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

Università degli Studi di Trento

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

ASIACCS'12, Seoul, Korea, 02 – 04 May, 2012

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

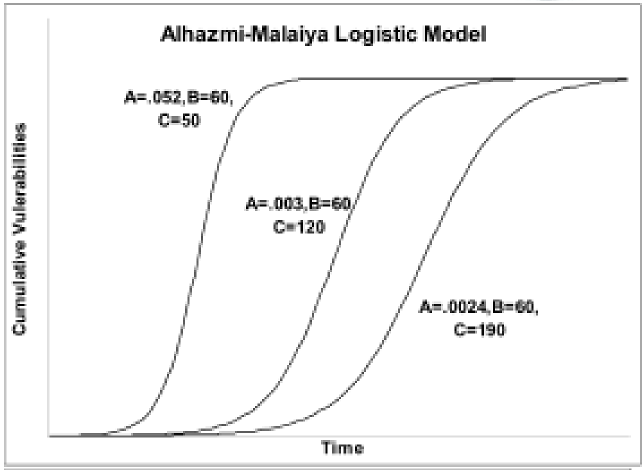
Basic Concepts

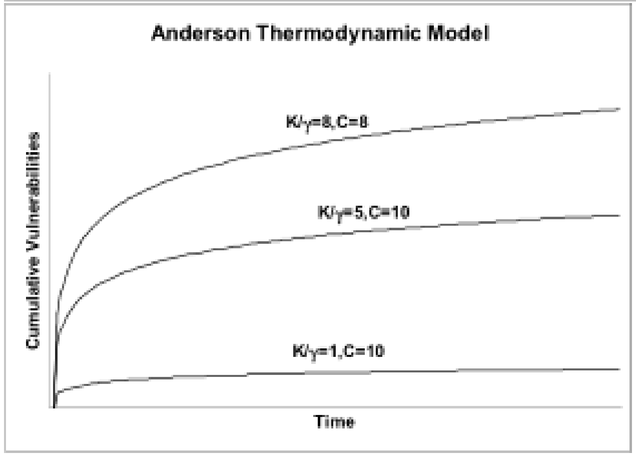
- Vulnerability
  - An instance of human mistake in specification, development, or configuration of software such that its execution can violate the security policy [Krsul98]
- Vulnerability Discovery Model (VDM)
  - A post-release stage where people identify and report security flaws of a released software
  - Usually represented as mathematic curves

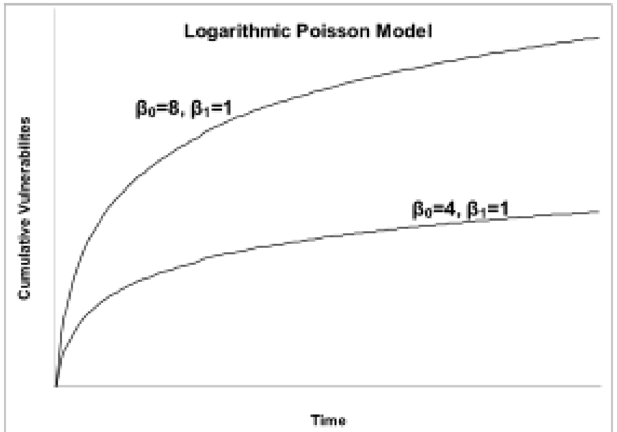
[Krsul98] Krsul I.V, Software Vulnerability Analysis, PhD Thesis, Perdue University, 1998

