Experimental Challenges in Cyber Security:

A Story of Provenance and Lineage for Malware

Tudor Dumitraș
Symantec Research Labs

Iulian Neamtiu
UC Riverside

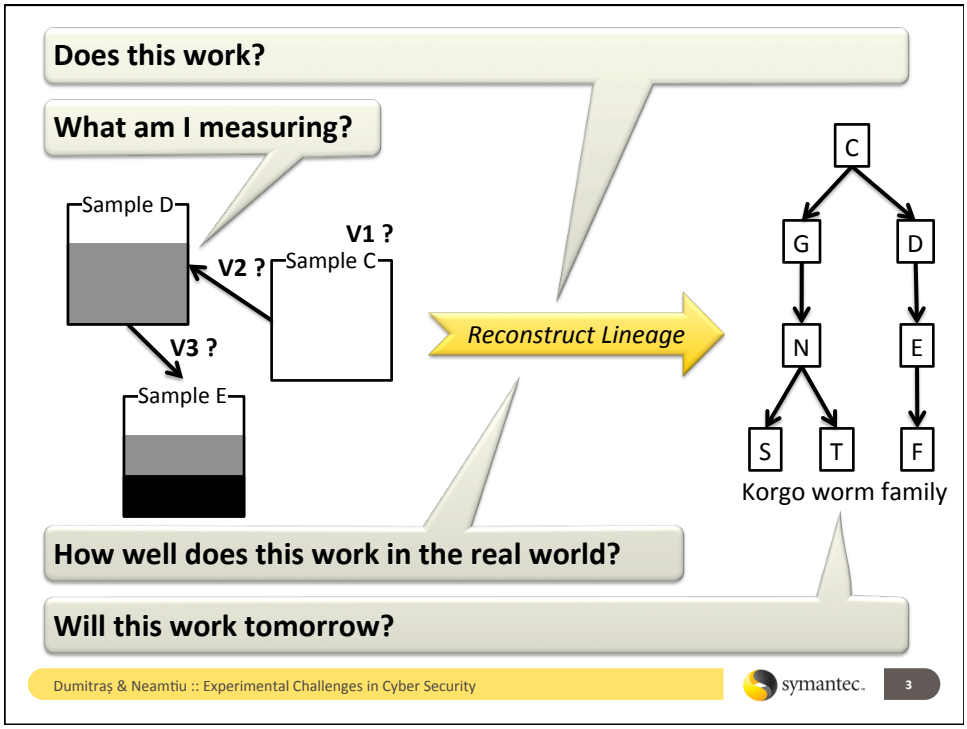
symantec.

UC RIVERSIDE UNIVERSITY OF CALIFORNIA

The IROP Keyboard

[Zeller, 2011]

To prevent bugs, remove the keystrokes
that predict 74% of failure-prone modules in Eclipse



Goal

Overcome common *threats to validity* in cyber security experiments

Our Approach

- Provenance and lineage reconstruction through *static*, *dynamic*, and *contextual analysis*

- **Key idea**:
  Evolution of open source binaries = Training data

  Linux:
  - 20 years of evolution
  - 70+ versions

  FreeBSD:
  - 18 years of evolution
  - 22+ versions

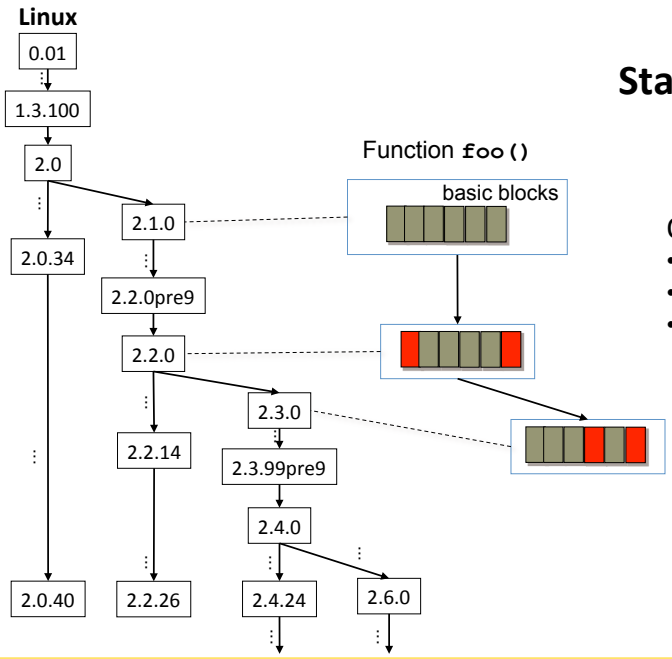
- Use the WINE benchmark for validation

Lineage and Provenance

Lineage: establish the ancestors and descendants of a binary artifact

Provenance: determine the compiler, development environment, testing methods, release schedule

Static Analysis



Other traits
- system calls
- dataflow
- memory accesses

Contextual Analysis

- Obfuscation techniques (e.g., *packing, randomization*) reduce effectiveness of static / dynamic analyses
- Idea: use contextual information
  - Network traces
  - Infection reports
- Answers questions such as
  - ***When*** has a malware artifact first appeared?
  - ***Where*** has it spread?
  - ***How*** has gained access?

