Measuring and Modeling the Label Dynamics of Online Anti-Malware Engines

Shuofei Zhu[1], Jianjun Shi[1,2], Limin Yang[3]
Boqin Qin[1,4], Ziyi Zhang[1,5], Linhai Song[1], Gang Wang[3]

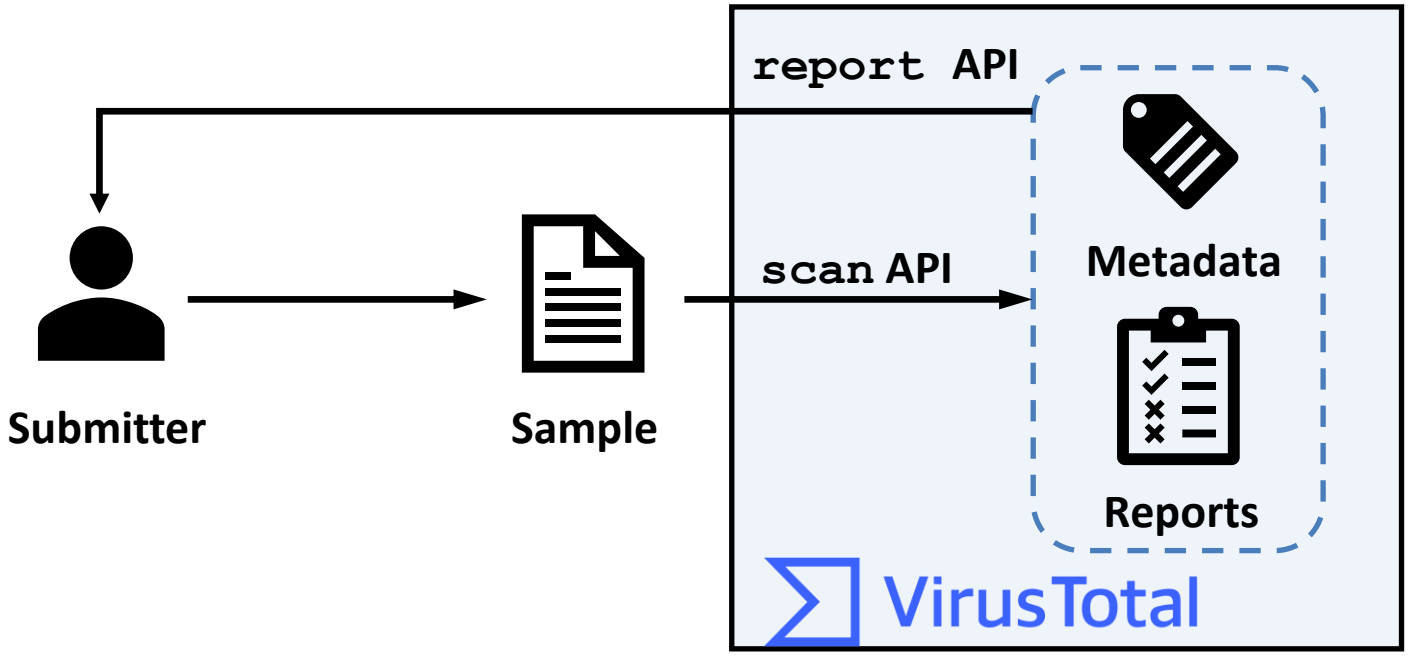
[1]*The Pennsylvania State University*
[2]*Beijing Institute of Technology*
[3]*University of Illinois at Urbana-Champaign*
[4]*Beijing University of Posts and Telecommunications*
[5]*University of Science and Technology of China*

PennState College of Information Sciences and Technology

ILLINOIS Computer Science

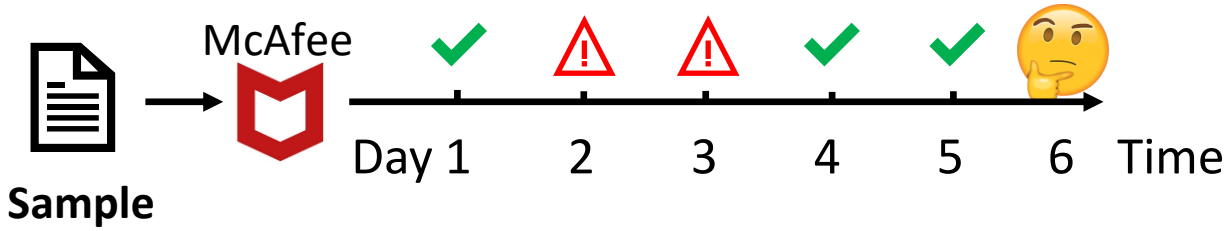
VirusTotal

- The largest online anti-malware scanning service
  - Applies 70+ anti-malware engines
  - Provides analysis reports and rich metadata
- Widely used by researchers in the security community



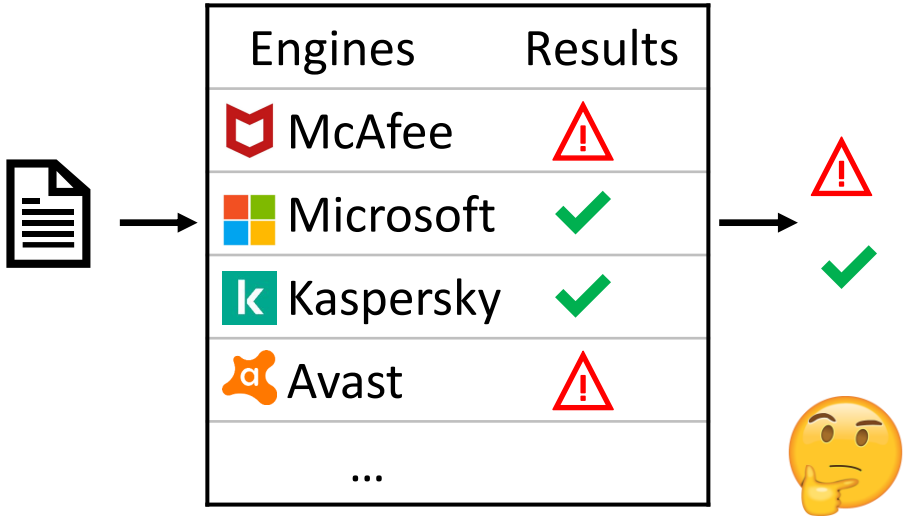
Challenges of Using VirusTotal

- Q1: When VirusTotal labels are trustworthy?



