The Risk of a Cyber Disaster

Estimating the Exceedance Probability Function of a Global Computer Virus

Vaughn H. Standley
National Defense University
Fort Lesley J. McNair
Washington D.C. 20319 USA
email: vaughn.standley@nnsa.doe.gov

Roxanne B. Everetts
National Defense University
Fort Lesley J. McNair
Washington D.C. 20319 USA
email: roxanne.everetts@ndu.edu

***Abstract*— Rigorous assessment of disaster risk requires an exceedance probability function relating the probability that 'S', a random variable representing the severity of the disaster, exceeds some threshold 's' above which destruction is expected. Calculating a valid exceedance probability function for disasters is not straightforward. The Power Law has served as a panacea for this difficulty, often erroneously. Here, an alternative approach is demonstrated using empirical data for interstate war, the coronavirus pandemic, and identity theft. The method relates the frequency distribution of severity S (deaths or failures per state) to the product of frequency distributions for vulnerability V (deaths or failures per case or combatant), exposure E (cases or combatants per capita), and population P (population per state). The probability density function for S, from which the exceedance probability function is derived, may then be computed using obtainable distributions for V, E, and P if data for S is not directly available. The method is used to estimate the risk of a global cyber disaster. Results suggest that the probability density functions for this situation follow log-gamma distributions. The fits can be used in stochastic decision formulae enabling authorities to optimally choose among alternative cyber preparedness or resilience measures to minimize overall risk.**

Keywords- catastrophe theory; military; power law; risk analysis.

I. INTRODUCTION

"What is cyber risk?" – "What are the costs and detrimental effects caused by cyber risk?" – "Where do we find data on cyber risk?" – "How can we model cyber risks?" These first four of ten key questions posed by The Geneva Association [1] suggest that as recently as 2016 very few of the technical fundamentals of cyber risk are understood. Fast forward five years; a global biological virus – not a digital virus – will help to answer these critical questions.

Malicious, replicating digital software is called a "computer virus" because it is characterized by rapid proliferation and high unpredictability, just like a biological virus. One phenomenon has real-life implications for the other. These implications ought to be studied and applied for common benefit, such as in decision formulae used to minimize risk. Informed by empirical data, these equations can help optimally choose long-term investments to mitigate cyber threats, be integrated into operational software to defend against cyber-attacks in real-time or be used by actuarial scientists to determine insurance premiums when cyber-defenses fail, among other applications.

Network epidemiology holds that the spread of disease can be modelled with network theory [2]. Social and commuter networks, modelled as nodes and segments in a matrix, approximate disease transmission. Similarly, in the case of a computer virus it is the internet cables, servers, and client computers that form the network. If the impacts of a virus across a network are great, sudden, and unforeseen, a disaster ensues. Catastrophe theory was developed to address the stochastic nature of these events so that logical investments into preparedness and resilience measures could be made.

In this study, we extend previous catastrophe theory that has been applied to interstate war [3] to the coronavirus pandemic to develop a method for characterizing the magnitude and uncertainty of the severity of a worst-case computer virus that spreads to Internet-connected computers. The results are expected to help inform the development and implementation of cyber preparedness and resilience measures.

Section II provides additional background on exceedance probabilities and why use of the Power Law is not valid. Section III describes a method to estimate the exceedance probability for a global computer virus. Section IV reports the results. Section V is a summary of conclusions.

II. BACKGROUND

Probability distributions of severity embody the highly unpredictable nature of catastrophic phenomena. A Probability Density Function (PDF) quantifies the relative likelihood that the value of a random variable '*S*' representing severity is equal to some severity '*s*'. The complement (i.e., subtracted from one) of the integral of the PDF from zero to s is the exceedance probability function, $P(S>s)$. The exceedance function relates the probability that *S* exceeds a threshold s above which destruction is expected [4]. For example, to construct a building to survive earthquakes, the architect is concerned with the probability that the earthquake will be less than some specified Richter value, such as "9". Above this severity, destruction of the building cannot be reasonably avoided. We would write this exceedance probability as $P(S>9)$.

