Compromised or Attacker-Owned: A Large Scale Classification and Study of Hosting Domains of Malicious URLs

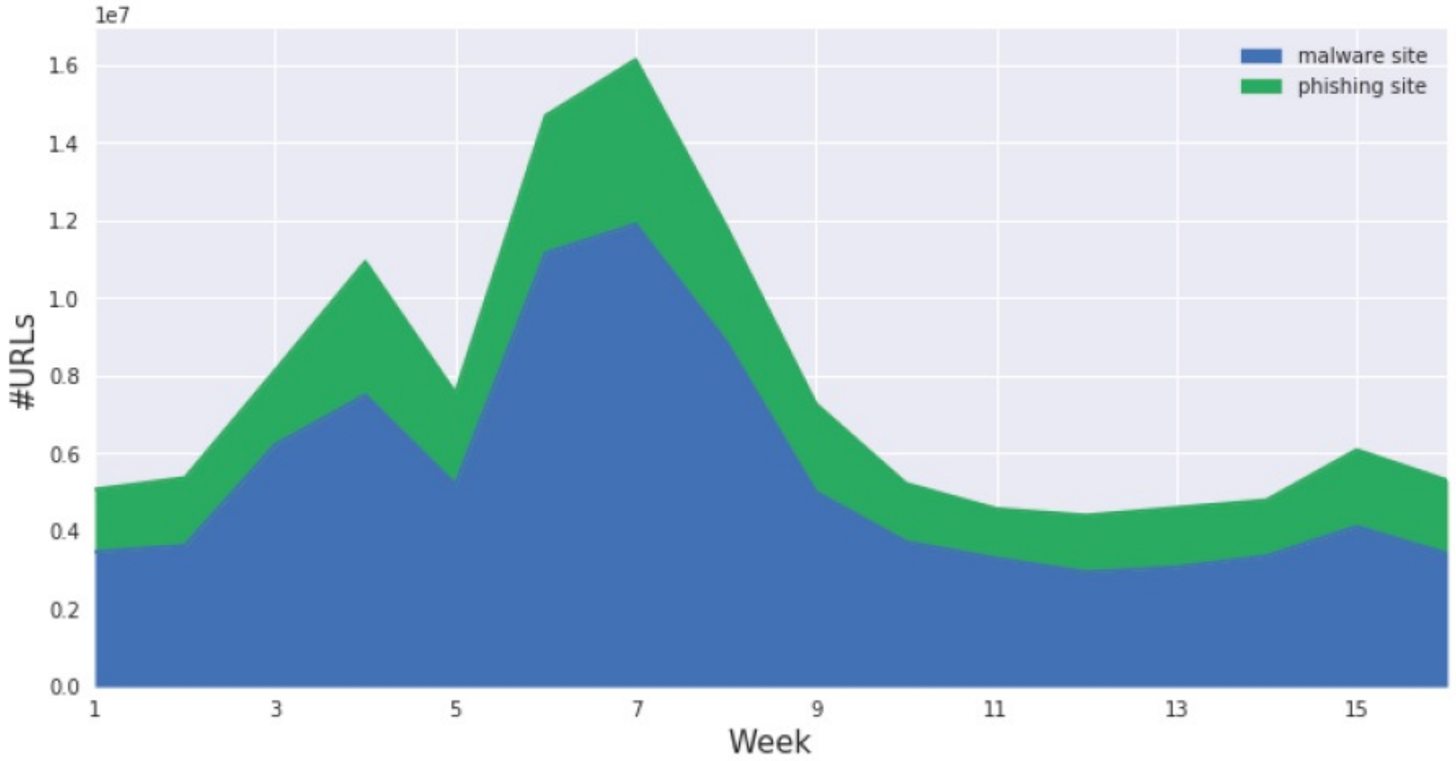
Ravindu De Silva, Mohamed Nabeel, Charith Elvitigala, Issa Khalil, Ting Yu,Chamath Keppitiyagama

