Symantec.

Benchmarking Computer Security

Through The Worldwide Intelligence Network Environment (WINE)

Tudor Dumitraş
Symantec Research Labs

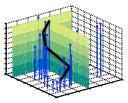
CCS Tutorial
October 2011

My Background

Network-on-chip protocols
[ASP-DAC'03 (**Best Paper Award**), DATE'03, ASP-DAC'04, VLSI Design'07]

Protocol-level fault tolerance


Source: Intel



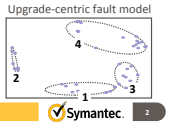
Fault-tolerant middleware
[WADS'04, Concurr&Comput'05, Middleware'05, Middleware'07]

Transparency and adaptation

Dependable, end-to-end software upgrades
(**J. Vlissides Award, A.G. Jordan Award**) [HotDep'07, Middleware'09, Onward!'10, OSR'10, MESOCA'11]

Benchmarking upgrade mechanisms

Upgrade-centric fault model



Benchmarks …

… allow apples-to-apples comparisons against the state of the art

… emphasize experimental design (hypothesis, metrics)

… point out what will keep working tomorrow

… have lasting impact on a field

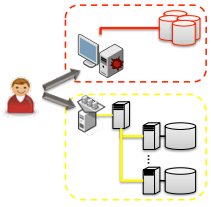
In cyber security: data sets not shared, experiments not repeated

Challenges for Cyber Security Experiments (CSET'11 Summary)

- Privacy is big challenge for collecting and sharing data
- How to ensure that data sets are relevant?
  - Dearth of metadata
- Why repeat experiments?
  - We're not funded to work on yesterday's problems
- Sharing algorithm implementations instead of data
  - High overhead to adapt to new data set
  - Code rot
- Cannot do meta-analysis
  - Lack of structured abstracts

WINE: Benchmark for Computer Security

`http://www.symantec.com/WINE`



Symantec's worldwide sensors

Platform for experimental reproducibility

The Worldwide Intelligence Network Environment (WINE)

- Goal: *repeatable cyber security experiments at scale*
- Field data collected on millions of *end-hosts*
- Data sampled from Symantec's *operational data* sets
- Access WINE on SRL site: *Culver City, CA* or *Herndon, VA*
  - Fee required
- Store *reference data* sets used in prior experiments
- Maintain *lab book*

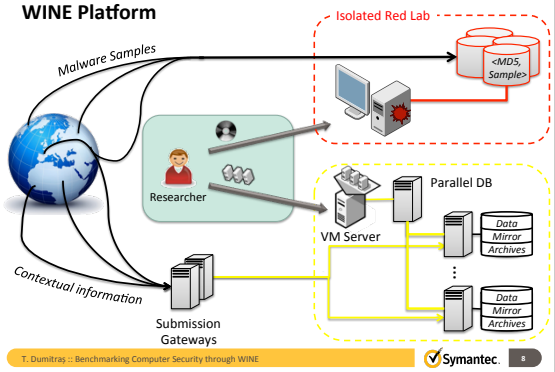
WINE Data

- Sampled field data, representative for what Symantec collects
  - Up to 20 TB
  - Over 1M end-hosts
  - Goes back to 2008
- Five data sets, initially:
  - Malware samples
  - Binary reputation (file downloads)
  - A/V and IPS telemetry
  - URL reputation
  - Spam

In response to expressed data needs of research community [NSF'10]

More data, in the future

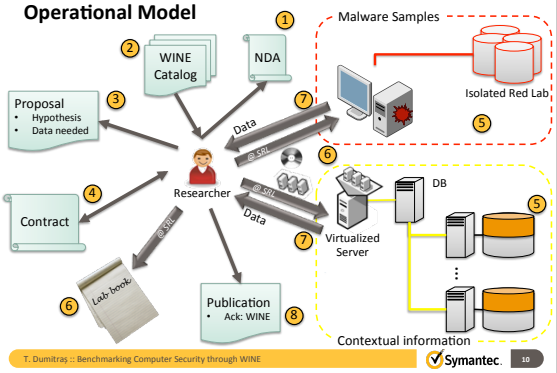
WINE Platform



What WINE is not …

- … a definitive benchmark suite
- … a data set that can be copied outside of SRL
- … a system that can be accessed remotely
- … a repository for all the data that Symantec collects
- … an effort targeted exclusively at cyber security

