MATHEMATICAL MODELING OF CYBER ATTACKS: A LEARNING MODULE TO ENHANCE UNDERGRADUATE SECURITY CURRICULA*

A. Herath, S. Herath, R. Goonatilake[**] S. Herath and J. Herath[***] Richard Stockton State College of New Jersey, NJ 08330 [**]Texas A&M International University, TX 78041[***]St. Cloud State University, MN 56301

ABSTRACT

We have been experimenting with methods for improving undergraduate curricula and research experiences as well as the security courses for Computer Science and Information Systems students. Recent surges in attacks and intrusions reflect vulnerabilities in computer networks and emerge of sophisticated new attacks are always associated with determining evolving risks and preventing them. Innovative methods and tools can help attack defenses, prevent attack propagations, detect and respond to such attacks. Mathematical or statistical modeling of attacks is an important concept to introduce to students in security courses as they help assess vulnerability under a variety of conditions that predetermine the extent of damages to a system and, thereby, predict possible losses. This paper describes a learning module developed to help students understand basic statistical and mathematical modeling of unscrupulous attacks.

Key Words: Worms, Epidemics, Networks, Attacks, Modeling, Risk, Poisson.

1. INTRODUCTION

The rapid growth of attacks and security measures degrades not only information system infrastructure and assurance, but also the performance of computers, networks and wireless devices. As soon as a computer starts to share the resources available on the web or local network, it immediately becomes vulnerable to attacks or infiltration. Every year more complicated variants of worms are released via the web [1 & 2]. In recent years,

---

* Copyright © 2007 by the Consortium for Computing Sciences in Colleges. Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the CCSC copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Consortium for Computing Sciences in Colleges. To copy otherwise, or to republish, requires a fee and/or specific permission.

focus has shifted to computer vulnerabilities, hardening infrastructure as well as information assurance and security. Consequently, intrusion, detection, prevention and response issues have been addressed at great length by many researchers, apart from the discussions of software and hardware capabilities. Intrusion and detection are to be considered separately as some intrusions cannot be detected. Various techniques such as encryption, hashing, Kerberos, IPSec, SSL, SET, firewalls and honey-pots are popular among network users, and analyzing the potential vulnerability of computer networks is essential to safeguard public and private interests from threats, attacks and damages.

This paper illustrates some models adapted for classroom application that make mathematical and probabilistic analysis practical for network security. These sample models foster an intuitive feel for the subject in students which enables them to think probabilistically and statistically about attack epidemics. Any realistic model of real-world phenomena must take into account the possibility of randomness. In the sections to follow, security breaches from external sources will be addressed mathematically. Section 2 presents the importance of network modeling. Section 3 discusses the mathematical background used to justify Poisson, binomial and multinomial distributions. Section 4 describes the management of cyber attacks and section 5 provides a brief summary, conclusions and suggestions for future work.

2. NETWORK MODELING - WORM EPIDEMICS

A simple example relates to a recent Federal criminal case in which documents provided by the prosecution counted the number of e-mail messages generated by a worm transmission. The payload of this worm was a 4kb packet. The network of the institution where the defendant worked could process packets at the rate of 1 Gb per second. To saturate the network, the worm would have to produce a minimum of 250,000 emails per second. According to the prosecution's data, only 261 e-mails were released into the network during the 3 hours the worm was active. This number was not sufficient to degrade the performance of the network and the defendant was cleared. The spreading rate of this worm was hindered due to the strange file name given to the attachment, which was not attractive enough for many users to open. This knowledge can be expanded by examining epidemiological models.

Epidemics can be represented using a linear equation with a constant propagation rate, a non-linear equation with exponential growth rate, a differential equation or a difference equation. To obtain a useful prediction model, one should record the observations of all variables that may significantly affect the response to the epidemic. By virtue of its wide applicability, the linear model plays a prominent role in this process. Mathematical models express the laws that govern described actions and assumptions together with hypotheses relevant to the equations established within a study. For example, the uninfected computer components in a network under virus attack can be represented using [a linear equation or] a differential equation and the initial condition:

$$\left\{ \begin{array}{l} \dfrac{dy(t)}{dt} = -ry(t), t > 0, \\ y(0) = y_0 \end{array} \right\}$$

The general solution of this differential equation is $y(t) = y_0 e^{-rt}$, where the initial uninfected computer population is y_0 with a constant negative growth rate *-r*. The geometric representation of this general solution is an infinite family of integral curves, one similar to Figure 2-1.