UCSC
UNIVERSITY OF COLOMBO
SCHOOL OF COMPUTING

CORE
Sustainable Computing Research

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

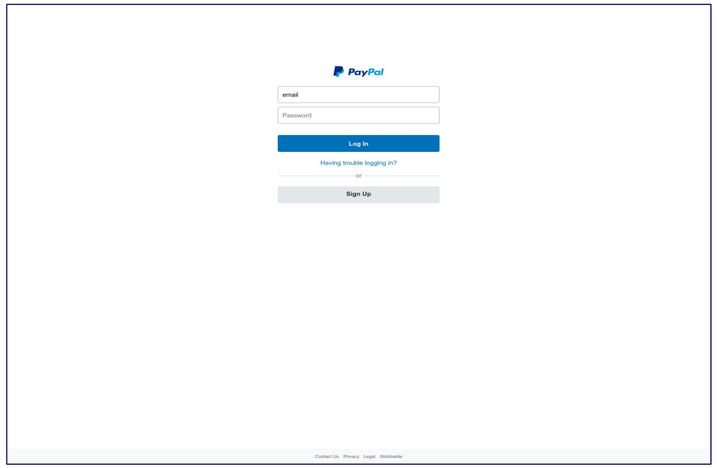
Malicious URLs from VirusTotal Feed



276K Malicious URLs/Week

How are these malicious sites hosted?

Compromised vs. Attacker-Owned



paypnl-son.shop (attack)

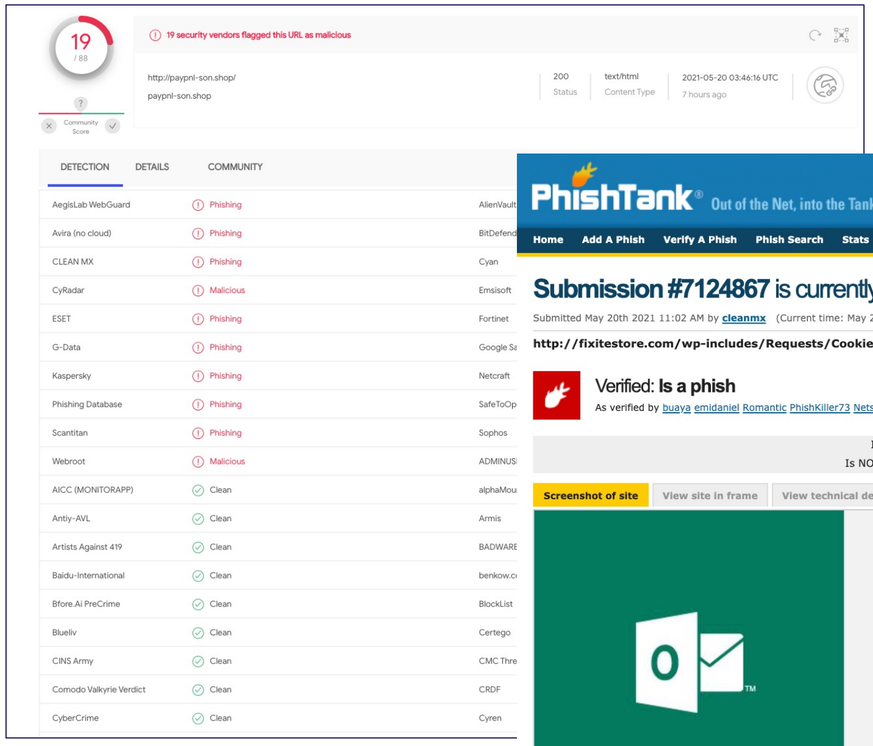
app-garden.com (bengin)

aaa.app-garden.com (compromised)

