Bucks for Buckets (B4B): Active Defenses Against Stealing Encoders

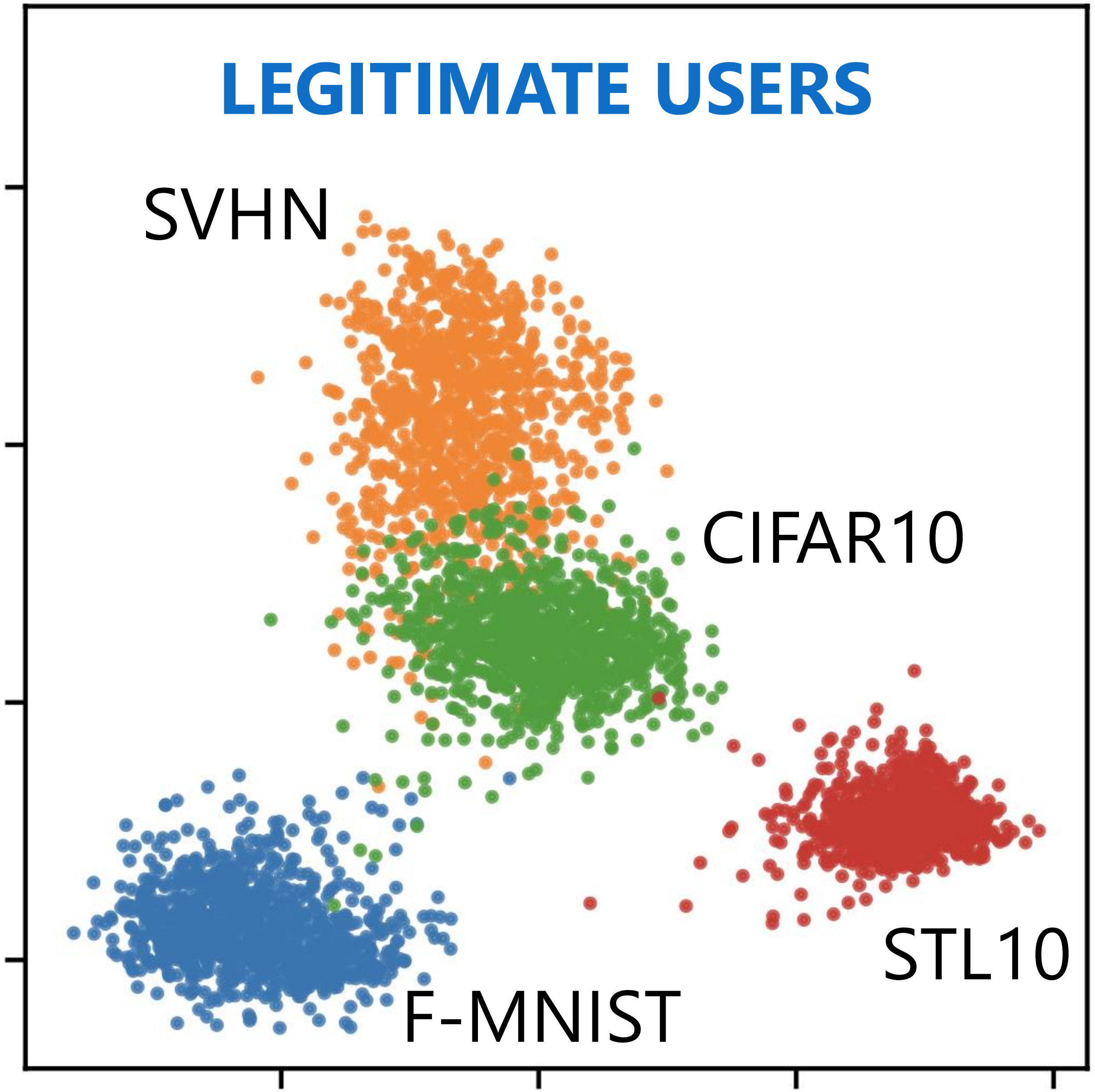
Jan Dubiński, Stanisław Pawlak, Franziska Boenisch, Tomasz Trzciński, Adam Dziedzic