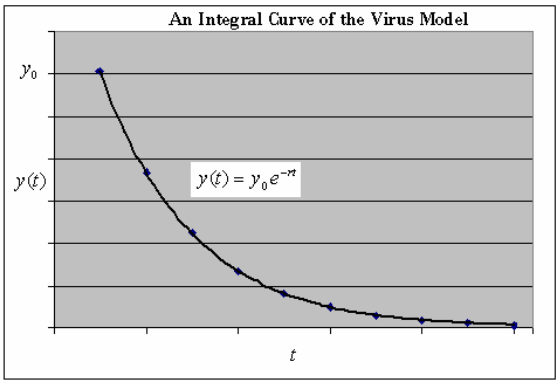

Figure 2-1: Decay of Un-Infected Computers Due to a Virus Epidemic

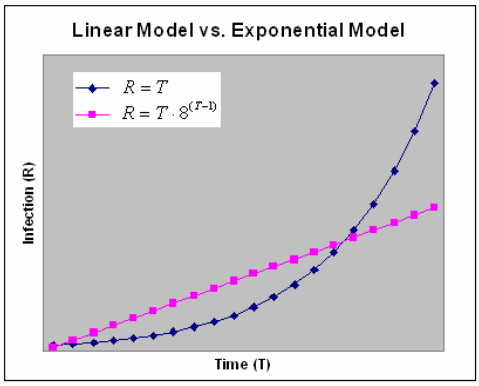

Figure 2-2: Growth of Infected Computers Due to Virus Propagations

A computer worm that randomly scans new hosts to attack and infect may follow the simple epidemiological model [3 & 4]. There are other applications that have been used in the past, stemming from pre-existing epidemiological, biological, and physical models. Figure 2-2 shows the number of infected computers over time due to two different types of attacks. At the initial detection stages, for smaller values of time *T*, the linear virus growth model, $R = T$, shows more infections than the exponential growth $R = T \cdot 8^{(T-1)}$. This false appearance leads us to make the wrong decision to allocate resources and assign a higher priority to suppressing the linear attack over the eventually more severe exponential attacks.

In many instances, there is a limit to the possible growth of epidemics. The logistic function models the restricted growth phenomenon as an exponential growth. As shown in Figure 2-3 below, the infection growth rate slows due to the limited capacity of the network, resulting in a logistic curve. The red curve to the left with the faster infection rate saturates the network sooner than the blue curve to the right. Figure 2-4 depicts the infection during an epidemic attack and the recovery due to the response. The growth of the number of infected computers begins in an exponential manner, levels out and eventually decays with the proper injection of response.

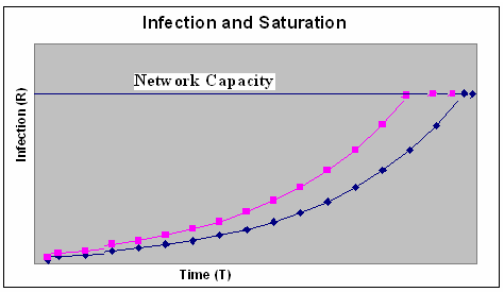

Figure 2-3: Infection and Saturation

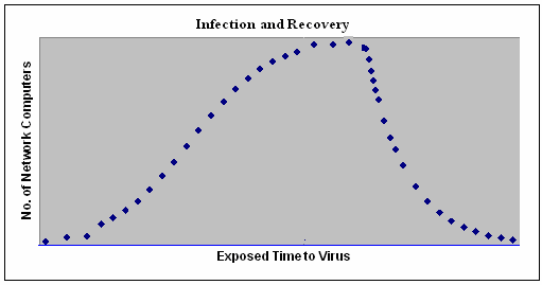

Figure 2-4: Infection and Recovery

Non-linear differential equations are used to model attack dynamics. The Epidemiological model derived from non-linear differential equations [4] can be represented by the equation:

$$n(t) = \frac{n_0(1-(d/\beta))}{n_0 + (1-(d/\beta) - n_0)e^{-(\beta-d)t}},$$

where $n(t)$ is the fraction of infected nodes, d is the infection or "death" rate, β is the recovery or "re-birth" rate, n_0 is the total number of vulnerable machines.

