Adaptive Semantics-Aware Malware Classification

Bojan Kolosnjaji, Apostolis Zarras, Tamas Lengyel, George Webster, Claudia Eckert
{kolosnjaji|zarras|tklengyel|webstergd|eckert}@sec.in.tum.de

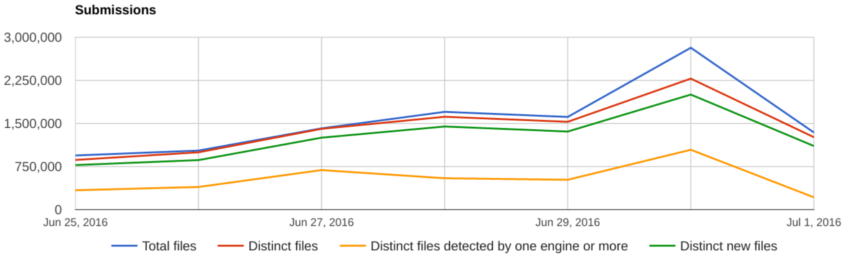
Lehrstuhl für Sicherheit in der Informatik
Prof. Dr. Claudia Eckert
Technische Universtität München

08.07.2016

What is the problem?

- **Millions** of newly discovered malware samples per day

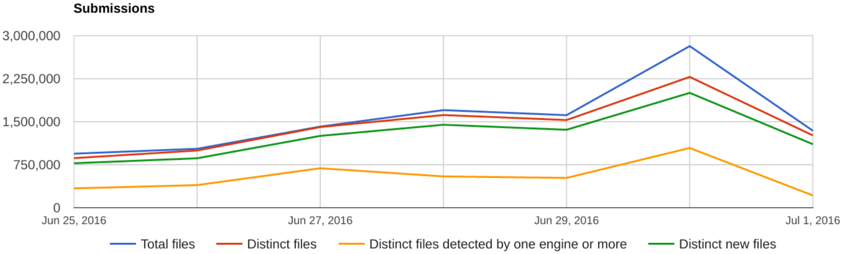
(Graph from: https://www.virustotal.com/en/statistics/)



What is the problem?

- **Millions** of newly discovered malware samples per day

(Graph from: https://www.virustotal.com/en/statistics/)



- Signature-based systems are not enough, **variance** between samples

What do we need?

- We need **statistical data-driven** approaches
- We must use **information retrieval** methods to leverage data
- We have to make analysis methods **adaptive** and **scalable**

What approaches exist?

- Multiple research efforts in malware detection
    - Modeling static code features
    - Sequencing behavioral traces
- One-class, multiclass classification, anomaly detection, clustering
- SVM, KNN, LDA, Neural Network...

What approaches exist?

- Many platforms for big data processing
  - Polonium (SIGKDD 2010)
  - BitShred (CCS 2011)
  - BinaryPig (BlackHat USA 2013)
  - ...
- Focuses on big data infrastructure and less on modeling

Our approach

We combine

- **Semantics-awareness**
  - We use *topic modeling* in order to extract high-level information from system call sequences and characterize malware behavior
- Semi-supervised Learning
- Nonparametric Learning
- Combination of static and dynamic data

Our approach

We combine

- Semantics-awareness
- **Semi-supervised Learning**
  - We combine a small amount of labeled data with a large set of unlabeled samples
- Nonparametric Learning
- Combination of static and dynamic data

Our approach

We combine

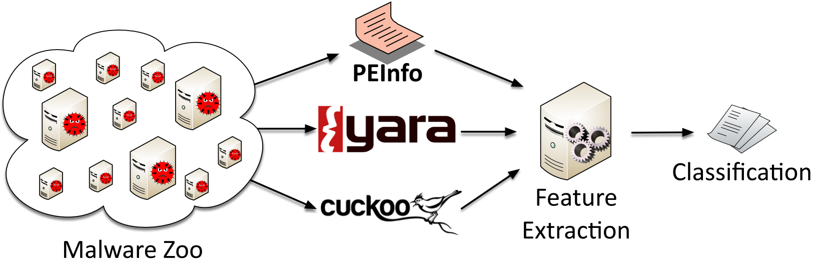
- Semantics-awareness
- Semi-supervised Learning
- **Nonparametric Learning**
    - We maintain the accuracy of our model during large malware influxes
- Combination of static and dynamic data