December 15th, 2023

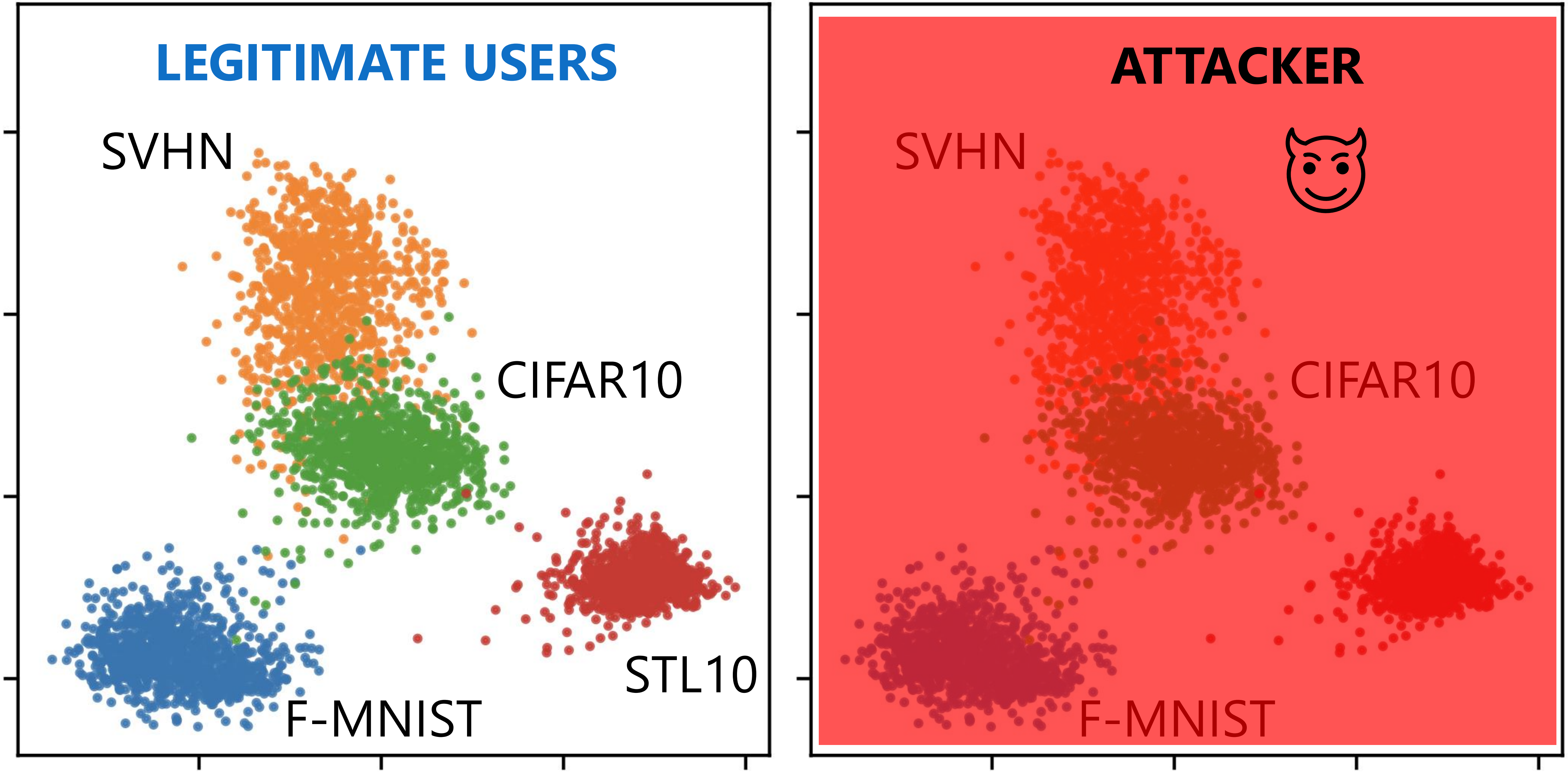
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