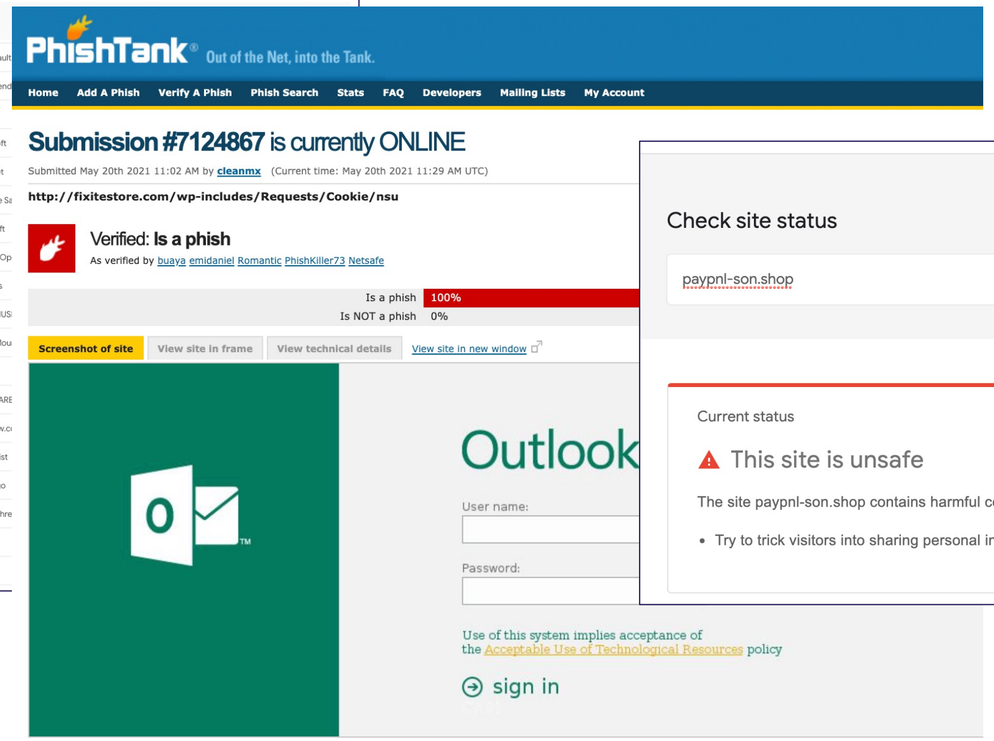
Why Distinguish Hosting Types?

- Different mitigation actions by different intermediaries
- Fine-grained blocking with minimal collateral damage
- Towards building better domain reputation systems