 ISBN: 978-1-61208-893-8

War is a man-made disaster that may be characterized by an exceedance probability. Beginning in 1960 with Lewis Fry Richardson's famous "The Statistics of Deadly Quarrels" [5] the Power Law has been widely used to model the exceedance probability of war. A phenomenon may be probabilistically distributed according to the Power Law if the logarithm of the exceedance probability plotted against the logarithm of severity s appears as a straight line with slope $-q$. This is written as $P(S>s) = s^{-q}$ and is quantified by grouping data according to consecutive ranges of severity and examining the frequency that wars fall into these groups. It turns out that the Power Law may be applied to many phenomena [6]. An example of the Power Law applied to cyber-crime data [7] is illustrated in Fig. 1. It is the straight line with an approximate slope of -0.7 (Note that variable b in the figure is negated in the Power Law formula).

Fig. 1 is an example of how the Power Law is often misapplied. For identification (ID) theft in the U.S., Circle A shows that the Power Law misrepresents data from 1 to 10,000 or about half of the entire range of the graph. Circle B shows that the data does not match the Power Law fit for greater than 10^7. Critically, the Power Law fails as a probability when $q<1$ unless the use is properly qualified. A proper qualification will recognize that "data held to be power-law distributed represent samples from some underlying population. As these samples often cover a narrower scale range than that of the population as a whole they are truncated" [8]. A q value less than one indicates that the exceedance probability decreases slower than the increase in severity. In such "long-tail" cases, the severity increases arbitrarily, causing the mean to become mathematically divergent. The slope associated with the Power Law fit to identification theft data in Fig. 1 is less than one, meaning that it is invalid as a probability distribution for the range indicated and, therefore, cannot be used in mathematical decision criteria (e.g., a likelihood ratio test) that may be used to compute minimum risk.

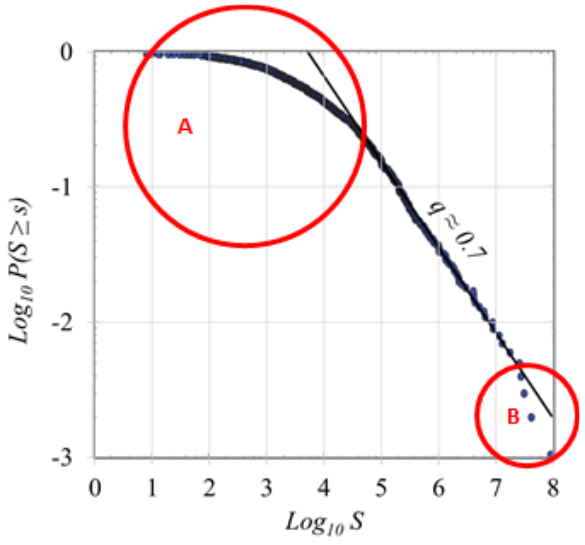

Figure 1. Normalized ID theft data and Power Law fit reported by [7].

Curvature in log-log data, such as that observed at both ends of the data in Fig. 1, suggest the applicability of logarithmic distributions other than the Power Law, like the log-normal distribution [9]. A Log-Normal (LN) distribution is a normal distribution applied to the logarithm of the statistic. Curvature in the integral of LN data is evident in many plots meant to demonstrate the applicability of the Power Law. The LN is a symmetric distribution, but often data will appear non-symmetric. A non-symmetric distribution skewed toward higher statistics is the log-gamma (LG). Conspicuously absent from disaster modelling literature is application of the LG to severity, except for one [10] linking LG and LN distributions of combat deaths to economic theory [11]. When plotted in a log-log graph, the middle section of the integral of the LN and LG PDFs will always appear somewhat straight, explaining why the Power Law is so often misapplied. Application of the Power Law in these cases is not only mathematically invalid, but it fails to reveal the true nature of the underlying phenomena.

