Hacking-Vigilance Distribution with Application to Assess Cyber Insecurity Level

Ramalingam Shanmugam

***Abstract*—Hacking into other's computers for a variety of reasons is a serious concern and nightmare to government agencies, private and public institutions in almost all parts of the world in this fast communication age of 21st century. The owners of confidential and sensitive files are consciously and constantly undertaking vigilances to protect their files from *hackers*. They are refereed here *vigilantes.* The intrusions are increased by the *hacker's efforts,* ρ when the vigilances are weaker. When the intrusion rate, θ is more, the *vigilante's efforts,* τ are bumped up. Both the hacker's and vigilante's efforts are latent, non-observable and hence are treated in this article as *parameters*. Only the number of intrusions is observable and hence is treated as a *random variable, Y.* This article introduces a new probability model connecting $Y\ \rho\ \tau$ and names it as *Hacking-Vigilance Distribution (HVD).* After deriving the properties of HVD, this article demonstrates their use to analyze and interpret computer intrusions data.**

***Index Terms*—Computer intrusion, mean-variance relation, poisson, cyber intrusion, count and chi-squared distributions.**

I. Introduction

Since September 11th, 2001, not only the United States but also many other nations became alert to tackle the invisible dangers due to cyber terrorism. A major weapon of mass disruption is the cyber-attack. The motives might include but are not limited to premeditated revenge not necessarily out of rage, evil intensions to damage the peaceful civic life in a nation, rivalries to sabotage political or defense structures. The attacks involve sending out "malicious software virus" to the computers. Some viruses take permanent residence in the receiving computer while others are periodically entering. There had been so many incidences of cyber-attacks around the world. The Honker Union of China and the Chinese Red Guest Network Security Technology Alliance orchestrated cyber-attacks on approximately 1,200 US web sites. Consequent to the North Atlantic Organization (NATO)'s bombing of Kosovo; more than 100 NATO's web sites were infected by the hackers. The Israel's web pages of Knesset (parliament), Defense and Foreign Ministries were cyber-attacked in the year 2000 after Israel attacked Palestinian houses. Over the Kashmir dispute, more than 400 Indian web pages received cyber-attacks from the Pakistan based hackers club. An American hacker group called PoizonBox claimed it had defaced more than 100 Chinese web sites. See [1] for details about the cyber threats. See [2]-[4] for important security issues in computers.

The cyber insecurity is a serious nightmare in this advanced communication age of 21st century to those who have to safeguard their confidential and sensitive files in their computers. They are in government agencies, private and public institutions. The cyber intrusions are done for a variety of reasons. The hackers might be: adversaries, thrillers, dissatisfied employees, terrorists, technical mischief makers, smugglers, money launders etc. The cyber intrusion rate is increased by the hackers whenever the vigilance level is weaker. When the traffic of cyber intrusions is heavily voluminous, the concerned people with responsibility to safeguard the files increase their vigilance level. The vigilances might be: *periodic security risk analysis* and *vulnerable components.* Still, there occurs an ongoing game between hackers and vigilantes. Their efforts real but are not directly observable. On the contrary, the only observable is the number of intrusions to sabotage cyber security. From such data on the number of cyber intrusions, the intrusion rate, the hacker's offensive efforts level for more cyber-attack and the vigilante's efforts level to protect the files with secured computer need to be estimated. For this purpose, a line of analysis has to be constructed. A necessity for the analysis is an underlying probability model for the collected intrusion data. The literature does not have a suitable model or an appropriate methodology for such data analysis. In Section II, this article introduces a new probability model and names it *Hacking-Vigilance Distribution (HVD).* Its statistical properties of HVD are derived. These results are demonstrated with data in Section III. Some conclusive thoughts are summarized in Section IV.

II. Derivation of Hacking-Vigilance Distribution and Its Properties

Suppose that there are Y number of independent intrusions at the end of a time, *t*. In an infinitely small next duration of time Δt, let the chance for one additional intrusion is $\theta\Delta t$ and the chance for two or more intrusions is zero, where the parameter $0<\theta<\infty$ is an unknown *intrusion rate*. Because of this scenario, the *random variable*, Y follows a *Poisson distribution (PD)*

$$\begin{aligned} &Pr[Y=y|\theta]=e^{-\theta}\theta^{y}/y!,\\ &y=0,1,2,...,.,0<\theta<\infty \end{aligned} \tag{1}$$

The mean, $E[Y|\theta]$ and dispersion, $D[Y|\theta]$ of the Poisson distribution in (1) are the same and equal to θ. A larger dispersion signifies more volatile occurrences. An implication is that whenever there is a high voluminous

Manuscript received January 25, 2013; revised March 25, 2013.