Operational Model



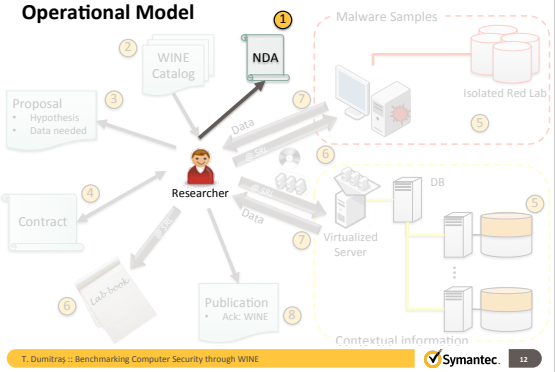
WINE Use Case

- Hypothetical experiment

  Evaluate a technique for detecting zero-day attacks that combines static / dynamic analysis of malware samples with data on their propagation patterns.

- Example of zero-day attack: *Stuxnet*
- Illustrative labels
  - Action done by visiting researcher
  - Action done by WINE team

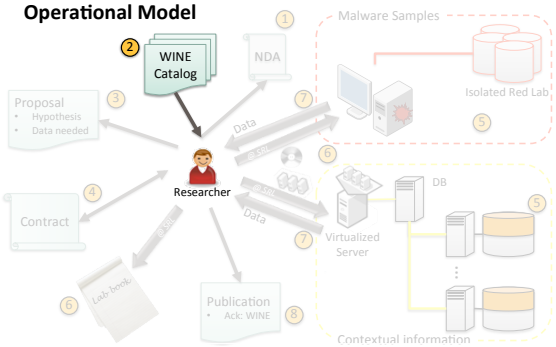
Operational Model



Step 1: Sign Non-Disclosure Agreement

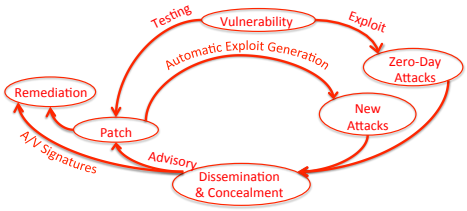
- Request from Darren Shou <darren_shou@symantec.com>
  - Copy of NDA
  - Fee schedule
- Return signed NDA
- NDA does not prevent publication
  - Collaboration agreement signed before site visit
- NDA provides access to the WINE data catalog

Operational Model



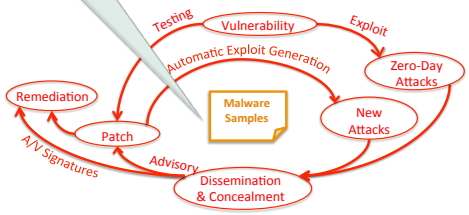
Step 2: Provide WINE Data Catalog

Describes data sets that cover the lifecycle of cyber attacks



WINE Data Set: *Malware*

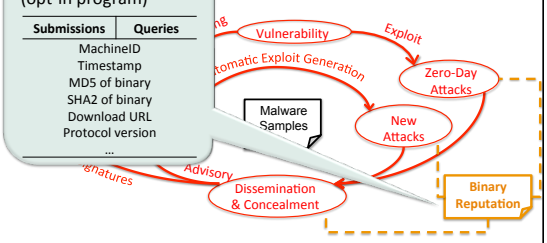
Packed and unpacked malware binaries



WINE Data Set: *Binary Reputation*

Norton Insight (opt-in program)

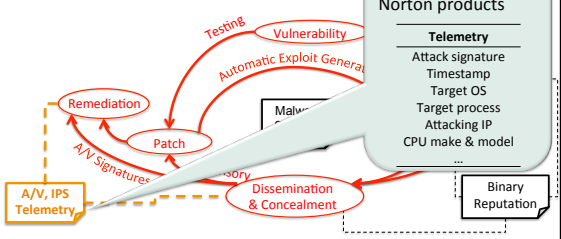
Submissions	Queries
MachineID	
Timestamp	
MD5 of binary	
SHA2 of binary	
Download URL	
Protocol version	
...	



