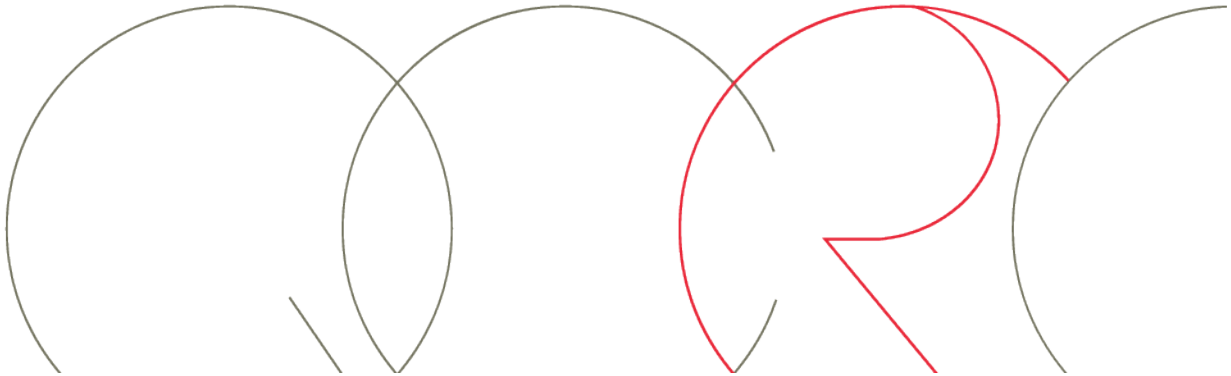
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

Member of Qatar Foundation عضو في مؤسسة قطر

AV-Meter: An Evaluation of Antivirus Scans and Labels

Omar Alrawi (Qatar Computing Research Institute)
Joint with Aziz Mohaisen (VeriSign Labs)



Overview

- Introduction to problem
- Evaluation metrics
- Dataset gathering and use
- Measurements and findings
- Implications
- Conclusion and questions

Example of labels

- ZeroAccess known labels by vendors and community:
  - Zeroaccess, Zaccess, 0access, Sirefef, Reon

Kaspersky	Backdoor.Win32.ZAccess.aer
Kingsoft	Win32.Malware.Heur_Generic.A.(kcloud)
Malwarebytes	Trojan.Agent
McAfee	W32/Sirefef.b
McAfee-GW-Edition	W32/Sirefef.b
MicroWorld-eScan	Trojan.Generic.KD.352199
Microsoft	TrojanDropper:Win32/Sirefef.B

Applications

- Anti-virus (AV) independent labeling and inconsistency
  - Heuristics, generic labels, etc.

- Machine learning (ground truth learning set and verification for classification)

- Incident response, mitigation strategies

- “Elephant in the room”
  - Symantec finally admits it!

Approach

- Contribution
  - Provide metrics for evaluating AV detection and labeling systems
  - Use of highly-accurate and manually-vetted dataset for evaluation
  - Provide several directions to address the problem

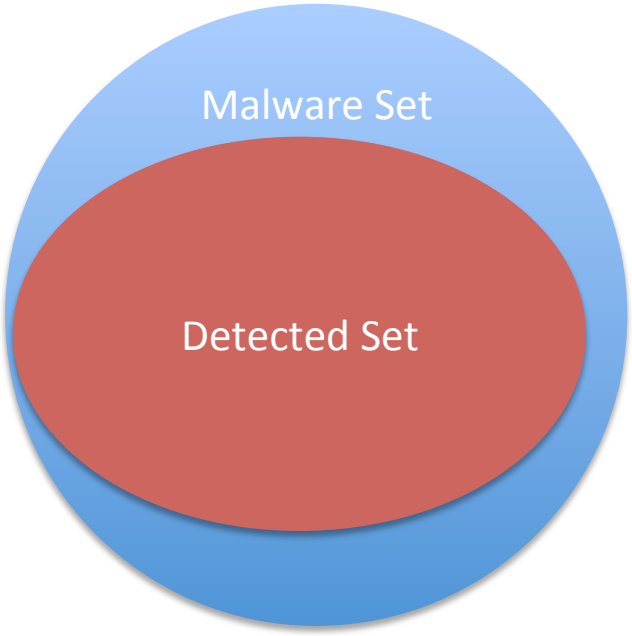
- Limitations
  - Cannot be used to benchmark AV engines
  - Cannot be generalized for a given malware family