The Analytical Active Worm Propagation, AAWP, [4], gives a better estimate than Epidemiological modeling and can be mathematically represented as a difference equation:

$$n_{t+1} = sn_t + \left[N - n_t\right]\left[1-(1-1/T)^{sn_t}\right],$$

where n_t is the fraction of infected nodes at time t, s is the scan rate, N is the total number of vulnerable machines and T is the address space the worm scans. The infection rate of the worm is proportional to the scanning rate. Random scanning has the searching ability to identify IP addresses to attack in a random order, localized scanning has the ability to attack targets that reside on the same subnet, hit list scanning has a list of targets to attack, permutation scanning uses a fixed pseudorandom permutation of IP addresses to attack and topological scanning uses the information stored in the victim's machine to attack the next target [4 & 5]. Worms that use localized scanning, like the Code Red II worm, spread at a slower rate than a worm that employs random scanning but has the capacity to penetrate firewalls.

3. STATISTICAL AND PROBABILISTIC BACKGROUND

Other useful concepts for calculating probabilities and expectations in the classroom, based on some appropriate random variable with available partial information, include conditional probability and conditional expectation. The expected value of a terminal location in the network that is vulnerable to an attack can also be computed using elementary probabilistic analysis. Such work will eventually pave the way for analyzing more complicated network systems. Applications pertaining to the theory of random graphs and their calculations presented in [6] will facilitate in assessing computer vulnerability.

Each unauthorized user that approaches a computer network will be able to attack a computer with probability p. If the number of users approaching the computer network is a Poisson distribution with mean λ, the probability that the unauthorized user is unable to compromise the computer network can be calculated as follows: Let X be the number of computers compromised, and N denote the number of unauthorized users that approach a given network. By conditioning on N and using the calculations in [6], we see that:

$$\Pr\{X=0\} = \sum_{n=0}^{\infty} \Pr\{X = 0 | N = n\} \cdot \Pr\{N = n\}$$

$$= \sum_{n=0}^{\infty} \Pr\{X = 0 | N = n\} \cdot \frac{e^{-\lambda} \lambda^n}{n!}.$$

Therefore, if a given number of n hackers approach the network, the probability that the network is not compromised is just $(1 - p)^n$. That is, $\Pr\{X=0 \mid N=n\} = (1 - p)^n$. Therefore:

$$\Pr\{X=0\} = \sum_{n=0}^{\infty} (1 - p)^n \cdot e^{-\lambda} \lambda^n / n!$$

$$= e^{-\lambda p}.$$

The last expression provides the probability that the network is not compromised by any unauthorized users as a function of λ and p. This case is obvious from the result $\Pr\{X=k\}$ for $k = 0$, which evaluates the probability for a vulnerability-free, absolutely perfect system and provides an intuition for subsequent cases.

Next, consider that only k number of computers is compromised. Then, the probability that only k number of computers have been compromised by attackers can be calculated as follows: For a given $N = n$, X has a binomial distribution with parameters n and p. Hence,

$$\Pr\{X = k \mid N = n\} = \left\{ \begin{array}{l} \binom{n}{k} p^k (1 - p)^{n-k}, n \geq k \\ 0, n < k \end{array} \right\}$$

so that

$$\Pr\{X = k\} = \sum_{n=k}^{\infty} \binom{n}{k} \frac{p^k (1 - p)^{n-k} e^{-\lambda} \lambda^n}{n!} = e^{-\lambda p} \frac{(\lambda p)^k}{k!}.$$

This implies that X has a Poisson distribution with mean λp. A Poisson distribution is often used to model the frequency with which a specified event occurs over a particular period of time and the Poisson variable has infinitely many possible values. Determining λ and p is necessary before putting this model to work. Maximum likelihood estimation (MLE) [7] can be used to determine the parameters λ and p that maximize the probability (likelihood) of the sample data.

