UMBC

LPS The Laboratory for Physical Sciences

Booz | Allen | Hamilton®

MalDICT: Benchmark Datsets on Malware Behaviors, Platforms, Exploitation, and Packers

Robert J. Joyce

Edward Raff

Charles Nicholas

James Holt

Malware Classification Tasks

- Nearly all research on malware classification focuses on benign/malicious classification and family classification

- We identify four other classification tasks that we believe are important but under-explored:
  - Behavior or malware category
  - File format, target OS, or programming language
  - Vulnerability that is exploited
  - Malware packer

Prior Labeled Malware Data

- Labeled benchmark datasets for these four uncommon tasks have limitations or are nonexistent
  - SOREL has 10 million files and 11 behavioral tags
  - Malicia has 11,363 files and 172 exploit tags but is 9 years old and no longer distributed

- No publicly-available malware datasets are labeled by platform, vulnerability, or packer

Prior AV-Based Malware Taggers

- Prior tools support tagging malware according to one or more of these tasks
- Only AVClass2 supports all four tasks

	Behavior	Platform	Exploit	Packer
EUPHONY	X			
SMART	X			
AVClass2	X	X	X	X
Garcia-Teodoro *et al*.	X			

Tagging Malware Using AV Scan Reports

- Scan a file with a collection of AV products
- Each one outputs an AV label that describes the file
- Can include behavior, platform, exploit, and packer information

```
1. TR/Andromeda.B
2. Trojan.Win32.Andromeda.xyz
3. Backdoor.Androm.99
4. Win32/Gamarue.1234
5. Trj.Gamarue!1.23W
6. W32.TrojanDownloader.Wauchos.A
7. Trojan.TR/Backdoor.Gen
8. Malware (ai Score=99)
9. BehavesLike:W32/Zbot-abc
```