Challenges of Using VirusTotal

- Q1: When VirusTotal labels are trustworthy?
- Q2: How to aggregate labels from different engines?
- Q3: Are different engines equally trustworthy?



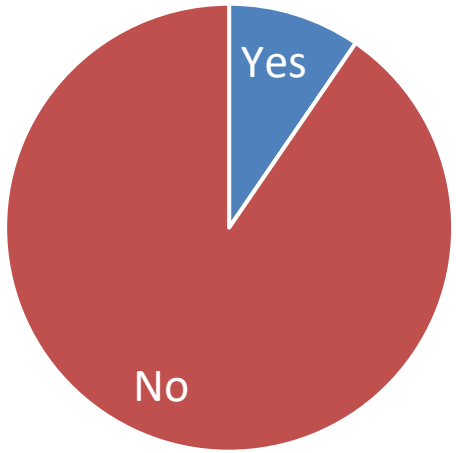
Challenges of Using VirusTotal

- Q1: When VirusTotal labels are trustworthy?
- Q2: How to aggregate labels from different engines?
- Q3: Are different engines equally trustworthy?

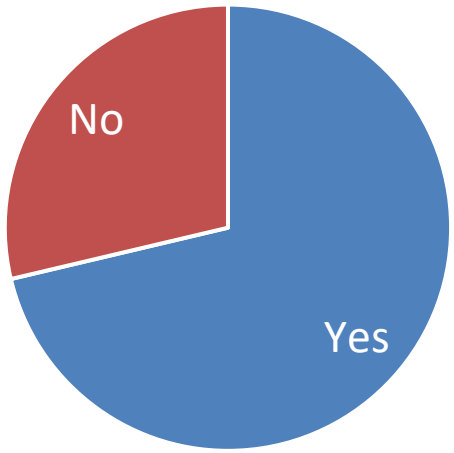
Literature Survey on VirusTotal Usages

- Surveyed 115 top-tier conference papers that use VirusTotal
- Our findings:
  - Q1: rarely consider label changes
  - Q2: commonly use threshold-based aggregation methods
  - Q3: often treat different VirusTotal engines equally

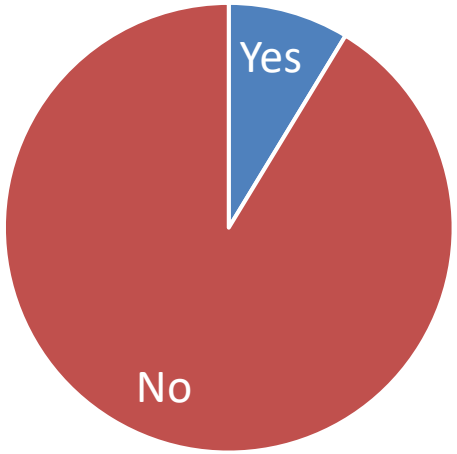
Consider Label Changes



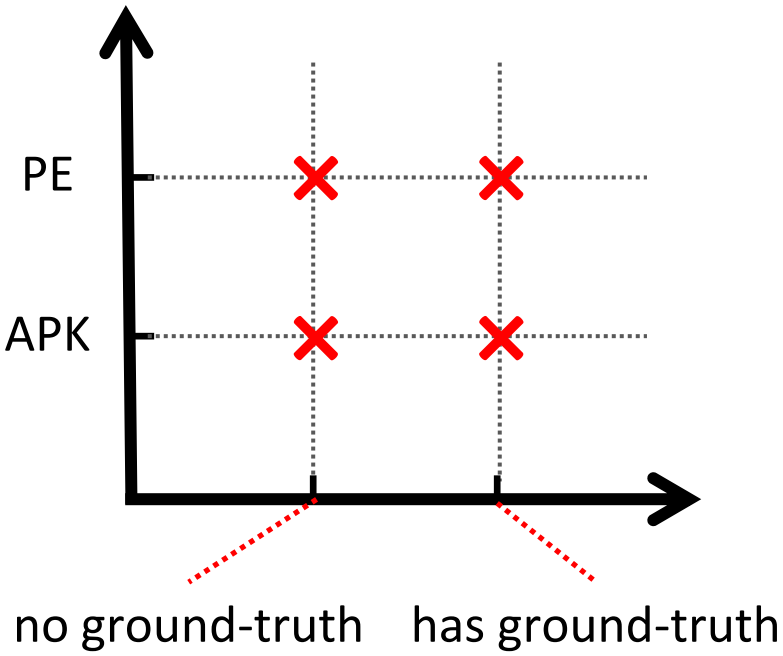
Threshold-Based Method



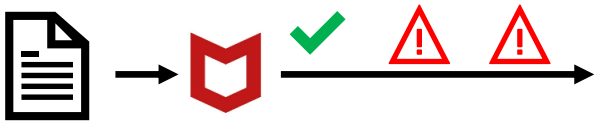
Reputable Engines



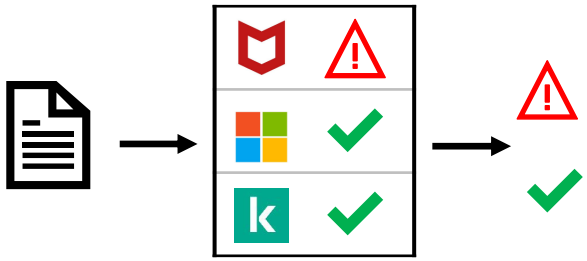
Overview



- Q1: the impact of label changes (label flips)

