December 2005

An Extreme Value Approach to Information Technology Security Investment

Jingguo Wang
State University of New York, Buffalo

Abhijit Chaudhury
Bryant University

Raghav Rao
State University of New York, Buffalo

AN EXTREME VALUE APPROACH TO INFORMATION TECHNOLOGY SECURITY INVESTMENT

Jingguo Wang
School of Management
State University of New York, Buffalo
Buffalo, NY U.S.A.
wang7@buffalo.edu

Aby Chaudhury
CIS Department
Bryant University
Smithfield, RI U.S.A.
achaudhu@bryant.edu

H. Raghav Rao
School of Management
State University of New York, Buffalo
Buffalo, NY U.S.A.
mgmtrao@buffalo.edu

Abstract

Information technology security investment is receiving increasing attention in recent years. Various methods have been proposed to determine the effective level of security investment. In this paper, we introduce an extreme value approach to address the issues of effective budgeting and investing in IT security. In our model, the security status of a system depends on two factors: system security level, which is measured by the level of security investment, and system attack level, which reflects the security risk with which the system is confronted. Security investment level is endogenous to the system, while attack level is exogenous. Extreme value analysis is used to characterize the stochastic behavior of high-level attacks based on the historical data and to make inferences on future attacks. Based on these inferences, we determine the effective security solutions and the level of security investment to modulate the likelihood of system failure. For illustration purposes, we use an extreme value approach to analyze a set of traffic data collected from a regional bank.

Keywords: Information assurance, security investment, two-factor model, extreme value theory, denial of service (DoS)

Introduction

The importance of effective management of information technology security has increased in recent years due to the increasing frequency and cost of security breaches (Gordon et al. 2005). While high-risk organizations may adopt security at any price, most commercial organizations have to consider the cost-benefit tradeoff for such an investment. How to efficiently invest in IT security is a big challenge. In the Ernst & Young Global Information Security Survey (Ernst & Young 2003, 2004), budget constraints are listed as one of the main obstacles to effective information security. Quantification tools, if applied prudently, can assist in the anticipation and control of direct and indirect computer security cost (Geer et al. 2003; Mercuri 2003).

In this paper, we propose an approach based on extreme value theory (Gumbel 1958) for IT security investment. In our model the security status of a system depends on two factors: system security level, which is measured by the level of security investment, and system attack level, which reflects the security risk with which the system is confronted. Attack level is treated as an exogenous variable that causes system failure, while security investment level is endogenous, preventing system failure, and is determined by organizations. The difference between security level and the attack level measures the vulnerability of the system.

Instead of calculating the expected loss value, we apply extreme value theory to study the extreme attack behavior and address the issues of effective budgeting and investing in IT security. Extreme value theory is one of the most important statistical disciplines for the applied science, and has found applications in engineering (Castillo 1988), insurance and finance (Embrechts et al. 1997), and management strategy (Dahan and Mendelson 2001), as well as in environment and biomedical research. Extreme value theory quantifies the stochastic behavior of a process at unusually large (or small) levels. It is concerned with probabilistic and statistical questions related to those extreme events. To our knowledge, this is the first paper to apply the extreme value theory in security investment decisions. With the application of extreme value theory, we attempt to address the following issues:

1. What is the probability distribution of high-level attacks (i.e., what is the probability that an attack over a given level will occur during a given year)?

2. What security investment is needed so that the probability of potential system failure is below a certain threshold?

3. What are the factors affecting the behavior of high-level attacks? Are the nature and causes of high-level attacks changing over time? Is it a seasonal phenomenon?

By answering these questions, we make inferences on future attacks, thus determining the effective security solutions and investment level to modulate the likelihood of system failure.

Consider defending against denial of service (DoS) as an example. A DoS attack is an incident in which an organization is deprived of the services of a resource they would normally expect to have. A Web site can occasionally be forced to temporarily cease operation when accessed by millions of people. High-level traffic is always regarded as a signal of forthcoming DoS attacks. Suppose that, as part of our design criteria to defend against DoS, the Web server is required to be able serve all traffic that it is likely to experience within its projected life span, say 1 year (or more years). Daily traffic is monitored and historical data might be available for the last 2 years. The challenge is to estimate the traffic level that might occur over the next 1 year given the 2-year history. Extreme value theory provides a framework enabling such extrapolations. Using extreme value theory, we may not only estimate the distribution of high-level traffic and the occurrence probability of the traffic over a given level during a given year, but also answer such questions as what level of traffic will be exceeded with probability 1/365 in a given day. In addition, we may identify factors that influence the behavior of high-level traffic with proper regression analyses, thus helping us predict the trend of traffic with the change of environment and time. Based on the characterization of extremely heavy traffic, we make inferences on future attacks and determine a proper security solution and the level of investment.

The organization of this paper is as follows. In the next section, after a literature review, we introduce our two-factor security model. We then present the extreme value theory. We apply extreme value analysis to a set of daily internal traffic data collected from a regional bank for illustrative purposes. Finally, we summarize our study and discuss ideas for future research.

Security Risk and Security Investment

Related Literature

Several models have been proposed to determine the effective level of security investment. There are basically two approaches (Cavusoglu 2004).

1. Using traditional risk or decision analysis framework. Generally these models apply a standard result in optimal-control theoretic certainty equivalence, which implies that only the mean values (probability-weighted average outcomes) of target variables matter for an optimal policy setting. Gordon and Loeb (2002) proposed an expected benefits of investment in information security (EBIS) model. Hoo (2000) used a decision analysis approach to evaluate different policies for IT security. Longstaff et al. (2000) proposed a hierarchical holographic model (HHM) to assess security risks and provide a model for assessing the efficacy of risk management.