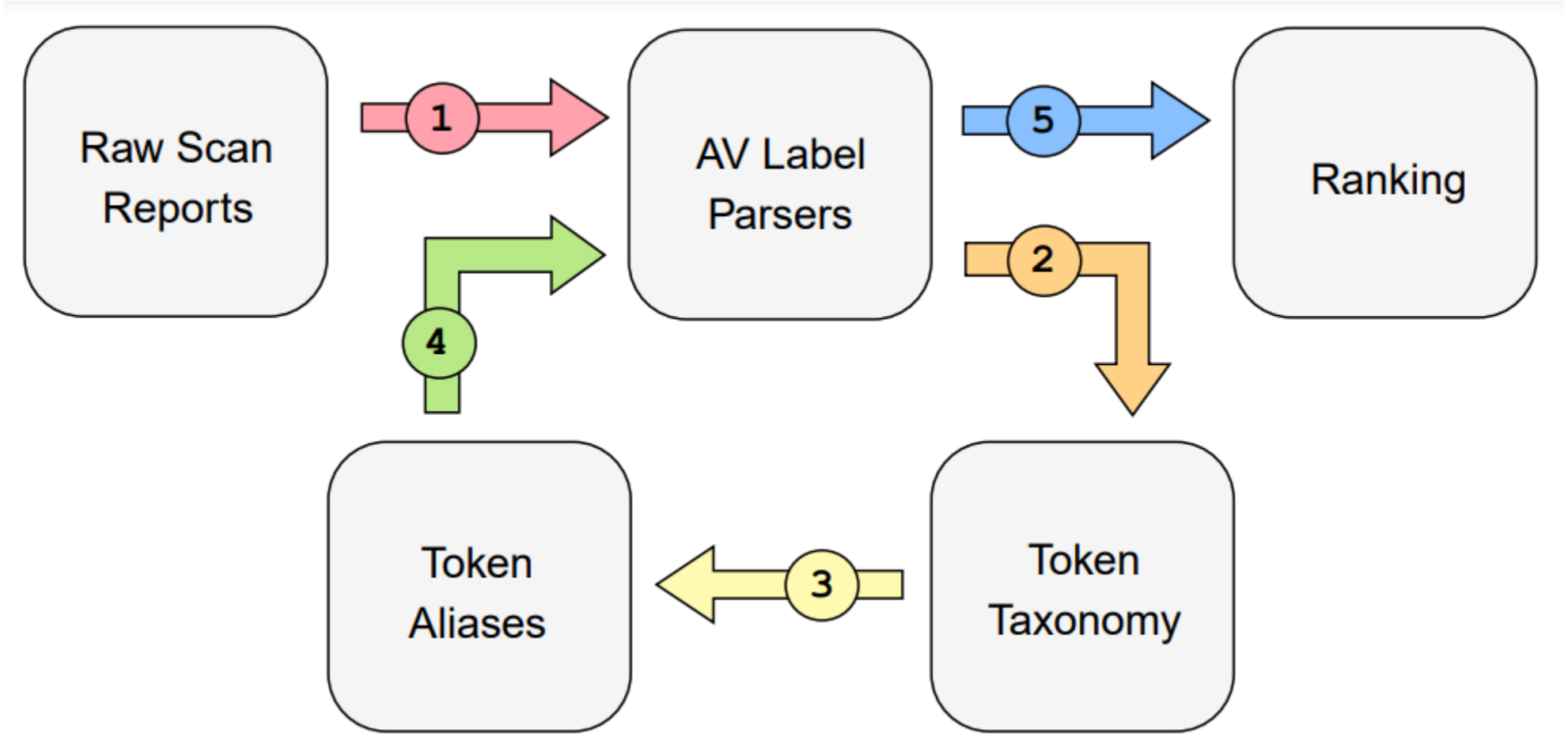
Improving AV-Based Tagging

- We created a new AV-based malware tagger called **ClarAVy**
  - “Clarifying noise in AV scan data”

- It offers improvements in parsing AV