when classifying Facet and DeltaShaper traffic. Search (min, max, step): OCSVM (ν(0.1, 1, +0.1), γ(0.01, 1, +0.01)); Autoencoder (hidden_layers(4,512,*2), compressed_representation(4,512,*2), learning_rate[0.001,0.01], epochs[1000]); Isolation Forest (n_trees(50,200,*2), n_samples(64,512,*2))

Autoencoders have potential to improve.

ex) Use autoencoders with more sophisticated structures

Web Security & Privacy Lab

KAIST

Findings

Multimedia Protocol Tunneling System	OCSVM Max AUC	OCSVM Avg AUC	Autoencoder Max AUC	Autoencoder Avg AUC	Isolation Forest Max AUC	Isolation Forest Avg AUC
Facet (s=50%)	0.631	0.576	0.702	0.638	0.561	0.551
Facet (s=25%)	0.629	0.580	0.700	0.650	0.528	0.519
Facet (s=12.5%)	0.639	0.584	0.706	0.647	0.536	0.520
DeltaShaper $\langle 320 \times 240, 8 \times 8, 6, 1 \rangle$	0.567	0.531	0.662	0.574	0.580	0.557
DeltaShaper $\langle 160 \times 120, 4 \times 4, 6, 1 \rangle$	0.548	0.518	0.576	0.544	0.553	0.532

Table 5: Maximum and average AUC of OCSVM, Autoencoder and Isolation Forest when classifying Facet and DeltaShaper traffic. Search (min, max, step): OCSVM (ν(0.1, 1, +0.1), γ(0.01, 1, +0.01)); Autoencoder (hidden_layers(4,512,*2), compressed_representation(4,512,*2), learning_rate[0.001,0.01], epochs[1000]); Isolation Forest (n_trees(50,200,*2), n_samples(64,512,*2))

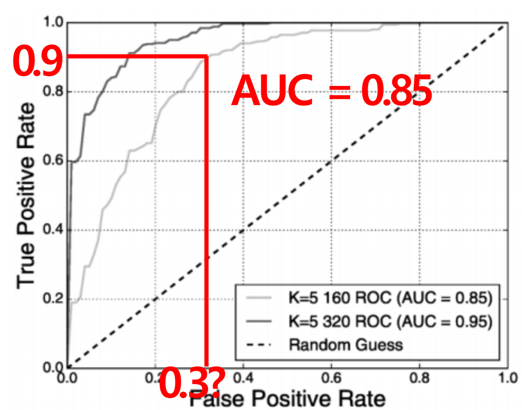
Using Isolation Forest has no advantage for detecting covert traffic.

Web Security & Privacy Lab

KAIST

Discussion

- Unobservability claims of existing multimedia protocol tunneling systems were flawed.
- Supervised ML algorithm can detect covert traffics.
- Does not mean multimedia protocol tunneling is inviable.
  - With some configuration, it is hard for adversaries to detect covert channels with low false positive rate. (ex: DeltaShpaer <160 X 120>)



(f) XGBoost – DeltaShaper.

Web Security & Privacy Lab

KAIST

Discussion

- Adversary cannot collect real world legitimate traffic dataset properly because of the multimedia protocol tunneling tools.
  - How can we know which stream is legitimate in advance?
- Adversary can construct dataset with their own traffic (like in this paper).
  - May fail to capture the underlying distribution in wild.

Questions