Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits

Work with Carl Sabotke and Octavian Suciu

Tudor Dumitraș
Assistant Professor, ECE & MC2
University of Maryland, College Park

Something About Me First

Tudor Dumitraș
Web: www.umiacs.umd.edu/~tdumitra
Email: tdumitra@umiacs.umd.edu
• Ph.D., Carnegie Mellon University
• 2.5 yr. at Symantec Research Labs
• Joined UMD in 2013
More and more software vulnerabilities are discovered ...

2014 CVE IDs format change: no longer limited to 9,999 vulnerabilities/year

OSVDB

Vendors sure like to wave the “coordination” flag… (revisiting the ‘perfect storm’)

By jerichoatthlon on February 2, 2015

Oh, did I forget to mention that kicker about all of this? October 14, 2014 has 254 vulnerabilities disclosed. On the same day that the [stressed POODLE vulnerability was disclosed](http://blog.osvdb.org/2015/02/02/vendors-sure-like-to-wave-the-coordination-flag-revisiting-the-perfect-storm/), impacting thousands of different vendors and products. That same day, OpenSSL, perhaps the most oft used SSL library released a patch for the vulnerability as well, perfectly “coordinated” with all of the other issues.
Research Questions

• How to prioritize the response to vulnerability disclosures?
• Can forecast vulnerabilities exploited in the wild?
  – ... earlier than existing data sources?
  – ... and with fewer false positives?

**Our approach: Twitter analytics**

Demonstration

**http://ter.ps/sec15demo**
System Design
[USENIX Security’15]

Machine Learning

Precision: one order of magnitude better than CVSS

Ground Truth Features

Predictions

Detection: median of 2 days ahead of existing data sets

Adversarial Interference

- Twitter is free and open to all users
- Could an adversary post false information in order to trick a detector?

Tudor Dumitraș :: Vulnerability Disclosure in the Age of Social Media
Talk Outline

• Design and implementation of a technique for early exploit detection using social media

• Performance evaluation for detecting exploits found in the wild

• Analysis of system robustness to adversarial interference

• Security implications
Twitter Dataset

- Twitter Public Stream
  - February 2014 - January 2015
  - 1.1 billion tweets
- Tracking the CVE keyword
- Collected unsampled corpus
  - 287,717 tweets
  - 5,865 vulnerabilities
Detecting Exploits in the Wild

Classifier Evaluation

- A classifier can make two kinds of errors
  - **False Positive** = marked as exploited but not exploited in the wild
  - **False Negative** = not marked as exploited, but exploited in the wild

- **Precision** = fraction of vulnerabilities marked as exploited that are **actually exploited**
  - False positives hurt precision

- **Recall** = fraction of exploited vulnerabilities that are **marked as exploited**
Baseline Classifier

- Using CVSS Score as indicator of an exploit
- CVSS marks many vulnerabilities as **exploitable**

\[ \{ \text{Precision} \} < 9\% \]

Detecting Exploits in the Wild

- CVSS Score: very low precision, high recall
Detecting Exploits in the Wild

- Database Information: High recall, low precision

[Graph showing precision and recall for different features]

Detecting Exploits in the Wild

- Twitter Word features: low recall, high precision

[Graph showing precision and recall for different features]
Detecting Exploits in the Wild

• Twitter Traffic features: higher recall, lower precision

Combining all features: variable regularization results in a precision/recall tradeoff
Improving the Performance

• Filtering based on ground truth coverage and tweet volume

Early Prediction of Exploits

<table>
<thead>
<tr>
<th>National Vulnerability Database</th>
<th>Features</th>
<th>Ground Truth</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>Training</td>
<td>Linear SVM</td>
<td>Symantec</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Microsoft Security Advisories</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prediction Threshold</th>
</tr>
</thead>
</table>

Tudor Dumitras: Machine Learning Techniques for Preventing the Global Malware Dissemination
Tweets Before Signatures

- Tradeoff between precision and detection lead time
  - Median detection: 2 days ahead of Symantec signatures
  - 45% classification precision

Adversarial Interference
Attacks Against the Exploit Detector

• Can we prevent the adversary from poisoning the training dataset?
  – No. Twitter is a free and open service.

• Can we keep the features secret?
  – No. Our ground truth comes from public sources.

• Is the adversary resource bound?
  – Yes. Adversary must control the properties of multiple accounts.

Adversary Model

• Adversary’s goal: to introduce false positives

• Simulation of causative attacks
  – 3 adversary types

• Adversaries cannot prevent benign users from posting
**Blabbering Adversary**

- Randomly posts tweets, no knowledge about features

![Blabbering Adversary Graph]

Random noise affects the system minimally

---

**Word Copycat Adversary**

- Mirrors the statistics of words corresponding to exploited vulnerabilities

![Word Copycat Adversary Graph]

Damage is bound due to other features (e.g. Traffic, CVSS, Databases)
**Full Copycat Adversary**

- Sybil-like: controls multiple accounts
  - Manipulates all Twitter features except account creation date and account verification

For resilience, need list of trusted users
Most informative tweets come from ~4,000 users

Damage is bound only by non-Twitter features

---

**Talk Outline**

- Design and implementation of a technique for early exploit detection using social media
- Performance evaluation for detecting exploits found in the wild
- Analysis of system robustness to adversarial interference
- Security implications
Security Implications

- Fighting exploits with machine learning
  - Can forecast some vulnerability exploits
  - High precision and recall for problems that already have good predictors (e.g. MS exploitability index)
  - Challenges: concept drift, adversarial interference
  - Models have more potential applications (e.g. cyber insurance)

- Few vulnerabilities are exploited in the wild
  - Exploit scarcity felt in the underground economy
    - Blackhole exploit kit (2013): $100,000 budget for purchasing 0-day exploits
    - 0-day exploit for CVE-2013-3906: both targeted attacks and botnet-based malware
  - Challenge: Poor ground truth coverage
    - Better information sharing would improve the detectors

Things I Haven’t Told You About

- Mining downloader graphs to detect malware
  [CCS’15]

- How we measured the patching rate of 1,593 vulnerabilities
  [Oakland’15]

- How we measured the duration an prevalence of zero-day attacks
  [CCS’12]

- Certificate reissues and revocations in the wake of Heartbleed
  [IMC’14]

- Security metrics based on field data
  [RAID’14]
Students

Octavian Suciu

Yantao Zhang

Ziyun Zhu

BumJun Kwon

Collaborators: Leyla Bilge, Petros Efstathopoulos, Daniel Marino (Symantec Research Labs), Jiyong Jang (IBM Research), Aaron Schulman (Google), Juan Caballero (IMDEA Software), Polo Chau (Georgia Tech), David Choffnes, Alan Mislove, Christo Wilson (Northeastern University), Amol Deshpande, David Levin, V.S. Subrahmanian (University of Maryland), Christos Faloutsos (Carnegie Mellon University), Iulian Neamtu (NJIT), Aditya Prakash, Eli Tilevich (Virginia Tech)

Thank you!

Tudor Dumitraș

tdumitra@umiacs.umd.edu

http://www.umiacs.umd.edu/~tdumitra

@tudor_dumitras

Paper and detailed feature list: http://ter.ps/sec15exploit

Demo: http://ter.ps/sec15demo