SprintML

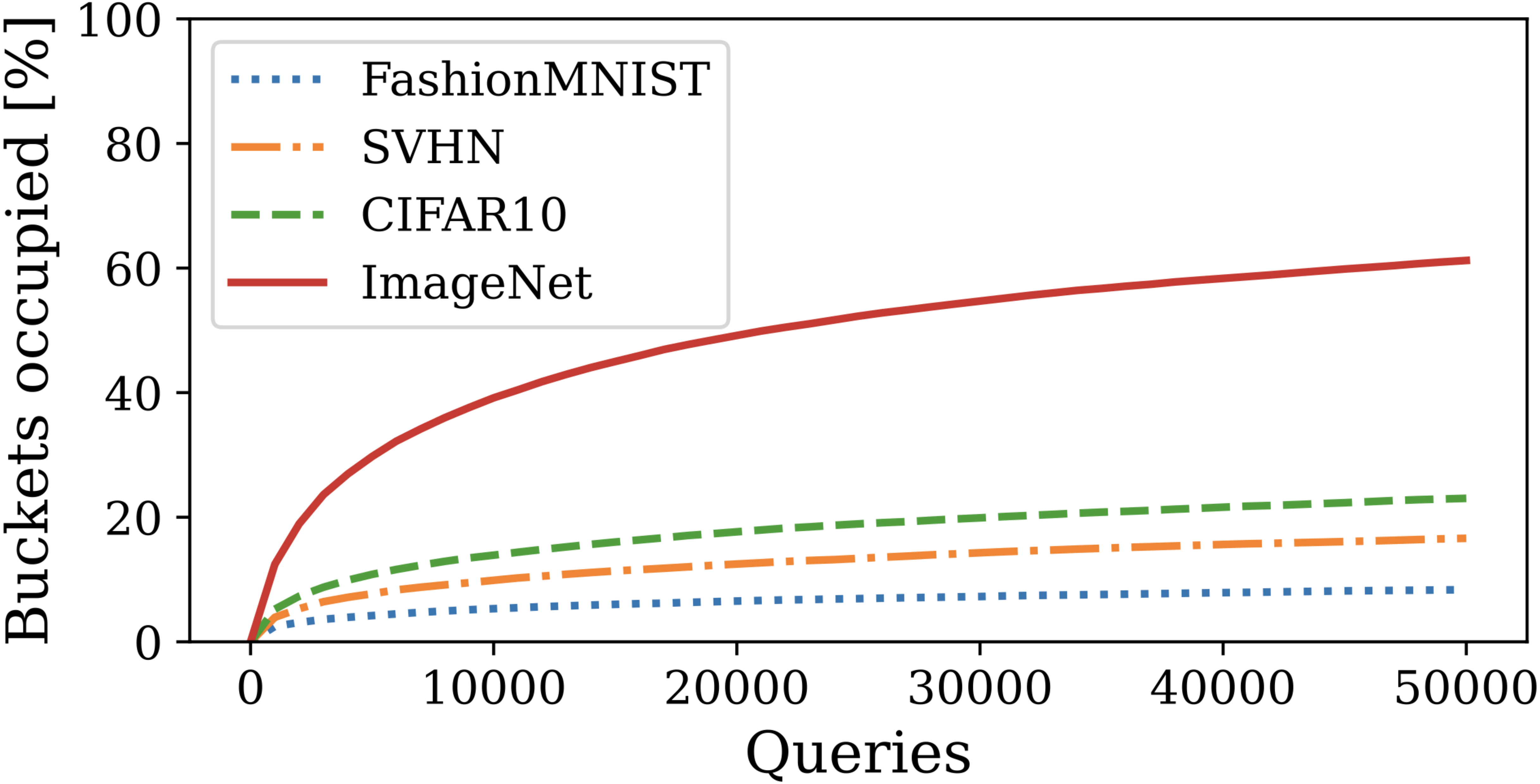
Active Defenses against Stealing SSL Models



