AndroOBFS: Time-tagged Obfuscated Android Malware Dataset with Family Information

Saurabh Kumar[1], Debadatta Mishra[1], Biswabandan Panda[2], and Sandeep K. Shukla[1]

[1]Indian Institute of Technology Kanpur

[2]Indian Institute of Technology Bombay

MSR 2022

Motivation

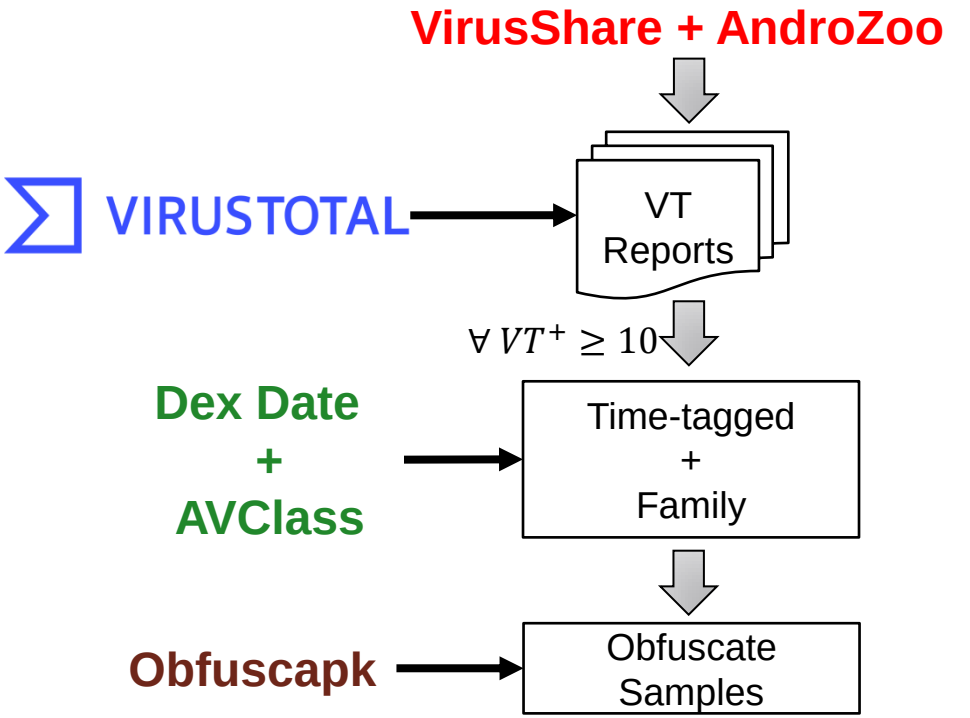
- Malware analysis system require labeled dataset
  - Non-obfuscated (Drebin, AMD, RmvDroid, etc..)
  - Obfuscated (PRAGuard)
- PRAGuard dataset
  - Outdated samples till March 2013
  - Does not contains family information
- Needed a new obfuscated malware dataset
  - Tagged with time
  - With family information

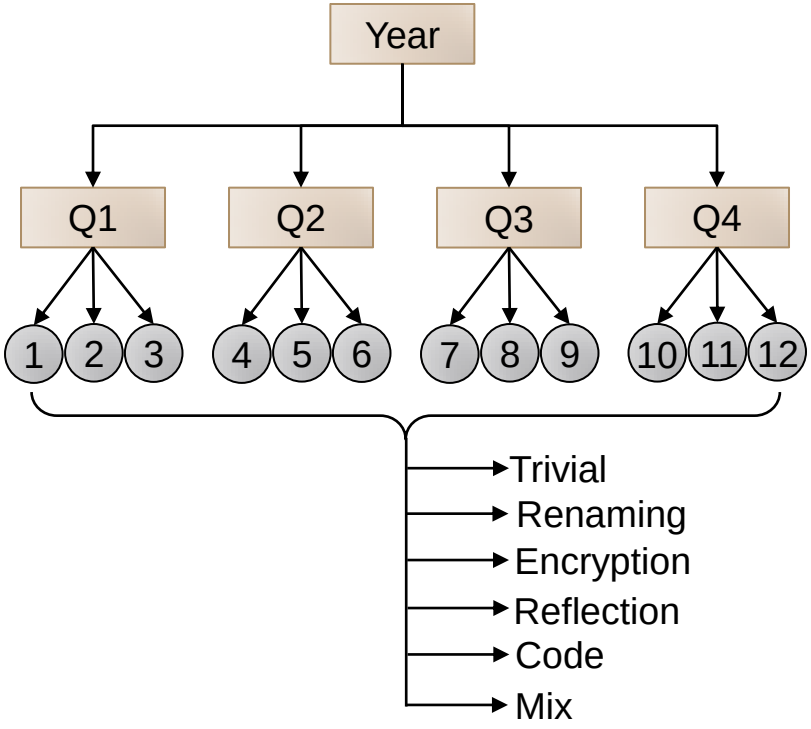
AndroOBFS Dataset

- Samples collected from
  - AndroZoo and VirusShare
  - For year 2018, 2019 and 2020
- Tools used
  - AVClass: Labeling with family
  - Obfuscapk: To obfuscate
- 16279 obfuscated samples in six categories
  - 14579 Familial
  - Spread across 158 families

Sources	Year	#Samples	
		Obfuscated	Obfuscated (#Family$_{\geq 2}$)
AndroZoo + VirusShare	2018	6525	5794 (136)
	2019	8313	7505 (94)
AndroZoo	2020	1441	1280 (44)
Total		**16279**	**14579 (158)**

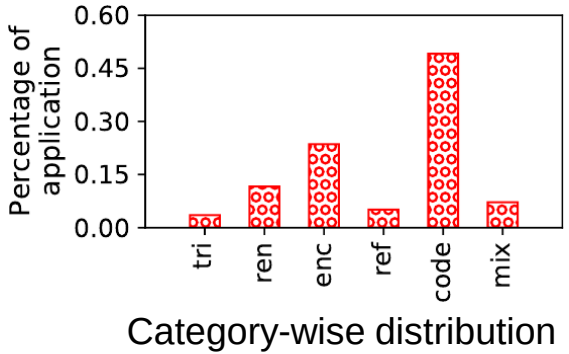
Dataset Creation Process





Malware Distribution

- 16279 samples distributed among six categories
  - ~50% fall under code obfuscation
  - ~23% in encryption



Usage Scenarios

- Designing robust malware detection and classification system
- Studying the efficacy of existing analysis systems
- Temporal/longitudinal study of an analysis system

Conclusion

- 16279 obfuscated samples in six categories
- 14579 familial obfuscated samples distributed among 158 families
- Tagged with time
- Spanning over three years from 2018 to 2020

Thank You