2. Using game theory to model the strategic interactions between the organizations and attackers. Some researchers argue that IT security can be treated as a kind of game between organizations and attackers. While the organizations try to cover vulnerabilities in their systems, attackers race in an effort to exploit them. Security investments not only prevent security breaches by reducing vulnerabilities that attackers can exploit but also act as a deterrent for attackers by making attacks less attractive (Schechter and Smith 2003). Longstaff et al. (2000)argued that investment in system risk assessment can reduce

the likelihood of intrusions, which yields benefits much higher than the investment. Cavusoglu, et al. (2004) constructed a game tree to describe the interaction between organization and hackers.

There are two main issues in these models.

1. The expected loss value or benefit value cannot fully characterize security failures. Usually security failures are low-probability events, but once realized, failures can bring huge loss. The loss may be intangible and not amenable to accurate estimation.

2. Rationality of hackers is hard to capture as they may be motivated by a different value system. They may be rational, but not in our terms. They may be driven by motivations other than money. It is hard for us to know their cost function for attacking the system.

A Two-Factor Security Model

Security risk assessment determines the level of security risk that exists within the organization. Farahmand et al. (2005) presented a subjective analysis and probability assessment with a damage evaluation of information security incidents. Geer et al. (2003) introduced a technique called business-adjusted risk (BAR) for classifying security defects by their vulnerability type, degree of risk, and potential business impact. In this paper, we define *attack level* ***a*** as a metric, which reflects the threats that an organization confronts, in the same manner as the temperature reflects the relative warmth and cold of a day and the Dow Jones Index reflects the healthiness of the stock market. Attack level may be evaluated daily or monthly based on the information on hackers, worms, virus, and other attack incident. A similar idea is used in the Homeland Security Advisory System (**http://www.dhs.gov/dhspublic/display?theme=29**). On a daily basis of monitoring and analyzing threat information, the government may issue a threat level to reflect the current situation (severe—red; high—orange; elevated—yellow; guarded—blue; and low—green). In defending against DoS, organizations may monitor the daily traffic, and regard the level of traffic as the attack level on systems.

We define *security level* ***s*** as the ability of an organization to defend its IT systems from failure resulting from a security attack, such as the capability that an organization has in defending the system again a DoS attack. The system's security level is converted from the organization's *security investment* ***i***. By investing in IT security (training security staff; buying new technologies such as an intrusion detection system, a firewall, etc.; timely installation of software patches; and increasing the system capacity), the organization improves its system security level. For simplicity in our discussion, we assume that the level of IT security investment is equivalent with the security level system in our model.

Schechter (2004) argued that when attacking a software system is only as difficult as it is to obtain a vulnerability to exploit, the security strength of that system is equivalent to the market price of such vulnerability. He suggested that the strength of a system's security should be quantified from the viewpoint of the attacker rather than the defender, and introduces an approach that security strength can be measured using a market mechanism. In our paper, we use the difference between security level and the attack level, which is the term *i-a*, to measure the vulnerability of the system. We assume that both the investment level ***i*** and the attack level ***a*** are continuous.

The security status of information systems is affected only by these two factors. We define a system survival function (the probability function of system failure) (F) depending on the probability that $i - a \leq \upsilon$, where υ is a certain threshold of vulnerability; that is,

$$\mathsf{F} = \text{probability of failure} = \text{prob.}\ (i - a \leq \upsilon) \tag{1.1}$$

where ***i*** is the investment level and ***a*** is the attack level. F increases when a increases, and decreases when i increases. When ***(i – a)*** increases, F decreases. If we assume $\upsilon = 0$, the probability of system failure depends on the probability that $i \leq a$ (i.e., the probability that investment level is less than equal to attack level). a (attack level) is exogenous in the function F, while ***i*** (investment level) is endogenous. An organization follows a dynamic investment strategy, in which it makes investment decisions based on attack level a and the status of the system F; that is,

$$i = i(a, \mathsf{F}) \tag{1.2}$$

Table 1. Notations	
Two-Factor Security Model	
F	System survival function (the probability function of system failure)
i	Security investment level
a	System attack level
Extreme Value Theory	
X	Random observations, x is one observation from X
F	Common cumulative distribution function F of random observations
M_n	Block maxima, $M_n = \max\{X_1, X_2, ...X_n\}$, and n is the number of observations.
$H(x)$	The limiting distribution of exetrma
μ, σ, α	Distribution Parameters of Frechet, Weibull, Gumbel, as well as Generalized Extreme Value distribution (GEV)
ξ	Shape Parameter of GEV
x_p	Return Level
$\hat{x}_p$	Estimated Return Level
$1/p$	Return Period, p is a probability
u	High threshold
Y	$Y = X - u > 0$, y is an observation from Y
G	Generalized Pareto Distribution (GPD)
ξ, σ, ψ	Distribution Parameters of GPD
Z_t	The maximum level in time period t

In the example of defending against DoS, the system status depends on the system capacity and the daily traffic experienced by the system. The probability of system failure is determined by the probability that the system capacity is less than the daily traffic. Based on the observed daily traffic, the organization determines proper security solutions.

One of the key requirements for such a dynamic investment strategy is to accurately capture and model the dynamic behavior of attacks. With our two-factor security model, it is important for us to know the behavior of extreme attacks for an effective investment. To defend against DoS, we need to understand the behavior of high-level traffic so that we can make inferences on future attacks and design proper defense solutions to prevent the system failure caused by extremely heavy traffic. In the following section, we introduce extreme value theory, which we use to characterize the stochastic behavior of high-level attacks and to identify the factors (including time) that may influence high-level attacks. Table 1 lists the notations we use in the analysis.

