Large Scale Studies: Malware Needles in a Haystack

Giovanni Bertão[1,3], Marcus Botacin[2], André Grégio[2], Paulo Lício de Geus[1]

[1]University of Campinas (UNICAMP)
{bertao, paulo}@lasca.ic.unicamp.br

[2]Federal University of Parana (UFPR)
{mfbotacin, gregio}@inf.ufpr.br

[3]Bolsista PIBIC-CNPq

Topics

1. Introduction
2. Motivation & Related Work
3. Methodology
4. Analysis Results
5. Discussion
6. Final Remarks

Topics

1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. Final Remarks
   - Conclusion
   - Acknowledgments
   - Questions

Malware Analysis Approaches

Coarse-Grained

- Highlight major aspects.
- Discard sample details.

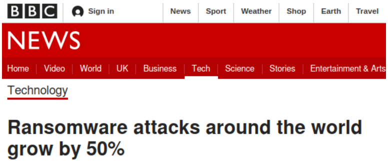
Fine-Grained

- Focus on implementation details.
- Don't state the risk of such sample in the overall scenario.

Topics

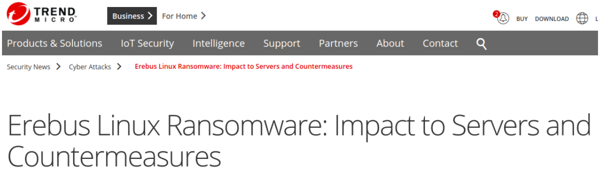
1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. Final Remarks
   - Conclusion
   - Acknowledgments
   - Questions

Malware Studies



Ransomware attacks around the world grow by 50%

Figure: **Coarse-Grained**
BBC: `https://www.bbc.com/news/technology-39730407`



Erebus Linux Ransomware: Impact to Servers and Countermeasures

Figure: **Fine-Grained**
Trend Micro: `https://bit.ly/2PaSPDC`

Topics

1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. Final Remarks
   - Conclusion
   - Acknowledgments
   - Questions

Related Work

Bayer (Windows)

- A view on current malware behaviors.

Lindorfer (Android)

- Andrubis – 1,000,000 apps later: A view on current android malware behaviors.

Topics

1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. Final Remarks
   - Conclusion
   - Acknowledgments
   - Questions

Building the Dataset

Dataset Composition

- Malware repositories and blacklists crawled daily.
- 135,000 unique malware samples collected from Malshare database.
- Only Windows samples.
- Samples submitted to static, dynamic and network analysis procedures.

Topics

1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. Final Remarks
   - Conclusion
   - Acknowledgments
   - Questions

Analysis Method

Static Analysis

- Presence of packers.
- Anti-analysis techniques.
- Anti-virus detection.

Dynamic and Network Analysis

- Logs from BehEMOT, an internal sandbox solution.

Topics

1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. Final Remarks
   - Conclusion
   - Acknowledgments
   - Questions

Large Scale Support

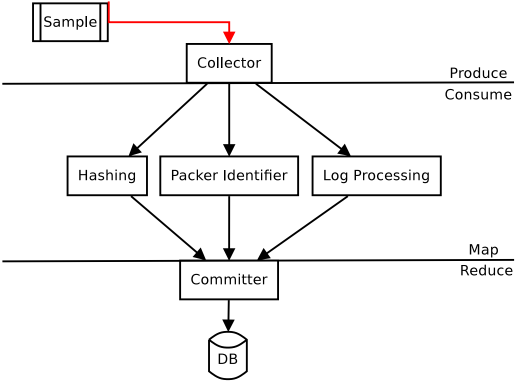


Figure: Parallel Processing Architecture. Samples are independently analyzed and their results are stored on a centralized database.

Time Elapsed

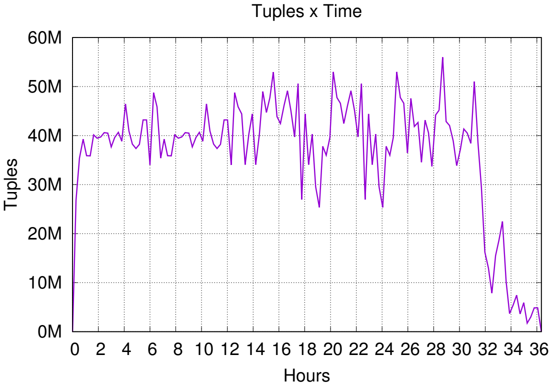


Figure: Parallel Processing Time. The complete analysis took 36 hours.