Metrics (4Cs)

- Completeness (detection rate)
- Correctness (correct label)
- Consistency (agreement among other Avs)
- Coverage

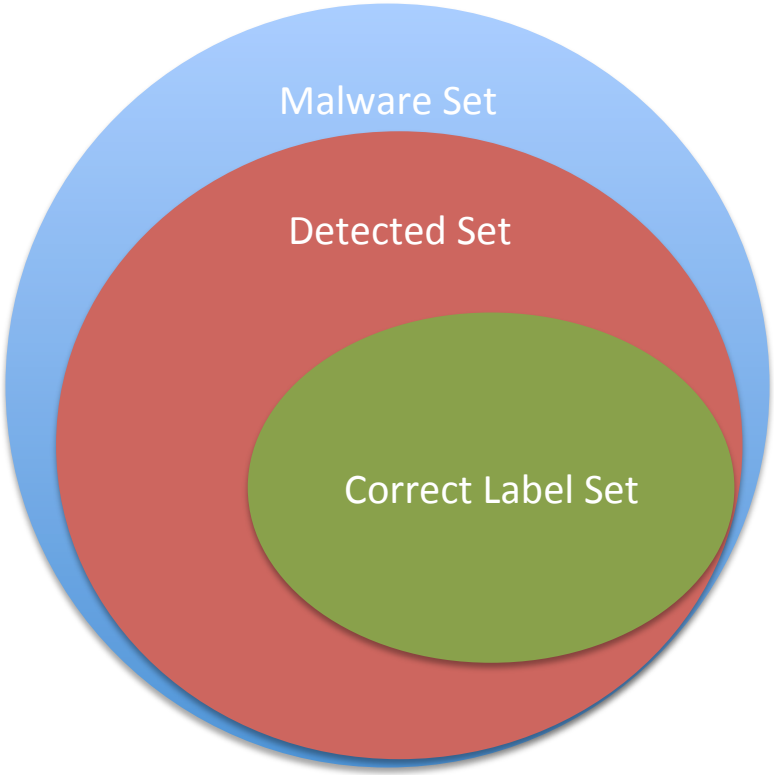
Completeness (detection rate)

- Given a set of malware, how many are detected by a given AV engine
- Normalized by the dataset size; value in [0-1]



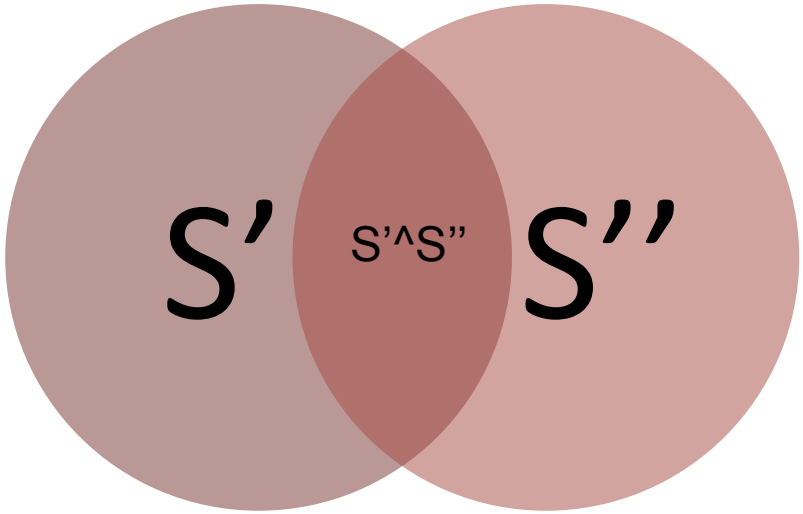
Correctness

• Score based on correct label returned by a given AV engine; normalized by the set size