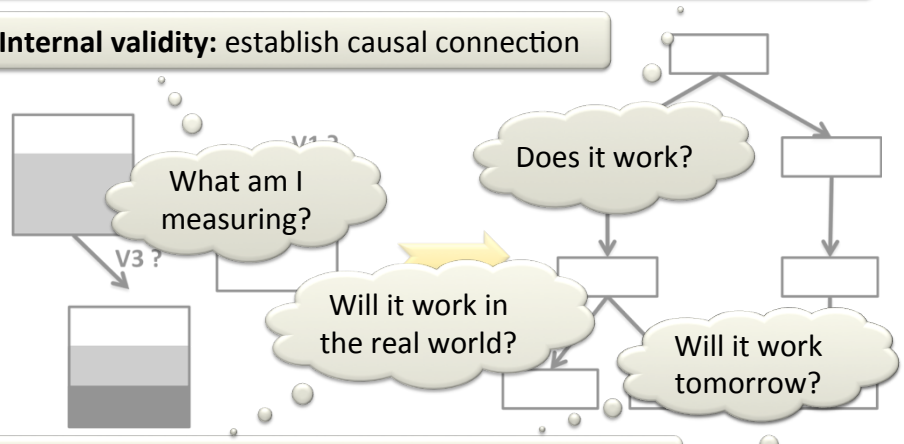
Threats to Validity

- Lineage
  - Lack of ground truth on malware families
  - Lack of contextual data: e.g., date and time of appearance
- Provenance
  - Different information provided by compilers and assemblers
  - Lack of contextual data: e.g., origin geolocation, dissemination patterns

Illustrate general threats to validity in experimental cyber security

Construct validity: use metrics that model the hypothesis

Internal validity: establish causal connection



Content validity: include only and all relevant data

External validity: generalize results beyond experimental data

Candidate Approaches

Validity

- Testbeds for repeatable experimentation (Emulab, DETER)
  - Representative data sets are also needed
- Synthetic test data generation [Lippmann, 2000]
  - Short-lived relevance

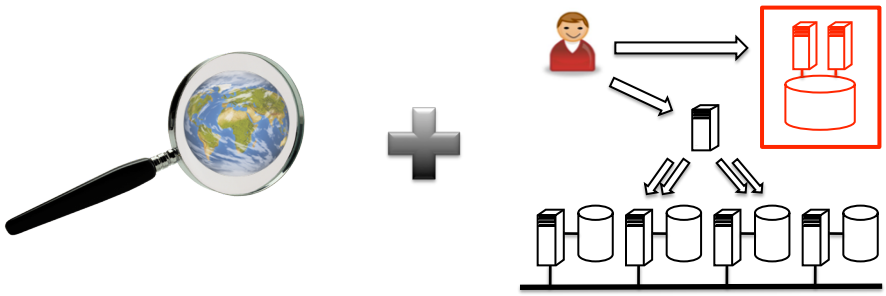
~~Content~~

- Field-gathered data [DHS PREDICT; Leita, 2010; Bilge, 2011]
  - Honeypots: instrumentation could alter results
  - Network traces: reveal only part of malware behavior
  - Lack of metadata on collection process [Camp, 2009; CSET 2009]