Extreme Value Analysis

Classic Extreme Value Theory

The principal results of extreme value theory concern the limiting distribution of sample extrema (maxima or minima). Since in our model the probability of system failure depends on the probability of the exceedance of attack over investment and we are concerned with the behavior of extreme large attacks, such as the distribution of high-level traffic, we will only discuss sample maxima here. Suppose that $X_1, X_2, \ldots X_n$ is a sequence of independent, identically distributed observations, such as n-day daily traffic in DoS, with a common cumulative distribution function F, which is not necessarily known. Let the sample maximum be denoted by $M_n = \max\{X_1, X_2, \ldots X_n\}$ (M_n is also referred as block maxima). We are interested in the stochastic behavior of M_n. We know

$$\Pr\{M_n \le x\} = F(x)^n \tag{1.3}$$

Result (1.3) is of no immediate interest, since it simply says that for any fixed x for which $F(x) < 1$, we have $\Pr\{M_n \le x\} \to 0$ as n goes to infinity. For nontrivial limit results we must renormalize: find $a_n > 0$, b_n such that

$$\Pr\left\{\frac{M_n - b_n}{a_n} \le x\right\} = F(a_n x + b_n) \to H(x) \tag{1.4}$$

where $H(x)$ is the limiting distribution of $F(a_n x + b_n)$. The fundamental theorem of the extreme value theorem provides three possible distributions for $H(x)$ as follow:

Theorem 1 (Fisher and Tippett 1928): The only three types of non-degenerate distributions $H(x)$ satisfying Equation (1.4) are

$$\text{Frechet: } H(x) = \begin{cases} e^{-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}} & if\ x \ge \mu \\ 0 & otherwise \end{cases} \tag{1.5}$$

$$\text{Weibull: } H(x) = \begin{cases} 1 & if\ x \ge \mu \\ e^{-\left(\frac{\mu-x}{\sigma}\right)^{\alpha}} & otherwise \end{cases} \tag{1.6}$$

$$\text{Gumbel: } H(x) = e^{-e^{-\left(\frac{x-\mu}{\sigma}\right)}} \quad -\infty \le x \le \infty \text{ and } \sigma > 0 \tag{1.7}$$

The three extreme-value distributions, normalized to mean and unit variance, are shown in Figure 1 ($\alpha = 3.9$ for Frechet, and $\alpha = -2$ for Weibull). The stars indicate the 99th percentile of the distributions respectively.

Frechet: This is the "long-tailed" case. The underlying attack-level distribution $H(x)$ for Frechet distribution has a fat tail (e.g., $1 - H(x)$ declines as $x^{-\alpha}$). The attack level confronted by organizations has great upside uncertainty.

Gumbel: This is the "medium-tailed" case for which $1 - H(x)$ decreases exponentially for large x. For this type of attack distribution, there are no specific limits on the attack level, but the attack level is not likely to be too high or too low. Most attack levels are distributed in a central range.

Weibull: This is the "short-tailed" case in which the distribution has a finite endpoint. Organizations face predictably finite attack levels.

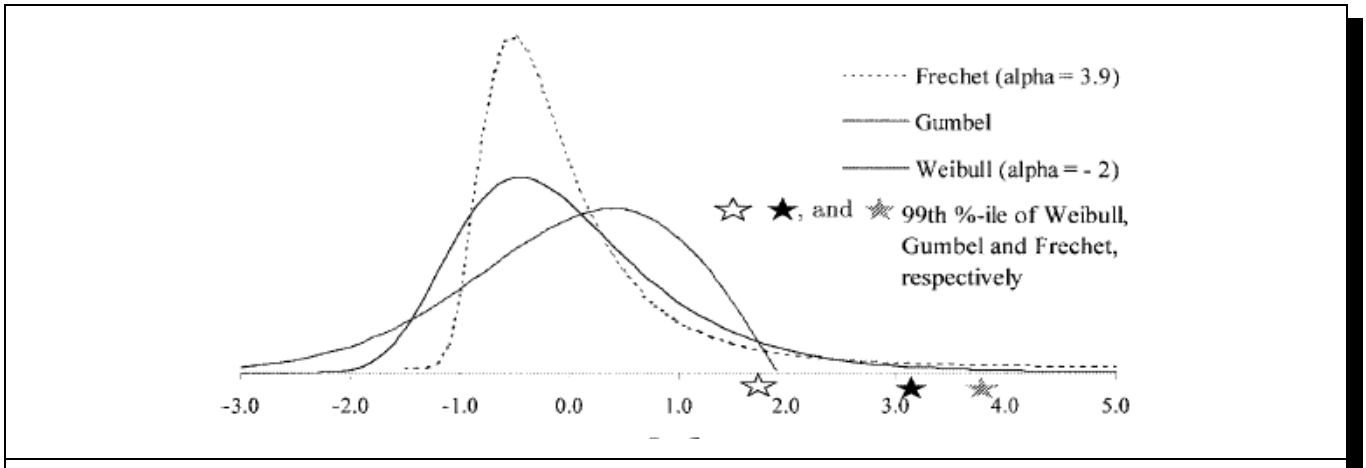

Figure 1. Densities for the Three Extreme Value Distributions ($\mu = 0$, $\sigma^2 = 1$)

There are several reasons why there may be different probability distributions for the high-level attack. First, it may be due to the nature of digital assets: the perceived value of digital assets and their criticalness to the public and organizations. Terrorism or engineered attacks are more attracted to high-valued digital assets, or the digital assets that can cause high-level damage to the public and/or organizations. Second, different attack types may have different distributions. For example, a violation that is initiated from a finite number of internal users in an organization is likely to differ from threats from viruses or worms, which can originate anywhere in the world. Third, the exposure range of the digital assets may also result in different distributions of the attack level. The digital assets connected with the Internet are more likely to come under high-level attack, while application systems having limited access in an isolated environment are less likely to be exposed to same level attack. Fourth, due to negative externality of attack (Camp and Wolfram 2004), the size of the organization and its network becomes a factor. A larger organization is more likely to suffer an attack than a smaller one. However, empirical exploration of these hypotheses is needed.