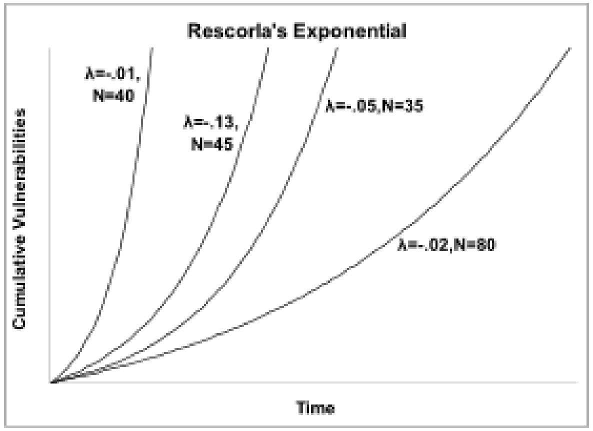
Existing VDMs

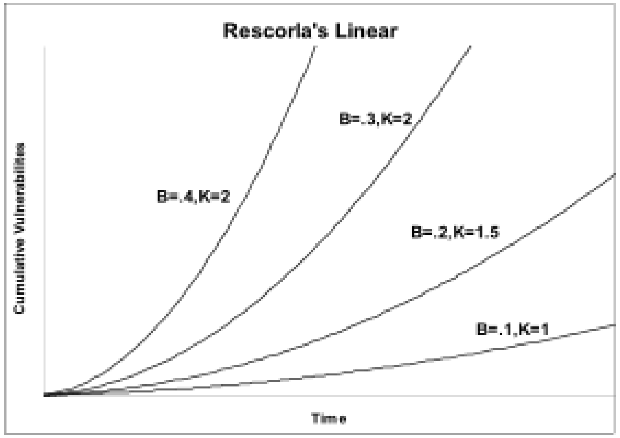
* Alhazmi-Malaiya Logistic (AML)
* Anderson Thermodynamic (AT)
* Linear (LN)
* Logarithmic Poisson (LP)
* Rescolar's Exponential (RE)
* Rescolar's Quadratic/Linear (RQ)







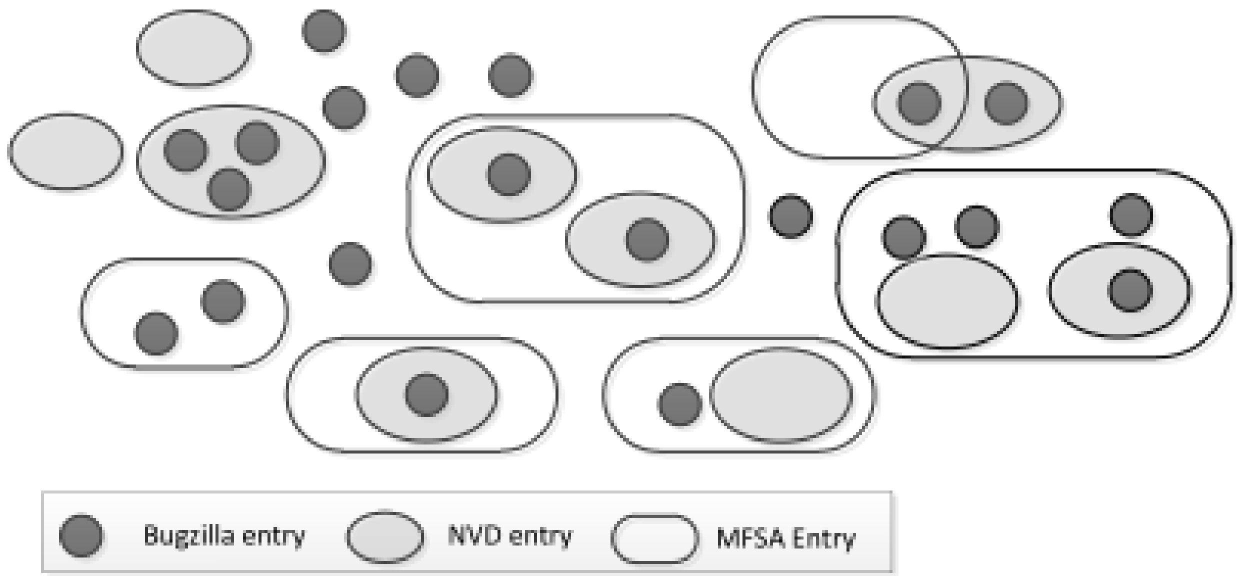




The Fallacy of Measurement

- 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.