R. Shanmugam is with the School of Health Administration, Texas State University - San Marcos 78666, USA (e-mail: rs25@txstate.edu).

DOI: 10.7763/IJIET.2013.V3.285

intrusion, they occur with volatility also.

The Poisson distribution is suitable for an ideal scenario in which there is no hacking or no defensive vigilantism to protect the computer security. This lack of enough vigilantism is taken advantage by the hackers. The hackers may put in extra efforts to increase their cyber-attacks and such efforts should result in more mean intrusions. Let the hacker's efforts be an unknown parameter, $\rho \geq 0$. To include the *hacker's efforts,* the PD in (1) needs to be expanded to a *spinned Poisson distribution (SPD)* in (2)

$$\begin{aligned} &Pr[Y = y|\rho,\theta] \\ &= [1+\rho y]e^{-\theta}\theta^{y}/[1+\rho\theta]y!, \\ &y = 0,1,2,...,.,0 < \theta < \infty, \rho \geq 0. \end{aligned} \quad (2)$$

The mean, $E[Y|\rho,\theta]$ and dispersion, $D[Y|\rho,\theta]$ of the SPD in (2) are respectively

$$E[Y = y|\rho,\theta] = \theta[1+\frac{\rho}{1+\rho\theta}] \quad (3)$$

and

$$D[Y = y|\rho,\theta] = \theta[1+\frac{\rho(1+\theta)}{(1+\rho\theta)^{2}}\{1+(1+\rho)\theta^{2}\}] \quad (4)$$

In the absence (that is, $\rho = 0$) of the hacker's efforts to intensify the intrusions, the probability mass function in (2), the mean in (3) and dispersion in (4) become the PD in (1), its mean and dispersion respectively. The SPD was introduced in [5] to comprehend the functioning of a health mechanism. A reason for choosing the SPD for our purpose is that the mean of the SPD in (2) is more than the mean of PD in (1). The mean number intrusions under the presence of the hacker's efforts are more than under their absence. The extra amount $[\frac{\rho\theta}{1+\rho\theta}]$ in the mean number of intrusions is due to the offensive nature of the hacker's efforts. Also, the number of intrusions in the presence of the offensive hacker's efforts is more volatile than under their absence. The extra volatility $[\frac{\rho\theta(1+\theta)}{(1+\rho\theta)^{2}}\{1+(1+\rho)\theta^{2}\}]$ is also due to the hacker's efforts.

Realizing the voluminous hacking activities or even otherwise, the owners of the files in a computer system bumps up their vigilances to counter the cyber insecurity. Let $\tau \geq 0$ be the unknown impact of the vigilante's efforts. In the presence of the vigilante's efforts, the intrusion rate of PD in (1) is $\theta/(1+\tau)$. That is,

$$\begin{aligned} &Pr[Y = y|\tau,\theta] = e^{-\theta/(1+\tau)}[\theta/(1+\tau)]^{y}/y!, \\ &y = 0,1,2,...,.,0 < \theta < \infty, \tau \geq 0. \end{aligned} \quad (5)$$

The mean, $E[Y|\tau,\theta]$ and dispersion, $D[Y|\tau,\theta]$ of the PD in (5) are the same and equal to

$$E[Y = y|\tau,\theta] = \frac{\theta}{(1+\tau)} = D[Y = y|\tau,\theta]. \quad (6)$$

In the absence (that is, $\tau = 0$) of the vigilante's efforts to reduce or eliminate the intrusions, the probability mass function in (5), the mean in (6), and dispersion in (7) become the PD in (1), its mean and dispersion respectively.