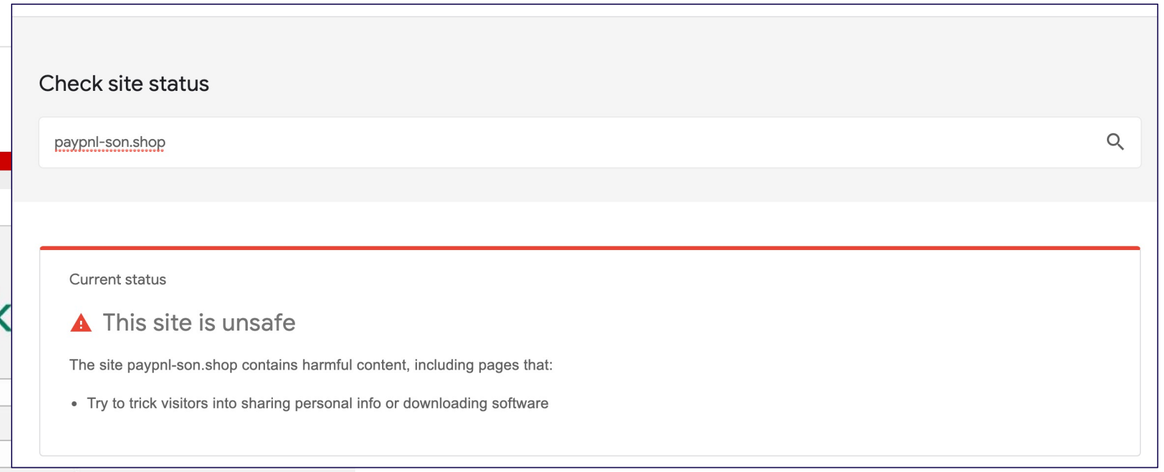
Current Status of Blacklists



VirusTotal

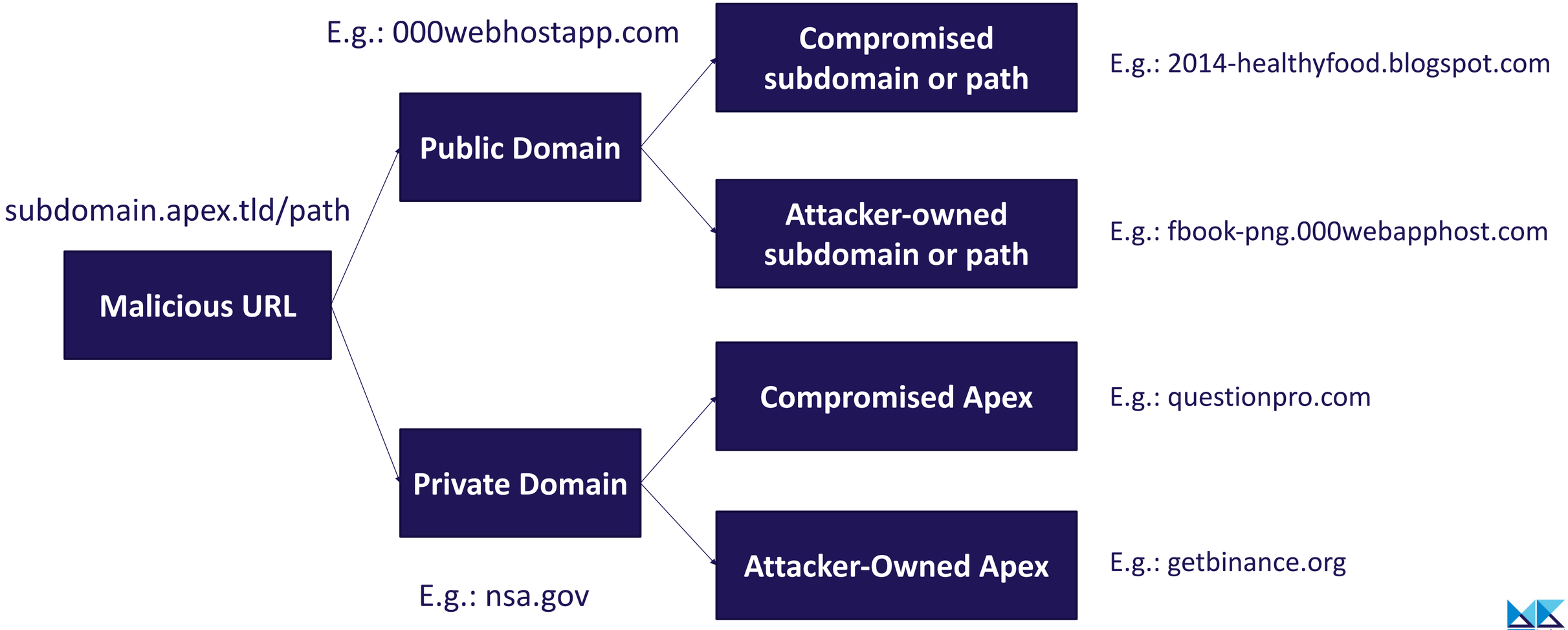


PhishTank

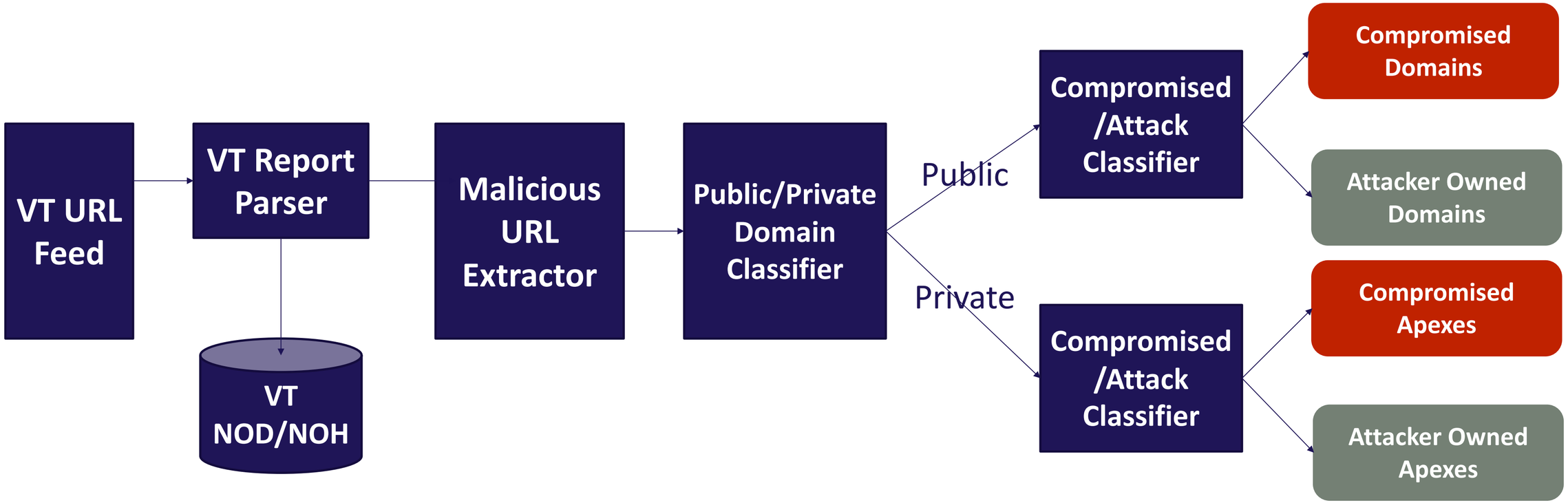


Google Safe Browsing

Malicious Hosting Types



Machine Learning Pipeline Design



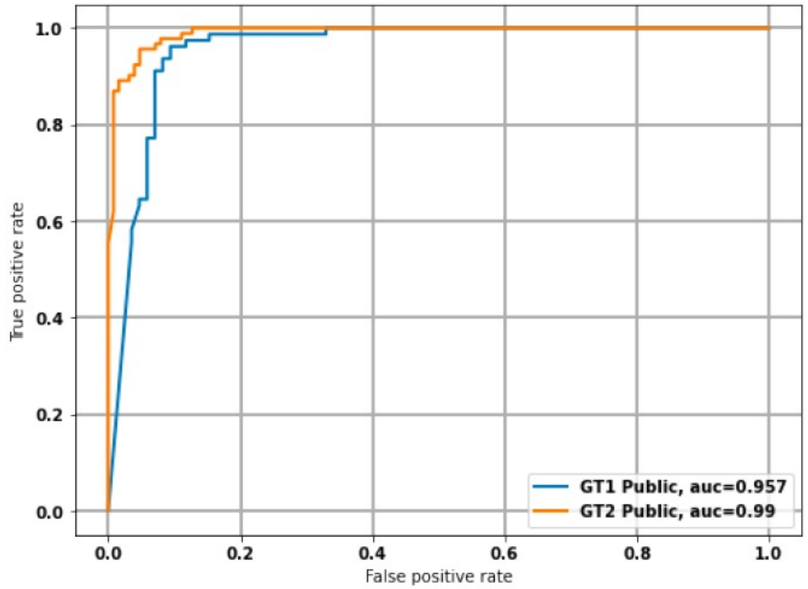
Datasets – VT URLs

- Study Period: Aug. 2019 to Nov. 2019

- Total number of URLs: 800+ million