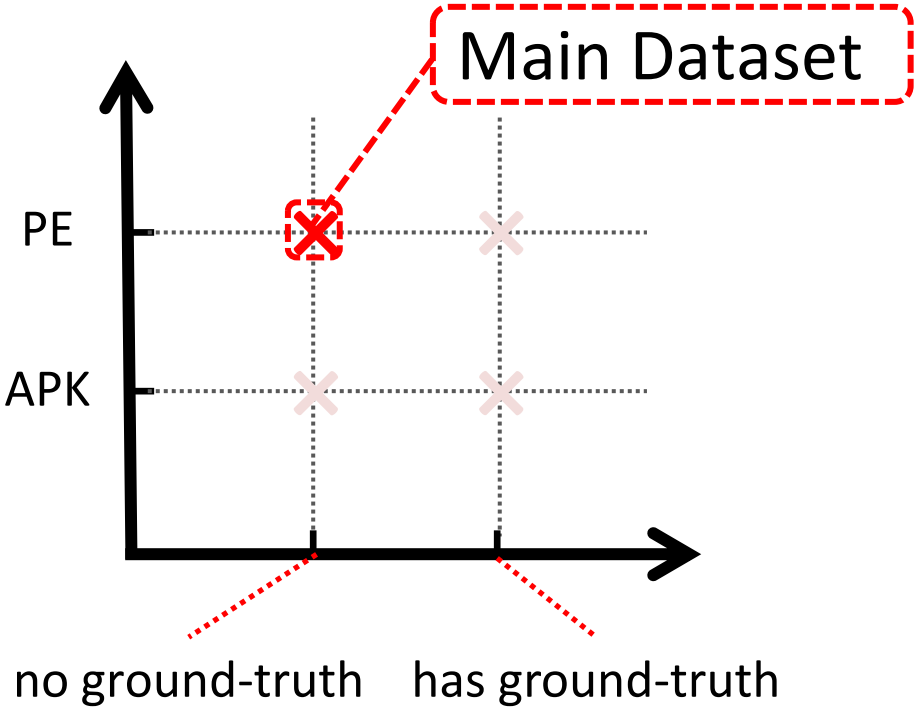


- Q2: threshold-based label aggregation methods



- Q3: the correlation between VirusTotal engines

Outline



- Q1: the impact of label changes (label flips)

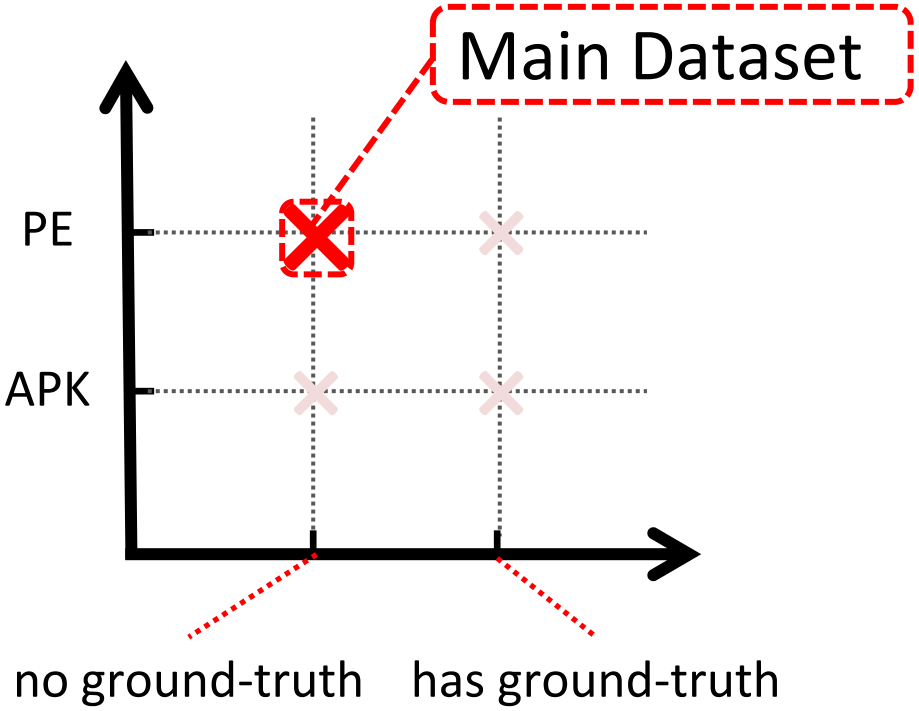
- Q2: threshold-based label aggregation methods

- Q3: the correlation between VirusTotal engines

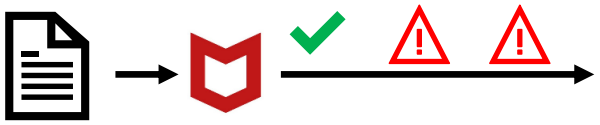
Data Collection of the Main Dataset

- We chose "**fresh**" files without prior VirusTotal history
  - Sampled 14,423 files submitted for the first-time on 08/31/2018
  - Roughly half were labeled as "benign" by all engines on day-1
  - The rest were labeled as "malicious" by at least 1 engine on day-1
- We collected "**daily**" VirusTotal labels over one year
  - Use `rescan` API to force VirusTotal to scan the samples everyday
  - Data collection window: 08/31/2018 – 09/30/2019
- Data Preprocessing
  - 341+ million data points from 65 engines

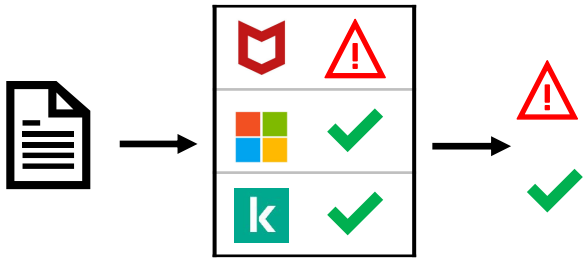
Outline



- Q1: the impact of label changes (label flips)



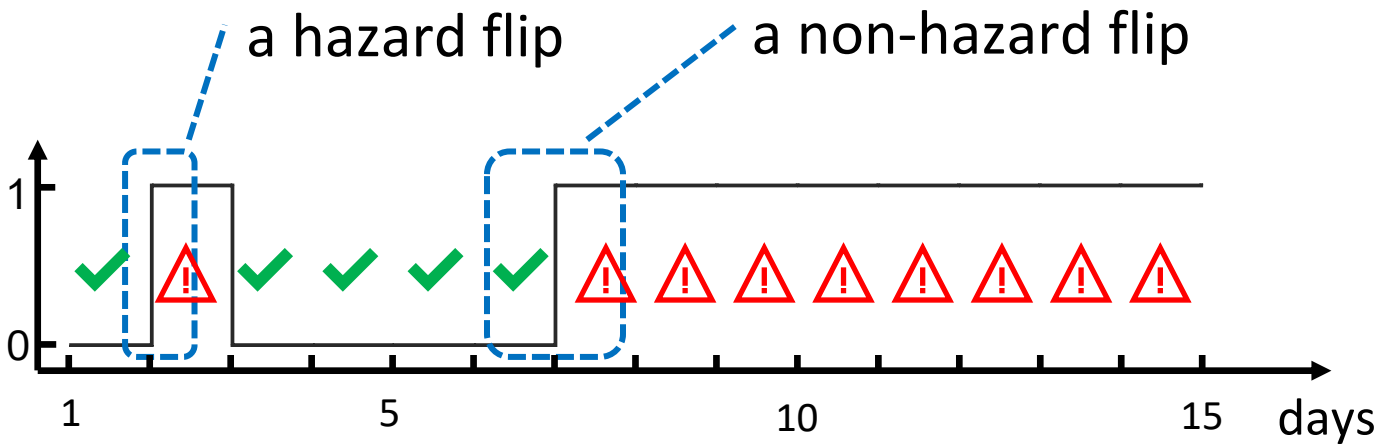
- Q2: threshold-based label aggregation methods



