Vulnerability Discovery Models: Which works, which doesn’t?

Viet Hung Nguyen and Fabio Massacci
University of Trento, Italy
{vhnguyen, massacci}@disi.unitn.it

The Roadmap

Targets of Analysis
- Precondition & applications for the study

Data Collection
- For each target, collect all available data sets

Data Fit
- Fit data to vulnerability discovery model

Analysis
- Perform analysis on result
Basic Concepts

* Vulnerability
  ✓ An instance of human mistake in specification, development, or configuration of software that its execution can violate the security policy [Krsul98]

* Vulnerability Discovery Model (VDM)
  ✓ A post-release stage where people identify and report security flaws of a released software
  ✓ Usually represented as mathematic curves

Existing VDMs

- Alhazmi-Malaiya Logistic (AML)
- Anderson Thermodynamic (AT)
- Linear (LN)
- Logarithmic Poisson (LP)
- Rescolar’s Exponential (RE)
- Rescolar’s Quadratic/Linear (RQ)
The Fallacy of Measurement

*How to measure vulnerabilities?*

- Different definitions/sources of vulnerabilities
- Eg. Firefox:
  - Mozilla Bugzilla (only security-relevant bugs)
  - Mozilla Foundation Security Advisory (MFSA)
  - National Vulnerability Database (NVD)
- What is the number of vulns?
  - 6 MFSA, 10 NVD, 14 (security) Bugzilla.
Research Questions

• RQ1: which VDM works, which doesn’t?
  ✓ Do the existing VDMs work?
  ✓ Which VDM is the best one?

* RQ2: How do different ways of counting vulns impact to the performance of VDMs?
  ✓ Do VDMs behave differently with different types of data set?
Types of Vulnerability Data Set

✓ Release X (eg. FF3.0)

✓ NVD(X) : 1 vuln is 1 NVD entry which mentions X

✓ NVD.Advice(X) : 1 vuln is 1 NVD entry which mentions X, and has a reference to an advisory confirmed by X’s vendor

✓ NVD.Bug(X) : 1 vuln is 1 NVD entry which mentions X, and has a reference to a bug confirmed by X’s vendor

✓ NVD.Nbug(X) : 1 vuln is 1 bug confirmed by X’s vendor, and is referred to by 1 NVD entry mentioning X

✓ Advice.Nbug(X) : 1 vuln is 1 bug confirmed by X’s vendor, and is directly or indirectly referred to by an NVD entry mentioning X
Targets of Analysis

★ Targets of Analysis: 17 releases of Browsers !!
  ✓ IE: v4 - v8
  ✓ Firefox: v1.0 - v3.6
  ✓ Chrome: v1.0 - v6.0

★ Why should they be browsers?
  ✓ Complex enough (like a small operating system)
  ✓ Quickly evolve
  ✓ Targets of many attacks

★ Why should they be IE, Firefox and Chrome?
  ✓ Top three most popular browsers in the world
Data Collection

**Data sources**
- IE : NVD
- Firefox : MFSA, Bugzilla, NVD
- Chrome: ChromeIssue, NVD

**Data collection**
- 58 data sets of 17 releases

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chrome</td>
<td>●</td>
<td>●</td>
<td>—</td>
<td>●</td>
<td>—</td>
<td>6 (v1.0–v6.0)</td>
</tr>
<tr>
<td>Firefox</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>6 (v1.0–v3.6)</td>
</tr>
<tr>
<td>IE</td>
<td>●</td>
<td>—</td>
<td>●</td>
<td>—</td>
<td>—</td>
<td>5 (v4.0–v8.0)</td>
</tr>
</tbody>
</table>

Bullets (●) indicate enabled data sets. Dashes (—), otherwise, mean there is no data sources available to collect the data sets.
Goodness of Fit (GoF) Analysis

* Fit data to VDMs
  * Non-linear regression method, implemented in R (www.r-project.org)

* Chi-square test for Goodness-of-Fit (GoF)
  * $O_i$ - observed values
  * $E_i$ - expected values

\[ \chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \]

* The meaning of Chi-square test
  * Measure the difference between observed and expected values
  * Use p-value of the chi-square test to know whether VDM works or not
RQ1: Which VDM works, which doesn’t?

RQ1: Which VDM works, which doesn’t?

**Intuitive conclusions**

- **p-value > 0.95** FIT (X)
- **0.05 <= p-value < 0.95** INCONCLUSIVE (?)
- **p-value < 0.05** NOT FIT (-)

NVD Data set

<table>
<thead>
<tr>
<th>Firefox</th>
<th>Chrome</th>
<th>IE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1.0 1.5 2.0 3.0 3.5 3.6</td>
<td>1.0 2.0 3.0 4.0 5.0 6.0</td>
</tr>
<tr>
<td>AT</td>
<td>-     -     X   -    X    ?</td>
<td>-    -    ?    -    -</td>
</tr>
<tr>
<td>LN</td>
<td>-     X     -    X    ?    ?</td>
<td>-    -    ?    -    -</td>
</tr>
</tbody>
</table>

The goodness of fit of a VDM is based on p-value in the $\chi^2$ test. p-value < 0.05: not fit (-), p-value $\geq$ 0.95: good fit (X), and inconclusive fit (?) otherwise.
RQ1: Which VDM works, which doesn’t?

NVD Data set

The goodness of fit of a VDM is based on $p$-value in the $\chi^2$ test. $p$-value $< 0.05$: not fit (-), $p$-value $\geq 0.95$: good fit (X), and inconclusive fit (?) otherwise.
RQ2: The Impact of Types of Data Set

Opposite results are obtained from different data sets

- Same model
- Same target (i.e., same software release)
- But different counting methods (different types of data set)
RQ2: The Impact of Data Sets

<table>
<thead>
<tr>
<th>VDM</th>
<th>Firefox</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>v1.0</td>
<td>v2.0</td>
<td>v3.0</td>
<td>v3.5</td>
<td>v3.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VDM</th>
<th>Google Chrome</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>v1.0</td>
<td>v2.0</td>
<td>v3.0</td>
<td>v4.0</td>
<td>v5.0</td>
<td>v6.0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VDM</th>
<th>IE</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>v4.0</td>
<td>v5.0</td>
<td>v6.0</td>
<td>v7.0</td>
<td>v8.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AML</td>
<td>X</td>
<td>?</td>
<td>X</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td></td>
</tr>
</tbody>
</table>

Each column has five cells corresponding to Advice.Nbug, NVD, NVD.Advice, NVD.Bug, NVD.NBug

Opposite results for the same models

* Different types of data set would strongly impact to VDM’s GoF
Conclusion and Future Work

🌟 Summary

✓ 6 VDMs are analyzed in 58 data sets of 17 browser releases

🌟 The findings

✓ VDM doesn’t work: AT (for browsers)
✓ VDMs (probably) don’t work: LN, RQ (for browsers)
✓ VDMs (probably) work: AML, LP, RE (for browsers)
✓ Different types of data set would strongly impact to VDM’s GoF

🌟 Future work

✓ How well could VDM predict the trend of vulns?
  • Split data sets in 2 parts: 1 for generating curves, 1 for testing
✓ How VDM’s GoF evolve overtime?
✓ Which types of data set are more appropriate for VDM?
  • In terms of yielding more stable results
Thank you

Viet Hung Nguyen – vhnguyen@disi.unitn.it
Fabio Massacci – massacci@disi.unitn.it