labels, detecting aliases, and accounting for correlated AV products

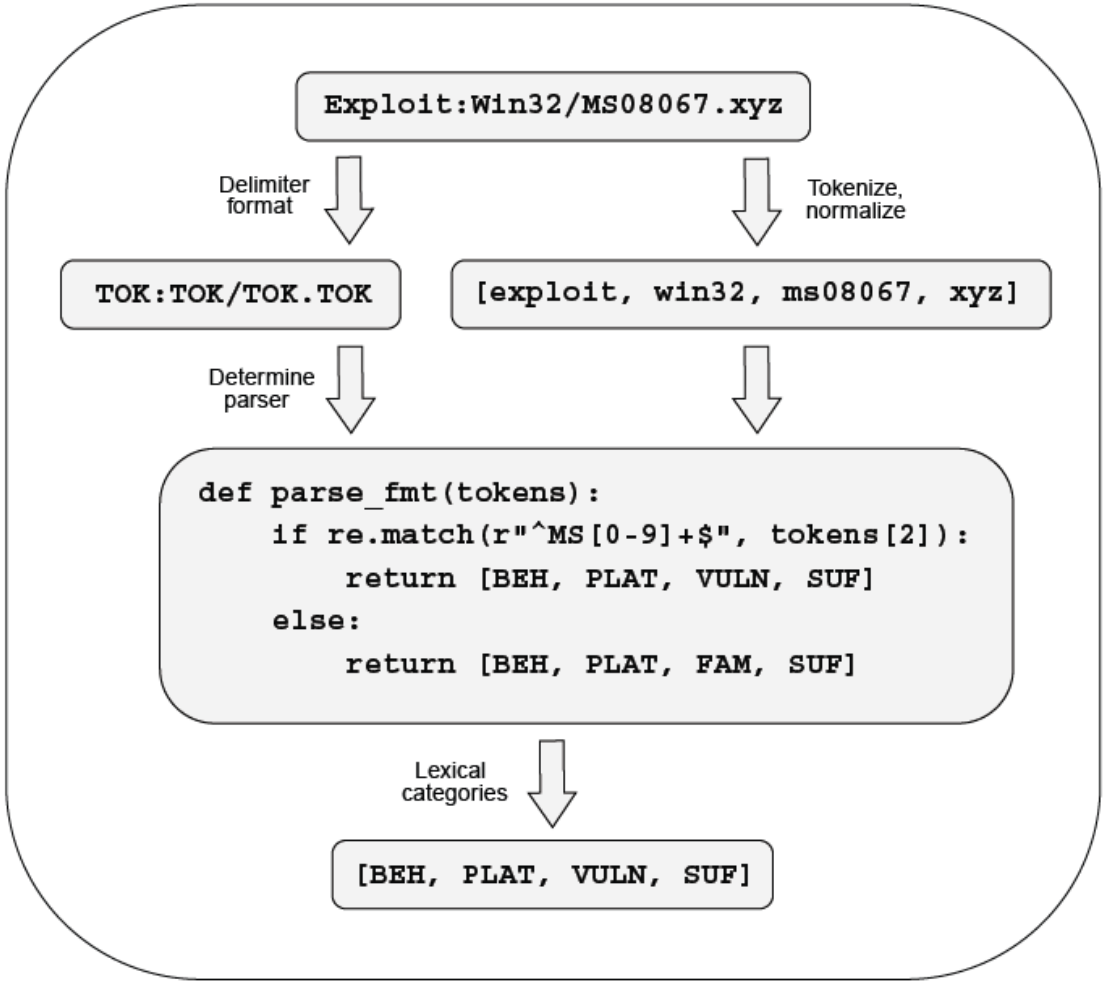
- We used ClarAVy to create benchmark datasets tagged by category, platform, vulnerability, and packer