The three types may be combined into a single generalized extreme value (GEV) distribution (Coles 2001a, p. 48).

$$H(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{1.8}$$

where $\left\{x : 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$, μ is a location parameter, $\sigma > 0$is a scale parameter and $\xi$is a shape parameter. The limit $\xi \to 0$ corresponding to the Gumbel distribution, $\xi > 0$ to the Frechet distribution with $\alpha = \frac{1}{\xi}$, $\xi < 0$ to the Weibull distribution with $\alpha = -\frac{1}{\xi}$.

By inversing the equation (1.8), we obtain

$$x_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left\{1-\left[-\log(1-p)\right]^{-\xi}\right\}, & for\ \xi \neq 0 \\ \mu - \sigma\log\left\{-\log(1-p)\right\}, & for\ \xi = 0 \end{cases}$$

where $G(x_p) = 1 - p$. In common terminology, x_p is the **return level** associated with the **return period** $1/p$, since, to a reasonable degree of accuracy, the level x_p is expected to be once every $1/p$ periods. In other word, x_p is exceeded by the period maximum in any particular period with probability p.

Exceedances over Thresholds

Extremes are scarce, so model estimations of block maxima have a large variance (Coles 2001a, p. 66). Modeling block maxima is a wasteful approach to extreme value analysis especially if one block happens to contain more extreme observations than another. If an entire time series of, say, hourly or daily observations are available, the data may be better used by avoiding the procedure of blocking. Exceedances over thresholds provide a alternative way to model extreme value by characterizing an observation as extreme if it exceeds a high threshold.

> **Theorem 2** (Coles 2001a, p. 75; Smith 2003): Consider the distribution of X conditionally on exceeding some high threshold u , and let $Y = X - u$, and $Y > 0$. We know
>
> $$F_u(y) = \Pr\{Y \le y \mid Y > 0\} = \frac{F(u+y) - F(u)}{1 - F(u)}$$

As $u \to \omega_F = \sup\{x : F(x) < 1\}$, we found a limit distribution,

$$F_u(y) \approx G(y; \sigma_u, \xi)$$

where G is generalized Pareto distribution (GPD)

$$G(y; \sigma, \xi) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi} \tag{1.9}$$

defined on $\{y: y > 0$ *and* $(1 + \xi y/\sigma) > 0\}$

The rigorous connection between exceedances over thresholds and the classic extreme value theory was established by Pickands (1975). Similar with GEV, GPD has three cases depending on the value of the parameter ξ:

- The case $\xi > 0$ is the long-tailed case, for which $1 - G(x)$ decays at the same rate as $x^{-1/\xi}$ for large x. This is reminiscent of the usual Pareto distribution, $G(x) = 1 - cx^{-1/\xi}$.
- For $\xi = 0$, we have the exponential distribution with mean σ as the limit

$$G(y; \sigma, 0) = 1 - \exp\left(-\frac{y}{\sigma}\right)$$

- For $\xi < 0$, the distribution has finite upper endpoint at $-\frac{\sigma}{\xi}$.

Replacing $Y = X - u$ into (1.9), now we have

$$\Pr\{X > x \mid X > u\} = \left(1 + \xi \frac{x - u}{\sigma}\right)^{-1/\xi} \tag{1.10}$$

It follows that

$$\Pr\{X > x\} = \varsigma_u \left(1 + \xi \frac{x - u}{\sigma}\right)^{-1/\xi} \tag{1.11}$$

where $\varsigma_u = \Pr\{X > u\}$. By inversing the equation (1.11), we obtain

$$x_p = \begin{cases} u + \frac{\sigma}{\xi}\left[\left(\frac{\varsigma_u}{p}\right)^{\xi} - 1\right], \ for\ \xi \neq 0 \\ u + \sigma \log \frac{\varsigma_u}{p}, \ for\ \xi = 0 \end{cases} \tag{1.12}$$

x_p is the **(1/*p*)-observation return level**. In other word, the level x_p is expected to be once every 1/*p* observations to a reasonable degree of accuracy , or the probability of an observation to exceed x_p is *p*. Suppose that we have one observation for each day. Then a 365-observation return level is the same as a 1-year (or 365-day) return level, which is the level expected to be exceeded once every 365 observations (or in a year).

Factor Analysis

In the above discussion, we do not consider that high-level attacks may systemically change through time, or be influenced by the changes of other environmental factors. In the context of DoS, the network traffic or server load may increase over time,

because the Internet is expanding and the e-business is maturing. The organization's internal traffic may be affected significantly by the number of employees and the number of enterprise applications. The activities of worms, viruses, or hackers may vary seasonally. In the following discussion, we introduce the models that capture these changes and influences.

Let $GEV(\mu, \sigma, \xi)$ denote the GEV distribution with parameters μ, σ, and ξ. Let Z_t denote the maximum attack level in time period t. To examine whether the maximum attack level changes linearly over the observation periods, a suitable model for Z_t is (Coles 2001a, p. 107)

$$Z_t \sim GEV(\mu(t), \sigma, \xi)$$

where

$$\mu(t) = \beta_0 + \beta_1 t$$

for coefficients β_0 and β_1.

To identify other factors that might have significant impact on the maximum attack levels, the model can be extended into a general form

$$\mu(t) = [1, z_1(t), ..., z_n(t)] \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_n \end{bmatrix}$$

where $z_i(t)$ are the factors to be examined (e.g., the number of employees and the number of enterprise applications in different time periods).