Our approach

We combine

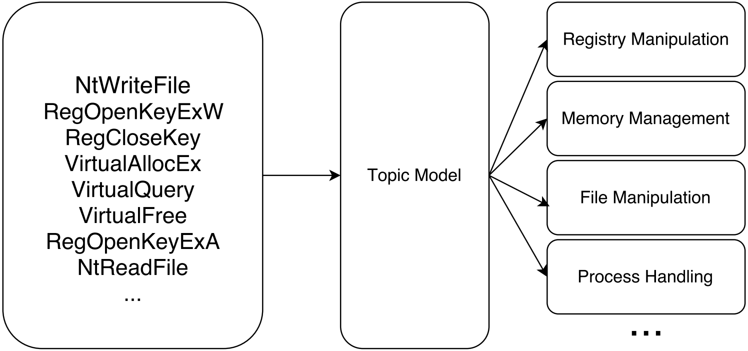
- Semantics-awareness
- Semi-supervised Learning
- Nonparametric Learning
- **Combination of static and dynamic data**
    - Separate machine learning methods on static code properties and behavioral sequential data

System Description



Topic Model

- Topic model assumption: Most of the information corresponds to a small number of topics



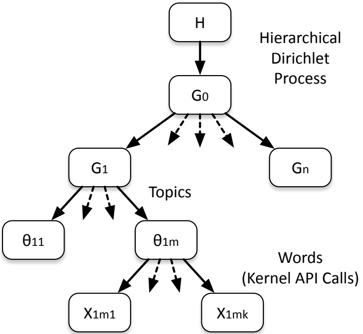
NtWriteFile
RegOpenKeyExW
RegCloseKey
VirtualAllocEx
VirtualQuery
VirtualFree
RegOpenKeyExA
NtReadFile
...

Topic Model

Registry Manipulation
Memory Management
File Manipulation
Process Handling
...

Topic Model

- Hierarchical Dirichlet Process[1]: nonparametric, **flexible** (adaptive) for retraining

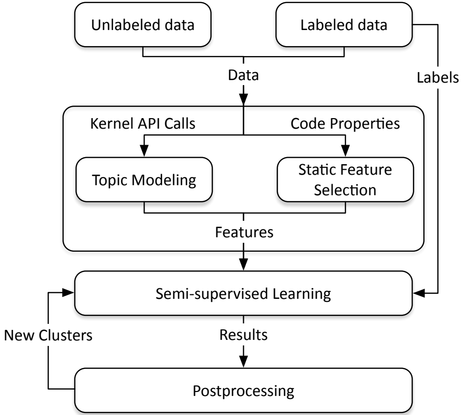


---

[1] Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M. (2006). Hierarchical dirichlet processes. Journal of the american statistical association.

Semi-supervised Learning

- Label propagation: propagate labels to unlabeled samples



Evaluation

- Sample set: 2k labeled, 15k unlabeled samples

- We create 10 classes based on AV signatures from VirusTotal

- 3-fold Crossvalidation

Performance

- High improvement with respect to parametric modeling (LDA), automatic determination of the number of topics (up to 50% improvement)
- Over 4% improvement when combining the topic model with static features, compared to using single data sources
- (97.5%) precision and (97.2%) recall using a semi-supervised approach
- Better average results than in related approaches

Performance

- Open world vs. closed world - small drop in accuracy (less than 10%)
- Linear growth in training time using approximate inference
- Topics with semantic relevance

Registry manipulation	**Memory management**	**File manipulation**	**Process Handling**
NtWriteFile	VirtualAllocEx	NtReadFile	OpenProcess
RegOpenKeyExW	VirtualQueryEx	NtWriteFile	ReadProcessMemory
RegCloseKey	VirtualQuery	NtDelayExecution	WriteProcessMemory
RegEnumValueW	VirtualFreeEx	LdrGetProcedureAddress	CloseHandle
RegQueryValueExW	VirtualFree	NtSetInformationFile	LocalAlloc
LdrGetProcedureAddress	LdrGetProcedureAddress	NtCreateFile	LocalFree
RegOpenKeyExA		NtQueryDirectoryFile	

Limitations

- Model more complex hierarchy of topics
- Include system call arguments and sequence-aware information
- Expand to more features and malware samples

Conclusion

- We create a machine learning-based **malware classification** model that is:
  - Semantics-aware
  - Semi-supervised
  - Nonparametric
  - Multi-view (static+dynamic data)
- We capture the essential properties of malware behavior
- We obtain improvements in classification performance