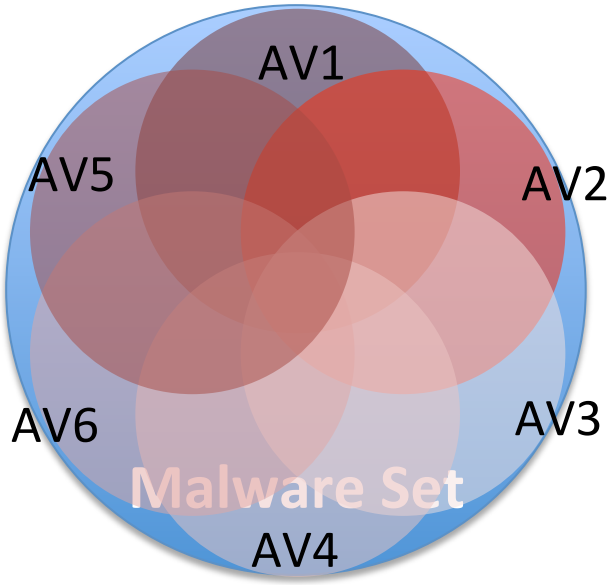
Consistency

- Agreement of labels (detections) among vendors
  - Completeness consistency
  - Correctness consistency
  - (S'^S'')/(S'vS'') for both measures
- Normalized by the size of the union of S' and S''



Coverage

- Minimal number of AV engines required to detect a given complete set of malware
- Normalized by the size of set; value in [0-1]



Data

- Eleven malware families
  - Zeus, ZeroAccess, Getkys, Lurid, DNSCalc, ShadyRat, N0ise, JKDDos, Ddoser, Darkness, Avzhan
  - Total of about 12k pieces of malware

- Three types of malware
  - Trojans
  - DDoS
  - Targeted

Data Vetting

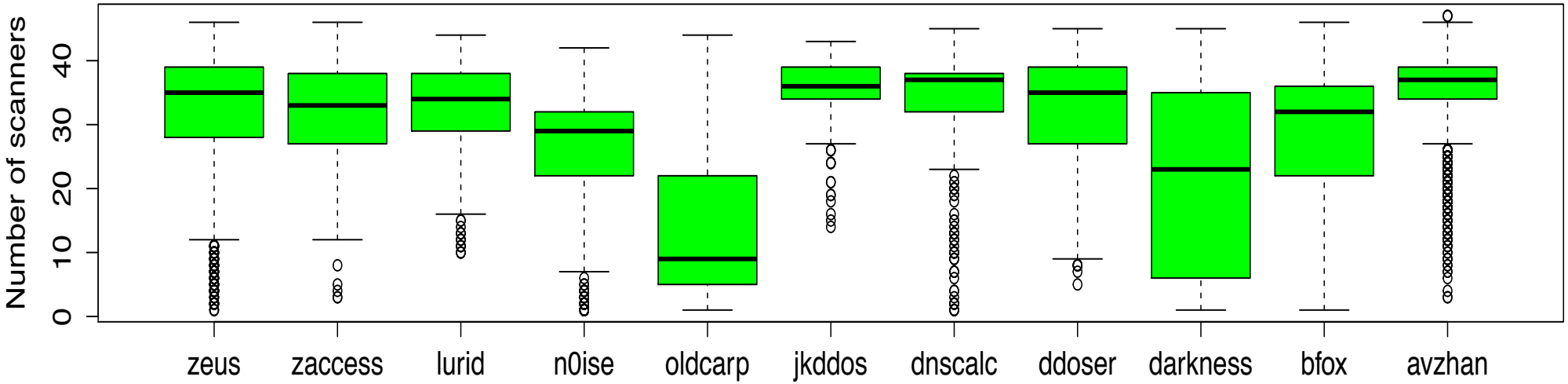
- Operational environment
  - Incident response
  - Collected over 1.5 years (2011-2013)

- Malware labels
  - industry, community, and malware author given labels (Zbot, Zaccess, cosmu, etc.)