In multinomial distribution [8 & 9] each user introduces a probability function $p_i = \Pr\{X=i\}$ on the set of operations, O, which are mutually exclusive. If we consider that each intruder attacking a network can be classified as successful, unsuccessful, or of unknown success then let us assume that the number of successful attacks in time interval $(0, t)$ is a Poisson random variable X_1 with parameter $\mu_1 t$. Similarly, assume the numbers of unsuccessful attacks or unknown attacks are also independent Poisson random variables X_2, X_3 with parameters $\mu_2 t$, $\mu_3 t$, respectively. The joint probability function [9] for X_1, X_2, X_3 is the product of their marginals represented by:

$$\frac{\mu_1^{x_1}\mu_2^{x_2}\mu_3^{x_3}}{x_1!x_2!x_3!}e^{-(\mu_1+\mu_2+\mu_3)t}.$$

The probability distribution for the total number of attacks is Poisson with parameter $(\mu_1 + \mu_2 + \mu_3)t$ and function $\dfrac{((\mu_1 + \mu_2 + \mu_3)t)^y}{y!}e^{-(\mu_1+\mu_2+\mu_3)t}$.

Given that $Y = n$ total number of attacks taken place during this time period, the conditional probability function for X_1, X_2, X_3 is:

$$\frac{y!}{x_1!x_2!x_3!}\left(\frac{\mu_1}{\mu_1 + \mu_2 + \mu_3}\right)^{x_1}\left(\frac{\mu_2}{\mu_1 + \mu_2 + \mu_3}\right)^{x_2}\left(\frac{\mu_3}{\mu_1 + \mu_2 + \mu_3}\right)^{x_3},$$

where $x_1 + x_2 + x_3 = y$. This conditional probability function, a multinomial distribution with parameters y, $\mu_i/(\mu_1 + \mu_2 + \mu_3)$, $i = 1, 2, 3$, is the ratio of the two probability functions described above and is called a Poisson sampling.

4. CYBER ATTACK MANAGEMENT

This section provides two examples, one for statistical consideration and the other to provide a probability consideration for determining the extent of cyber attack management needed, by using the single server Poisson attacks and detection rate. To evaluate the system's accuracy, two sets of measurements were obtained from an example in [10] for a detection rate and a false positive rate for varying thresholds. The detection rate is the percentage of attack records that have been correctly identified. The false positive rate is the percentage of normal records that have been mislabeled as anomalous. The threshold determines whether a record is normal or an attack. Regression analysis is used to find possible relationships between the Detection Rate vs. False Positive Rate, the

Varying Threshold vs. False Positive Rate and the Varying Threshold vs. Detection Rate. Regression analysis is used to obtain these models as the correlation coefficient provides a measure of fit associated with determination of the extent of the relationship.

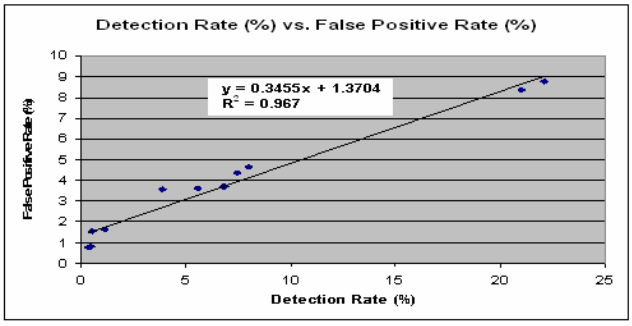

Figure 4-1: Detection Rate (%) vs. False Positive Rate (%)

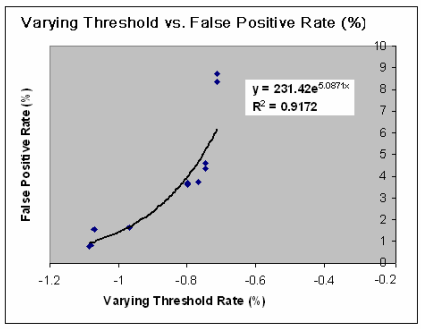

Figure 4-2: Threshold vs. False Positive Rate

Accordingly:

False Positive Rate = 0.3455 × (Detection Rate) + 1.3704

with a strong positive correlation. This model will allow us to determine the False Positive Rate from the corresponding Detection Rate.