Topics

1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. Final Remarks
   - Conclusion
   - Acknowledgments
   - Questions

Malicious Behavior (Static Analysis)

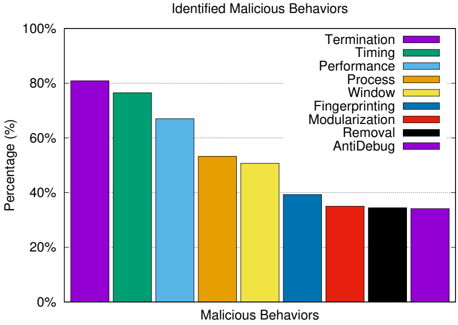


Figure: Identified Malicious Behaviors. We identified multiple, distinct malicious behaviors during samples executions, such as AV analysis evasion using `Timing` delays, measuring `Performance` overhead due to monitoring and the `Termination` of security solutions to avoid detection.

Prevalent Signatures

Table: **Most Prevalent Signatures.** Compilers are more prevalent than packers.

Signature	Type	Occurrence (%)
Microsoft	Compiler	37.95%
Nullsoft PIMP	Installer	25.51%
Borland Delphi	Compiler	15.06%
UPX	Packer	4.23%
MSLHR	Packer	2.25%
PEcompact	Packer	1.66%

Malicious Behavior (Dynamic Analysis)

Table: **Dynamic Analysis.** Identified Malicious Behaviors.

Subsystem	Operation	Samples (%)	Target	Samples (%)
File Subsystem	Create Files	91.56	Internet Explorer	10.14%
	Read Files	89.18%	.DLL	86.80%
			Internet Explorer	7.01%
			.SYS	1.26%
	Write Files	81.74%	.EXE	46.20%
			.DLL	31.62%
			Internet Explorer	<0.01%
			Host	0.00%
	Delete Files	62.45%	Internet Explorer	0.00%
Process Subsystem	Create Process	22.84%		
	Delete Process	23.38%		
Registry Subsystem	Set Registry Values	74.73%	Proxy	68.36%
			Autorun	5.66 %
	Delete Registry Values	55.43%		

Anti-Analysis Techniques

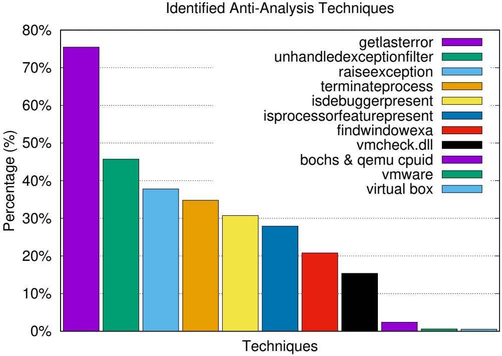


Figure: Identified Anti-Analysis Techniques. Samples employ anti-analysis techniques to avoid being inspected during sandbox execution. We identified techniques aimed to detect the presence of debuggers and virtual-machines.

Network Analysis — Protocols Distribution

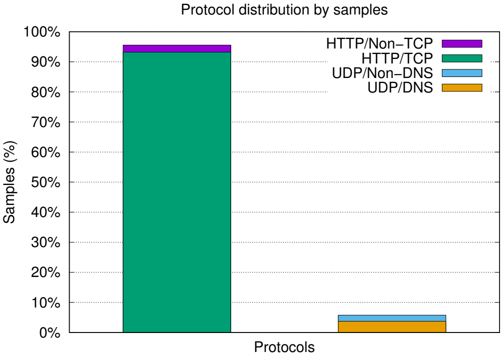


Figure: **Protocol usage distribution by sample.** HTTP over TCP and DNS over UDP are the prevalent communication channels.

Network Analysis — Domains Contacted

Table: **Most Contacted Domains**. We observe the presence of cloud providers among the most accessed domains.

Domain	**Accesses (%)**
Cloudfront	4.39%
Amazonaws	3.32%
Kirov	3.20%
Kerch	3.11%
Comcast	1.75%
Akamaitechnologies	1.48%
Sbcglobal	1.10%
Broadband	1.08%

Network Analysis — TLDs

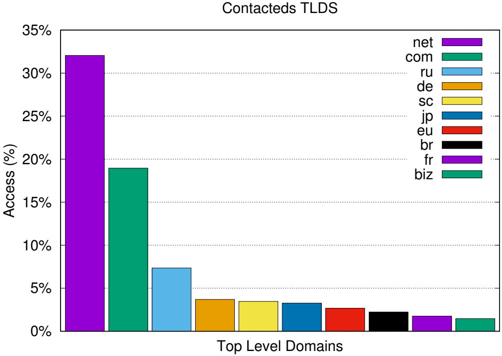


Figure: **Most Accessed TLDs distribution.** Generic domains are prevalent and country-specific domains are well distributed, thus showing that malware creators are ready to exploit vulnerabilities in multiple countries.