Finding a valid exceedance probability is not straightforward. The Power Law erroneously serves as a panacea for this difficulty. The deficiencies noted in Fig. 1 illustrate how the Power Law is misapplied to cyber-risk. A better quantitative method is needed to estimate exceedance probabilities. An approach based on the spread of known computer viruses would be the best way to proceed. However, the data needed for such an approach is not available and/or public. Another method is needed.

III. METHOD

A novel alternative to the Power Law is demonstrated here with empirical interstate war and coronavirus pandemic fatality data that relates frequency distribution for severity S (deaths per state) to frequency distributions for vulnerability V (deaths per case or combatant), exposure E (cases or combatants per capita), and population P (population per state) by (1). Because all three of these variables are found to conform to parametric distributions associated with random variables, each may be viewed as a random variable.

$$S = VEP \tag{1}$$

In war, deaths are "transmitted" from one combatant to another by contact following geographic movement, much like how an airborne biological virus is transmitted. Similarities between interstate war and a global pandemic, including the finding that war is a network phenomenon [12], lead us to posit that the statistics of interstate war are representative of these and similarly networked phenomena. War data used in this study are from the Correlates Of War (COW) Project. Combat death statistics were obtained from the COW War Data, 1816 - 2007 (v4.0) [13]. Population and military personnel (i.e., combatant) statistics were obtained from the COW National Material Capabilities (NMC) (v5.0) dataset [14]. The two datasets were combined manually for this study. The data involves 93 wars. However, proper

 ISBN: 978-1-61208-893-8

application of game theory [15] requires that these wars be differentiated by participating nations, of which there are 337. Due to missing military personnel data for some of these wars, the number of states is reduced to 250. Other defects further reduce the set to 236 states. Moreover, it is reported that 25 of these warring states lost more combatants than reported in the NMC database. In these cases, we limit the number of combat dead to be 100% of the combatants.

The magnitude and variability of *S* for interstate war, measured in terms of combat dead, is represented by the red lines in Fig. 2(a). The solid red line with square markers indicates combat deaths taken directly from the COW War Data set. The thick semi-transparent red line with no markers indicates combat dead computed using (1). The solid green line with circle markers is the distribution of state populations taken from the NMC dataset, which represents P in the equation. The solid blue line with solid diamond markers indicates the distribution of combatants per capita, also taken from the NMC dataset, which represents exposure *E*. The distribution of vulnerability *V*, or deaths per combatant, taken from the interstate war dataset, is indicated by the solid orange line marked by triangles. In all graphs, solid lines with solid markers indicate empirical data, whereas dotted lines with no markers indicate parametric fits to data.

Curves in Fig. 2 appearing to the right of zero on the logarithmic axis are greater than one, while those to the left of zero are numbers between zero and one. The calculated deaths per nation *S*, a number greater than one, is the product of a random number *P* and likewise greater than one, with two random variables *E* and *V* that are both fractions. For this reason, the red curves are situated between the green curve and zero on the x-axis. The fact that the thick semi-transparent red curve overlaps the solid line with square markers is a good indication that estimation of deaths using *S=VEP* accurately reflects what is reported in empirical data. Small mismatches are mainly attributed to inaccurate army sizes reported in the COW NMC dataset.

Parametric fits are important because they help to determine if the data are mathematically well-behaved, discern what processes underly the phenomena, and applicable to risk-minimizing formulae. The distribution of state populations *P* follows a negatively Skewed LN (SLN) distribution. The *E* and *V* curves follow an LG distribution, described by (2). The distribution of combat deaths *S* follows an LG distribution.

