Modeling Skewness in Vulnerability Discovery Models in Major Operating Systems

HyunChul Joh and Yashwant K. Malaiya
Computer Science Department
Colorado State University
Fort Collins, CO, USA
{dean2026, malaiya}@cs.colostate.edu

***Abstract*—A few vulnerability discovery models have been proposed recently. Studies have shown that the S-shaped AML vulnerability discovery model generally performs better than other models. The AML model assumes a symmetrical Logistic discovery pattern. This work examines the cases when discovery pattern is not symmetrical; thus potentially an asymmetrical discovery model might perform better. Here, new vulnerability discovery models based on asymmetrical S-shaped distributions are proposed, and their fit and prediction capabilities are compared. The results show that all the right skewed datasets are represented better with the Gamma distribution based model.**

Keywords-Vulnerability discovery model (VDM); risk assessment; security; S-shaped distribution; Operating Systems

I. INTRODUCTION

A vulnerability discovery model (VDM) can be used to assess the discovery process and for predicting future trends. A few VDMs have been proposed recently and evaluated using field data. The VDMs include the Alhazmi-Malaiya Logistic (AML) model which is the only S-shaped model. Studies [1][2] have shown that it generally performs better than other models. However, since the AML model assumes a symmetrical Logistic discovery process around the peak discovery rate value, it might not perform well when discovery behavior is asymmetrical. Barua and Srinivasan [3] have shown that a risk analysis ignoring the skewness can lead to inaccurate results. This study examines whether the performance of S-shaped VDMs is related to their underlying probability density functions (pdf) and skewness in the target vulnerability datasets.

II. DATASET & SKEWNESS

Skewness characterizes the degree of asymmetry of a distribution around its mean value. The vulnerability datasets, obtained from NVD [4], investigated here are Red Hat Linux (RHL), Red Hat Enterprise Linux (RHEL), Windows XP, and Windows Server 2003. As Fig. 1 shows, a right skewness indicates a distribution with a tail on its right side while a left skewness has a tail extending toward more negative values. TABLE I shows the calculated skewness for each dataset with the number of vulnerabilities; only Windows XP is skewed left, and others are skewed right.

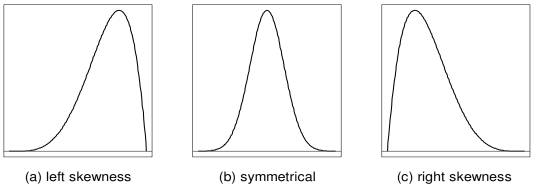
(a) left skewness (b) symmetrical (c) right skewness

Figure 1. When a tail is located at the left (right) side, it is called skewed left (right). Left (right) skewness is also termed negative (positive).

TABLE I. SKEWNESS AND TOTAL NUMBER OF VULNERABILITIES

	RHL	RHEL	Win XP	Server 2003
Skewness	0.4539	0.6918	−0.1394	0.2636
# of Vuln.	228	159	302	219

III. S-SHAPED VULNERABILITY DISCOVERY MODELS

Four S-shaped VDMs are examined here in addition to the AML model, which is based on the observation that the attention given to an operating system increases as it gains market share, it peaks at some time and then drops when a newer competing version is introduced. Equation (1) gives the model, where A and C are empirical parameters and B represents the total number of vulnerabilities.

$$\Omega_{AML}(t) = \frac{B}{BC^{-ABt}+1} \tag{1}$$

$$\Omega_{Weibull}(t) = \gamma\left\{e^{-\left(\frac{t}{\beta}\right)^{\alpha}}\right\} \tag{2}$$

$$\Omega_{\beta}(t) = \gamma\int_{\tau=0}^{t}\frac{\Gamma(\alpha+\beta)\tau^{\alpha-1}(1-\tau)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}d\tau \tag{3}$$

$$\Omega_{\gamma}(t) = \gamma\int_{\tau=0}^{t}\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\tau^{\alpha-1}e^{-\frac{\tau}{\beta}}d\tau \tag{4}$$

$$\Omega_{Normal}(t) = \left[1+\mathrm{erf}\left(\frac{t-\mu}{\sqrt{2s^2}}\right)\right] \tag{5}$$

$$Average\ bias\ error = \frac{1}{n}\sum_{t=1}^{n}\frac{\Omega_t-\Omega}{\Omega} \tag{6}$$