The seasonal model with k seasons $s_1, s_2, \ldots, s_n$ takes the form

$$\mu(t) = [I_1(t), I_2(t), ..., I_k(t)] \begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_k \end{bmatrix}$$

where $I_j(t)$ is the dummy variable having

$$I_j(t) = \begin{cases} 1, \textit{if } s(t) = s_j \\ 0, \textit{otherwise} \end{cases}, \quad j = 1, ..., k$$

Using these regression models, we are able to identify whether high-level attacks are changing over time, and/or whether there is any seasonal effect. We also can identify factors that may influence the maximum attack level. Following the same logic, we can also test the factors that might have an impact on the parameters α and ξ. The information helps us understand the trend of attacks and thus make strategic investment in IT security more effectively with the change of environment.

An Empirical Analysis

Malicious traffic from self-propagating worms and denial of service attacks constantly threatens everyday operation of an organization's Internet systems. Defending networks from these threats demands appropriate tools to conduct comprehensive

vulnerability assessments of networked systems (Sommers et al. 2004). The high-level traffic above a certain threshold is perceived as a signal of attack to IT systems. High-level traffic causes network outage and denial of service. In this section, we analyze daily internal traffic collected from a large regional bank situated in New York state. With 1 year as the return period, we estimate the return level of traffic. The return level of traffic provides valuable information, enabling us to design proper defense strategies and adjust the investment level to prevent network outage and denial of service.

We record the internal traffic from January 16, 2004, to March 20, 2005, daily (see Figure 2). The traffic is comprised by a number of activities, including

- employee login/logout, file/printer access, or any other activity done at network level
- inter-server communication (most of which happens automatically or is scheduled)
- application access information (only some applications are monitored)

Since a series of daily data is available, "exceedances over thresholds" is employed in our extreme value analysis. We use the maximum-likelihood method to estimate the distribution parameters of generalized Pareto distribution (GPD) with the S-PLUS functions obtained from the website (Coles 2001b). To use exceedances over thresholds, a proper threshold must be selected. The mean residual life plot (Coles 2001a, p. 78) is a diagnostic plot drawn before fitting any model and gives guidance about what threshold to use. Figure 3 shows the mean residual life plot with approximate 95 percent confidence intervals for the daily traffic. The plot is initially linear, but shows substantial curvature in the range of $1.1 \times 10^6 \leq u \leq 1.3 \times 10^6$. For $u > 1.3 \times 10^6$, the plot is reasonably linear when judged relative to confidence intervals, suggesting we set $u = 1.3 \times 10^6$. (We also did a sensitivity analysis with $u = 1.5 \times 10^6$, the results do not have much difference.) The choice leads to 149 exceedences in the series of length 430. Thus $\hat{\varsigma}_u$ =149/430=0.347 with $\text{var}(\hat{\varsigma}_u) = 5.27 \times 10^{-4}$. The maximum likelihood estimators of GPD parameters are $(\hat{\sigma}, \hat{\xi})$ = (281972.2, 0.09), with standard error 33901.08 and 0.09 respectively. The 95 percent confidence interval for ξ is [-0.09, 0.26] (Figure 4) . Therefore, the maximum likelihood estimate corresponds to an unbounded distribution, although the evidence is not overwhelming and 0 lies inside the 95 percent confidence interval.

Diagnostic plots for the fitted GPD are shown in Figure 5. Both the set of the probability plot and of the quantile plot are near linear, showing the validity of the fitted model. The return level curve asymptotes to an infinite level. The corresponding density estimates are roughly consistent with the histogram of the data, but not perfect.