$$f(x;\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \qquad (2)$$

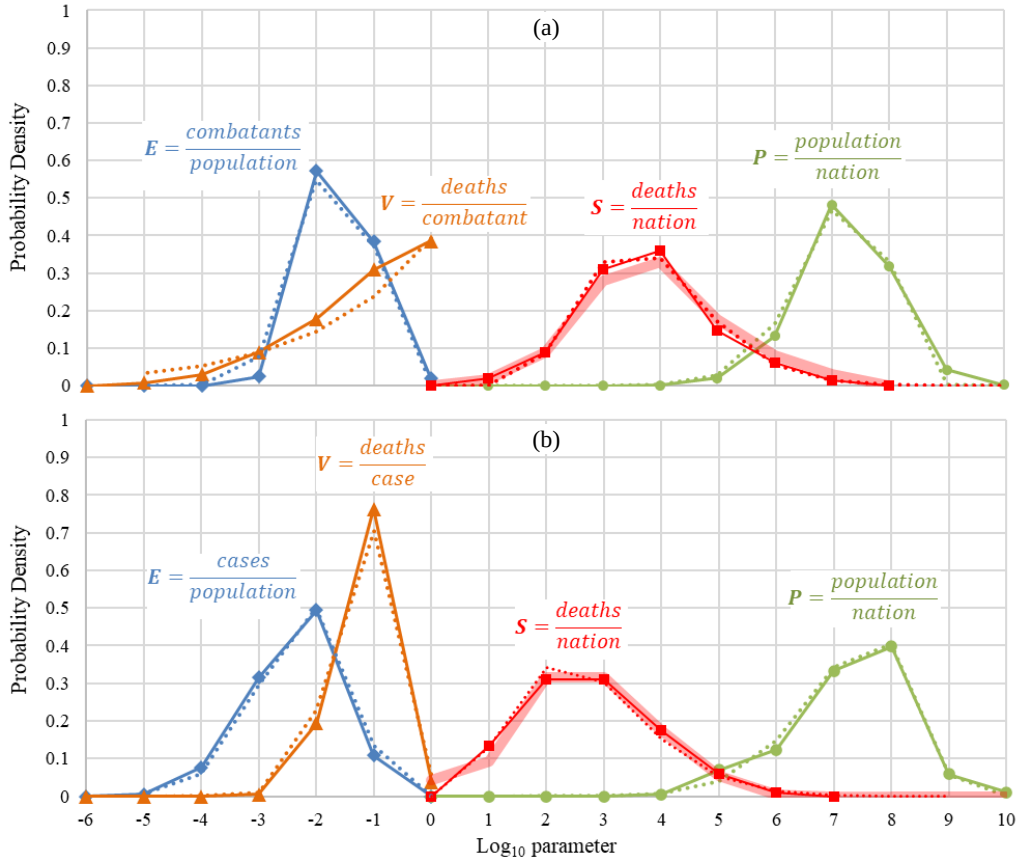

Figure 2. PDFs for (a) interstate war (top) and (b) the COVID-19 pandemic (bottom). Solid lines with solid markers indicate empirical data, dotted lines with no markers indicate parametric fits to data, and the thick semi-transparent line indicates the curve computed using the relation *S* = *VEP*.

 ISBN: 978-1-61208-893-8

Fig. 2(b) reports the *S*, *V*, *E*, and *P* curves associated with the coronavirus pandemic derived from Our World In Data (OWID) statistics [16]. Of the 217 nations reporting data, only 199 are used because zero values reported by 18 nations cannot be included in a logarithmic graph. The coronavirus graph is presented just below and in alignment with the interstate war graph using the same scales to help the reader compare and contrast the two sets of curves. The meaning of the solid and dashed lines is the same as for Fig. 2(a). It turns out that the same parametric functions fit the coronavirus data, except with different parameters.

Similarities and differences between phenomena are more evident when their data are separated into constituent random variables in this way. The similarities between the cases of interstate war and COVID-19 appear to be mostly a result of similar population data. The only difference between the population distribution for these cases is that the interstate war data spans 191 years from 1816 to 2007, whereas the COVID-19 pandemic population data is taken only from 2020. The most striking difference between the two are their vulnerability curves. For interstate war, there is a 40% chance that a nation loses all its combatants in a war. Compare this to the COVID-19 pandemic, where there is a zero probability that all exposed to the virus will die, but an 80% chance that 10% of those exposed will die. At first glance, these curves appear to be associated with two different parametric distributions because the V distribution for interstate war looks like an exponential. This difference is resolved by the fact that an exponential distribution is a gamma distribution for certain combinations of parameters. In other words, they both can be considered gamma distributions applied to the logarithm of the statistic.