Now, consider the scenario in which both the offensive nature of the hacker's efforts and the defensive nature of the vigilante's efforts exist in a realistic sense. The collected data on the number of intrusions do not identify how many occurred due to the hacker's efforts and how many due to vigilante's efforts. Hence, the chosen underlying model for the data needs to take care of it. In this sense of a mixed situation, let $H = 0, 1, 2, ...,$ and $V = 0, 1, 2, ...,$ denote the unobserved number of intrusions because of the hacker's and vigilante's efforts respectively. Then, the needed model is for their sum $Y = H + V$. Assume that H and V are independent *random variables*. Also, assume that H follows a spinned Poisson probability pattern (that is, $H \sim P(i|\rho,\theta)$) in (2) and V follows a Poisson probability pattern(that is, $H \sim P(y-i|\tau,\theta)$) in (5). Then, their sum Y = H+V follows a probability pattern in (7) below. That is,

$$\begin{aligned} &Pr[Y = y|\rho,\tau,\theta] = \sum_{i=0}^{\infty} Pr(H = i|\rho,\theta)Pr(V = y-i|\tau,\theta) \\ &\approx \frac{e^{-(\tau+2)\theta/(\tau+1)}\theta^{y}}{(1+\rho\theta)y!}\sum_{i=0}^{\infty}(1+\rho i)\binom{y}{i}(1+\tau)^{i} \\ &= \frac{e^{-(\tau+2)\theta}[(\tau+2)\theta]^{y}[1+\frac{\rho(\tau+1)y}{(\tau+2)}]}{[1+\rho(\tau+1)\theta]y!} \\ &y = 0,1,2,...,.,0 < \theta < \infty, \rho \geq 0, \tau \geq 0. \end{aligned} \quad (7)$$

The result in (7) is new to the literature and hence, it is now named as *Hacker-Vigilante Distribution (HVD),* where $\rho \geq 0$, $\tau \geq 0$ and $\theta \geq 0$ are the *hacker's efforts,* the *vigilante's efforts* and the *intrusion rate* respectively.

Of course, the owners of the confidential and sensitive files in a computer system desire to have an intrusion free situation. Could such an ideal intrusion free situation happen? What are its odds? The odds are the ratio of the chance for an intrusion free over the chance for no intrusion free situations to occur. The *odds* is then

$$\text{Odds}_{\rho\neq0,\tau\neq0} = \frac{Pr[Y = 0|\rho,\tau,\theta]}{Pr[Y \geq 1|\rho,\tau,\theta]} = [e^{(\tau+2)\theta}\{1+\rho(\tau+1)\theta\}-1]^{-1}. \quad (8)$$

Realize that there could be *four* mutually exclusive scenarios. The *first scenario* is realistic where the hacker's efforts and the vigilante's efforts prevail and the odds of intrusion free to occur in the scenario is (8). The *second scenario* is an ideal type where both the hacker's and vigilante's efforts are absent and the odds of intrusion free is

$$\text{Odds}_{\rho=0,\tau=0} = [e^{2\theta}-1]^{-1}. \quad (9)$$

The *third* scenario is one in which the hacker's activities are absent but the vigilante's efforts exist and the odds of intrusion free is

$$\text{Odds}_{\rho=0,\tau\neq0} = [e^{(\tau+2)\theta}-1]^{-1}. \quad (10)$$

The *fourth* scenario is one in which the hacker's efforts exist in the absence of the vigilance and the odds of intrusion free is

$$\text{Odds}_{\rho\neq 0,\tau=0}=[e^{(\tau+2)\theta}-1]^{-1}. \qquad (11)$$

How are these odds inter-related? Substituting (9), (10) and (11) in (8), the odds in (8) becomes

$$\text{Odds}_{\rho\neq 0,\tau\neq 0}$$
$$=[(\frac{1+\text{Odds}_{\rho=0,\tau\neq 0}}{\text{Odds}_{\rho=0,\tau\neq 0}})\{\rho\tau\theta+(\frac{\text{Odds}_{\rho=0,\tau=0}}{1+\text{Odds}_{\rho=0,\tau=0}})(\frac{\text{Odds}_{\rho\neq 0,\tau=0}}{1+\text{Odds}_{\rho\neq 0,\tau=0}})\}-1]^{-1}.$$