Vulnerability space of Firefox



Research Questions

* RQ1: which VDM works, which doesn't?
  - ✓ Do the existing VDMs work?
* RQ2: how do different ways of counting vulns impact to the performance of VDMs?
  - ✓ Do VDMs behave differently with different types of data set?
* RQ3: in which definition of vuln, VDMs yield more stable results?
  - ✓ Which type of data set is most appropriate for VDM study?
* RQ4: which VDM is globally superior?
  - ✓ Which VDM yields better results during software's lifetime?

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

Data Collection

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

- Data collection
  - ✓ 58 data sets of 17 releases

	nvd	nvd.Bug	nvd.Advice	nvd.Nbug	advice.Nbug	#Releases
Chrome	•	•	—	•	—	6 (v1.0–v6.0)
Firefox	•	•	•	•	•	6 (v1.0–v3.6)
IE	•	—	•	—	—	5 (v4.0–v8.0)

Bullets (•) indicate enabled data sets. Dashes (—), otherwise, mean there is no data sources available to collect the data sets.

Goodness of Fit (GoF) Analysis

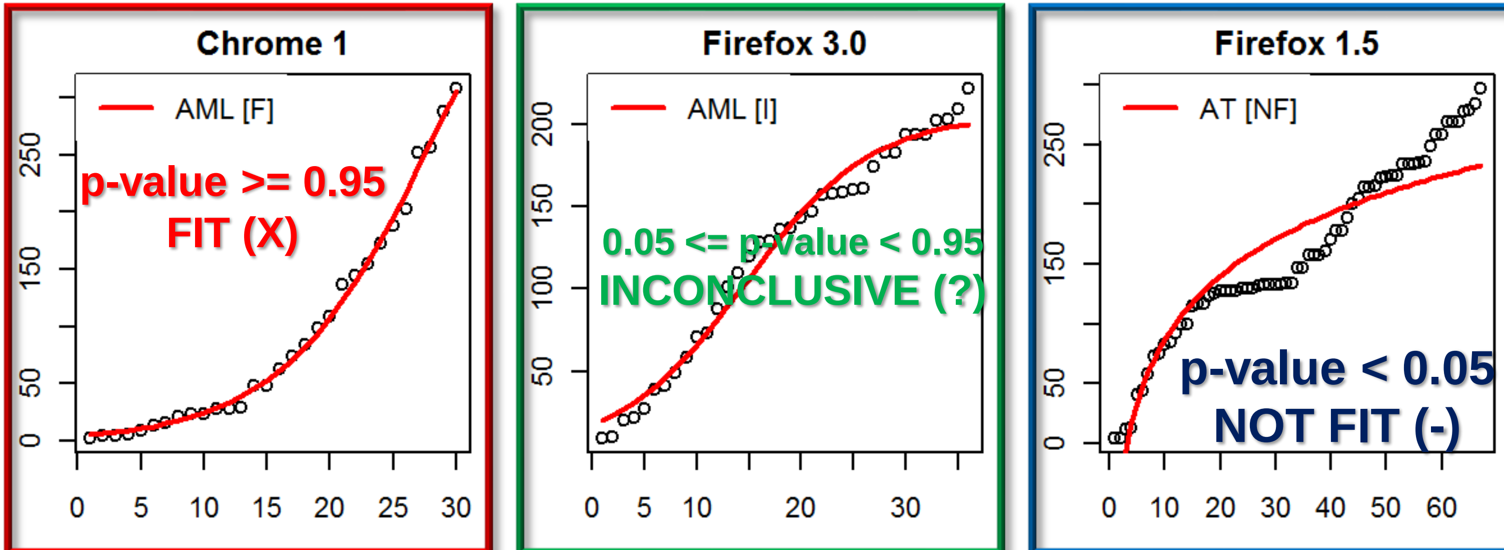
- Fit data to VDMs
  - ✓ Non-linear regression method, implemented in R (www.r-project.org)

- 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?