- Virus scans
  - VirusTotal
  - Multiple occurrence of vendors, use best results

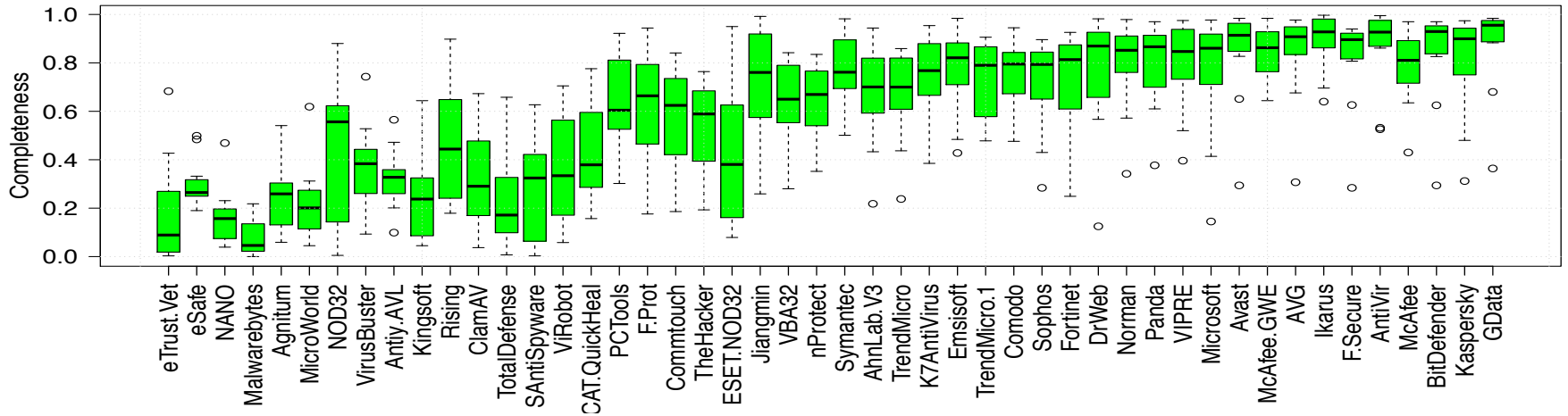
Experiment - Completeness

- More than half of AV engines detect our pool of samples (positive outcome!)
- These samples contribute to the high detection rates seen across AV engines



Experiment - Completeness

- Completeness score for each AV for all 12k samples
- Maximum completeness provided is 99.7%
- Average completeness provided is 59.1%



Experiment - Completeness

- Completeness versus number of labels
  - On average each scanner has 139 unique label per family and median of 69 labels

- Completeness versus largest label
  - We see an average largest label is 20%
    - Example: if largest label 100, then average AV has 20 labels per family
  - AV with smaller labels can be deceiving regarding correctness
    - Example: Norman has generic label (ServStart) for Avzhan family covering 96.7% of the sample set