Now, the statistical properties of the HVD in (7) are explored. The mean, $E[Y|\rho,\tau,\theta]$ and dispersion, $D[Y|\rho,\tau,\theta]$ are respectively

$$E[Y|\rho,\tau,\theta]=[1+\tau+\frac{\rho}{1+\rho\theta}]\theta \qquad (12)$$

and

$$D[Y|\rho,\tau,\theta]=[1+\tau+\frac{\rho(1+\theta)\{1+(1+\rho)\theta^{2}\}}{(1+\rho\theta)^{2}}]\theta. \qquad (13)$$

The mean-variance relation in the HVD is

$$D[Y|\rho,\tau,\theta]$$
$$=[E(Y|\rho,\tau,\theta)+\frac{\rho\{(1+\theta)\{1+(1+\rho)\theta^{2}\}-(1+\rho\theta)\}}{(1+\rho\theta)^{2}}]\theta.$$

Now, a procedure to estimate the three parameters of the HVD in (7) has to be worked out. Three *equations* are needed to estimate the parameters with a given data. The mean in (12), dispersion in (13) and the zero probability, $Pr[Y=0|\rho,\tau,\theta]$ will suffice for this purpose. The maximum likelihood estimation is more efficient but will be computationally nonlinear and cumbersome. Their approximate estimates can be sequentially obtained using

$$\hat{\rho}\approx s_{y}^{2}-\bar{y} \qquad (14)$$

$$\hat{\tau}\approx\hat{\rho}-\frac{\ln\hat{P}r(Y=0)}{\bar{y}}-2 \qquad (15)$$

and

$$\hat{\theta}\approx\frac{\bar{y}}{(1+\hat{\tau}+\hat{\rho})}. \qquad (16)$$

III. Illustration Using Cyber-Attacks Data

In this section, the HVD and their properties are illustrated using the number of intrusions over eight causes to a bank's computer in Austin, Texas during six months in Table I below.

Using the estimators in (14), (15) and (16), the estimates of the intrusion rate, hacker's efforts and the vigilante's efforts in the six months are captured and displayed in the Fig. 1, Fig. 2 and Fig. 3 respectively. The intrusion rate (see Fig. 1) is more to begin with, declines later until March and then moves up. The hacker's efforts have been oscillating (see Fig. 2). The vigilante's efforts have also been oscillating parallel to the hacker's efforts (see Fig. 3). The odds of intrusion free situation improves until March but then slides deep down to get better later (see Fig. 4).

TABLE I: Y = # Cyber-Attacks (in 1,000)

Cause of intrusions	M1	M2	M3	M4	M5	M6
Denial of service	2	4	0	3	6	8
Phishing scams	3	0	1	8	1	6
Online Trojans	8	2	0	4	0	0
Cyber stalking loading	0	5	3	8	3	2
Fraud & Stealing	1	3	2	0	8	1
Password sniffing	3	0	5	5	9	8
Rootkit diluting security	7	8	0	9	9	0
Cyber espionage	6	9	9	2	5	9
Mean	3.8	3.88	2.5	4.88	5.13	4.25
Dispersion	8.5	11.3	10	10.4	12.4	15.1

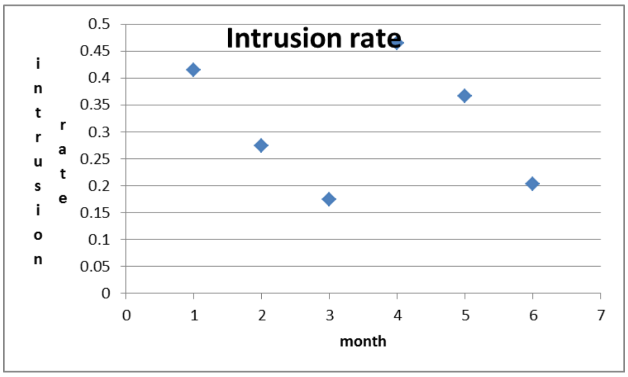

Fig. 1. The intrusion rate.