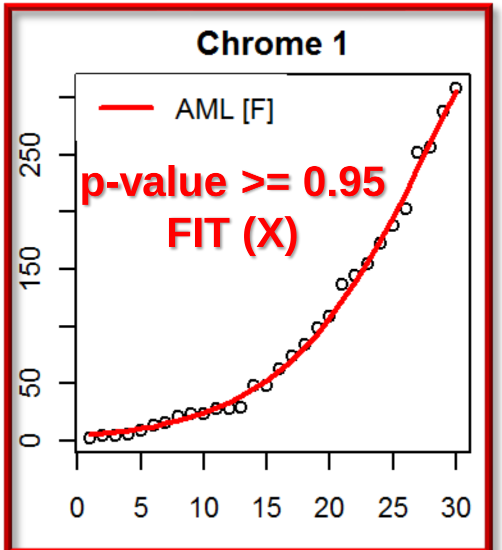


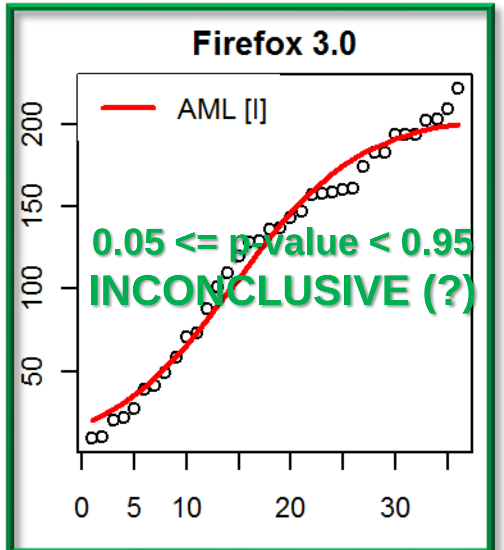
NVD Data set

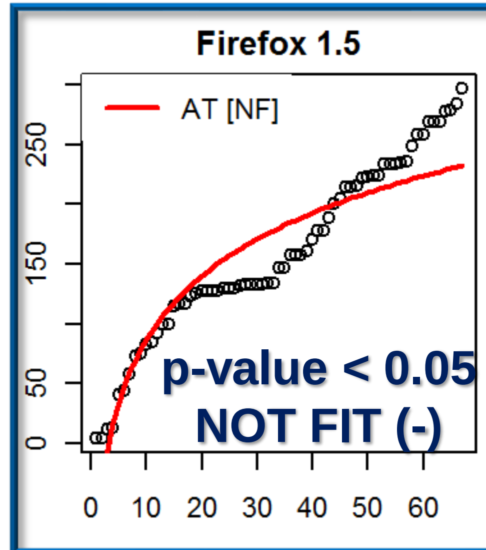
	Firefox						Chrome						IE				
Model	1.0	1.5	2.0	3.0	3.5	3.6	1.0	2.0	3.0	4.0	5.0	6.0	4.0	5.0	6.0	7.0	8.0
AML	–		?	?	?	?	X	?	?	?	?	?	X	?	?	–	X
AT	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	–
LN	–	–	X	–	X	?	–	–	–	?	–	–	–	–	–	?	?
LP	–	–	X	?	X	X	–	–	–	–	?	?	–	X	–	X	?
RE	–	–	X	?	X	X	–	–	–	–	?	?	–	X	–	?	?
RQ	–	–	–	?	?	X	–	–	?	?	?	?	–	–	–	–	X

The goodness of fit of a VDM is based on *p-value* in the χ^2 test. *p-value* < 0.05: not fit (–), *p-value* ≥ 0.95: good fit (X), and inconclusive fit (?) otherwise.

RQ1: Which VDM works, which doesn't?