Network Analysis — DNS Resolvers

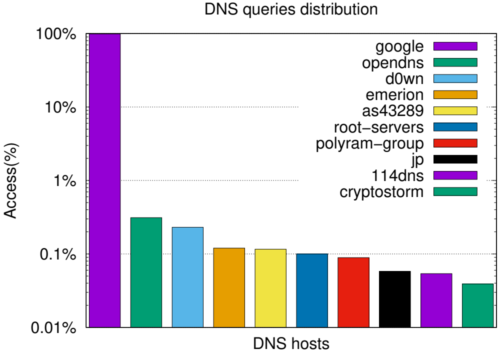


Figure: **DNS Resolvers Distribution.** Default sandbox DNS (`google`) was the most used one.

Topics

1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. Final Remarks
   - Conclusion
   - Acknowledgments
   - Questions

Keys Modification

Table: **Modified AutoRun Keys.** Whereas coarse-grained analysis identified that $\approx 5.66\%$ of samples write in an `Autorun` key, the fine-grained analysis specified their location.

Key	**Samples(%)**
HKCU\ID\Software\Microsoft\Windows\CurrentVersion\Run	46.33%
HKCU\.DEFAULT\SOFTWARE\Microsoft\Windows\CurrentVersion\Run	$< 0.01\%$

Contacted IPs

Table: **Contacted IPs by Sample.** Coarse-grained analysis showed that, in average, each sample contacts 2 distinct IP addresses. Fine-grained analysis revelead that ransomware samples which spread through scanning contact many more IPs.

MD5 Hash	Number of Distinct IPs	Label
c1abb496deb7bd51a4ad2f8a43113b13	16386	Ransomware.Cerber
bc88096e7cc09f02f11deec35f84d5cd	16385	Ransomware.AWA
a801cdef09a61d3ba7969015a8bffec0	1	Ransomware.VirLock

HTTP requests

Table: **Prevalent HTTP Requests.** Coarse-grained analysis shows that HTTP payloads dominate TCP traffic. Fine-grained analysis shows that this is due to `Downloader` samples.

MD5 Hash	**Total of Distinct HTTP Request**	**Label**
ede13f40a96a8b6e5de1029200c0b15e	394	Downloader
e5f4116d08c343623d5ee3af5553cbee	353	Downloader
47a328b0b903bb68147facc3a084172c	310	Downloader
28c4e2a48d9ddfffa01a943ca1ba1262	304	Downloader

DNS Resolvers

Table: **DNS queries resolvers distribution.** Coarse-grained analysis shows that DNS queries dominate UDP traffic. Fine-grained analysis reveals that this is due to Bot samples.

Query	Contacted(%)	Description
bmp.pilenga.co.uk.	12.29%	Hijacked Subdomain - Andromeda Botnet
tgr.tecnoagenzia.eu.	5.96%	Hijacked Subdomain - Andromeda Botnet
tds.repack.it.	2.48%	Andromeda Botnet
rxxl.tecnoagenzia.eu.	2.06%	Andromeda Botnet
and31.blllaaaaaazblaaa1.com.	0.92%	Andromeda Botnet

Topics

1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. Final Remarks
   - Conclusion
   - Acknowledgments
   - Questions

Drawing Panoramas

Dataset Characterization

- Mainly executables.
- Few libraries.
- Rely on system native libraries.
- Few external libraries.
- GUI usage.
- Strong presence of system interactions (file and registry creation/deletion).

Panorama Comparison

Brazilian Panorama

- Mix of binaries and DLL.
- Rely on system native libraries.
- Mostly as background activity.

Topics

1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. Final Remarks
   - Conclusion
   - Acknowledgments
   - Questions

Identifying project decisions

Coarse-Grained

- In average, each sample contacts 2 different IP address.

Fine-Grained

- **c1abb496deb7bd51a4ad2f8a43113b13** contacts 16386 IPs.
- **a801cdef09a61d3ba7969015a8bffec0** contacts only 1 IP.

Topics

1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. Final Remarks
   - Conclusion
   - Acknowledgments
   - Questions

Conclusion

- A coarse-grained analysis procedure is the only approach able to draw threat panoramas.
- A fine-grained analysis procedure is the only approach which enables individual samples characterization.
- Fine-grained and coarse-grained analysis approaches must be combined for increased threat understanding.

Topics

1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. Final Remarks
   - Conclusion
   - Acknowledgments
   - Questions

Acknowledgments

- CNPq
- Institute of Computing/Unicamp

LASCA
LABORATORY OF SECURITY AND APPLIED CRYPTOGRAPHY

Topics

1. Introduction
   - Analysis Type: Coarse-grained and Fine-grained
2. Motivation & Related Work
   - Motivation
   - Related Work
3. Methodology
   - Dataset
   - Processing the Data
   - Architecture
4. Analysis Results
   - Coarse-Grained
   - Fine-Grained
5. Discussion
   - Coarse-Grained
   - Fine-Grained
6. **Final Remarks**
   - Conclusion
   - Acknowledgments
   - **Questions**

Questions?

bertao@lasca.ic.unicamp.br