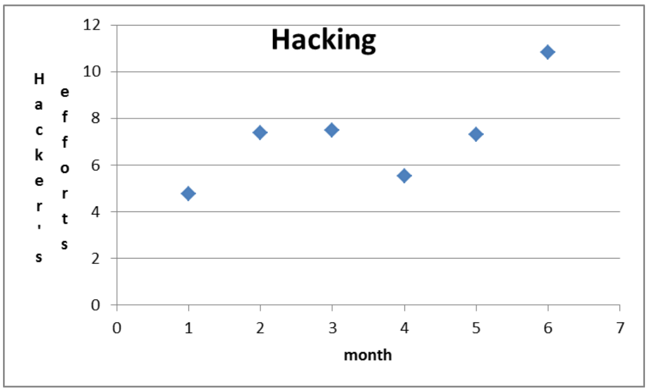

Fig. 2. The hacker's efforts.

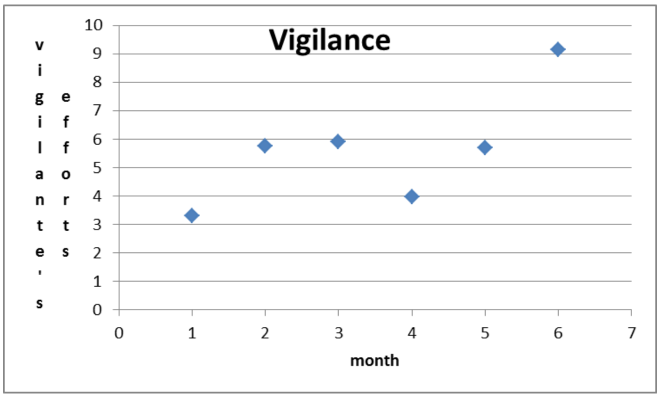

Fig. 3. The vigilante's efforts.

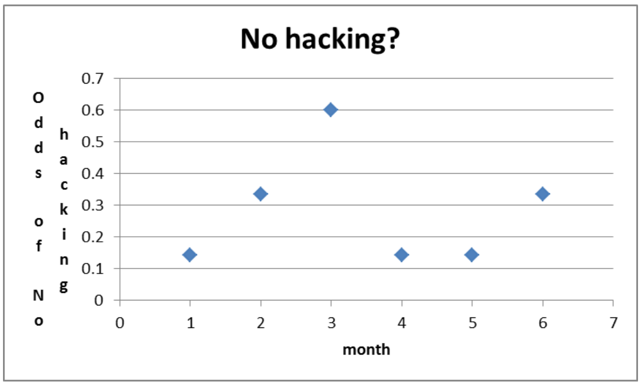

Fig. 4. The odds for hacking free situation.

IV. Conclusion

In conclusion, the model and methodology of this article help to identify, estimate and interpret the intrusion rate, the hacker's efforts and the vigilante's efforts out of the collected hacking data. The next need is to probe into the motives of the hackers and their statistical significance. Data on related covariates would be helpful to configure whether or not they control significantly the causation of hacking or the prevention of computer insecurity. For this purpose, a regression type statistical methodology will be constructed and reported in the future. Such methodologies will be helpful to institutions and government agencies in their goals of securing their confidential and important sensitive files from being stolen.

References

[1] J. J. Prichard and L. E. MacDonald, "Cyber terrorism: A Study of the extent coverage in computer security textbooks," *JITE*, vol. 3, pp. 279-289, 2004.

[2] I. Green, T. Raz, and M. Zviran, "Analysis of active intrusion prevention data for predicting hostile activity in computer networks using a generic and reliable model to anticipate future attack scenarios," *Communications of the ACM.*, vol. 50, no. 4, pp. 63-68, 2007.

[3] K. I. Choi, X. Chen, S. L. M. Kim, K. Chae, and J. C. Na, "Intrusion detection of NSM based DoS attacks using data mining in smart grid," *Energies*, vol. 5, pp. 4091-4109.

[4] A. M. Rushdi and O. M. Ba-Rukab, "Fault-free modeling of computer system security," *International Journal of Computer Mathematics*, vol. 82, no. 7, pp. 8054-819, 2005.

[5] R. Shanmugam, "Spinned Poisson distribution with health management application," *Health Care Management Science*, vol. 14, pp. 299-306, 2011.

Ramalingam Shanmugam received Ph.D. degree from the Statistics Department at Temple University. Currently, he is a professor in the School of Health Administration at Texas State University - San Marcos. His recent publications are on modeling infectious diseases, diagnostic methodology, and modeling computer viruses. He serves as Book Review Editor for the Journal of Statistical Computation and Simulation.