- Q3: the correlation between VirusTotal engines

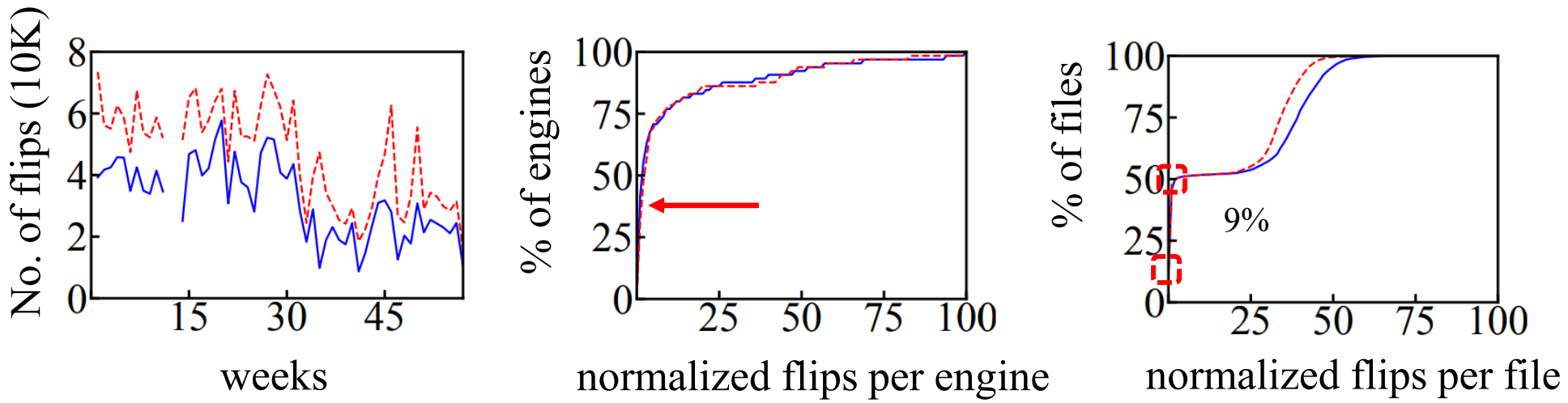
Label Change or Flip

- We model the label dynamics by sequences of “0” and “1”
  - ✔ (benign): 0 ⚠ (malicious):1



- A Flip: 0 → 1 or 1 → 0
  - hazard flip: temporary, lasts only one day
  - non-hazard flip: long term, lasts at least two days

Characteristics of Flips



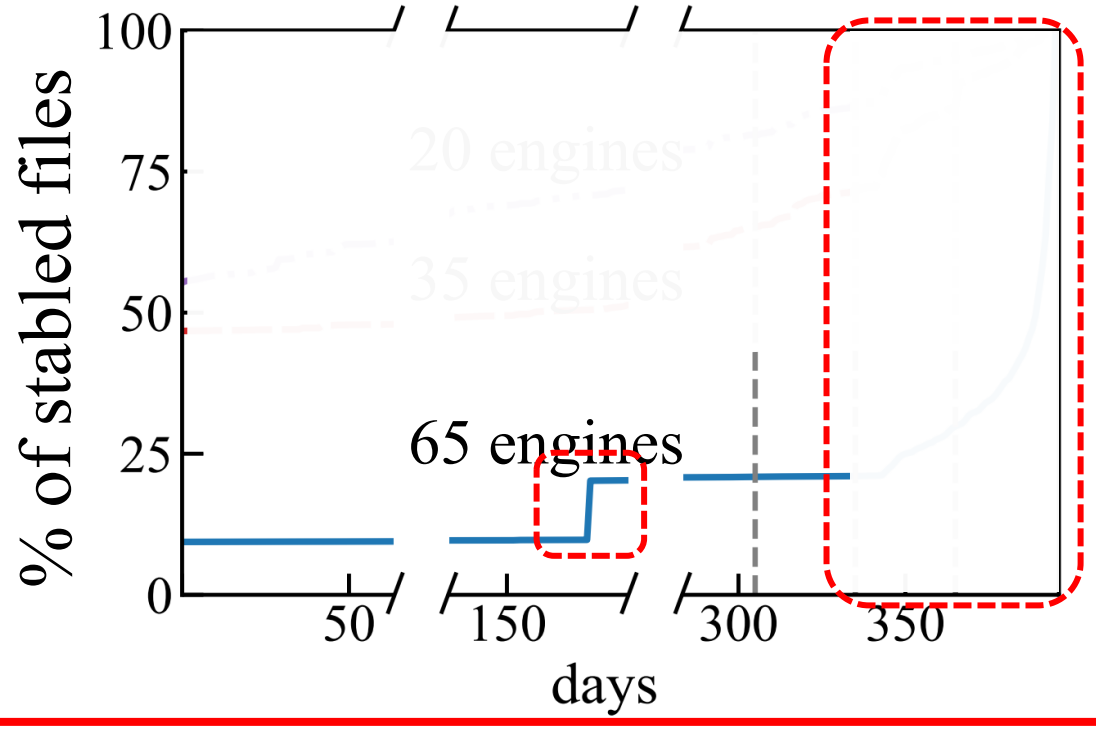
hazard flips

all flips

Both flips and hazard flips widely exist across scan dates, engines and files.