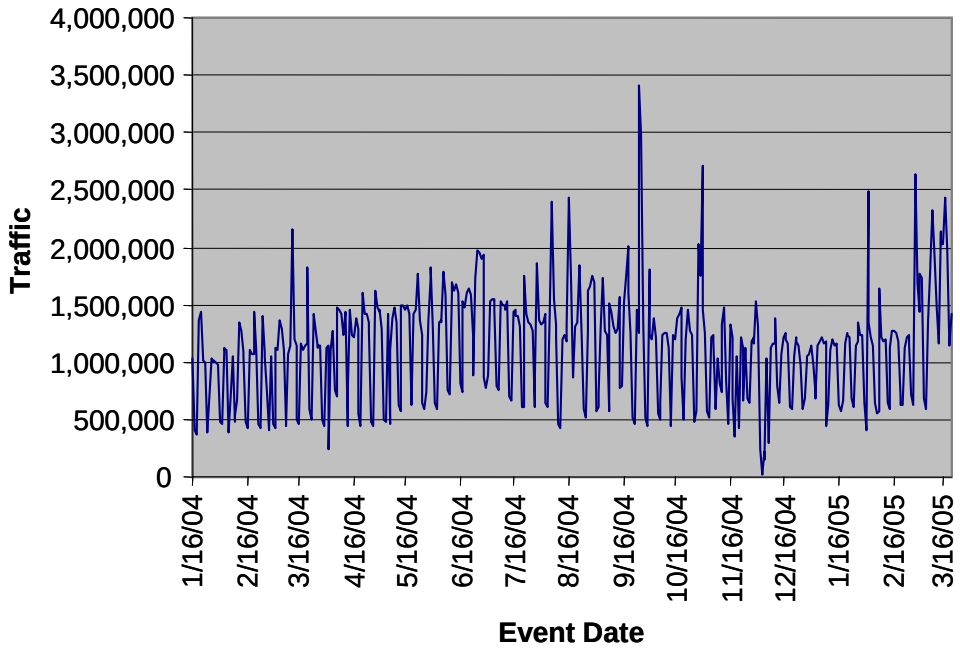

Figure 2. Daily Traffic from January 16, 2004, to March 20, 2005

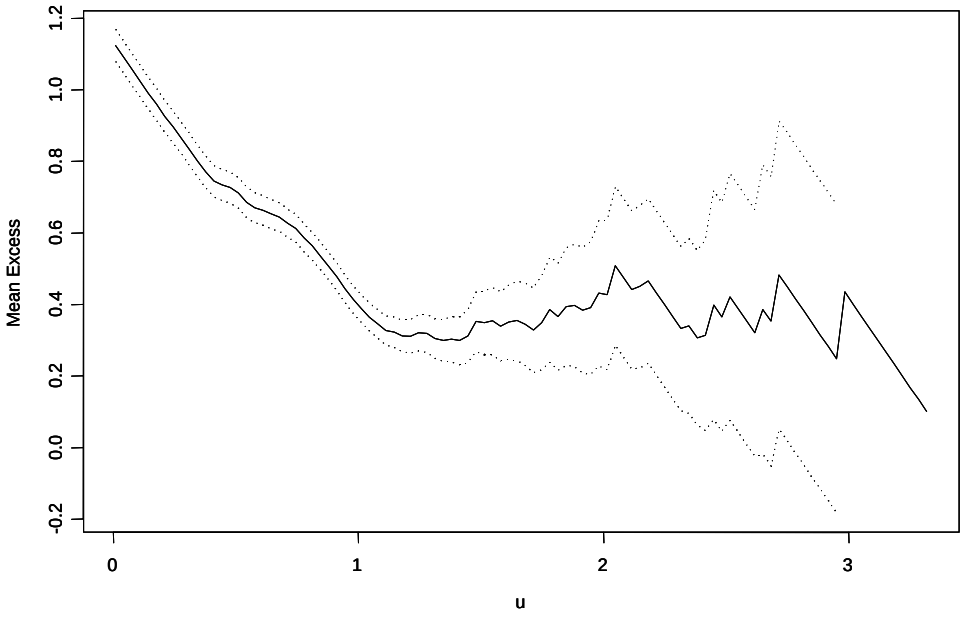

Figure 3. Mean Residual Life Plot for Daily Traffic

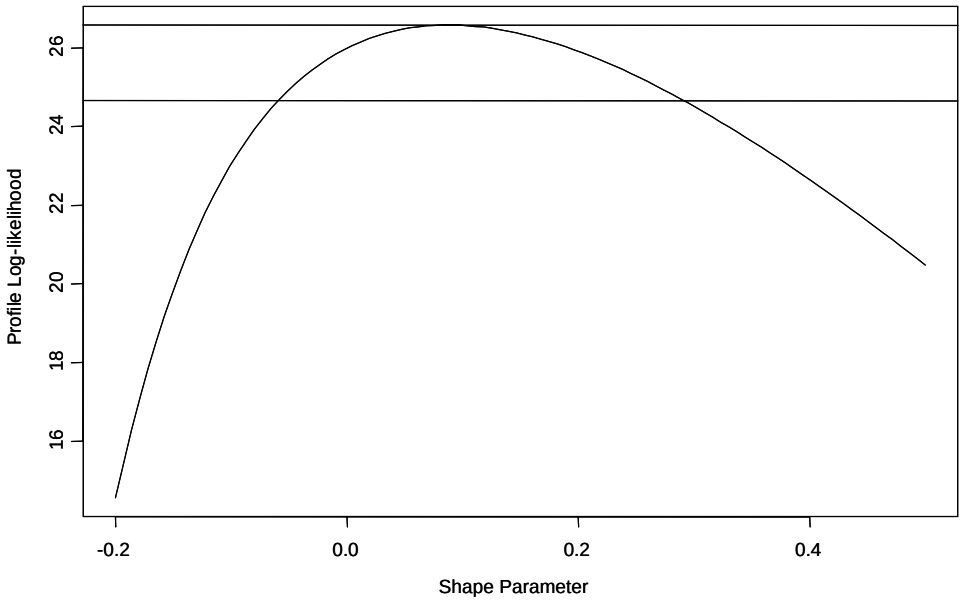

Figure 4. Profile Likelihood for Shape Parameter ξ in Threshold Excess Model of Daily Traffic