NVD Data set

	Firefox						Chrome						IE				
Model	1.0	1.5	2.0	3.0	3.5	3.6	1.0	2.0	3.0	4.0	5.0	6.0	4.0	5.0	6.0	7.0	8.0
AML	–	–	?	?	?	?	X	?	?	?	?	?	X	?	?	–	X
AT	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	–
LN	–	–	X	–	X	?	–	–	–	?	–	–	–	–	–	?	?
LP	–	–	X	?	X	X	–	–	–	–	?	?	–	X	–	X	?
RE	–	–	X	?	X	X	–	–	–	–	?	?	–	X	–	?	?
RQ	–	–	–	?	?	X	–	–	?	?	?	?	–	–	–	–	X

The goodness of fit of a VDM is based on *p-value* in the χ^2 test. *p-value* < 0.05: not fit (–), *p-value* ≥ 0.95: good fit (X), and inconclusive fit (?) otherwise.

RQ2: The Impact of Types of Data Set

Advice.Nbug, NVD, NVD.Advice, NVD.Bug, NVD.NBug

	Firefox					
VDM	v1.0		v2.0	v3.0	v3.5	v3.6
AML	- - - - -	- - - ? X	? ? ? X X	? ? - ? X	X ? ? ? X	X ? ? ? X
AT	- - - - -	- - - - -	- - - - -	- - - - -	- - - - -	- - - - -
LN	- - - - -	- - - - -	- X X - -	? - ? ? ?	? X X X X	? ? ? ? X
LP	- - - - -	- - - - -	? X X - -	? ? ? ? -	- X X X -	? X ? ? ?
RE	- - - - -	- - - - -	- X X - -	- ? ? ? -	- X X X -	? X ? ? -
RQ	- - - - -	- - - - -	- - X ? -	- ? ? ? -	- ? X X -	- X ? ? -

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

Opposite results for the same models

* Opposite results are obtained from different data sets
  - ✓ Same model
  - ✓ Same target (ie. same software release)
  - ✓ But different counting methods (diff. types of data set)

RQ2: The Impact of Data Sets

Advice.Nbug, NVD, NVD.Advice, NVD.Bug, NVD.NBug

VDM	Firefox v1.0	[illegible]	v2.0	v3.0	v3.5	v3.6
AML	– – – – –	– – – ? X	? ? ? X X	? ? – ? X	X ? ? ? X	X ? ? ? X
AT	– – – – –	– – – – –	– – – – –	– – – – –	– – – – –	– – – – –
LN	– – – – –	– – – – –	– X X – –	? – ? ? ?	? X X X X	? ? ? ? X
LP	– – – – –	– – – – –	? X X – –	? ? ? ? –	– X X X –	? X ? ? ?
RE	– – – – –	– – – – –	– X X – –	– ? ? ? –	– X X X –	? X ? ? –
RQ	– – – – –	– – – – –	– – X ? –	– ? ? ? –	– ? X X –	– X ? ? –

VDM	Google Chrome v1.0	v2.0	v3.0	v4.0	v5.0	v6.0	IE v4.0	v5.0	v6.0	v7.0	v8.0
AML	○ X ○ ? X	○ ? ○ ? X	○ ? ○ ? X	○ ? ○ ? X	○ ? ○ ? ?	○ ? ○ ? ?	○ X X ○ ○	○ ? X ○ ○	○ ? ? ○ ○	○ – – ○ ○	○ X X ○ ○
AT	○ – ○ – –	○ – ○ – –	○ – ○ – –	○ – ○ – –	○ – ○ – –	○ – ○ – –	○ – – ○ ○	○ – – ○ ○	○ – – ○ ○	○ ? – ○ ○	○ – – ○ ○
LN	○ – ○ – –	○ – ○ – –	○ – ○ – –	○ ? ○ – –	○ – ○ – ?	○ – ○ ? ?	○ – – ○ ○	○ – X ○ ○	○ – X ○ ○	○ ? ? ○ ○	○ ? ? ○ ○
LP	○ – ○ – –	○ – ○ – –	○ – ○ – –	○ – ○ – –	○ ? ○ ? ?	○ ? ○ ? ?	○ – – ○ ○	○ X X ○ ○	○ – X ○ ○	○ X ? ○ ○	○ ? ? ○ ○
RE	○ – ○ – –	○ – ○ – –	○ – ○ – –	○ – ○ – –	○ ? ○ ? ?	○ ? ○ ? ?	○ – ? ○ ○	○ X X ○ ○	○ – – ○ ○	○ ? ? ○ ○	○ ? ? ○ ○
RQ	○ – ○ ? –	○ – ○ ? –	○ ? ○ ? –	○ ? ○ ? –	○ ? ○ ? ?	○ ? ○ ? ?	○ – – ○ ○	○ – – ○ ○	○ – X ○ ○	○ – ? ○ ○	○ X X ○ ○