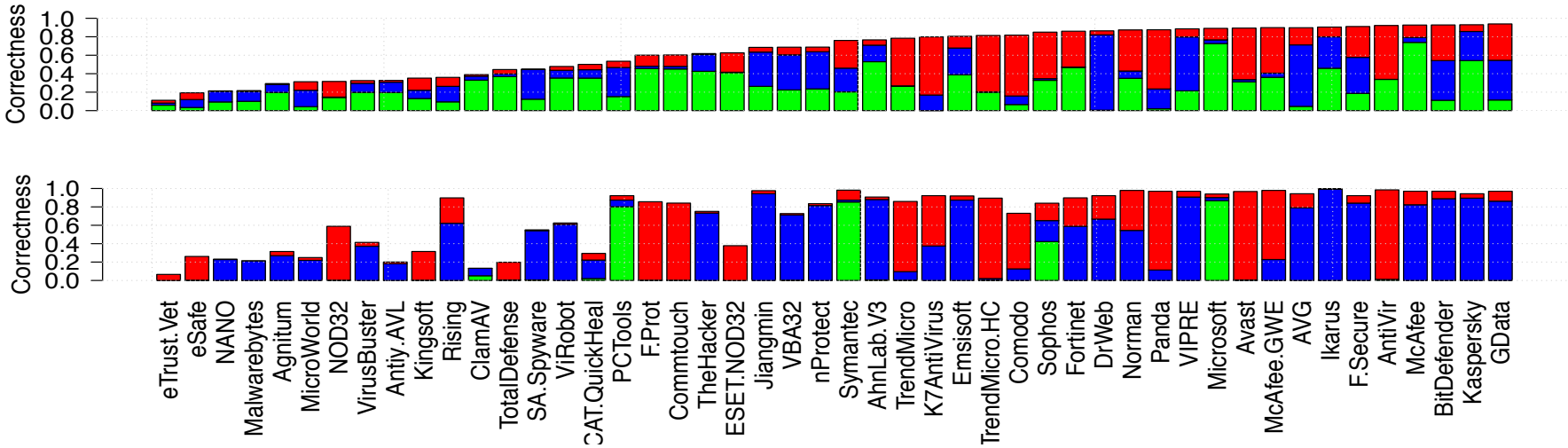
Experiment - Correctness

- Highest correct label is Jkddos (labeled jackydos or jukbot) by:
  - Symantec (86.8%), Microsoft (85.3%), PCTools (80.3%), with completeness close to 98%

- Others
  - Blackenergy (64%,)
  - Zaccess (38.6%)
  - Zbot (73.9%)

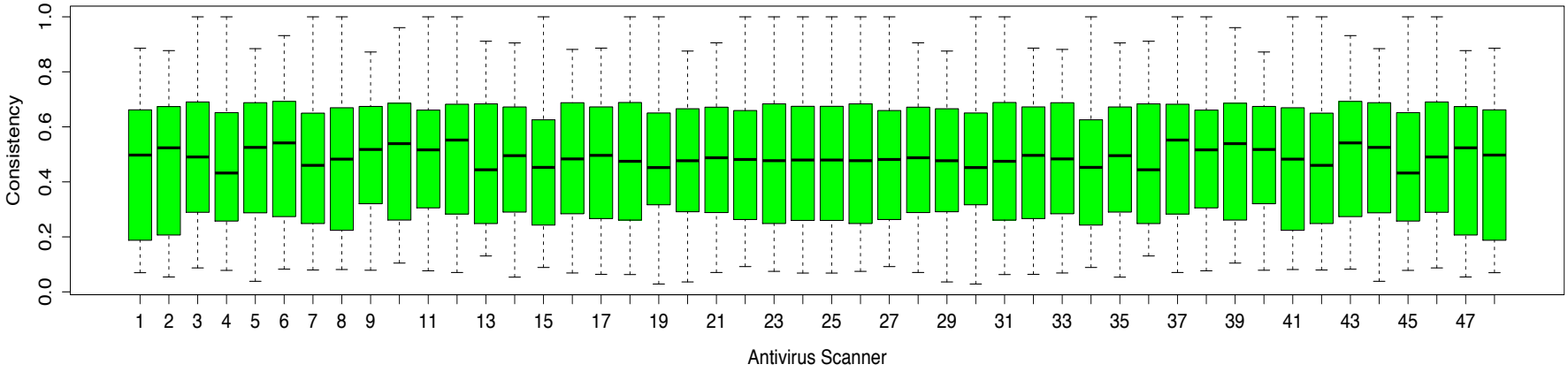
Experiment - Correctness

- Correctness - Zeus and JKDDoS
  - Static scan labels - green
  - Behavior labels (Trojan, generic, etc.) - blue
  - Incorrect labels (unique label) - red