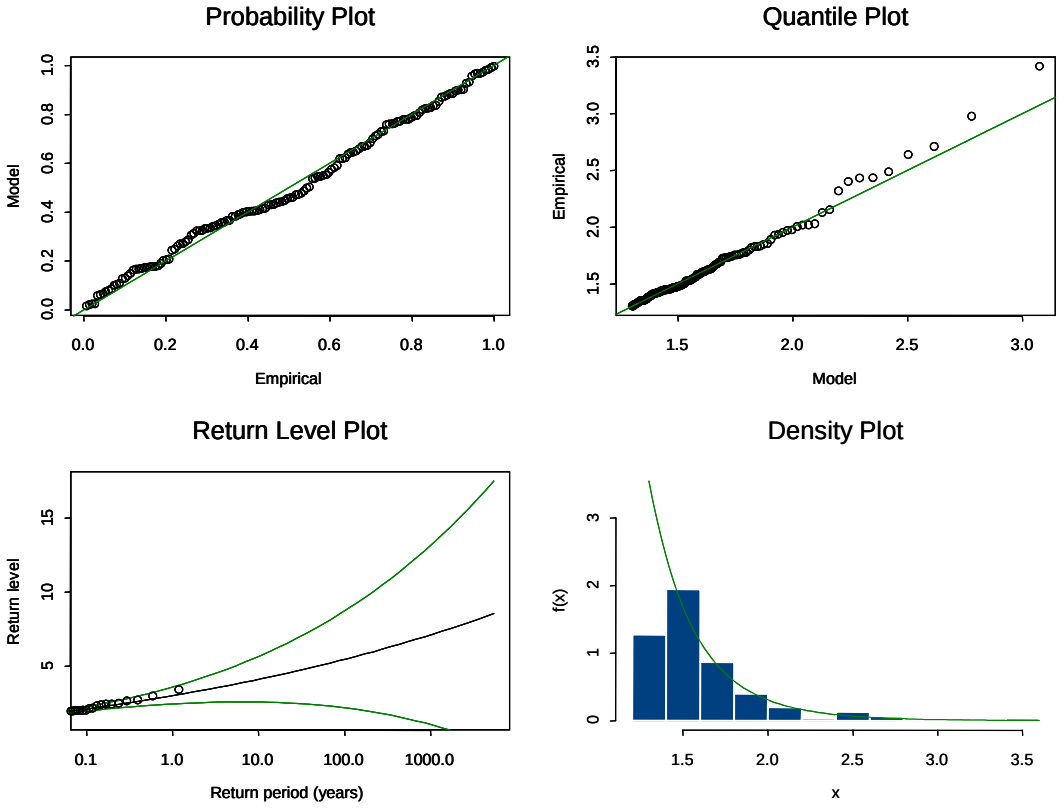

Figure 5. Model Diagnostic Plots for Threshold Excess Model Fitted to Daily Traffic Pattern

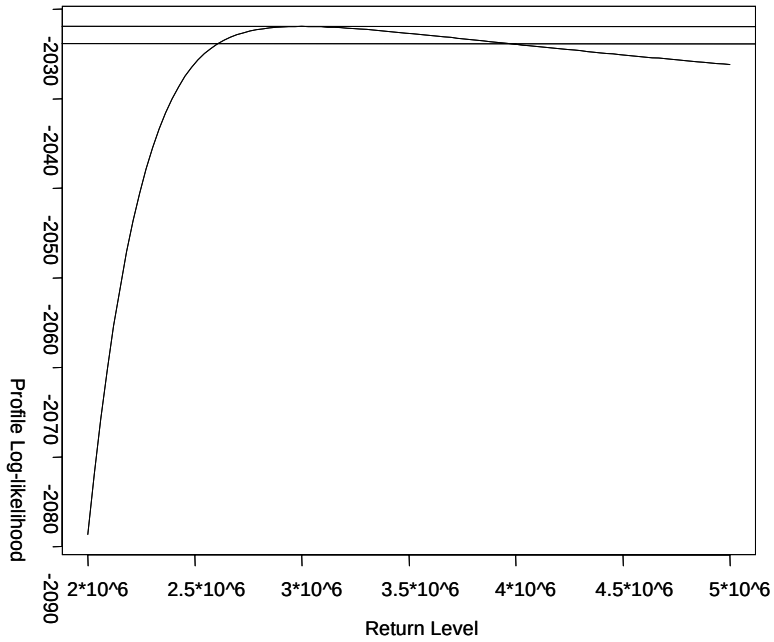

Figure 6. Profile Likelihood for 1-Year Return in Level in the Threshold Excess Model

The GPD model provides a direct method for risk estimation using return level. In our analysis, we use 1 year as our return period. Since we have one observation for each day, the 1-year return level corresponds to the 365-observation return level with $p = 1/365$. As $\hat{\xi} > 0$, we use the first part of equation (1.12) to calculate the return level and have $\hat{x}_p = 3.01 \times 10^6$. Figure 6 plots the profile log-likelihood for the 1-year return level. The 95 percent confidence interval is $[2.6 \times 10^6, 4.0 \times 10^6]$. The return level of traffic shows the level of investment we should match for the next year such that the exceeded probability of traffic at $\hat{x}_p = 3.01 \times 10^6$ in next year is less than 1/365 with 95% confidence.

In our model, we have (*i* - *a*) as a measure of vulnerability, where *i* is the investment factor and *a* is the attack factor. In the data analysis, *a* is the packet traffic rate, while *i* is the capacity rate in packets per second that the system can handle. Estimated return level of traffic $\hat{x}_p$ in the data analysis provides an extremal value estimate of packet traffic rate to which the system is subject, which in turn gives the value of attack factor *a*.

Taking this value as the system capacity *i*, and assuming that the security system is a simple serial system of three elements—connection to the Internet, firewall/router, and the server—we can then specify the bandwidth of the pipe that is connecting to the firewall, the firewall/router's filtration capacity, and the capacity of the server operating system to handle that TCP/IP traffic, all in terms of packets per unit of time. Given these capacities, we can estimate what the cost of such a security system is likely to be, and decide whether to increase or decrease investment by comparing with current configurations. We may also introduce intrusion detection systems (IDS) or reconfigure the existing IDS to properly protect the system.

Conclusion

In this paper, we introduce an extreme value approach for security investment. Compared with other methods on determining the effective level of security investment, our model does not need to calculate the expected loss due to system failure, nor make assumptions relating to hackers' behavior. It is a dynamic strategy for security investment. With extreme value analysis, the distribution of high-level attacks is estimated. We may then determine the return level of attacks for a certain return period. The return level of attacks provides important information for us to design a proper defense capability and make an investment decision. Using the daily traffic data collected from a large regional bank, we examine the distribution of high-level traffic. Using one year as the return period, we estimate the return level of traffic. The methodology provides many avenues of research in the future. First of all, using the extreme value approach, we can examine whether there is any difference in the distribution of high-level attacks from different types of attacks, such as denial of service, malicious code, etc, as well as from different initiators, such as internal employees, hackers, or competitors. Second, the time-effect on the attack level can be examined empirically. With the extreme value approach, we can answer whether the maximum attack level is changing over time, and whether there is any seasonal effect. Further we can identify factors that influence the maximum attack level. This information will help to make strategic investment in IT security more effective.