Comparison of *S*, *V*, *E*, and *P* data for interstate war and the COVID-19 pandemic appear to make clear that two very different phenomena have real-life implications for the other, possibly due to common or similar underlying phenomena (e.g., both involve networks), and that the same might be true for cyber phenomena. One may choose different populations to study, or the populations themselves may change. However, for a given threat (e.g., war, coronavirus, etc.), we suspect that distributions for *E* and *V* may be common to or similar for these different populations. Thus, we can use this knowledge to estimate underlying distributions for *E* and/or *V* and use them in (1) to determine *S* for a given population under consideration. In contrast to the threat, vulnerability, and consequence model used by the Federal Emergency Management Agency (FEMA) where *Risk=TVC* [17], ours is based on a population because the threat and severity both derive from the population itself and exposure expresses the population's ability to convey the threat.

IV. RESULTS

We apply the method to the case of a hypothetical computer virus that spreads like COVID-19 and inflicts a combat-like mortality rate on Internet-connected computers. For this case, the population distribution is taken to be the number of people per nation with access to the Internet. For all nations, OWID [16] reports the fraction of people with Internet connections, which is multiplied by the population of the respective state. For *E* and *V* distributions, we intend to use those associated with the coronavirus pandemic and interstate war, respectively. Before doing so, however, we would like some evidence that these distributions are appropriate for modelling a computer virus. Unfortunately, there is no quantitative data available that directly serves this purpose.

The Privacy Rights Clearinghouse (PRC) is one of the few organizations to publish an online database quantifying different types of cyber-crime [18]. However, this database does not provide any statistics about the number of attacks or infiltrations per capita or the number of records per attack or infiltration. That is, the database does not help quantify *E* or *V*. Only the distribution of *S* can be discerned from the PRC data. Our approach in this case is to "reverse-engineer" the vulnerability distribution using (1) by first positing that the exposure distribution is the same as for the coronavirus. We then adjust the vulnerability distribution until the relation *S=VEP* produces an *S* distribution that matches the distribution of the PRC data. State populations in the U.S. were taken from Wikipedia [19]. The result of this process applied to "datalossdb" records in the PRC ID theft database is reported in Fig. 3. As before, *S=VEP* is represented by the thick semi-transparent red line and the empirical data for S is represented by the solid red line with square markers. The resulting *V* distribution is more consistent with the vulnerability associated with interstate war than to vulnerability associated with the coronavirus, a finding that gives us some confidence that the respective *E* and *V* distributions for the Coronavirus and interstate war can be applied to our hypothetical computer virus.

Fig. 4 reports the exceedance probability for *S* computed using *P* and *E* distributions from the coronavirus pandemic and a reverse-engineered *V* distribution. Only a thick semi-transparent red line is reported (i.e., no solid line with square markers) because the severity distribution is based on *S=VEP* and there is no *S* data with which to compare directly. However, we can compare the *S=VEP* curve to the exceedance probability for the PRC ID theft data, which is indicated by the solid black line with solid square markers. The main difference between the curves is that they diverge beginning at a value of 5 (i.e., 100,000) on the x-axis. The Power Law approximation associated with the PRC is reproduced in Fig. 4 as the black dotted line. Our best fit of a power law to the PRC data is with a slope of -0.65, which rounds to -0.7, the value reported by Maillart and Sornette [7]. As with the Power Law fit in Fig. 1, the fit in Fig. 4 diverges from the data for s values less than 4.5 and greater than 7.5.