Dataset	Malicious URLs	Malicious Apexes
Dataset1: Aug. 2019	3,434,226	373,238
Dataset2: Oct. 2019	4,398,584	358,762

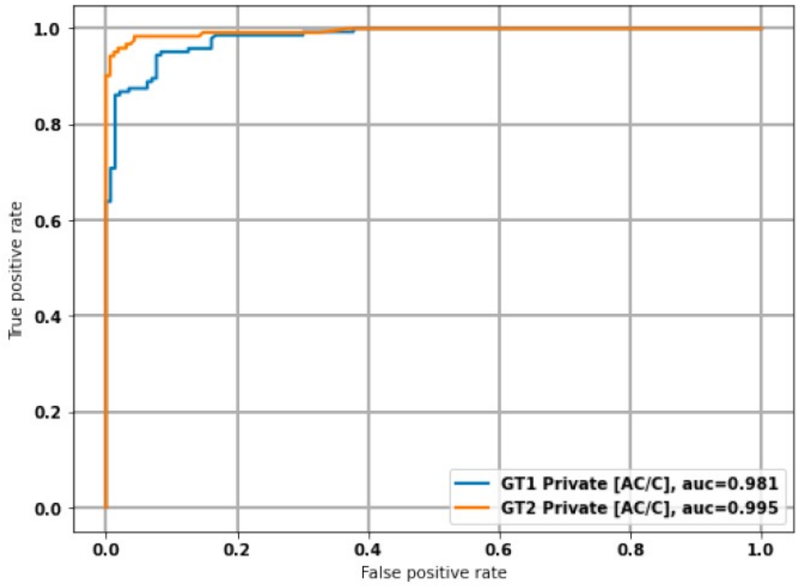
Public/Private Classifier



ROC Curve

- Key observations for feature engineering
  - Subdomain diversity
  - Variation and volume of scans
- Performance
  - Accuracy: 97.2%
  - Precision: 97.7%
  - Recall: 95.6%