WINE Data Set: *A/V & IPS Telemetry*

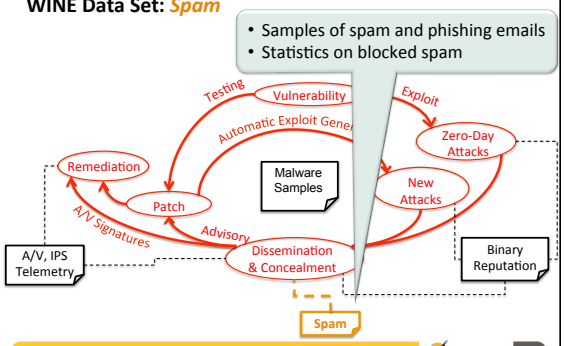
Threats detected by Norton products

Telemetry
Attack signature
Timestamp
Target OS
Target process
Attacking IP
CPU make & model
...



WINE Data Set: *Spam*

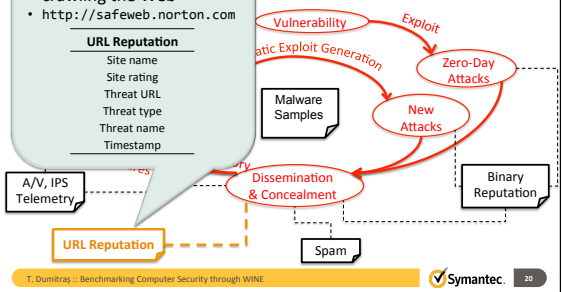
- Samples of spam and phishing emails
- Statistics on blocked spam



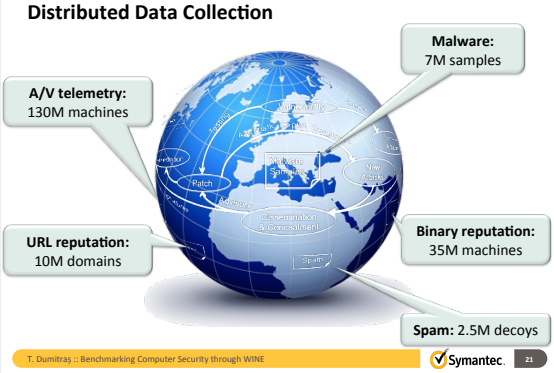
WINE Data Set: *URL Reputation*

- Data collected by crawling the Web
- `http://safeweb.norton.com`

URL Reputation
Site name
Site rating
Threat URL
Threat type
Threat name
Timestamp



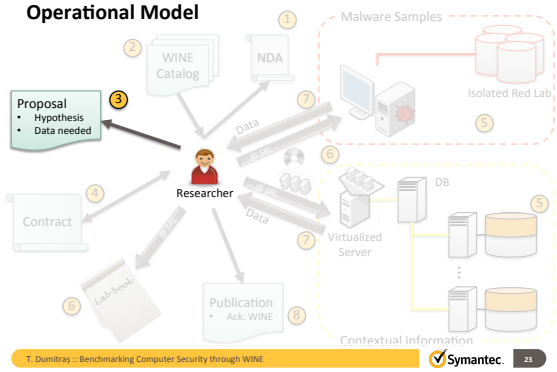
Distributed Data Collection



WINE Data Sets – Summary

- Choice of initial data inspired by needs of research community
  - Requested data, e.g. URLs from spam, representative malware samples [NSF'10]
  - Unique data, e.g. historical information on malware presence *before* the threat identification
- Representative samples of Symantec's operational data sets
- Relevant to many disciplines
  - Examples: machine learning, visual analytics, software reliability
- Additional data sets, in the future

Operational Model



Step 3: Write WINE Research Proposal

- One-page proposal
  - Problem studied
  - Proposed research approach
  - Data needed
  - Estimate of visit duration (min 2 weeks)
- Send proposal to Darren Shou `<darren_shou@symantec.com>`

Proposal Example

- Problem studied / hypothesis
  - "*My new approach can detect zero-day attacks, based on the following traits (…) of the binary samples and on their propagation patterns* "
- Proposed research approach
  - "*…* "
- Data needed
  - "*Binary samples of W32.Stuxnet (W32.Temphid). Counts of machines that downloaded these samples between April – July 2010* "
- Estimate of visit duration
  - "*2 weeks* "

Symantec virus names

Stuxnet discovered on June 17

Proposal Example: Input Data

- Detailed specification allowing us to assemble reference data set
  - Facilitates result reproducibility

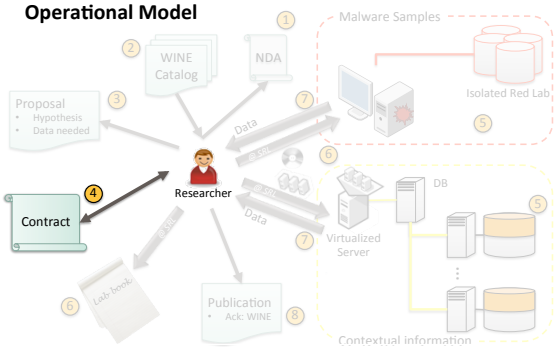
Example:
How to find Symantec virus names?