Figure 4-2 illustrates the relationship between the Threshold Value and the False Positive Rate (%). This statistical model has an exponential growth behavior with strong correlation coefficients between the two variables considered. This will allow us to determine the False Positive Rate from the Varying Threshold and vice versa.

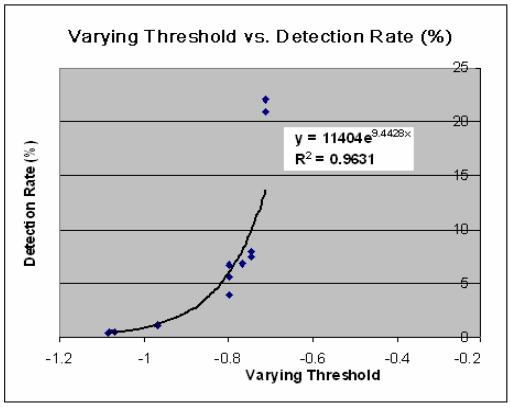

Figure 4-3: Threshold vs. Detection Rate

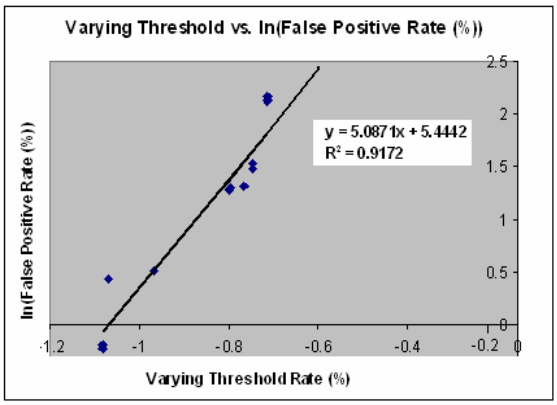

Figure 4-4: Threshold vs. ln(False Positives)

Figure 4-3 illustrates the relationship between the Threshold Value and the Detection Rate (%). It is important to note that this model has an exponential growth behavior with strong correlation coefficients among the two variables considered. This will allow us to determine the Detection Rate from the Varying Threshold and vice versa.

Some nonlinear exponential models can be transformed into linear models by applying the logarithm of variables. For example, Figure 4.4 illustrate the use of False Positive Rate) instead of False Positive Rate to obtain a linear relationship from Figure 4.2.

Let us consider a single-server network [6], in which attacks are made according to a Poisson process having rate λ. An attack is immediately made whether the server is vulnerable or not. If the server is not free, then the attacker waits for the server to be free. The successive attack times will be assumed to be independent and identically distributed random variables with mean $1/\mu$, where $\mu > \lambda$. Some machines are more vulnerable than others because of measures that have been taken during software development to secure the machine. The above process will alternate between busy periods when the server is working, and idle periods when the attacker is determined to attack. Hence, a cycle will consist of a busy period followed by an idle period. Thus, letting P_B denote the proportion of time that the server is busy, we can infer that:

$$P_B = \frac{E[\text{Length of a busy period}]}{E[\text{Length of a busy period}] + E[\text{Length of an idle period}]}.$$