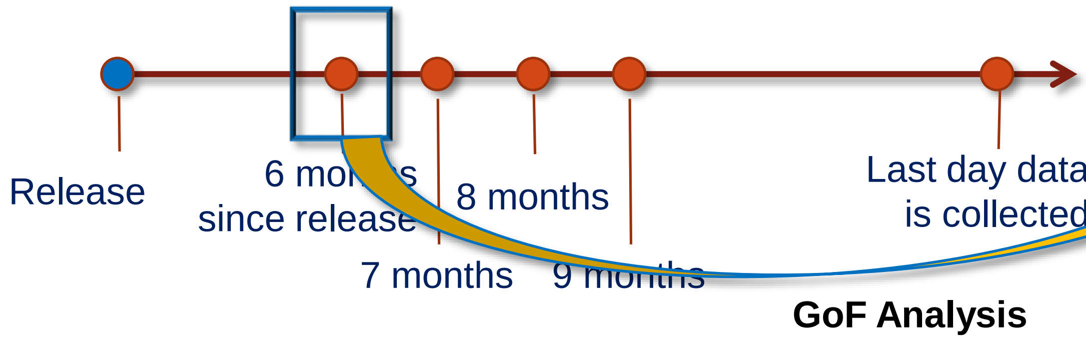
Each column has five cells corresponding to Advice.Nbug, NVD, NVD.Advice, NVD.Bug, NVD.NBug

Opposite results for the same models

* Different types of data set would strongly impact to VDM's GoF

Temporal Analysis on Goodness-of-Fit

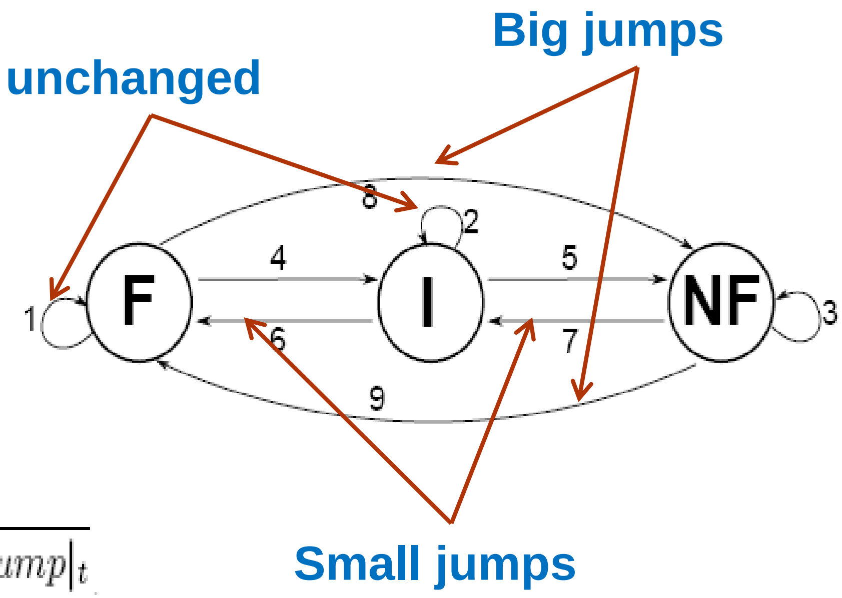
* Temporal Analysis on GoF



App.	Data Set	VDM	Time	GoF
X	**nvd**	**AML**		**NF**
X	nvd	AML		NF
X	nvd	AML		I
X	nvd	AML		F
...	…	…	...	…
X	nvd	AML		NF