Parametric fits to each of the four PDFs (i.e., *S*, *V*, *E*, and *P*) for each of the four phenomena (i.e., interstate war, the COVID-19 pandemic, U.S. ID theft, and hypothetical global computer virus) are recorded in Table 1. For the hypothetical virus, a non-skewed LN distribution ($\alpha=0$) fits the population of internet-connected devices, which is slightly different than the other populations fit by a SLN distribution. The curve for the computed severity of the hypothetical computer virus appears to follow an LG distribution, which is the same as for interstate war and the coronavirus pandemic.

 ISBN: 978-1-61208-893-8

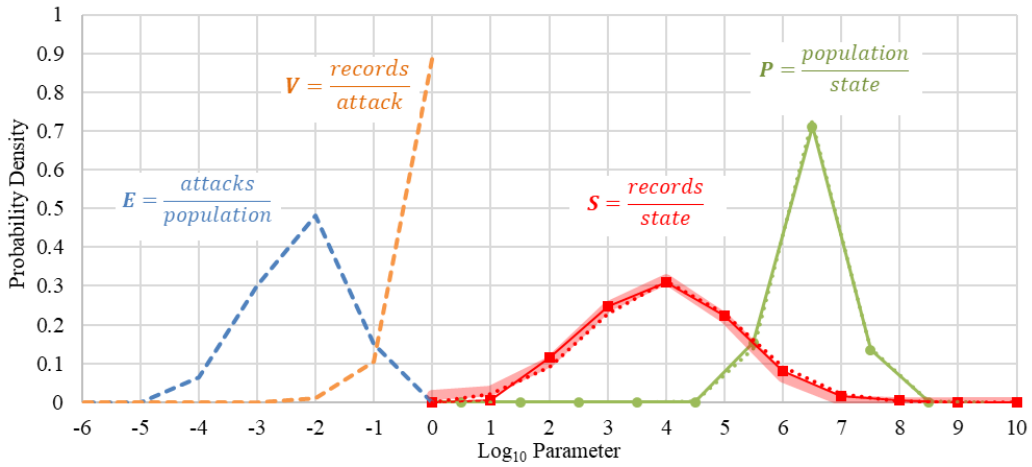

Figure 3. PDFs for ID theft in the U.S. The severity distribution is based on Privacy Rights Clearinghouse data [18]. The population distribution, which is Internet access per nation, is based on OWID [16]. Curve formats have the same meaning as for curves in Fig. 2.

Parametric distributions that mimic empirical data are valuable to decision formulae. What is particularly important in the case of logarithmic severity distributions, like those in this study, is that the parametric fits adequately model the high-severity portion of the data. Consider the data and fits to the data in Fig. 4. The data represented by the solid black line is turning downward, gaining a more negative slope whereas the black dotted line (i.e., the Power Law fit) is a straight line with negative slope 0.7. The Power Law approximation cannot be used in probabilistic decision formulae because it is divergent for slopes equal to or greater than negative one. Conversely, the LG fit represented by the dotted red line, which faithfully mimics the solid red data line, is valid for use in such formulae because it becomes increasingly negative. As can be seen in these exceedance probabilities, there is a portion in the middle that is approximately straight, which creates the temptation to report the distribution as a Power Law. This tendency is particularly prevalent for war statistics [20].

Results should not be overinterpreted. The method is not useful for investigating microscopic causes of cyber-risk, although it can be used to posit or confirm the macroscopic result of microscopic causes vis-à-vis parametric distributions. However, the method is a better bookkeeping and estimation tool for uncertainty in the constituents of risk when the threat is created and propagated by the population. FEMA's model is applicable to threats that are independent of the population (e.g., earthquakes).