~~Internal~~
~~External~~
~~Multiple~~

WINE: Benchmark for Computer Security

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



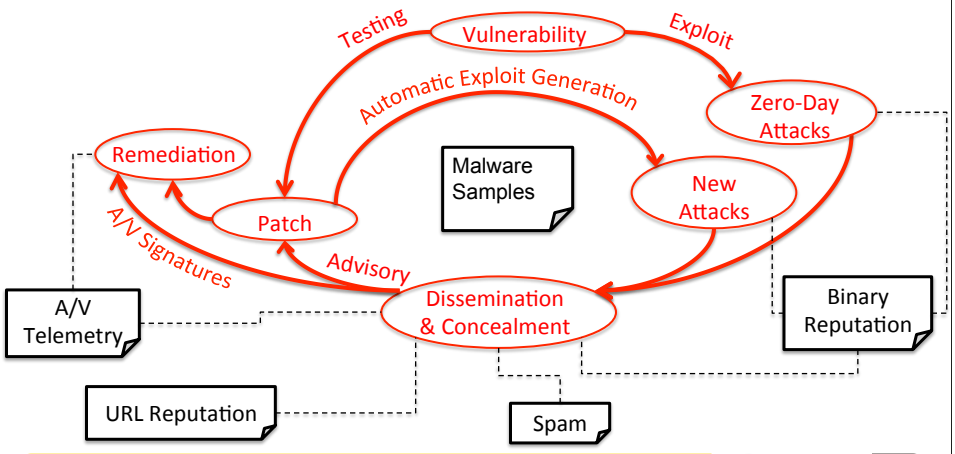
Symantec's worldwide sensors

Platform for experimental reproducibility

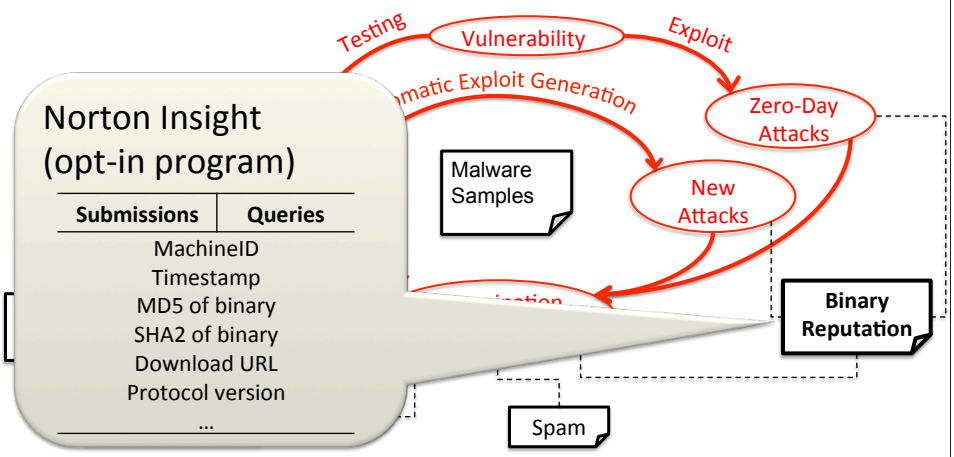
The Worldwide Intelligence Network Environment (WINE)

- Goal: *reproducible experiments* in cyber security
- 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*

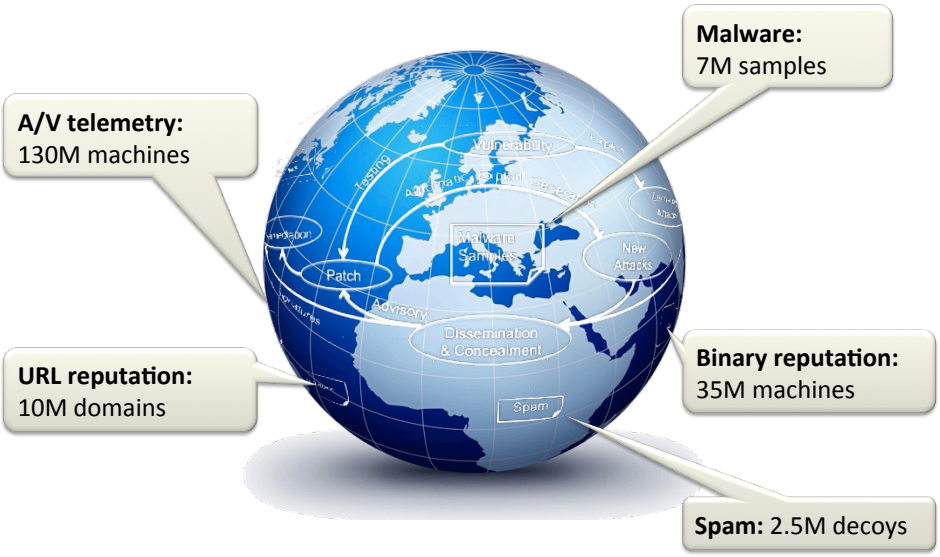
WINE: Cover all Phases of Cyber Threat Lifecycle



Traits for Contextual Analysis: *Binary Reputation*

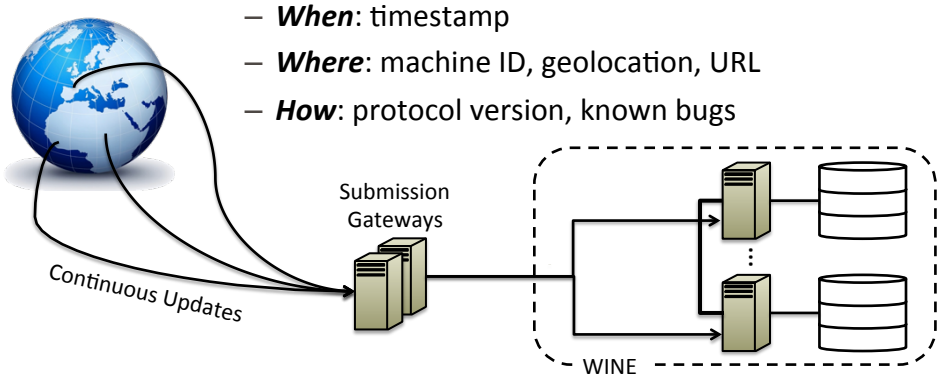


Distributed Data Collection



WINE Experiments: Threats to Validity (1)

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



- Enables defining *relevant* data sets (content validity)
- Enables reasoning if results *generalize* (external validity)

WINE Experiments: Threats to Validity (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 (internal and construct validity)

Conclusions

- Rigorous experiments give us an edge in the security arms race
  - Develop techniques that are likely to keep working tomorrow
- Challenges
  - Ground truth
  - Relevant data sets
  - Rigorous analysis
- WINE: first step toward a rigorous benchmark for cyber security

symantec.

Thank you!

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