The Weibull distribution based VDM model, given in Equation (2), has been proposed by Kim [5] who initially pointed out that the asymmetrical S-shaped VDM might be an alternative worth considering. Here α and β are shape and scale parameters respectively and γ signifies the total number of vulnerabilities that would eventually be found. The model can express both right and left skewness.

We propose the following three S-shaped models to model skews.[1] The Beta distribution, given by Equation (3), can represent both positive and negative skewness, it is frequently used for skewed data [6]. Here both α and β are positive shape parameters, and γ has the same meaning as Equation (2).

[1] [8] briefly mentions testing Gamma and log-normal distributions for Windows XP in the context of patch release time, but found other models performs better for the specific case.

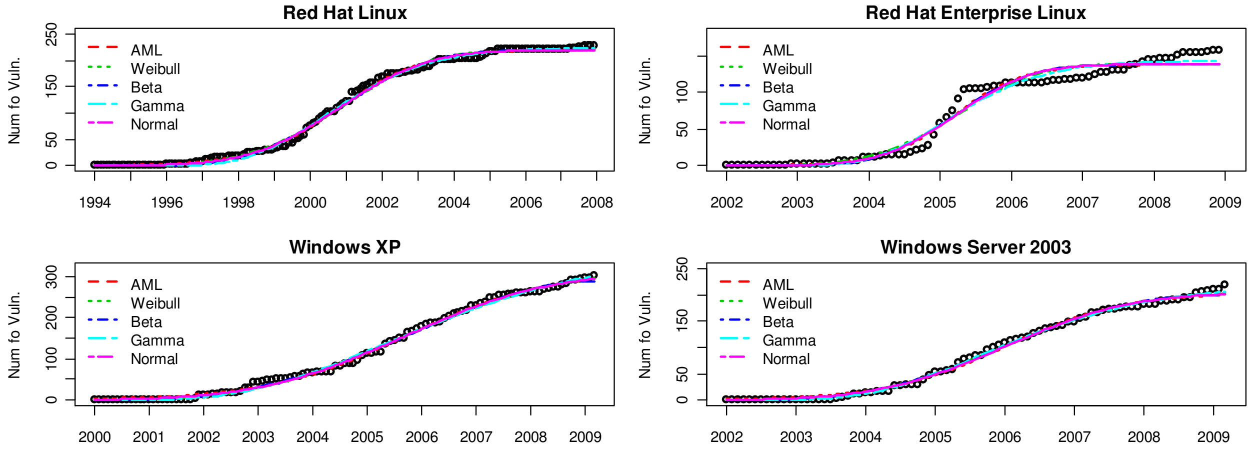

Figure 2. Fitting for the five VDM models. Datasets are obtained from NVD [4] on March 2009. Inspite of the different time ranges, all the five S-shaped VDMs fit very well except for RHEL.

TABLE II. χ^2GODDNESS OF FIT TEST P-VALUES (α: 0.05)

	AML	Weibull	Beta	Gamma	Normal
RHL	1.0000	0.9999	0.9992	0.9978	0.9998
RHEL	0.0019	0.0000	0.0001	0.0001	0.0001
XP	1.0000	0.9925	0.9941	0.9209	0.9785
2003	0.9217	0.9950	0.9953	1.0000	0.9500

TABLE III. AVERAGE BIAS ERROR (UNIT: %)

	AML	Weibull	Beta	Gamma	Normal
RHL	-6.7354	-8.8726	-7.6986	-4.1872	-7.4433
RHEL	-21.6734	-22.9155	-22.2010	-21.2352	-22.1414
XP	-0.8953	16.1234	17.0792	21.0189	2.8674
2003	-18.1470	-17.9880	-18.2970	-11.2597	-17.7703