V. CONCLUSION

Risk is the product of probability and severity. Exceedance probability is the mathematical object connecting both. The magnitude and variability of the severity *S* of a computer virus can be computed in terms of frequency distributions representing the subject population, *P*, that part of the population exposed to the risk, *E*, and the vulnerability of the exposed, *V*. Currently there is not enough cyber risk data to calculate S directly, so the advantage of

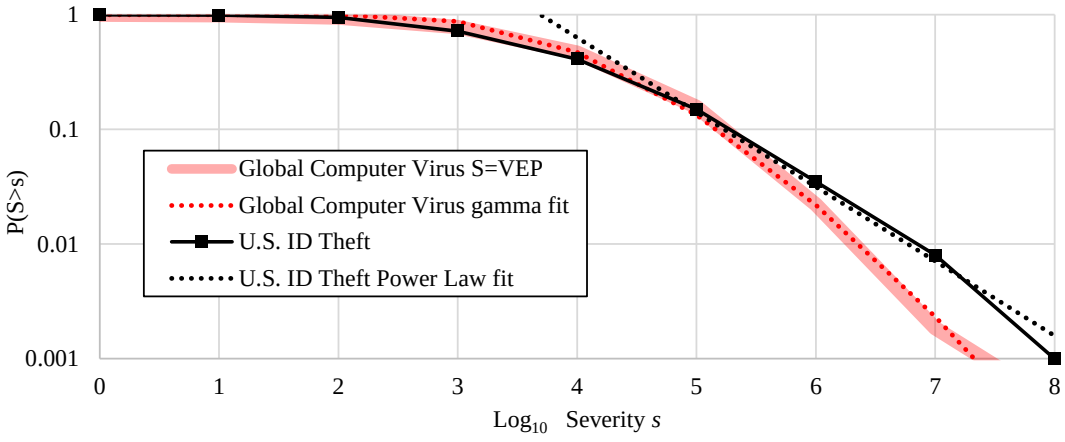

Figure 4. Exceedance probability functions, *P(S>s)*, for U.S. ID theft (solid black with square markers) and a highly "contagious" global computer virus (thick semi-transparent red) developed using the S=VEP relation, with power law fit to U.S. ID theft data (dotted black) and log-gamma fit to *S=VEP* curve for the global computer virus (dotted red).

TABLE 1. PARAMETRIC FITS TO PROBABILITY DENSITY FUNCTIONS.

	Interstate War	Coronavirus Pandemic	Records Theft	Computer Virus
P	SLN ξ=8, ω=1.2, α=-3	SLN ξ=8.4, ω=1.5, α=-3	SLN ξ=7.8, ω=1.2, α=-2	SLN ξ=7, ω =0.9, α=0
E	LG α=17, β=0.13	LG α=14, β=0.20	LG α=14, β=0.20	LG α=14, β=0.20
V	LG α=1.0, β=2.0	LG α=5.5, β=0.35	LG α=1.0, β=0.50	LG α=1.0, β=0.50
S	LG α=9.8, β=0.34	LG α=4.0, β=0.6	LG μ=3.5, σ=1.3	LG α=4.0, β=0.6

this method is that the PDF of S, from which the exceedance probability function is derived, may be computed indirectly using more readily obtainable or representative probability densities for *V*, *E*, and *P*. The Power Law is divergent when applied to the cyber-risk so it should be avoided for these purposes in favor of methods such as the one proposed here. The method was applied to a hypothetical computer virus given the propensity to spread like COVID-19, predicated on the hypothesis that the frequency distributions associated with interstate war, COVID-19, and computer viruses manifest similar network behavior. Results are consistent with this hypothesis.

The PDF associated with the logarithm of severity for a worldwide computer virus is fit by a gamma distribution. This parametric distribution can be used in operational computer software designed to detect and react to cyber threats in real-time, in stochastic decision formulae enabling authorities to optimally choose among alternative cyber preparedness or resilience measures, or in actuarial equations to determine insurance premiums for cyber risks.

Using data from Tab. 1, we compute the logarithmic variance of the computer virus to be 1.44 ($=\alpha\times\beta^2$) and the logarithmic standard deviation to be 1.2, which is equal to a factor of 16 ($=10^{1.2}$). For x-axis values greater than 6 in Fig. 4, the exceedance probability varies by an order of magnitude in one standard deviation, meaning that the risk of a global cyber disaster is associated with very high uncertainty. This finding is likely to hold for a real-live computer pandemic because it is rooted in empirical U.S. cyber-crime data that has been corrected in terms of its population, exposure, and vulnerability distributions.