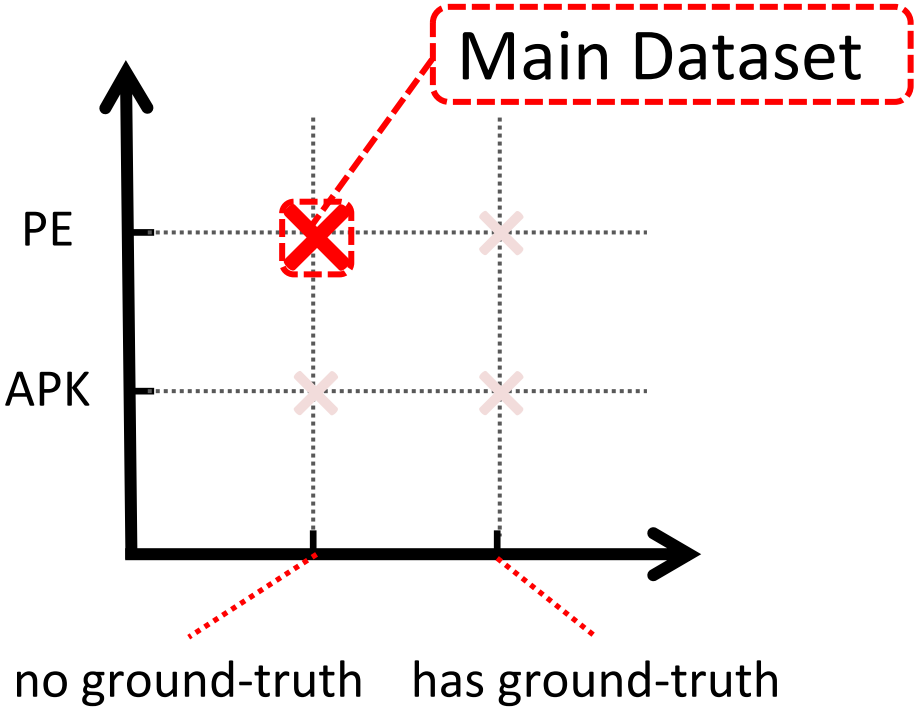
Individual Label Stabilization

- How long to wait for a file's labels to become stable?
- Stable file: all engines' labels on the file do not change any more



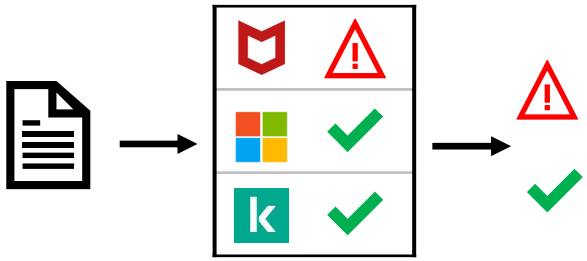
Waiting for longer time does not guarantee to have more stable files.

Outline



- Q1: the impact of label changes (label flips)

- Q2: threshold-based label aggregation methods



- Q3: the correlation between VirusTotal engines

Aggregated Label Stabilization

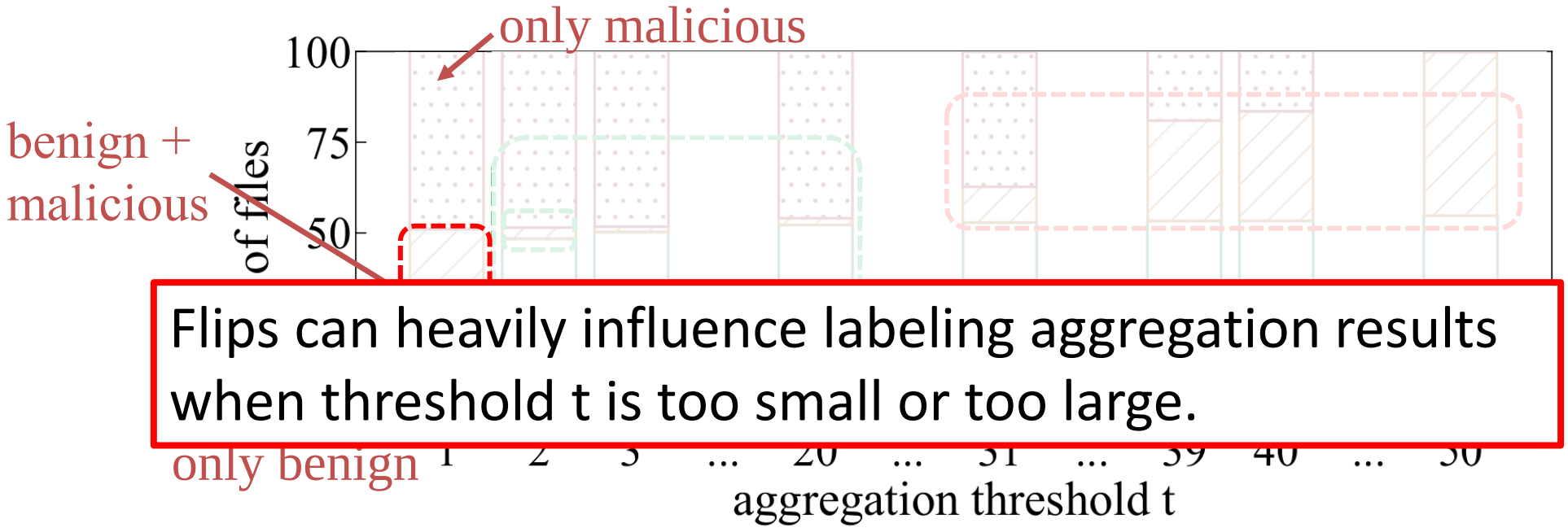
- Many researchers use a threshold (t) to aggregate engines' labels
  - A file is considered as malicious, when ≥ t engines detect the file
- How flips impact this aggregation policy?
  - Influenced files: files with both benign and malicious aggregated labels
  - Measure % of influenced files for different t

Day	1	2	3	4	5	…
McAfee	1	1	1	1	0	…
Microsoft	1	1	1	0	0	…
Kaspersky	0	1	0	0	0	…
… (62 engines)		… (all 0)				
Aggregated labels	1	1	1	0	0	…

(t = 2)