A similar analysis can also be done for spam e-mail where we can estimate the capacity of the e-mail system to handle the surge in traffic due to spams. There too, we will get a system size in terms of packet handling capacity, which in turn will suggest some dollars as investment. In future research, we propose to show how our methodology can help size a security system in terms of packet handling capacity systematically and thereby help estimate the dollar investment that may be required.

There are certain limitations of our paper. First, we only focus on the discussion using extreme value theory to characterize the behavior of attacks. We do not look at how to operationally decide a corresponding security investment level, nor do we convert it into a real protection level for a system through the combination of various technology and security policy. This is an interesting topic that needs further exploration. Second, in extreme value analysis we view the system attack as an exogenous variable. The causal issues of the attack are not explored. Third, extreme value analysis, being a statistical approach based on past data has limited application where the security scenario is evolving such that past data are no longer a reliable indicator of what future situations may entail.

Acknowledgements

The authors would like to thank the track chair, the associate editor, and the two referees for their comments, which have greatly improved the paper. This research has been funded in part by NSF under grant #0402388. The usual disclaimer applies.

References

Camp, L. J., and Wolfram, C. "Pricing Security," in *Economics of Information Security*, L. J. Camp and S. Lewis (Ed.), Kluwer Academic Publishers, Boston, 2004, pp. 17-34.

Castillo, E. *Extreme Value Theory in Engineering*, Academic Press, San Diego, 1988.

Cavusoglu, H. "Economics of IT Security Management," in *Economics of Information Security*, L. J. Camp and S. Lewis (Ed.), Kluwer Academic Publishers, Boston, 2004, pp. 71-83.

Cavusoglu, H., Mishra, B., and Raghunathan, S. "A Model For Evaluating IT Security Investments," *Communications of the ACM* (47:7), 2004, pp. 87-92.

Coles, S. *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag, London, 2001a.

Coles, S. "How to Use the S-PLUS Functions and Datasets," June 2001b (available online at **http://www.maths.bris.ac.uk/~masgc/ismev/summary.html**).

Dahan, E., and Mendelson, H. "An Extreme-Value Model of Concept Testing," *Management Science* (47:1), 2001, pp. 102-116.

Embrechts, P., Kluppelberg, C., and Mikosch, T. *Modeling Extremal Events for Insurance and Finance*, Springer, New York, 1997.

Ernst & Young. "Global Information Security Survey 2003," Ernst & Young LLP, 2003 (available online at **http://www.deloitte.com/dtt/cda/doc/content/Global%20Security%20Survey%202003.pdf**).

Ernst & Young. "Global Information Security Survey 2004," Ernst & Young LLP, 2004 (available online at **http://www.deloitte.com/dtt/cda/doc/content/dtt_financialservices_SecuritySurvey2004_051704.pdf**).

Farahmand, F., Navathe, S. B., Sharp, G. P., and Enslow, P. H. "A Management Perspective on Risk of Security Threats to Information Systems," *Information Technology and Management* (6:2-3), 2005, pp. 203-255.

Fisher, R. A., and Tippett, L. H. C. "Limiting Forms of The Frequency Distributions of The Largest or Smallest Member of a Sample," in *Proceedings of the Cambridge Philosophical Society* (24), Cambridge University Press, London, 1928, pp. 189-190.

Geer, D., Hoo, K. S., and Jaquith, A. "Information Security: Why the Future Belongs to the Quants," *IEEE Security & Privacy* (1:4), July/August 2003, pp. 32-40.

Gordon, L. A., and Loeb, M. P. "The Economics of Information Security Investment," *ACM Transactions on Information and Systems Security* (5:4), 2002, pp. 438-457.

Gordon, L. A., Loeb, M. P., Lucyshyn, W., and Richardson, R. "2005 CSI/FBI Computer Crime and Security Survey," Computer Security Institute, 2005 (available online at **http://www.cpppe.umd.edu/Bookstore/Documents/2005CSISurvey.pdf**).

Gumbel, E. J. *Statistics of Extremes*, Columbia University, New York, 1958.

Hoo, K. J. S. "How Much is Enough? A Risk-Management Approach to Computer Security," CISAC Working Paper, Stanford University, August 2000 (available online at **http://cisac.stanford.edu/publications/11900/**).

Longstaff, T. A., Chittister, C., Pethia, R., and Haimes, Y. Y. "Are We Forgetting the Risks of Information Technology?" *IEEE Computer* (33:12), 2000, pp. 43-51.

Mercuri, R. T. "Analyzing Security Costs," *Communications of the ACM* (46:6), 2003, pp. 15-18.

Pickands, J. "Statistical Inference Using Extreme Order Statistics," *Annals of Statistics* (3), 1975, pp. 119-131.

Schechter, S. E. *Computer Security Strength and Risk: A Quantitative Approach*, unpublished Ph.D. dissertation, Harvard University, 2004.

Schechter, S. E., and Smith, M. D. "How Much Security is Enough to Stop a Thief? The Economics of Outsider Theft via Computer Systems Networks," in *Proceedings of the 7th Financial Cryptography Conferences*, R. N. Wright (Ed.), Guadeloupe, French West Indies, January 27-30, 2003.

Smith, R. L. "Statistics of Extremes, with Applications in Environment, Insurance, and Finance," unpublished manuscript, Department of Statistics, University of North Carolina, 2003.

Sommers, J., Yegneswaran, V., and Barford, P. "A Framework for Malicious Workload Generation," in *Proceedings of the 4th ACM SIG COMM Conference on Internet Measurement*, J. Kurose (Ed.), Taormina, Sicily, Italy, 2004, pp. 82-87.