Active Defenses against Stealing SSL Models

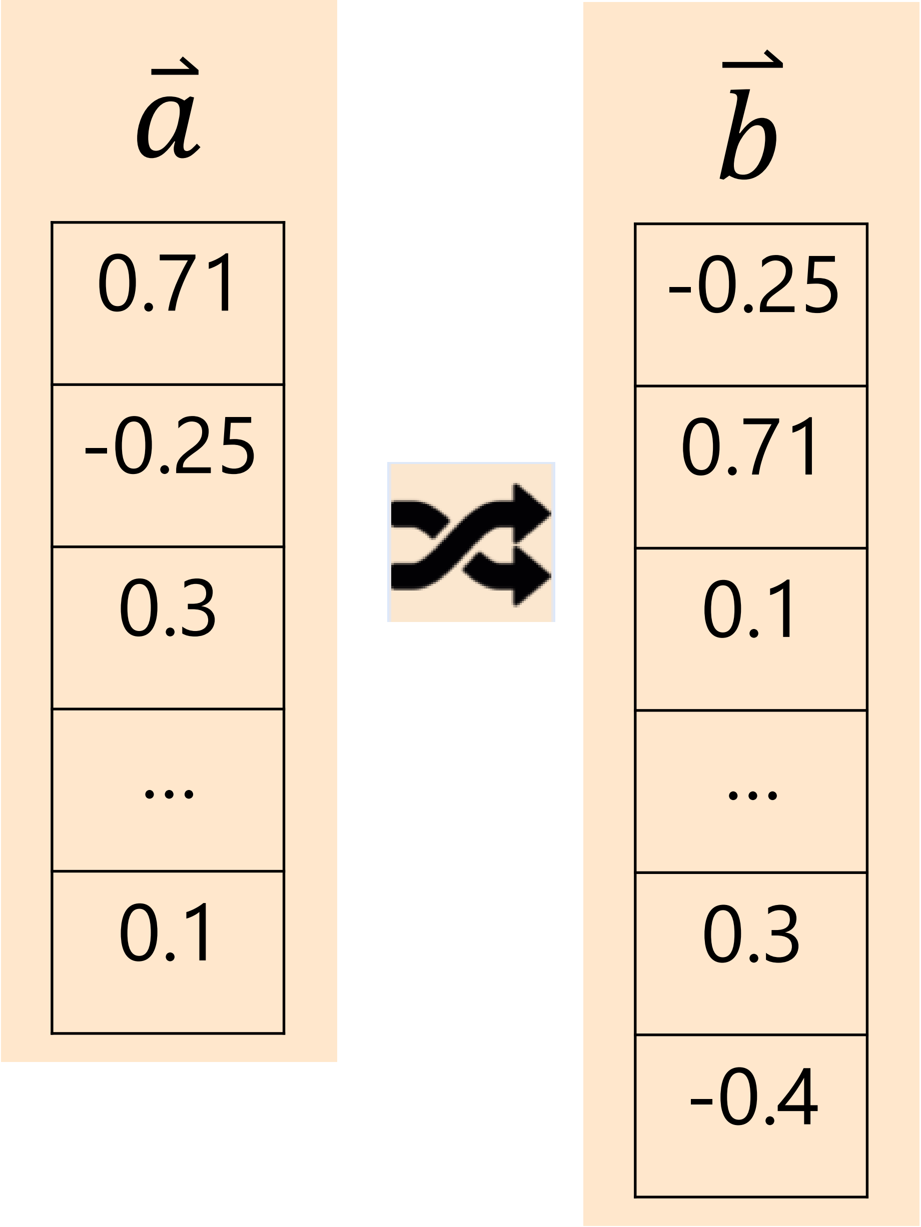


Fraction of Occupied Buckets



Output Transformations against Sybil Attacks

- Affine
- Pad+Shuffle
- Affine+Pad+Shuffle
- Binary



$\vec{a}$

0.71
-0.25
0.3
…
0.1

$\vec{b}$

-0.25
0.71
0.1
…
0.3
-0.4

Preserve performance for Legitimate Users

Defense	CIFAR10	STL10	SVHN	F-MNIST
None	90.41	95.08	75.47	91.22
B4B	90.24	95.08	74.96	91.7

Lowers Performance for Attackers

Defense	CIFAR10	STL10	SVHN	F-MNIST
None	68.1	63.1	61.5	89.0
B4B	12.01	13.94	19.96	69.93