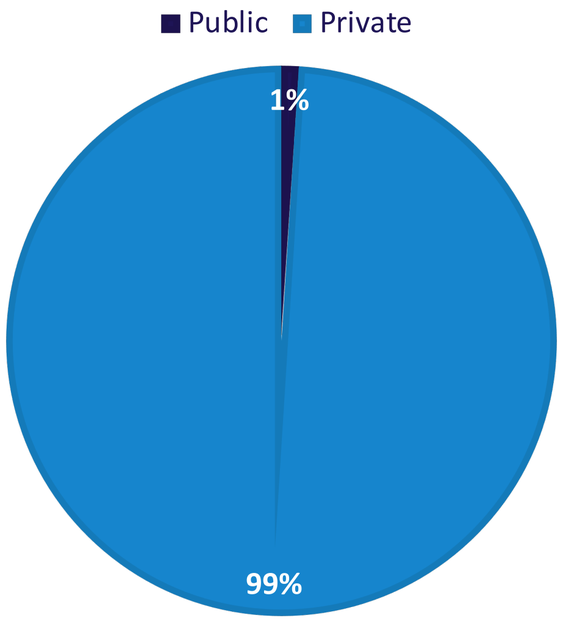
Compromised/Attacker-Owned Classifier



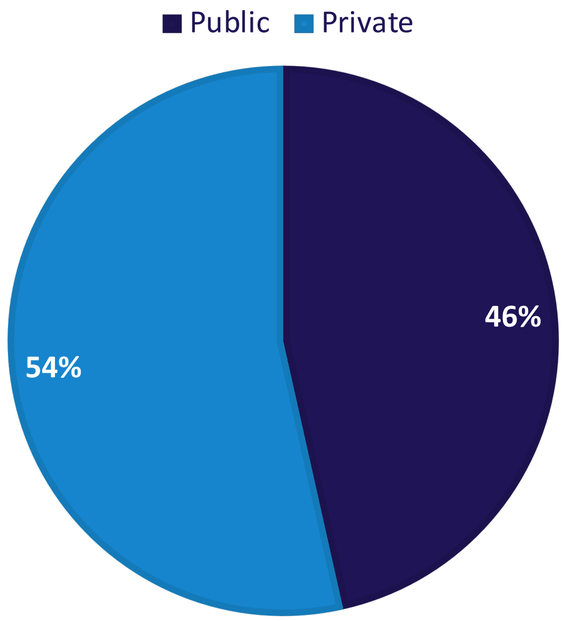
ROC Curve

- Key observations for feature engineering
  - Deviations in the hosting infrastructure
  - Lexical formation of URLs
  - Scan diversity
- Performance
  - Accuracy: 96.4%
  - Precision: 99.1%
  - Recall: 92.6%

