Prudent Practices for Designing Malware Experiments: Status Quo and Outlook

Rossow et al, IEEE S&P 2012

N. S. Rutherford
nathan.rutherford.2019@live.rhul.ac.uk

Information Security Group
Royal Holloway, University of London

Classic Paper Reading Group, 31st October 2019

Table of Contents

Terminology

Malware

- **Mal**icious Soft**ware**
- Cause deliberate harm to users, systems, and networks

Malware Families

- Way of classifying types of malware
- e.g:
  - Worms
  - Trojans
  - Spyware
  - Ransomware

Motivation

Malware-Execution [1] Datasets

- Observation of malware behaviour in a 'sandbox' environment
- Useful for generating ground truth characterisations of malware

These ground truths are regularly applied in scientific experiments

Data collection and analysis processes face many potential pitfalls

Lab results may not reflect real-world evaluations

[1] Also known as *Dynamic Analysis*

Contributions

1. Identify a number of pitfalls when designing malware experiments, and reason about the impact
2. Devised a number of guidelines that can control for these pitfalls
3. Assess these guidelines on a sample of 36 papers (40% of which are from top-tier venues)
4. Identify that these issues are universal across top-tier and less prestigious venues.

Table of Contents

Designing Prudent Experimentation

Group pitfalls that can arise into four categories

- Correct Datasets
- Transparency
- Realism
- Safety

Each of these "cornerstones of prudent experimentation" also outline more specific aspects to consider

Correctness

Goal

Ensure that datasets do not bias the results

1. Check if goodware samples should be removed from datasets
2. Balance datasets over malware families
3. Check whether training and evaluation datasets should have distinct families
4. Perform analysis with higher privileges than the malware's
5. Discuss and if necessary mitigate analysis artifacts and biases
6. Use caution when blending malware activity traces into benign background activity

Transparency

Goal

Provide detailed description of experiment setup and interpret results for objective replication

1. State family names of employed malware samples
2. List which malware was analysed when
3. Explain the malware sample selection
4. Mention the system used during execution
5. Describe the network connectivity of the analysis environment
6. Analyse the reasons for false positives and false negatives
7. Analyse the nature/diversity of true positives

Realism

Goal

Designing experiments that are reflective of how malware behaves *in the wild* to allow generalisation of findings

1. Evaluate relevant malware families
2. Perform real-world evaluations
3. Exercise caution generalizing from a single OS version, such as Windows XP
4. Choose appropriate malware stimuli
5. Consider allowing Internet access to malware

Safety

Goal

Mitigation from harm by highlighting the need for both implementation and discussion of containment policies within a malware sandbox

1. Deploy and describe containment policies

Table of Contents

Overview

Aim

To verify if these guidelines are useful, and analyse in which cases they would have significantly improved experiments in the existing literature

Surveyed 36 publications, 40% from 6 top-tier venues and the rest from less-prestigious venues

Defined subsets of *applicable papers*

Survey interpretation was split into three parts

- *Per-Guideline Analysis*
- *Per-Paper Analysis*
- *Top-Venue Analysis*

Survey Results

Criterion	Imp.	All Papers			Top-Venue Papers		
Correctness		App	Ukwn	OK	App	Ukwn	OK
Remove Goodware	●	9	0%	44%	4	0%	50%
Avoid Overlays	●	7	0%	29%	4	0%	0%
Balanced Families	◐	13	0%	54%	2	0%	50%
Separated Datasets	◐	8	0%	0%	1	0%	0%
Mitigated Artefacts/Biases	◐	36	0%	28%	14	0%	50%
Higher Privileges	◐	36	6%	75%	14	0%	86%
Transparency		App	Ukwn	OK	App	Ukwn	OK
Interpreted FPs	●	25	n/a	64%	9	n/a	89%
Interpreted FNs	●	21	n/a	48%	7	n/a	57%
Interpreted TPs	●	30	n/a	60%	11	n/a	55%
Listed Malware Families	◐	36	n/a	81%	14	n/a	86%
Identified Environment	◐	36	n/a	81%	14	n/a	79%
Mentioned OS	◐	36	n/a	64%	14	n/a	64%
Describe Naming	◐	32	n/a	50%	12	n/a	58%
Describe Sampling	○	16	n/a	81%	5	n/a	60%
Listed Malware	○	36	n/a	11%	14	n/a	7%
Describe NAT	○	30	n/a	10%	11	n/a	9%
Mentioned Trace Duration	○	36	n/a	64%	14	n/a	57%
Realism		App	Ukwn	OK	App	Ukwn	OK
Removed Moot Samples	●	16	0%	0%	5	0%	0%
Real-World FP Exp.	●	20	0%	50%	6	0%	67%
Real-World TP Exp.	●	20	0%	35%	6	0%	67%
Used Many Families	●	36	1/8/745		14	1/8/745	
Allowed Internet	◐	36	6%	75%	14	0%	79%
Added User Interaction	○	36	0%	3%	14	0%	0%
Used Multiple OSes	○	36	22%	19%	14	21%	29%
Safety		App	Ukwn	OK	App	Ukwn	OK
Deployed Containment	●	28	71%	21%	11	64%	27%

Figure: Overview of survey results; ● is a must, ◐ should be done, ○ is nice to have

Per-Category

Analysis

To what extent have specific guidelines been met?