It is assumed that if one attack takes place, no other attacks will be planned until the attacker is able to succeed subsequently. Because of the lack of memory in the Poisson process [6] it follows that an idle period will be exponentially distributed with a mean of $1/\lambda$ such that $E[\text{Length of an idle period}] = 1/\lambda$. We determine $E[L_B]$, the mean length of a busy period, by conditioning on both N, the number of attackers entering the network during the service time of the attacker initiating this vulnerability and S, the length of the attack: $E[L_B] = E[E[L_B \mid N, S]]$. This leads to $E[L_B] = \frac{1/\mu}{1 - E[N]}$ as in [6]. Now, we can determine $E[N]$, the expected number of attacks during the initial period as $E[N] = E[E[N \mid S]]$. However, since the attack process is a Poisson process with rate λ it follows that the expected number of attacks during an interval of length S is just λS and hence: $P_B = \frac{1/(\mu - \lambda)}{1/\lambda + 1/(\mu - \lambda)} = \frac{\lambda}{\mu}$ From this it follows that P_I, the proportion of time that the network server is prone to be attacked, is $P_I = 1 - \lambda / \mu$.

5. CONCLUSIONS AND FUTURE WORK

Modeling computer attacks and epidemics is a way to make people aware of the consequences and to prepare the public for these ultimate consequences. It is vital that security breaches be taken seriously at all levels of the decision making process. With secure networks, organizations can build e-systems with assurance, enhancing current market relationships by driving aggressively into the future.

This paper presented several mathematical and probabilistic models for worm epidemics and possible security breaches under different circumstances. These models can be employed in security curricula as an initial basis for discussions and used together with emergent technologies.

The determination of parameters for these models can be challenging. For example, a simulation is used for determining λ and p in the Poisson distribution in section 2. It is also possible to apply Monte Carlo simulations to these problems but their major disadvantage is running time. Except in extreme situations, this approach is computationally efficient and robust with respect to possible missing data and anomalies. Using the basic idea formulated from the Poisson distribution, chain, denial of service and distributed denial service attacks have been remodeled [11]. Expanding on probabilistic cryptography together with the probabilistic and statistical models presented in this article and [6] will help us lay a solid foundation for related other curriculum development that leads to applicable student research.

REFERENCES

[1] Niels Provos, Joe McClain, Ke Wang, *"Search worms"*, Proceedings of the 4th ACM workshop on Recurring Malcode WORM '06, November 2006

[2] Moore, David, Paxson, Vern, Savage, Stefan, Shannon, Colleen, Staniford, Stuart, Weaver, Nicholas, *"Inside the Slammer Worm"*, IEEE Security and Privacy, July 2003.

[3] Chen, Thomas M., Robert, Jean-Marc, *"Worm Epidemics in High Speed Networks"*, http://engr.smu.edu/~tchen/papers/Computer_Jun2004.pdf.

[4] Chen, Zesheng, Gao, Lixin, Kwiat, Kevin, *"Modeling the Spread of Active Worms"*, IEEE Infocom, http://www.ieee-infocom.org/2003/papers/46_03.PDF, 2003.

[5] Ellis, Daniel R., Aiken, John G., Attwood, Kira S., Tenaglia, Scott. D., *"A Behavioral Approach to Worm Detection"*, Proceedings of ACM Workshop on Rapid Malcode, 2004.

[6] Ross, Sheldon M., *"Introduction to Probability Models"*, Second Edition, Academic Press, Inc., New York, New York, 1981.

[7] Siekierski, K., *"Comparison and Evaluation of Three Methods of Estimation of the Johnson Sb Distribution"*, Biometrical Journal, Vol. 34, pp. 879–895, 1992.

[8] Elbaum, Sebastian, Munson, John C., *"Intrusion Detection: Through Dynamic Software Measurement"*, Proceedings of the Workshop on Intrusion Detection and Network Monitoring, The USENIX Association, April 9-12, 1999.

[9] Larson, Harold J., *"Introduction to Probability"*, Addison-Wesley Advanced Series in Statistics, Addison-Wesley Publishing Company, Reading, Massachusetts, 1995.

[10] Heller, Katherine A., Svore, Krysta M., Keromytis, Angelos D., Stolfo, Salvatore J., *"One Class Support Vector Machines for Detecting Anomalous Windows Registry"*, Columbia University, http://www1.cs.columbia.edu/~kmsvore/ocsvm.pdf, 2003.

[11] Goonatilake, R, Herath, A, *"Assessing Computer Network Vulnerabilities for Insurance Safeguards"*, International Journal of Effective Management (IJEM), December 2006.