Distribution of Attack Types – Malicious Apexes/URLs

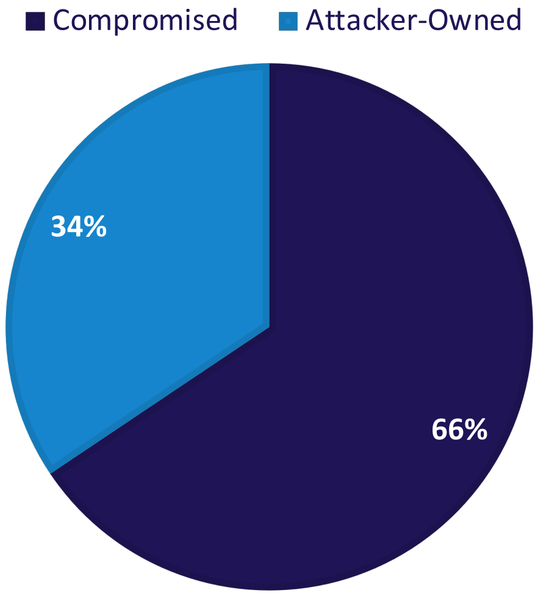


Malicious Apexes

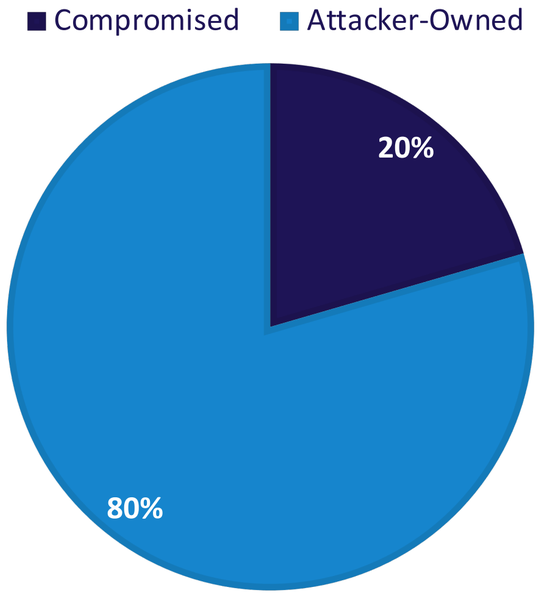


Malicious Websites

Distribution of Attack Types – Compromised/Attack

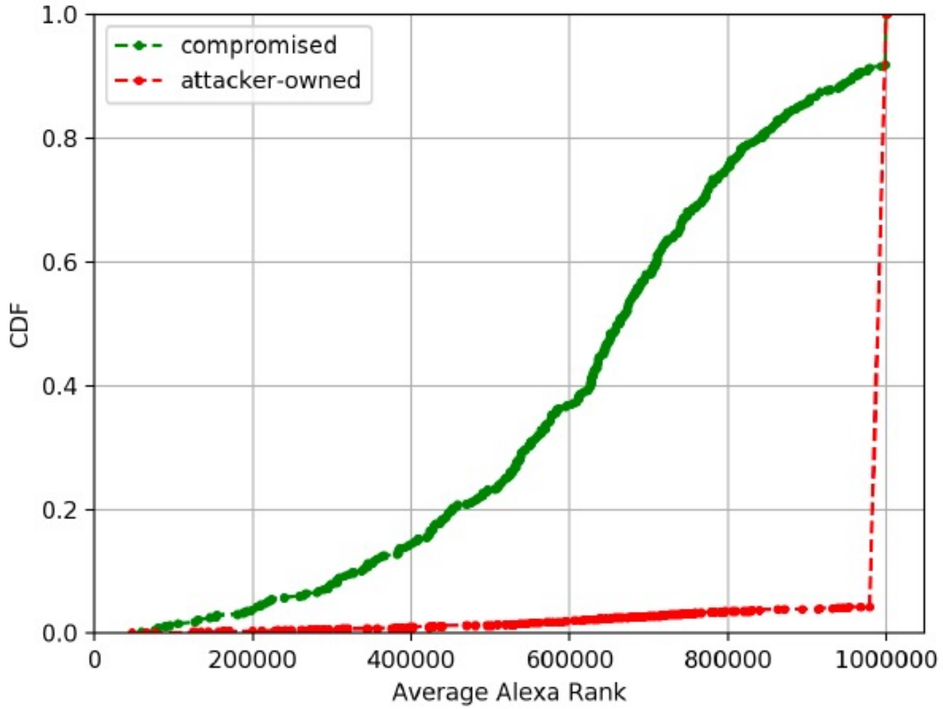


Private Malicious Websites

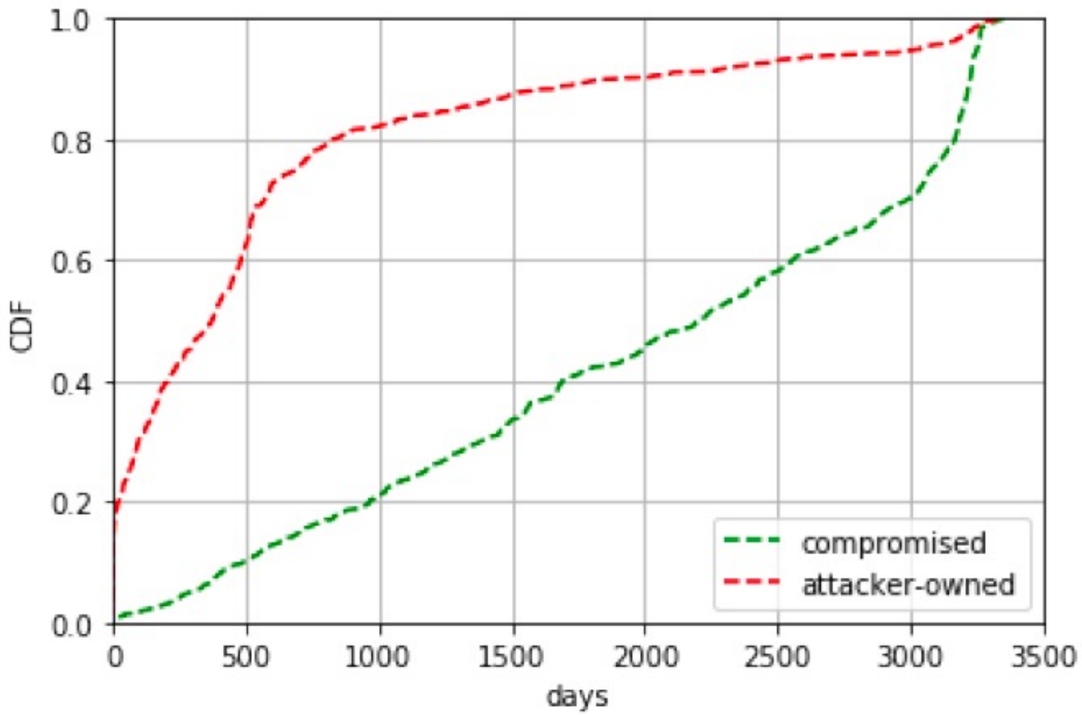


Public Malicious Websites

Properties of Compromised/Attack Websites

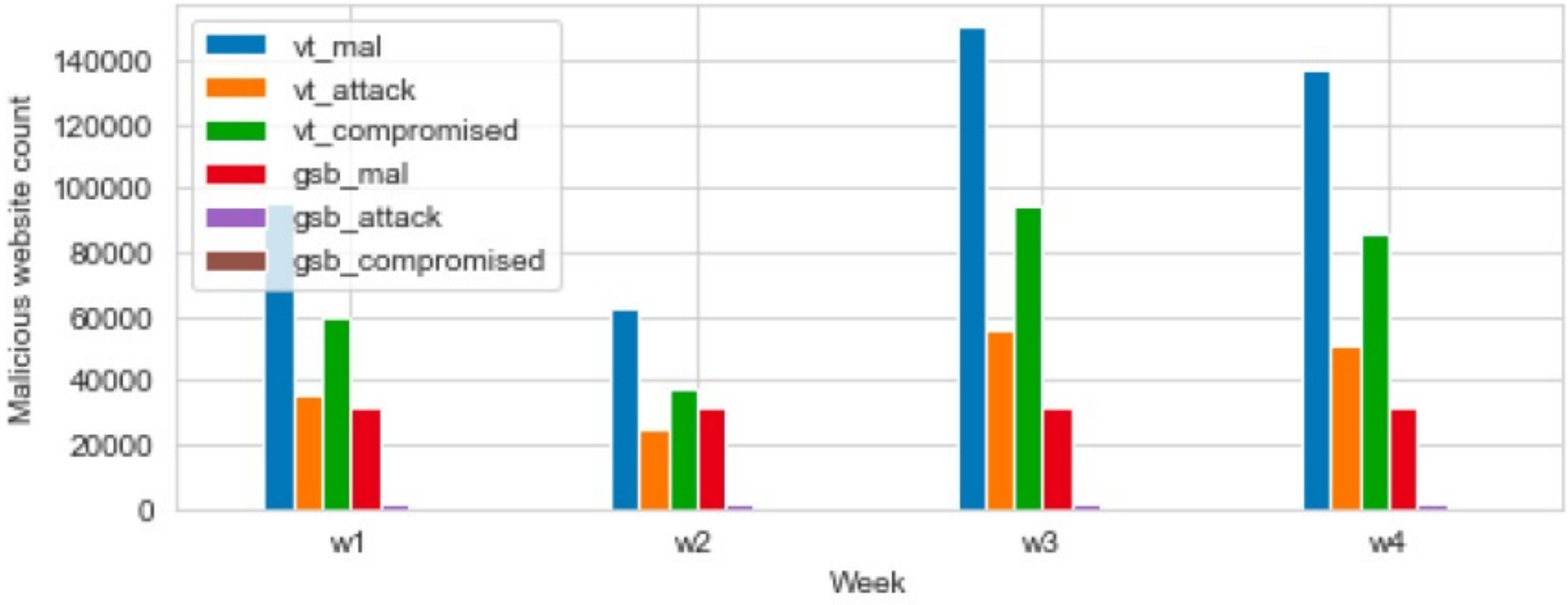


Popularity of Domain



Gap Between Registration and Attack Time

Comparing against GSB



Conclusions

- Content agnostic detection of various hosting types
  - Public vs. Private classifier
  - Compromised vs. Attacker-owned classifiers
- 81.7% malicious websites are hosted on apexes that attacks do not own
  - More needs to be done to secure benign websites and public domains
- Future work
  - Continuous evaluation of malicious URL types
  - Integration with existing blacklists and reputations systems

Thank You!

Mohamed Nabeel
mnabeel@hbku.edu.qa