DISCLAIMER

The opinions, conclusions, and recommendations expressed or implied are the authors' and do not necessarily reflect the views of the Department of Defense or any other agency of the U.S. Federal Government, or any other organization.

REFERENCES

[1] M. Eling and W. Schnell, "Ten Key Questions on Cyber Risk and and Cyber Risk Insurance," The Geneva Association, Zurich, Switzerland, 2016.

[2] L. Danon et al., "Networks and the Epidemiology of Infectious Disease," Hindawi Publishing Corporation, Interdisciplinary Perspectives on Infectious Diseases, Volume 2011.

[3] V. H. Standley, F. G. Nuño, and J. W. Sharpe, "Modeling Interstate War Combat Deaths," International Journal of Modeling and Optimization, vol. 10, no. 1, pp. 1-8, 2020.

[4] T. G. Lewis, Critical Infrastructure Protection in Homeland Security - Defending a Networked Nation, Hobeken, New Jersey: John Wiley & Sons, 2015.

[5] L. F. Richardson, The Statistics of Deadly Quarrells, Chicago: Quadrangle Books, 1960.

[6] L. Cederman, "Modeling the Size of Wars: From Billiard Balls to Sandpiles," The American Political Science Review, vol. 97.1 , no. April 2015, pp. 135-50, 2003.

[7] T. Maillart and D. Sornette, "Heavy-Tailed Distribution of Cyber-Risks," Physics of Condensed Matter, vol. 75, no. 3, pp. 1-16, 2008.

[8] G. Pickering, J. M. Bull, and D. J. Sanderson, "Sampling power-law distributions," Tectonophysics, vol. 248, pp. 1-20, 1995.

[9] L. Benguigui and M. Marinov, "A classification of natural and social distributions Part one: the descriptions," 2015. [Online]. Available: https://arxiv.org/abs/1607.00856 [retrieved: August, 2021]

[10] V. H. Standley, J. W. Sharpe, and F. G. Nuño, "Fusing attack detection and severity probabilities: a method for computing minimum-risk war decisions," Computing, 102, pp. 1385–1408 2020.

[11] J. von Neumann and O. Morgenstern, Theory of Games and Economic Behavior, 3rd ed., Princeton N.J.: Princeton University Press, 1953.

[12] M. O. Jackson and S. Nei, "Networks of Military Alliances, Wars, and International Trade," PNAS, vol. 112, no. 50, pp. 15277-15284, 2015.

[13] M. R. Sarkees and F. Wayman, Resort to War: 1816 - 2007, Washington DC: CQ Press, 2010.

[14] D. J. Singer, S. Bremer and J. Stuckey, "Capability Distribution, Uncertainty, and Major Power War, 1820 - 1965," in Peace, War, and Numbers, Beverly Hills, Sage, 1972, pp. 19-48.

[15] T. C. Schelling, The Strategy of Conflict, 1st ed., Cambridge: Harvard College, 1960.

[16] C. Appel et al. "Data on COVID-19 (coronavirus) by Our World in Data," The Oxford Martin Programme on Global Development, [Online]. Available: https://github.com /owid/covid-19-data/blob/master/ public/data/README.md. [retrieved: August, 2021]

[17] Analysis, Committee to Review the Department of Homeland Security's Approach to Risk, "Review of the Department of Homeland Security's Approach to Risk Analysis," The National Academies Press, Washington D.C., 2010.

[18] P. R. Clearinghouse, "Data Breaches," Privacy Rights Organization, 13 January 2020. [Online]. Available: https://privacyrights.org/data-breaches. [retrieved: August 2021].

[19] "List of states and territories of the United States by population," Wikipedia, 5 November 2020. https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population.

[20] R. González-Val, "War Size Distribution: Empirical Regularities Behind the Conflicts," Defence and Peace Economics, vol. 27, issue 6, pp. 838-853, 2014.

 ISBN: 978-1-61208-893-8