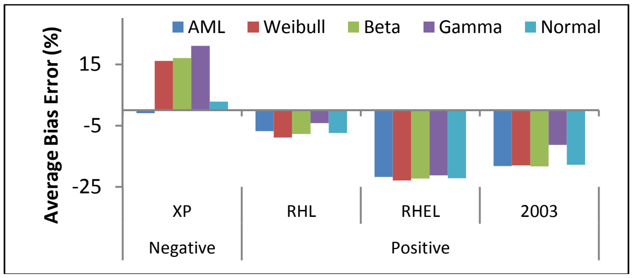

Figure 3. Average bias

The Gamma distribution based VDM, given by Equation (4), can only represent right skewed distributions. Here α, β and γ have the same significance as in Equation (2). Equation (5) shows the Normal distribution based VDM, where μ is a location parameter, s is a scale parameter, and γ has the same meaning with Equation (2). It is similar to the logistic distribution used by AML model in shape but has lighter tails in both sides at its pdf.

IV. MODEL COMPARISONS & OBSERBVATIONS

Fig. 2 shows the fitted data, and TABLE II gives the corresponding χ^2 values. P-values closer to 1 mean good fits and less than 0.05 imply that fits are not significant. All the five s-shaped VDMs fit equally well except for RHEL.

Fig. 3 and TABLE III give the average bias in prediction. Average bias assesses the general tendency of the model to overestimate or underestimate. In Equation (6), n is total time in months, and Ω is the actual number of total vulnerabilities, whereas Ω_t is the number of vulnerabilities estimated at time t. The results clearly show that all the positive skewed datasets are better represented with the Gamma VDM which confirms a strong relationship between the skewness presumed by the model and that inherent in a dataset. Results also show that Gamma VDM performs the worst with the negatively skewed dataset.

AML and Normal VDMs perform better with the left skewed dataset. Weibull and Beta VDMs, which can model both left and right skewed behavior, did not support a significant relationship between their presumed skew and the skewness in the datasets.

For all the right skewed datasets, the five VDMs always underestimate the number of vulnerabilities (below 0% error line at Fig. 3) suggesting that they are very good candidates for the recalibration which relies on the consistency of the bias for adjusting the future predictions [7].

The results suggest that the Gamma VDM is a better candidate for modeling the vulnerability discovery process with right skewed datasets. For other datasets, AML is generally a better choice. We plan to examine other types of software systems for further confirmation of the results.

REFERENCES

[1] Alhazmi, O. H. and Malaiya, Y. K., Application of Vulnerability Discovery Models to Major Operating Systems, *IEEE Trans. on Reliability*, 57(1), 2008, pp. 14-22.

[2] Alhazmi, O. H., Malaiya, Y. K., and Ray, I., Measuring, Analyzing and Predicting Security Vulnerabilities in Software Systems, *Computers & Security*, 26(3), 2007, pp. 219-228.

[3] Barua, S. K., and Srinivasan, G., Investigation of decision criteria for investment in risky assets, *Omega*, 15(3), 1987, pp. 247-253.

[4] National Vulnerability Database, http://nvd.nist.gov/

[5] Kim, J., *Vulnerability Discovery In Multiple Version Software Systems: A Open Source And Commercial Software System*, Thesis, CS Dept., Colorado State University, 2007.

[6] Moitra, S. D., Skewness and the Beta Distribution. *The Journal of the Operational Research Society*, 41(10), 1990, pp. 953-961.

[7] Brocklehurst, S., Chan, P. Y., Littlewood, B. and Snell, J., Recalibrating Software Reliability Models. *IEEE Trans. on Software Engineering*, 16(4) , 1990, pp. 456-470.

[8] Okamura, H., Tokuzane, M., and Dohi, T., Optimal Security Patch Release Timing under Non-homogeneous Vulnerability-Discovery Processes, 20th ISSRE 2009, pp.120-128.