- Symantec Threat Explorer
  - Symantec name
  - Names given by other vendors
  - Discovery date
  - Technical details
  - CVE references
  - …

Input Data Specification

- Symantec resources
  - A/V signatures (Threat Explorer): `http://www.symantec.com/business/security_response/threatexplorer/`
  - IPS signatures: `http://www.symantec.com/business/security_response/attacksignatures/`
- Other ways to specify the data
  - Data from prior experiment
  - List of SHA2 or MD5 hashes (e.g., from Anubis or VirusTotal)
  - Date range (e.g., files downloaded worldwide during 1st week of May'11)
  - Other well-defined criteria
- We can provide assistance for this step

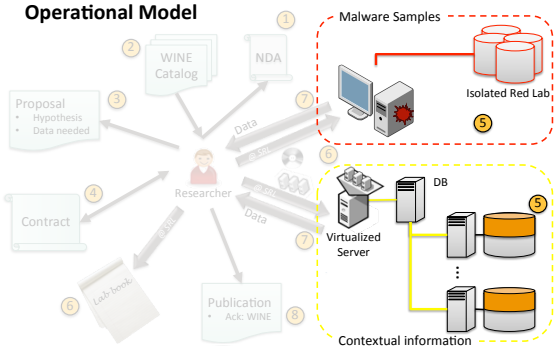
Operational Model



Step 4: Sign WINE Collaboration Agreement

- Signed by researcher's institution and Symantec
- Provision for publication
- Symantec retains ownership of data
- Symantec receives copies of all research products
- All right, title and interest belong to the researchers
  - Unless licensing exception is negotiated beforehand
  - Data set should be acknowledged in publications

Operational Model



Step 5: Assemble Reference Data Set

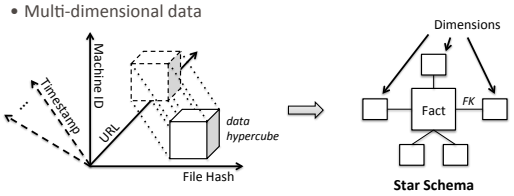
- Input data relevant to the experiment
  - For malware: a set of samples, identified by MD5 or SHA-2 hash
  - For other data sets: views over the existing WINE DB tables
- We preserve reference data sets for future experimenters

Example:
Prevalence of Stuxnet as a zero-day attack

1. Find Stuxnet's VID (Symantec internal *virus ID*)
2. List files associated with the VID, in the *A/V telemetry*
3. Search for their occurrences, in the *binary reputation* (focus on dates before Stuxnet's discovery)

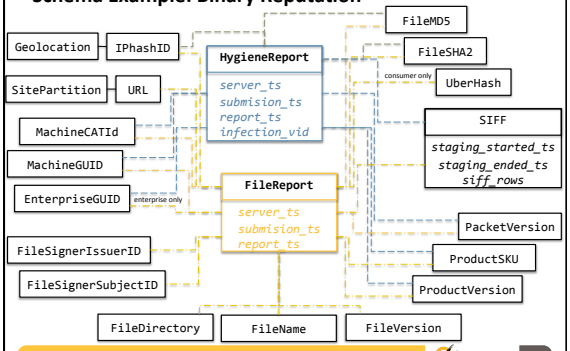
Database Schema

- Multi-dimensional data

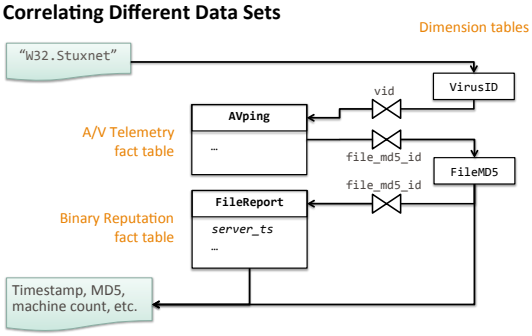


- Example: *Binary Reputation* submissions
  - HygieneReport: each fact corresponds to an infection detected
  - FileReport: each fact corresponds to a file detected
  - Dimension tables: FileMD5, FileSHA2, IPhashID, Machine GUID, …