14, 817 data points in total

Temporal Analysis on Goodness-of-Fit



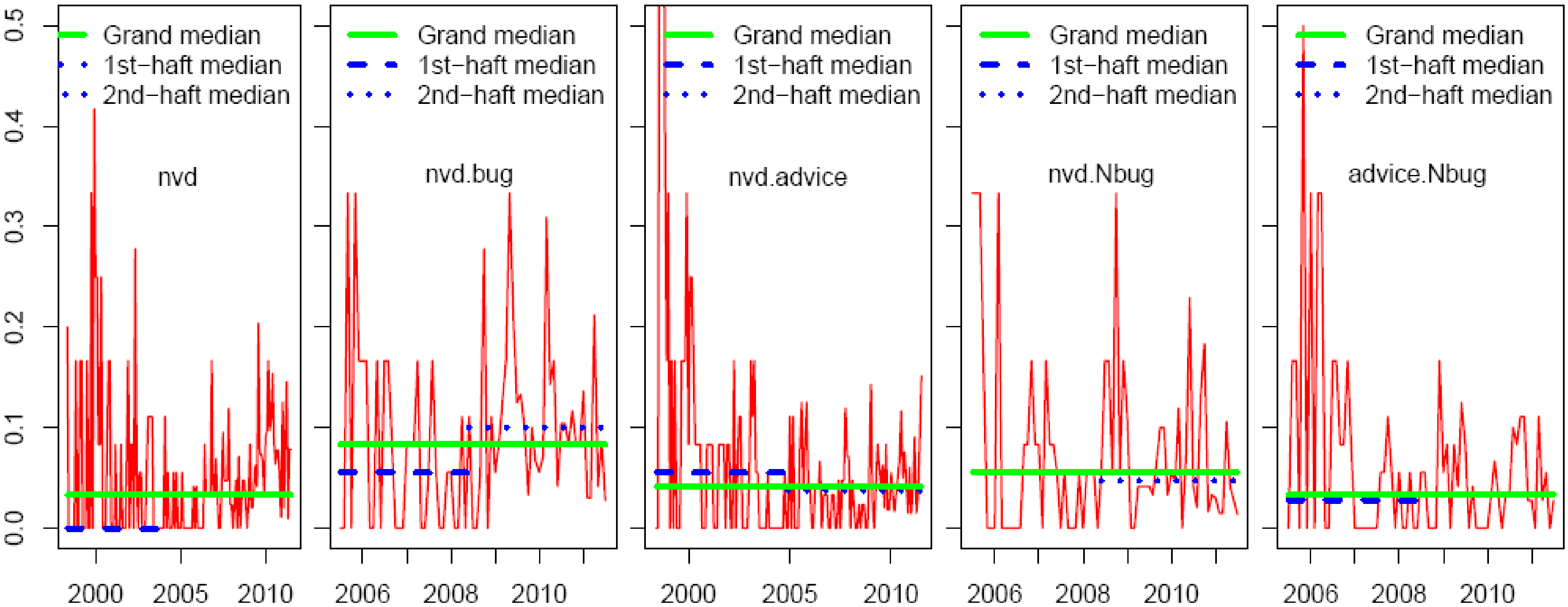
- The GoF Entropy of VDM
  - ✓ The chaotic of VDM's GoF from time t-1 to t
  - ✓ Measured by using the GoF transition diagram
  - ✓ Higher entropy, lesser stability

$$E_\beta(t) = \frac{|smalljump|_t + \beta \cdot |bigjump|_t}{|unchanged|_t + |smalljump|_t + \beta \cdot |bigjump|_t}$$

