Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

The sigmoidal growth of operating system security vulnerabilities: An empirical revisit

Jukka Ruohonen *, *Sami Hyrynsalmi*, *Ville Leppänen*

Department of Information Technology, University of Turku, FI-20014 Turun yliopisto, Finland

ARTICLE INFO

Article history:
Received 25 January 2015
Received in revised form 30 June 2015
Accepted 6 July 2015
Available online 13 July 2015

Keywords:
Software vulnerability
Growth curve
Gompertz
Replication
Operating system
Technology diffusion

ABSTRACT

Purpose. Motivated by the calls for more replications, this paper evaluates a theoretical model for the sigmoidal growth of operating system security vulnerabilities by replicating and extending the existing empirical evidence. *Approach*. The paper investigates the growth of software security vulnerabilities by fitting the linear, logistic, and Gompertz growth models with nonlinear least squares to time series data that covers a number of operating system products from Red Hat and Microsoft. *Results*. Although the fitted models are not free of statistical problems, the empirical results show that a sigmoidal growth function can be used for descriptive purposes. The paper further shows that a sigmoidal trend applies also to the number of software faults that were fixed in the Red Hat products. *Conclusion*. The paper supports the contested theoretical growth model. The few discussed theoretical problems can be used to develop the model further.

1. Introduction

This paper evaluates and extends the existing empirical results concerning the presumed sigmoidal growth of software security vulnerabilities in operating system software products. The original theoretical model and the initial empirical results were presented by Alhazmi and associates (Alhazmi et al., 2005; Alhazmi and Malaiya, 2008; Alhazmi et al., 2007). The model is an important contribution to the ongoing efforts to understand the discovery of software vulnerabilities by means of time series analysis. This provides the underlying rationale for the replication: a solid empirical foundation is a prerequisite for further theoretical advances and practical applications.

The empirical target in the model relates to the long-run software life cycles that are typical to operating systems. If the cumulation of vulnerabilities tends to follow a systematic sigmoidal pattern across these cycles, this information alone can be valuable when different software and release engineering decisions are made regarding further cycles. The presumed theoretical explanation is not explicitly related to software engineering, however.

The basic theoretical argument is that sigmoidal vulnerability growth trends tend to follow the popularity of operating systems: once the popularity reaches an inflection point, decreasing rate of adoption makes the given operating system less lucrative for exploitation development, which leads to fewer and fewer vulnerabilities (Alhazmi et al., 2005). Eventually a saturation point is reached; the operating system product has been substituted by a new product, and a new sigmoidal growth function takes over.

This theoretical background places the model into an interesting intersection of empirical software engineering research (Massacci and Nguyen, 2014). At the methodological front the model largely falls to the domain of software reliability modeling, but the theoretical argument is distinctively different.

* *Corresponding author*. Tel.: +358 (0)44 326 4270.
E-mail addresses: juanruo@utu.fi (J. Ruohonen), sthyry@utu.fi (S. Hyrynsalmi), ville.leppanen@utu.fi (V. Leppänen).
http://dx.doi.org/10.1016/j.cose.2015.07.001
0167-4048/

Accordingly, the amount of discovered vulnerabilities does not slow down because security issues would be harder to find as time passes, but rather because there is a decreasing interest to find vulnerabilities from old software products. Consequently, security bugs are different from normal bugs, and, by implication, vulnerability modeling is different from reliability modeling. Besides these software engineering aspects, the model contains also theoretical presumptions that can be related particularly to the different sigmoidal models for the progress of technology. This wider theoretical scope raises the relevance of the replication: provided that the foundation holds, further opportunities exist for interdisciplinary research.

The paper proceeds by first discussing the larger theoretical background behind sigmoidal growth models, connecting the vulnerability model to technology diffusion and software life cycle models. Three hypotheses are also postulated for contesting the model empirically: (a) vulnerability discovery follows a sigmoidal trend; (b) the same applies to fixed software faults; and (c) major and minor operating system releases do not systematically differ with respect to the presumed growth trends.