Schema Example: Binary Reputation

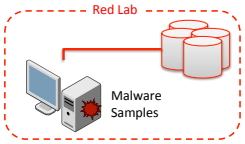


Correlating Different Data Sets



Malware Analysis

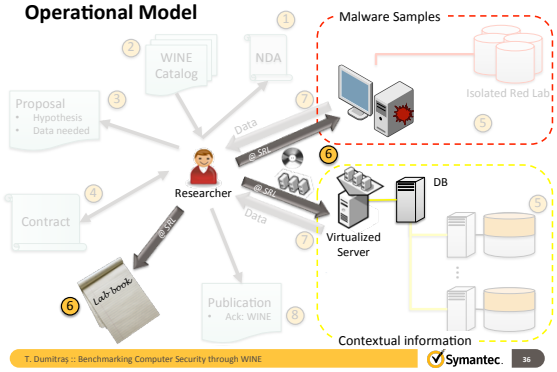
- Experiments conducted in an isolated *Red Lab*
  - No outbound network access



Example:
Analysis of Stuxnet samples

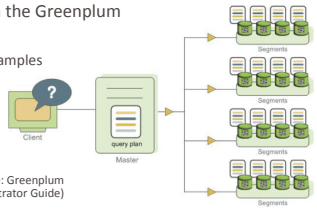
1. Search for Stuxnet samples in Symantec's malware collection
   - Use the list of *MD5 hashes* from A/V telemetry
2. Conduct static and dynamic analysis on these samples
   - *Traits*: n-grams, basic blocks, system calls, dataflow, memory accesses, etc.

Operational Model



Step 6: Site Visit and Experiments

- Conduct data analysis at scale
  - Up to *20 TB* data in WINE
  - Example: *1M hosts, 250M submissions/week* in binary reputation
- WINE stores data in the Greenplum parallel DB
  - Excluding malware samples



(source: Greenplum Administrator Guide)

Greenplum

- Compliant with most of SQL
  - Based on Postgres:
    CLI client: `psql` table definition: `\d table`
    tables organized in schemas (namespaces) get help: `\?` or `\h`
- Parallel database
  - One *master* (Postgres instance): accepts user queries
  - Multiple *segments* (degree of concurrency): transparent to users
  - Tables distributed on all segments based on hash of *distribution key* (DISTRIBUTED BY clause of table definition)
  - Table scans, joins, aggregations and sorts: execute in parallel
  - Single value of distribution key in predicate => query runs on one segment
  - *MapReduce* as alternative to SQL

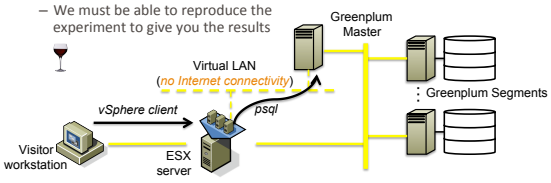
Greenplum

- Compliant with most of SQL
  - Based on Postgres:
    CLI client: `psql` table definition: `\d table`
    tables organized in schemas (namespaces) get help: `\?` or `\h`
- Parallel database
  - One *master* (Postgres instance): accepts user queries
  - *MapReduce* as alternative to SQL

In practice, for WINE:
- Use plain SQL
- Use MapReduce for analyses hard to express in a declarative manner
- Worry about distribution policy only if you need to optimize

Experimentation Platform

- Database accessed from a virtual machine
  - We provide a VM with the Greenplum DB (WINE schema instantiated)
  - You add the tools and external data needed for the experiment
- No mechanism for extracting data
  - We must be able to reproduce the experiment to give you the results



Find Stuxnet's VID

Use results in later queries

```
CREATE TEMPORARY TABLE stuxnet_vids AS
SELECT
       virus_name,
       vid                  AS stuxnetid
FROM
       dim.virusid_current
WHERE
       virus_name = 'W32.Stuxnet';
```

VID dimension table

Current virus name

Find MD5 File Hashes Associated with Stuxnet

For efficiency, slice fact table

```
CREATE TEMPORARY TABLE telemetry_slice AS
SELECT *
FROM          wine_telemetry.avping
       JOIN   stuxnet_vids          ON stuxnetid=virusid;
ANALYZE telemetry_slice;
```