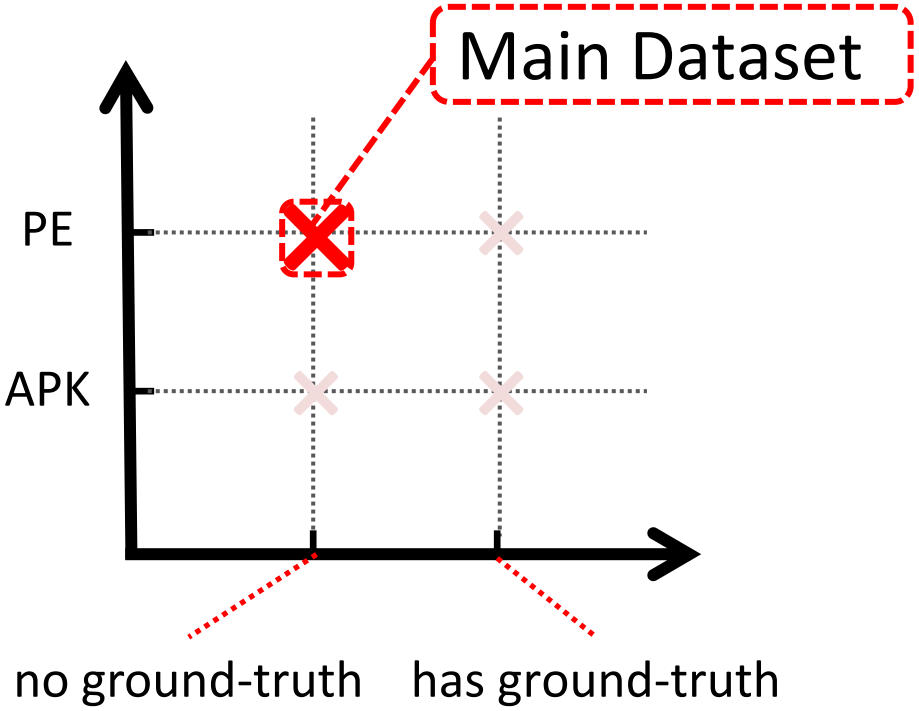
Aggregated Label Stabilization

- Many researchers use a threshold (t) to aggregate engines' labels
  - A file is considered as malicious, when ≥ t engines detect the file
- How flips impact this aggregation policy?
  - Influenced files: files with both benign and malicious aggregated labels
  - Measure % of influenced files for different t



Flips can heavily influence labeling aggregation results when threshold t is too small or too large.

Outline



- Q1: the impact of label changes (label flips)

- Q2: threshold-based label aggregation methods

- Q3: the correlation between VirusTotal engines

Equally trustworthy?

Temporary Labeling Similarity

- How to compute the similarity between engines A and B?
  - Compute the similarity between the two labeling sequences for each file
  - Compute the average sequence-level similarity over all the files
- An example for sequence-level similarity

engine A on file X: (0 1 0 0 0 0 0)(0 0 0 0 0 0 0)0 1 0 0 0 …

(1, 1, 1, 5, 0, 0, 0, 7, 7, . . .)

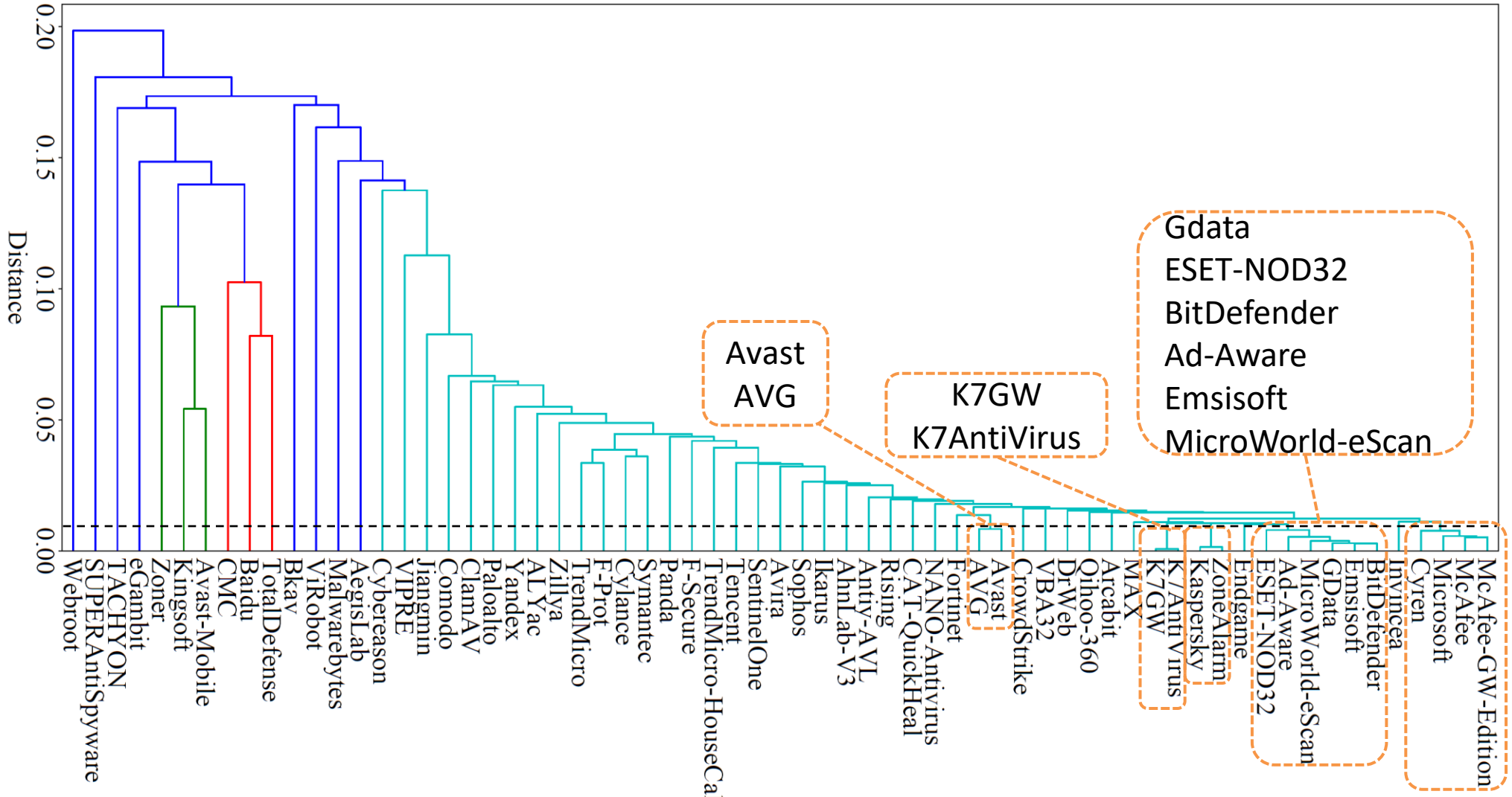
engine B on file X: 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 …