Correctness

- Lack of goodware removal (over 50%) which may bias results

Transparency

- Experiment setups lacked sufficient detail to ensure replicability
- Majority of papers failed to interpret numeric results

Realism

- Minority of papers included real-world evaluation

Safety

- Most papers did not deploy or adequately describe containment policy

Per-Paper

Analysis

How many papers could have benefited from the application of these guidelines?

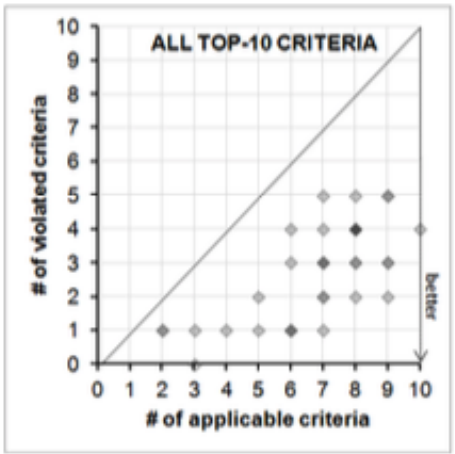

Figure: Guideline violations related to applicable criteria

- Shows a correlation between number of violated and number of applicable criteria
- Guidelines become increasingly important when designing more complex experiments

Top-Tier Venues

Analysis

Are experiments presented at top-tier venues more prudent than others?

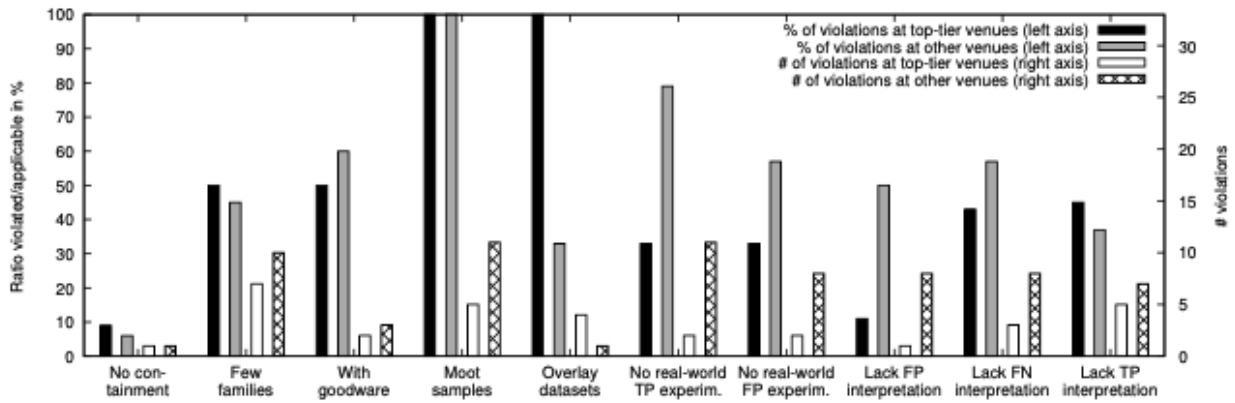

Figure: Violations at top-tier venues compared with other venues

- Violations generally remain comparable
- Top-tier papers present more real-world scenarios and interpretation of FPs

Table of Contents

Conclusions

Number of pitfalls in malware-execution driven experiments with relation to the scientific method

This could be improved with more effort to presentation (*transparency* guidelines)

Correctness and *realism* are more difficult to control

- Not always obvious that certain practices can lead to incorrect datasets or unrealistic scenarios

Guidelines outlined here help to establish a common set of criteria to ensure future prudent experimentation with malware datasets

Summary

Malware-Execution is a technique used for the generation of ground-truth characterisations of malware for scientific experiments.

- The use of these datasets in evaluation can result in pitfalls

Rossow et al propose four categories of guidelines to address these pitfalls

- *Correctness* to avoid biasing the experiment results
- *Transparency* when describing the experimental setup and interpreting results
- *Realism* to ensure that observed behaviour is indicative of what is observed 'in the wild'
- *Safety* by implementing and describing an appropriate containment policy

Survey assessment identified that these guidelines could have been utilised across the community for improving the prudence of malware experiments

For those interested

The Turing Way [2]

"a lightly opinionated guide to reproducible data science"

[2] https://the-turing-way.netlify.com/

Questions?

How does this description of the scientific method differ from others?