These three hypotheses are tested in the empirical part of the paper. The primary empirical dataset contains time series from 29 deprecated Red Hat operating system products. In order to assess the generalizability of the results, the presence of a sigmoidal trend is further evaluated with a secondary dataset that contains 40 operating system products from Microsoft. The time period covered spans from circa 2000 to 2014. Estimation is carried out by testing the linear, logistic, and Gompertz growth functions by nonlinear least squares. Model comparisons are carried out with an information criterion measure and different statistical tests. Given the general need for sensitivity (Höök et al., 2011; Suominen and Seppänen, 2014) and model assessments (López et al., 2004; Meade and Islam, 2006; Wang and Bushman, 2006) in growth curve modeling, the estimated models are exposed to a few conventional statistical tests over the basic time series characteristics. In general, however, the paper adopts a viewpoint that growth curves are essentially stylized and descriptive characterizations of the underlying time series trends.

Finally, a few brief remarks should be made to clarify the replication approach itself. Like any replication, the paper ultimately tests an existing theory and verifies existing empirical results. In contrast to the term reproduction, the term replication is, consequently, defined in this paper to mean that "*independent researchers are able to reach the same qualitative conclusions by repeating the original study using the same methods on different data* (Boylan et al., 2015, p. 81)". This fundamental goal can be narrowed with the different replication functions that were recently described by Gómez et al. (2014). Thus, the replication seeks: (1) to validate hypotheses; (2) to control potential sampling errors in the original empirical experiment; (3) to understand potential population limits regarding generalizability; and (4) to affirm that researcher bias (Shepperd et al., 2014) is not present. The four replication functions are equally important.

2. Theoretical background

The contested theoretical model is simple. In essence, the model states that (a) the logistic growth curve can be used to empirically describe the cumulative amount of security vulnerabilities that have affected a software product, and that (b) the observed sigmoidal growth pattern can be theoretically related to the popularity of the examined product (Alhazmi et al., 2005). These simple but fundamental presumptions are easy to relate to the general growth curve modeling literature and to a few specific topics in software engineering.

2.1. *Sigmoidal growth*

If growth is taken to follow a S-shaped curve that is determined by a given sigmoid function, there are only a limited number of theoretically meaningful parameters that can be derived from the curve. After a certain *lag phase* λ, the growth rate eventually leads to the upper *asymptote* α. The used functional forms imply also a certain point at which the growth attains its *maximum slope* μ. In essence, besides the actual statistical fit, these three parameters are the primary interest in many empirical growth curve modeling applications. The delivered theoretical message is consequently simple: although growth starts slow (λ), it accelerates rapidly until the maximum growth rate is reached (μ), after which it slows down but still approaches a saturation point asymptotically (α). The growth is thus bounded and the maximum growth rate separates two growth phases.

The interpretation given for the two phases varies from an application domain to another. It is still possible to reach the fundamental theoretical trait by considering the first phase as a replication process during which growth is proportional to an existing population, while maintaining that the second phase signifies an inhibiting process during which the population reaches its stable carrying capacity (Cunningham and Kwakkel, 2014). These dual premises were common in the early models for the growth of human populations. In other words, at any instant of time, the population growth rate was taken to be proportional not only to the already attained growth – the already existing population, but also to the unutilized potential in a given limited area to support the existing population (Pearl and Reed, 1920). The latter premise can be located also from many models for the progress of technology.

In particular, it has been presumed that advancements occur through competitive substitution of different technologies; coal for wood, mobile telephones for landlines, IPv6 for IPv4, and so forth. Once a substitution process has initially taken off, it is assumed to always reach the saturation point, but the rate of substitution of new for old is proportional to the remaining old technology to be substituted (Fisher and Pry, 1971). When a few more steps are taken along this theoretical path, more general technology life cycle models quickly emerge in the horizon. In these models a given S-shaped curve is theoretically divided into different stages such as emergence, growth, maturity, and eventually the saturation stage at which another substitution process may again begin (Gao et al., 2013). A comparable fivefold theoretical construct would be the hypothetical progress of a technology through basic research, applied research, development, application, and eventually the wider social impact (Suominen and Tuominen, 2010). These general models for the progress and adoption of technology fall under the label of technology diffusion models, which can be seen to differ from the noted substitution models in that the latter require an assumption of existing markets; there must be something old to be substituted for something new (Norton and Bass, 1987). These