- The Quality of VDM
  - ✓ How good a VDM is
  - ✓ Measured by the #GoF at time t

$$Q_\omega(t) = \frac{|Fit|_t + {}^1/_\omega \cdot |Inconclusive|_t}{|Fit|_t + |Inconclusive|_t + |NotFit|_t}$$

RQ3: The Stability of VDMs in Data Sets

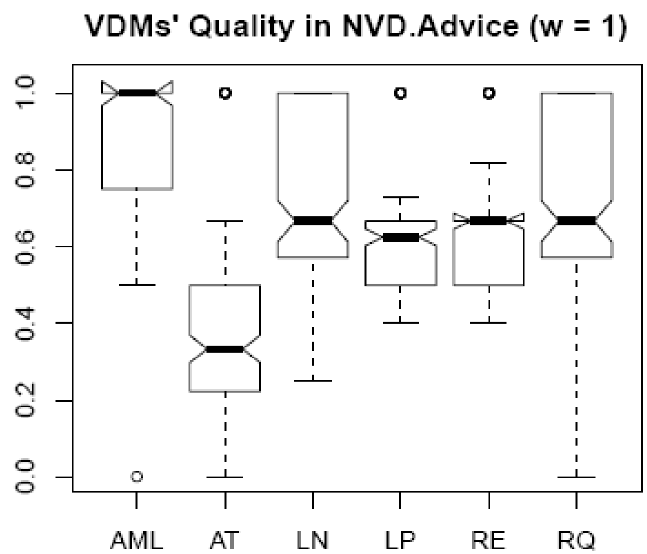


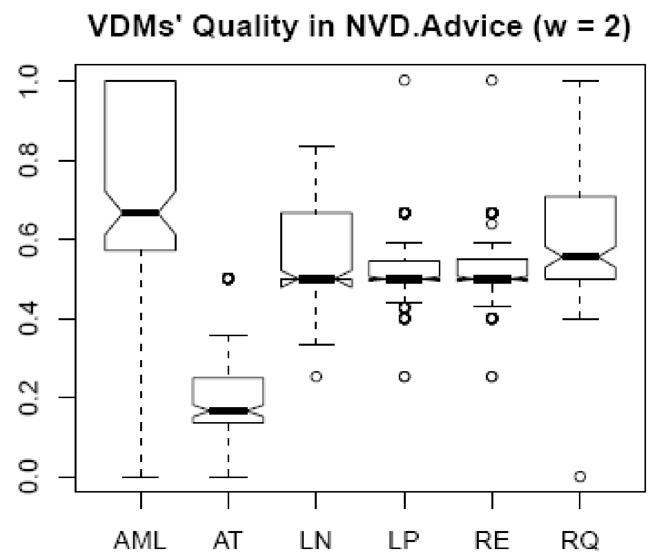
- The trend of GoF Entroy
  - ✓ VDM stability in NVD.Bug is likely the worst
  - ✓ VDM stability in NVD.Advice is likely the best

RQ4: The Quality of VDMs

* VDM Quality
  - ✓ AML is the winner
  - ✓ AT is the loser

$$Q_\omega(t) = \frac{|Fit|_t + {}^1/_\omega \cdot |Inconclusive|_t}{|Fit|_t + |Inconclusive|_t + |NotFit|_t}$$





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)
  - ✓ VDM (probably) work well: AML (for browsers)
  - ✓ VDMs might work: LN, LP, RE, RQ (for browsers)
  - ✓ Different types of data set would strongly impact to VDM's GoF
  - ✓ VDMs likely yield more stable result in Vulnerability-as-an-NVD entry confirmed by vendors' advisories data set (NVD.Advice)
- Future work
  - ✓ Replicate experiment in other types of application
    - E.g., Web Servers, Operating Systems,…

Thank you

Viet Hung Nguyen– vhnguyen@disi.unitn.it

Fabio Massacci – massacci@disi.unitn.it