A/V Telemetry

Only reports relevant to Stuxnet

```
CREATE TEMPORARY TABLE stux_md5 AS
SELECT DISTINCT
       file_md5_id,
       file_md5
FROM
              telemetry_slice
       JOIN   dim.filemd5          USING (file_md5_id);
```

Slice of MD5 dimension

Join with MD5 dimension table

Same column name (natural join)

Search for Stuxnet's Historical Presence

```
CREATE TEMPORARY TABLE binrep_slice AS
SELECT *
FROM   wine_binrep.filereportconsumer NATURAL JOIN stux_md5
WHERE  server_ts > '2010-04-01' AND server_ts < '2010-07-01';
ANALYZE binrep_slice;

SELECT
       encode (file_md5, 'hex')      AS md5,
       server_ts,
       machine_guid_id,
       url
FROM                                  binrep_slice
       NATURAL JOIN                   all_md5
       NATURAL LEFT OUTER JOIN        dim.url
ORDER BY
       server_ts;
```

Binary reputation

Date range, before discovery

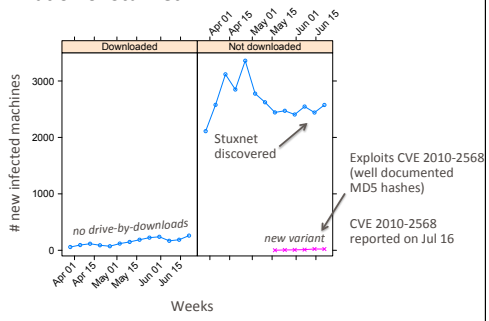
Timestamp

Download URL, if available

Join w/ MD5 slice & URL dimension

Include reports w/o URLs (not downloads)

Dissemination of Stuxnet



Conduct Data Analysis

- Test your algorithm on the reference data set
  - Example: *use Stuxnet dissemination to test zero-day detection technique*
- Use only tools in your VM and the Greenlpum facilities
  - VM does not have Internet connectivity during the experiment
  - Interactions with the outside world would prevent reproducibility, owing to the transient nature of Internet resources
- Prepare script to run experiment from end to end
- Update WINE lab book (on internal wiki)
  - How to reproduce the results?

A Note on Efficiency

Integer comparison	Byte array comparison	String comparison
`file_md5_id = 123456`	`file_md5 = decode('cc1db…', 'hex')`	`encode(file_md5, 'hex') = 'cc1db…'`

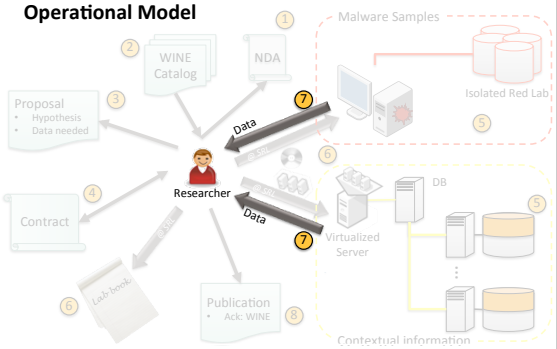
Efficiency

- Do not join full fact tables with dimensions
  - Instead, create *slice* with relevant reports from fact table
- If possible, avoid predicates referencing a single value of the *distribution key*
  - This prevents parallel execution

Other Ways to Analyze the Data

- MADlib analytics library: `http://madlib.net/` [Cohen'09]
  - Supervised learning (e.g., naïve Bayes, decision tree, SVM), clustering, sketch-based estimators, etc.
- Greenplum MapReduce
- For small scale experiments: copy data to file in experiment VM
- Hadoop cluster for malware analysis
  - In Culver City, CA red lab

Operational Model



Step 7: Produce Results & Archive Data

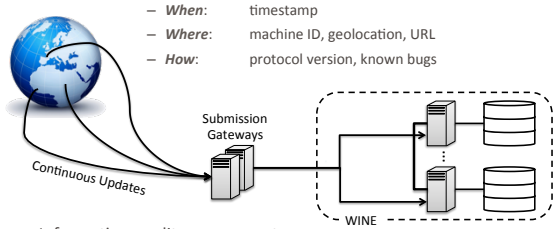
- We *run* the experiment
  - We use the information from the lab book (reference input data, script to invoke, output data desired)
- We provide the output data to the researcher
- We *archive the data and the VM* for future experiments

Ensures that the experiment is repeatable

Can an experimental result be reproduced, from the data collection to the final conclusion?