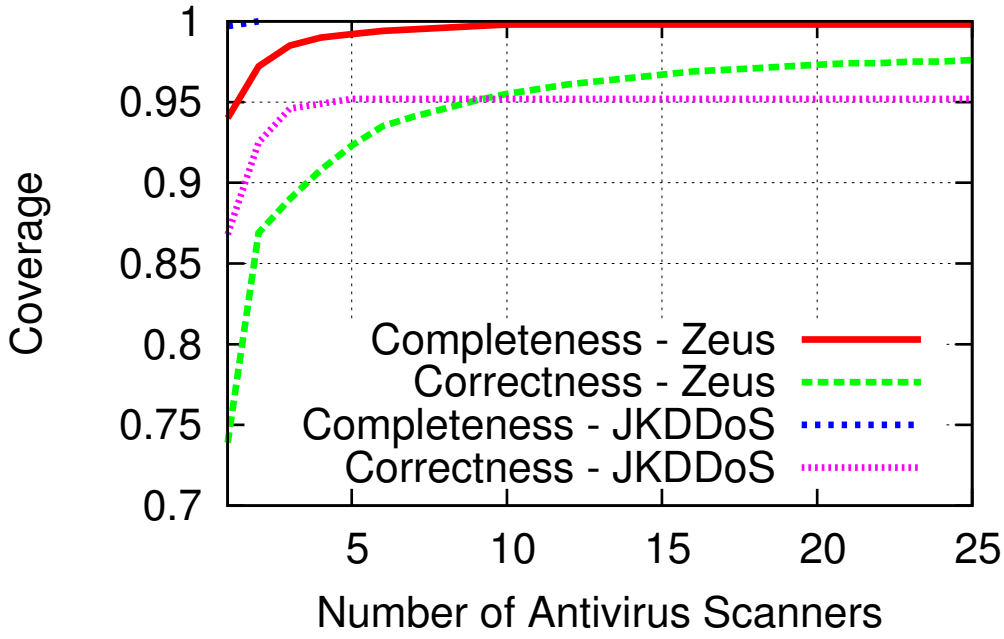
Experiment – Consistency

- Consistency of detection
  - Pairwise comparison for sample detection across two vendors
- On average 50% agreement
- 24 vendors have, almost, perfect consistency
  - AV sharing information is a potential explanation;
  - AV vendor 1 depends on vendor 2 detection (piggybacking)
- Example of one family (Zeus)



Experiment - Coverage

- JKDDoS and Zeus
- Coverage for
  - Completeness (3-10 AV engines) depending on family
  - Correctness (Never reached with all 48 engines)
  - Highest score observed for correctness is 97.6%



Implications

- Relying on AV labels to evaluate proposed approaches seems problematic at best;
  - Machine learning, classification and clustering

- Rapid incident response based on AV labels
  - Applying wrong remediation for incident based on incorrect label may cause long-lasting harm.

- Tracking and attribution of malicious code (Law enforcement)
  - Tracking inaccurate indictors due to incorrect label

Conclusion

- Proposed remedies
  - Data/indicator sharing
  - Label unification
  - Existing label consolidation
  - Defining a label, by behavior, class, purpose, etc.

- Future work
  - Methods and techniques to tolerate inconsistencies and incompleteness of labels/detection

- Full paper
  - http://goo.gl/1xFv93

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

Member of Qatar Foundation عضو في مؤسسة قطر

ص.ب. 5825
برج التورنيدو، الطابق 10
الخليج الغربي، الدوحة - قطر
هاتف 0629 4454 974+
فاكس 0630 4454 974+

P.O.Box 5825
Tornado Tower, 10th Floor
West Bay, Doha - Qatar
Tel +974 4454 0629
Fax +974 4454 0630

www.qcri.qa

Omar Alrawi
oalrawi@qf.org.qa
+974 4544 2955