(0, 0, 0, 7, 1, 1, 1, 4, …)

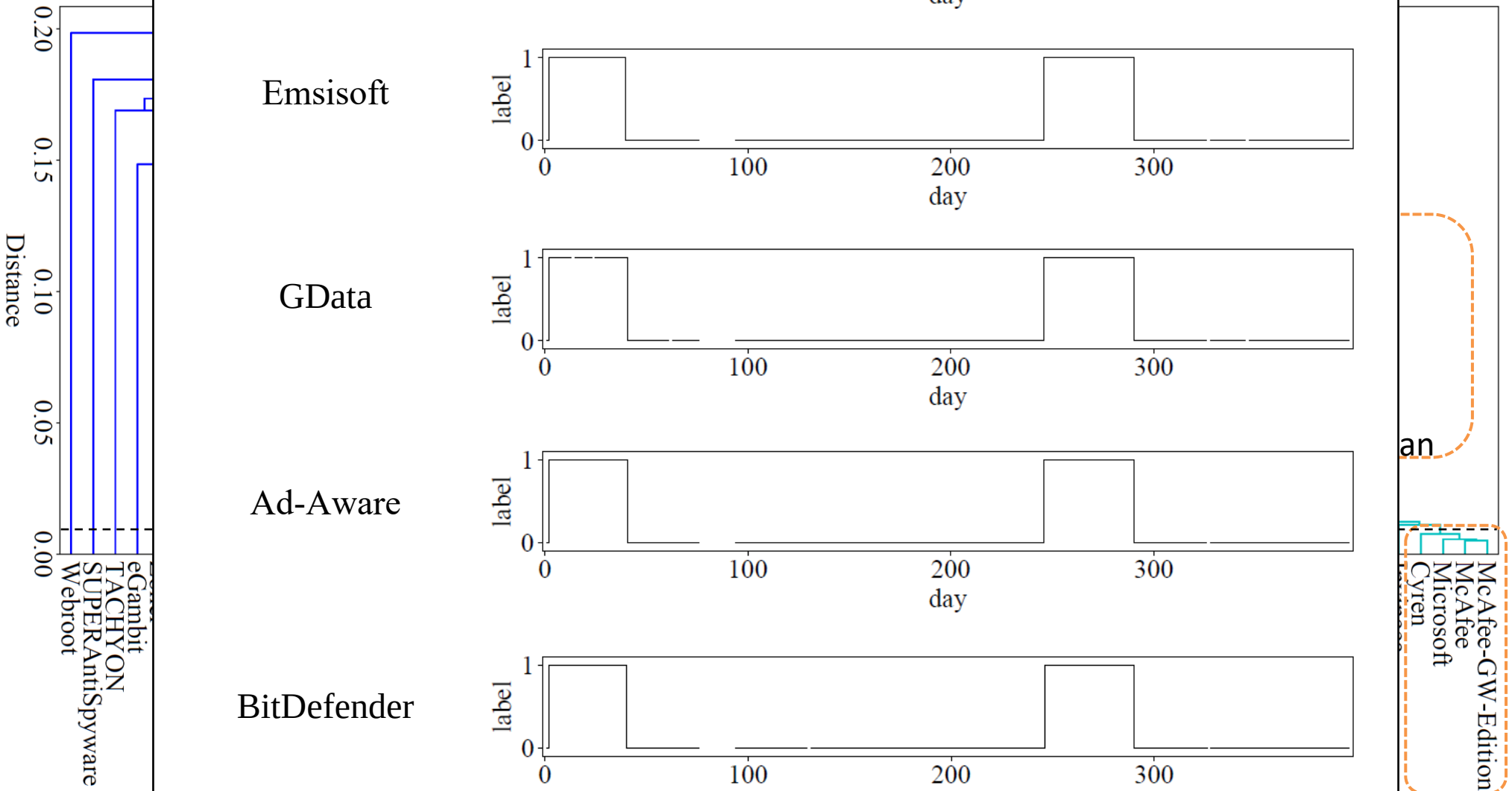
cosine

0.87

Label Correlations Between Engines

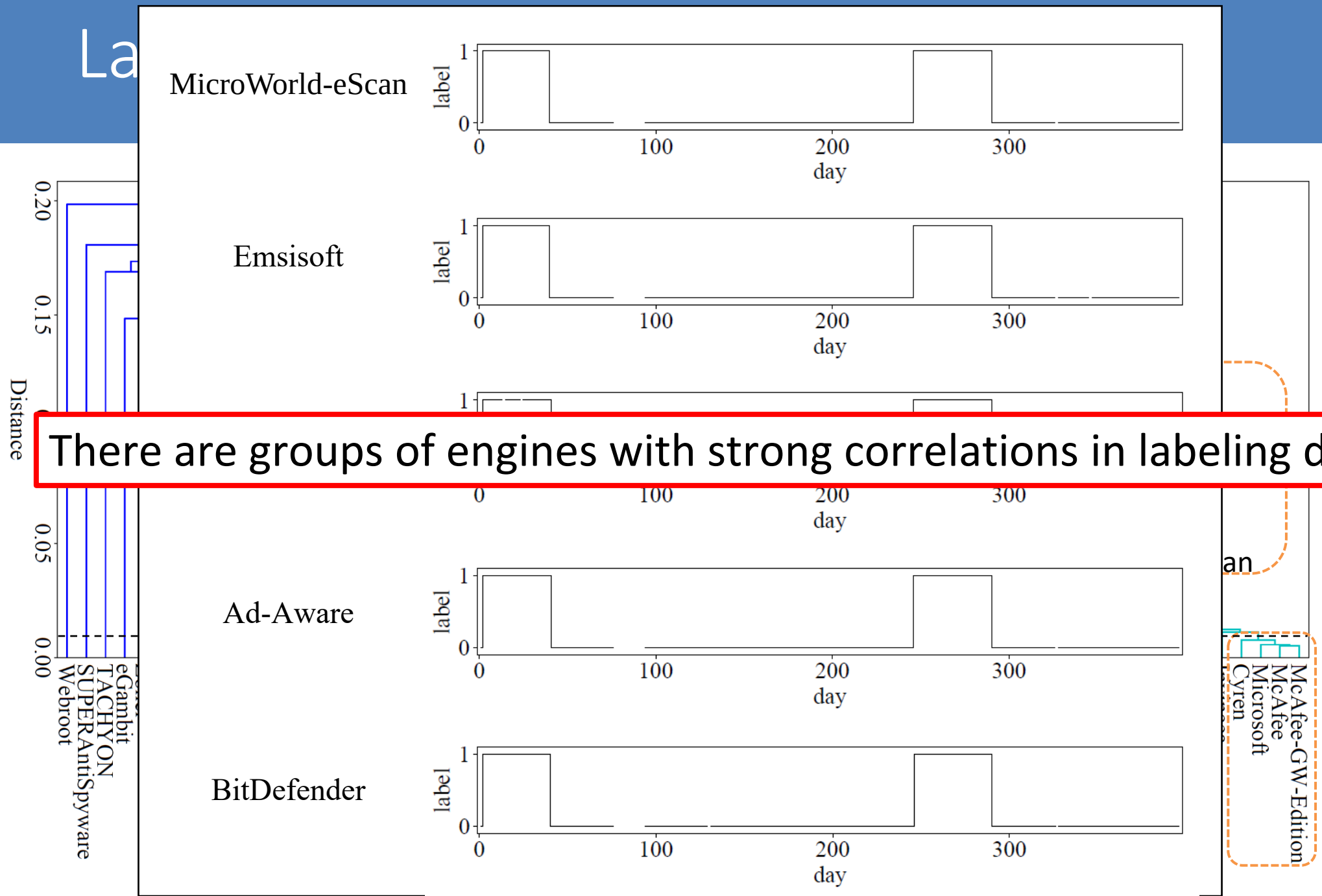


La


Distance
0.20
0.15
0.10
0.05
0.00
Webroot
SUPERAntiSpyware
TACHYON
eGambit
Cyren
Microsoft
McAfee
McAfee-GW-Edition

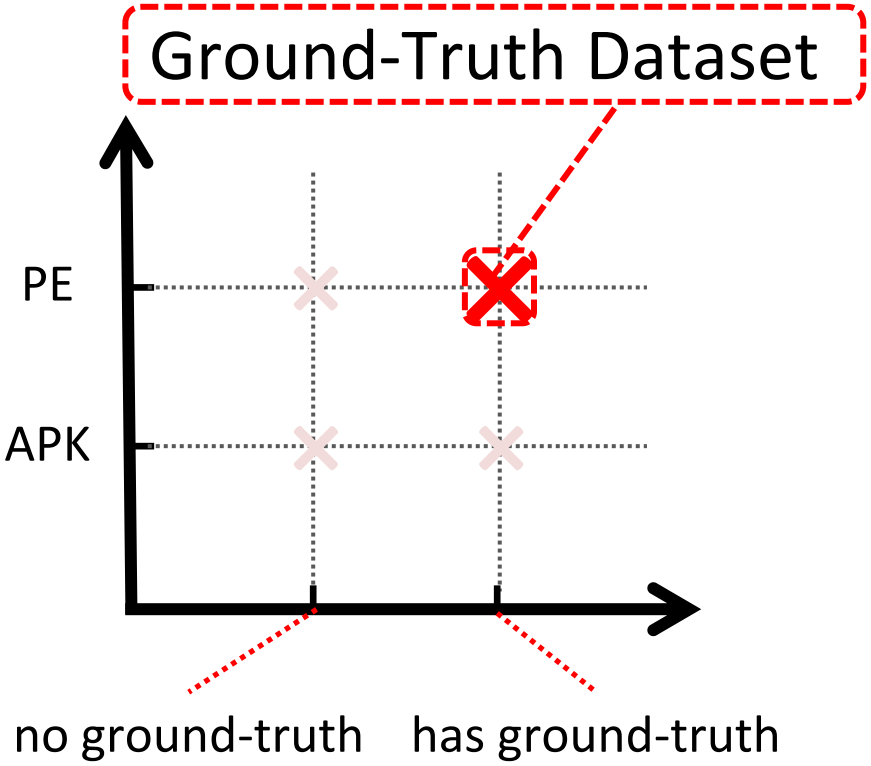
MicroWorld-eScan
Emsisoft
GData
Ad-Aware
BitDefender
label
1
0
0
100
200
300
day

La

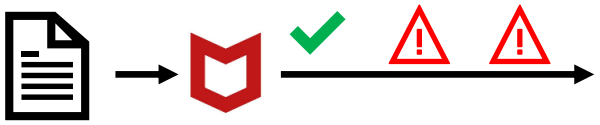


There are groups of engines with strong correlations in labeling decisions.

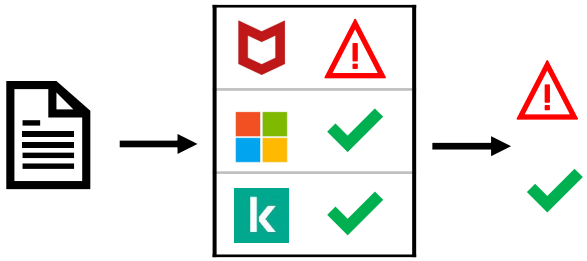
Outline



- Q1: the impact of label changes (label flips)



- Q2: threshold-based label aggregation methods



- Q3: the correlation between VirusTotal engines

Ground Truth Dataset

- How we create "**fresh**" ground-truth samples?
  - Obfuscating ransomware to create malware
  - Obfuscation + compiling open-source software to create goodware
- Findings:
  - Obfuscation brings many false positives
    - Even for high-reputation engines
  - $3 \leq t \leq 15$ can produce good aggregation results
    - As long as the benign files are not obfuscated
  - Inconsistency exists between the desktop and the VirusTotal versions

More results in our paper…

Conclusion and Takeaways

- A paper survey on how researchers use VirusTotal
- Data-driven methods to validate labeling methodologies
- Takeaways and suggestions
  - Data preprocessing
    - Submit the same files in 3 consecutive days to detect hazards
    - No need to wait over long time
  - Threshold-based label aggregation
    - Stable: when t is within a reasonable range (2-20)
    - Correctness: t = 3 to 15 when benign files are not obfuscated
  - Correlation and causality exists between engines
  - High-reputation Engines are not always accurate

Thank you!

- Also thanks to my collaborators
- Contact
  - sfzhu@psu.edu
- Artifact
  - https://sfzhu93.github.io/projects/vt/index.html