ClarAVy Architecture



ClarAVy AV Label Parsing

- ClarAVy supports parsing 882 AV label formats across 90 different AV products
- Parsers were developed with 40 million VirusTotal scans
- Coverage for 99.5% of our **1.1 billion** AV labels



ClarAVy Token Taxonomy

BEH	The malware category or behavior
PLAT	The OS, file format, or programming language
VULN	A vulnerability exploited by the malware
PACK	The packer used to pack the file
FAM	The malware family that the file belongs to
SUF	A suffix token at the end of the AV label
PRE	Ambiguous, but not a FAM or SUF token
UNK	A token whose lexical category cannot be determined

Handling Parsing Ambiguity

- ClarAVy tracks how frequently tokens are assigned to each category in its taxonomy

- It assigns “permanent” lexical categories for some tokens
  - Coverage for unsupported AV label formats
  - Resolves parsing ambiguities

Identifying Token Aliases

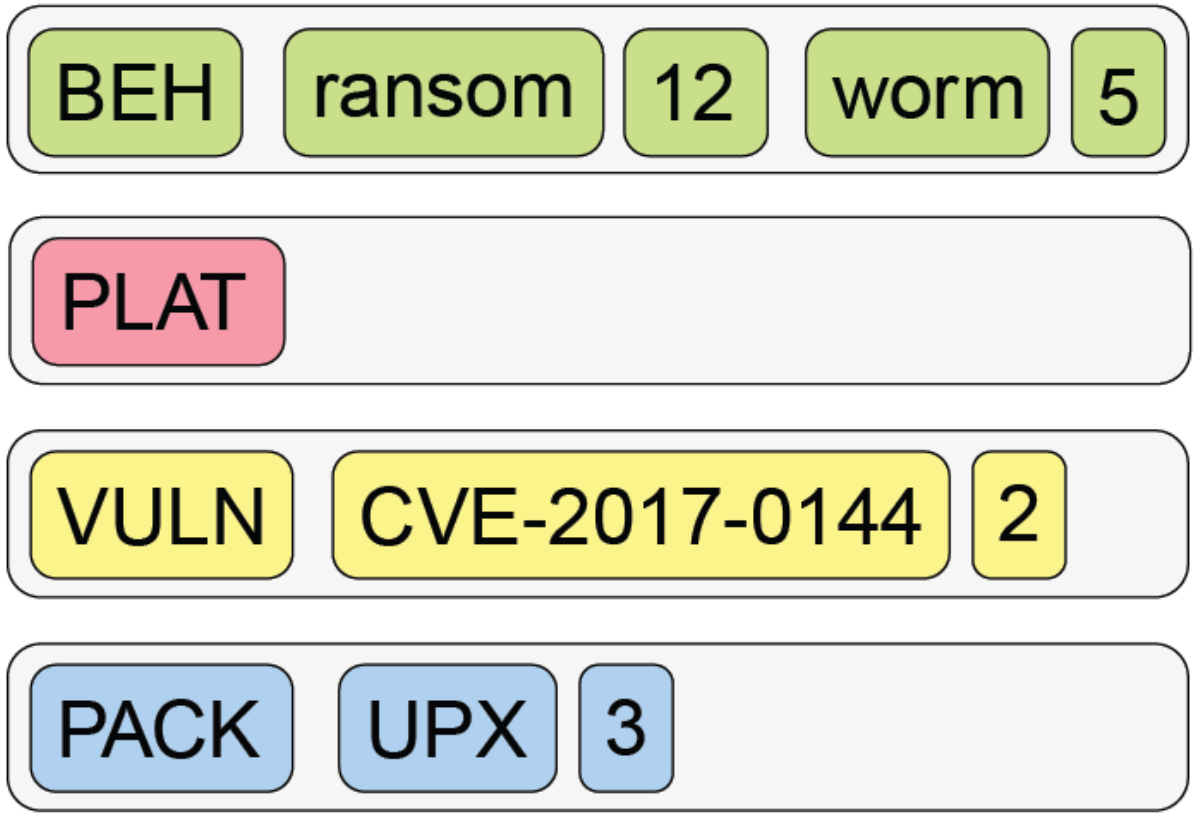
- Need to identify tokens used by different AV products that have identical meaning
  - ClarAVy can identify two different classes of token aliases:

- **Trivial aliases**: Two tokens are identical except for an extra character at the end

- **Parent/Child alias**: Two tokens co-occur frequently and have a low edit distance

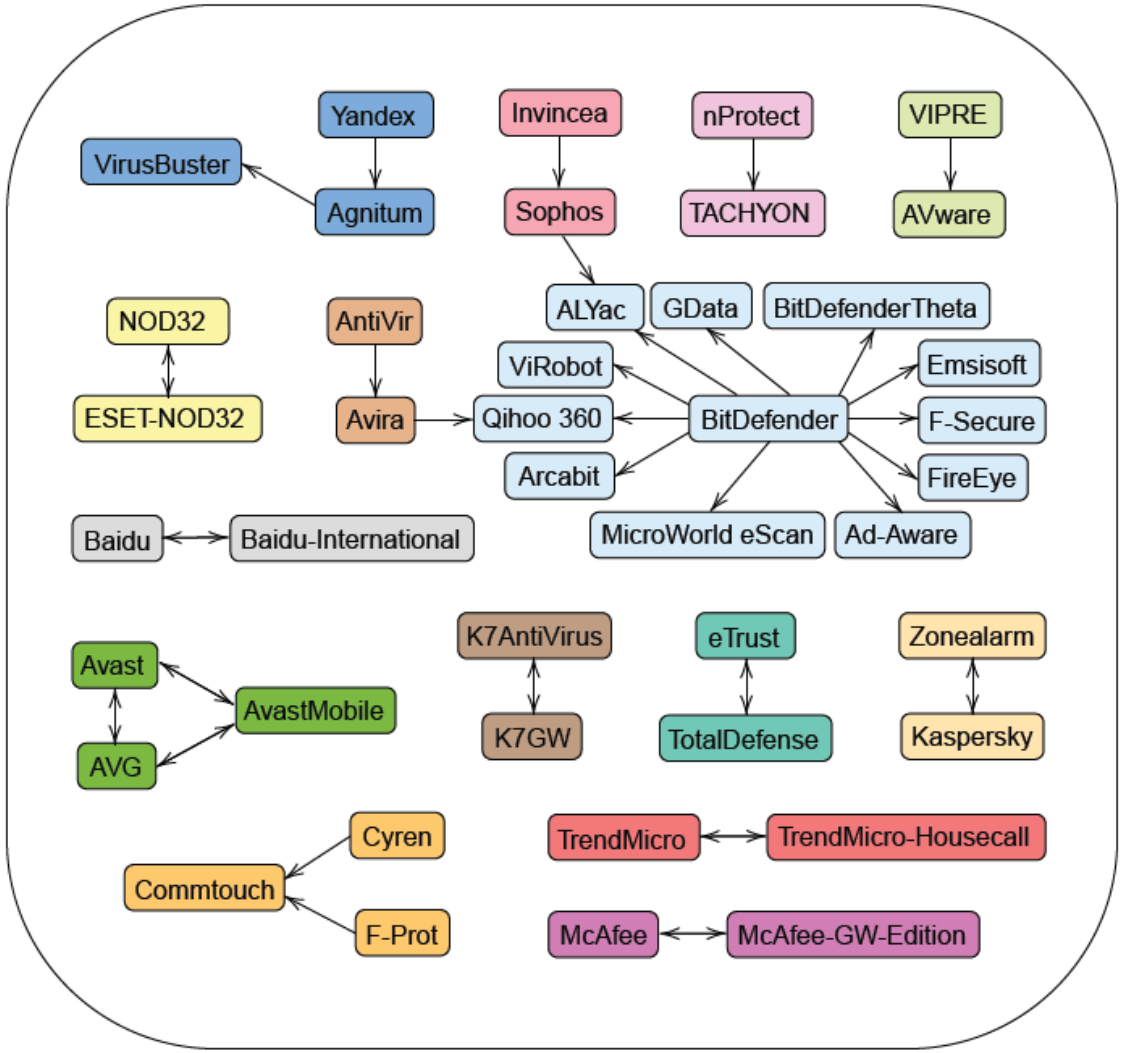
ClarAVy Voting

- A threshold parameter T controls the votes needed to assign a tag
- T defaults to 5 for BEH and PLAT tokens
- T defaults to 1 for VULN and PACK tokens



Correlations between AV Products

- Manually identified each AV product with a known relationship to another AV
- If 2+ correlated AVs output the same token, it is treated as a single “vote”



ClarAVy Validation

- Validated each parsing rule on 10,000 AV labels
- Manually reviewed and corrected lexical category assignments and token aliases
- Two more validation experiments in our paper
  - Consistency with SOREL behavioral tags
  - Comparison against AVClass2

Malware Datasets for Infrequent Classification Tasks

- We used the VirusTotal API to collect over 40 million AV scan reports for chunks 0 – 465 of the VirusShare dataset
- Then we tagged all of this malware using ClarAVy
- We discarded tags which were too rare and down-sampled tags which were too common
  - Different thresholds were selected for each dataset

MalDICT Train and Test Dataset Splits

- Behavior and Platform datasets use a temporal train/test split
- Exploit and Packer datasets use a stratified train/test split

	Total Files	Train Set	Test Set	Tags
Behavior	4,317,241	3,744,022	573,219	75
Platform	963,492	738,264	225,228	43
Vulnerability	173,886	136,467	37,419	128
Packer	252,148	201,392	50,756	79

MalDICT Benchmark Results

MalConv Evaluation (Micro Avg.)

	Behavior	Platform	Vulnerability	Packer
Precision	.651	.750	.926	.897
Recall	.492	.718	.888	.801
F1-Measure	.560	.733	.906	.846
ROC-AUC	.929	.965	.995	.987

MalConv Evaluation (Weighted Avg.)

	Behavior	Platform	Vulnerability	Packer
Precision	.617	.772	.926	.892
Recall	.492	.718	. 888	.801
F1-Measure	.512	.718	.903	.842
ROC-AUC	.896	.960	.995	.980

LightGBM OvR Evaluation (Micro Avg.)

	Behavior	Platform	Packer
Precision	.177	.682	.783
Recall	.555	.953	.948
F1-Measure	.268	.795	.857
ROC-AUC	.897	.958	.992

LightGBM OvR Evaluation (Weighted Avg.)

	Behavior	Platform	Packer
Precision	.177	.682	.783
Recall	.555	.953	.948
F1-Measure	.268	.795	.857
ROC-AUC	.897	.958	.992

Contributions

- We are open-sourcing ClarAVy on GitHub
  - https://github.com/NeuromorphicComputationResearchProgram/ClarAVy

- Releasing file hashes and ClarAVy tags for MalDICT
  - More than 5.5 million tagged malware samples

- Releasing disarmed executables and EMBER feature vectors for all PE files in MalDICT