Reproducibility of Experimental Results (1)

- *Collection metadata*: data is self-descriptive
  - ***When***: timestamp
  - ***Where***: machine ID, geolocation, URL
  - ***How***: protocol version, known bugs

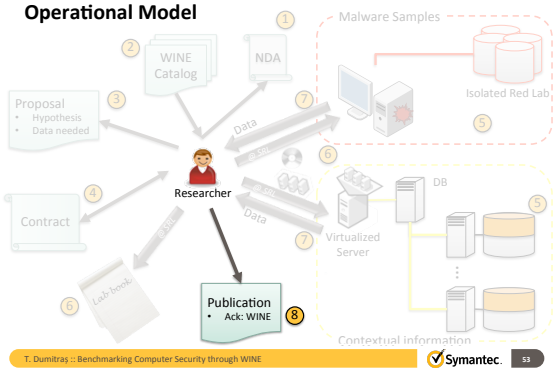


- Information quality assessment

Reproducibility of Experimental Results (2)

- *Experiment metadata*: recorded in lab book
  - External researcher describes experiment in proposal
    - Research hypothesis
    - Input/output data
  - Researcher develops script to run experiment from end to end
  - Hypothesis, data and script are documented on a wiki
- Enables independent verification of experimental design

Operational Model



Step 8: Acknowledge WINE in Publication

- Each reference data set in WINE will have a unique identifier
- The identifier must be mentioned in all publications reporting corresponding results
- Permission to reuse tools & reference data set must be stated explicitly in the acknowledgment:

> *"We [give | do not give] other researchers permission to repeat our experiments"*

Operational Model – Summary

1. Non-disclosure agreement
2. WINE data catalog
3. Research proposal
4. Collaboration contract
5. Reference data set
6. Site visit and experiments
7. Experimental results
8. Publication & WINE acknowledgment

Benchmarking Methods

- Evaluate multiple metrics
- Run multiple tests
  - Stuxnet is just one example; behavior of other viruses might also be relevant
  - Developing five tests takes nearly as much effort as developing two
- Address the threats to validity
  - Do the metrics used actually model the hypothesis? (*construct validity*)
  - Is there a causal connection between dependent & independent variables? (*internal validity*)
  - Have included all relevant data points & excluded the irrelevant ones? (*content validity*)
  - Can generalize results to data outside the scope of the study? (*external validity*)

Challenges for Cyber Security Experiments & WINE

- Privacy concerns for collecting/sharing data
  - **Controlled access to sensitive data**
- Dearth of metadata; how to ensure relevance?
  - **End-to-end control => high quality metadata**
- Share code vs. data: code rot, adaptation overhead
  - **Archive both data and VMs used**
- Attacks change; why repeat experiments?
  - **Easy to compare approaches**
- No meta-analyses owing to lack of structured abstracts
  - **Extract structured information from lab book**

Many Ways to Use the WINE Data

- Security
  - What are the sources and prevalence of zero-day attacks?
  - Malware detection: can we do better than signatures and heuristics?
  - Does Patch Tuesday make the world a safer place?
- Software engineering
  - How to prevent the bugs that matter?
- Machine learning
  - How to analyze billion-node graphs?

Collaboration and Funding Opportunities

- Symantec Fellowship
  - 3 Best Paper Awards over the past 4 years
- NSF support: Trustworthy Computing program
  `http://www.gtisc.gatech.edu/nsf_workshop10_data.html`
  - We provide letters of collaboration for proposals
- Joint proposals
  - IARPA STONESOUP, with Columbia, Stanford & GMU
  - HS-ARPA Cyber Security, with Georgia Tech and Imperial College
  - DARPA MRC, with Columbia

Conclusions

WINE: a step toward rigorously benchmarking cyber security

- Can analyze unique data sets
- Can correlate data collected from multiple observation perspectives
- Can conduct experiments at scale
- Can ensure the reproducibility of experimental results

Collaborators

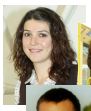
- At Symantec Research Labs:


Marc Dacier


Darren Shou

Petros Efstathopoulos

- In academia:


Leyla Bilge, EURECOM


Jiyong Jang, CMU


Iulian Neamtiu, UC Riverside

Symantec.

Thank you!

Tudor Dumitraș
tudor_dumitras@symantec.com
http://www.ece.cmu.edu/~tdumitra
@tudor_dumitras