Information Assurance and Security Letters.
ISSN 2150-7996 Volume 2 (2011) pp. 032-036

Modeling Estimated Risk for Cyber Attacks: Merging Public Health and Cyber Security

R. David Parker, PhD[1] and Csilla Farkas, PhD[2]

[1]Research Assistant Professor of Medicine
2 Medical Park, Suite 505 Columbia, SC, USA 29203
+1.803.540.1055 (p) +1.803.212.6596 (f)
david.parker@uscmed.sc.edu

[2]Associate Professor of Computer Science and Engineering
315 South Main Street, Columbia, SC, USA 29208
+1.803.576.5762(p) +1.803.777.3767
farkas@cse.sc.edu

Corresponding Author:

R. David Parker, PhD

***Abstract*: The proposed United States FY2013 budget requests $769 million dollars for information security initiatives under the Department of Homeland Security. Projected annual spending for information security exceeds $10 billion by 2015. A new method using validated processes to quantitatively identify risk of a successful cyber attack would allow targeted interventions and a more efficient use of funds.**

Estimating the risk of a successful cyber attack is an important objective in this field. Validated methods imported from other fields to information technology such as epidemiologic statistical modeling could be of substantial benefit . This modeling not only determines the direction of influence, whether increasing or mitigating, but allows a standard unit across multiple areas allowing experts to determine the areas at greater increased risk. Extant data imported into a Cyber Security Surveillance System (CS^3), statistical modeling would calculate the cumulative effects of multiple factors on specific risk factors. The model outputs statistics allowing easy interpretation of the potential risk introduced by factor. Measuring the estimated risk in multiple domains within a predetermined unit (whether organizational or geographic) would enable decision makers to intervene prior to an attack and implement preventative measures to improve system stability.
Keywords: risk estimation, risk prediction, information assurance

I. Introduction

Multiple issues confront a nation's ability to ensure the protection of its infrastructure. As global technology use increases, the risk of non-traditional military and terrorist events increases. [1] Intentional attacks that exploit vulnerabilities of new technology have been well documented with the first reported exploit of network vulnerability during the initial demonstration of the wireless telegraph in 1903. [2] As technology has evolved and international dependence upon information technology (IT) solutions for daily operations has increased, it is imperative that information assurance (IA) methods increase at a rate at least proportionate to the persons who would exploit system vulnerabilities.

When addressing vulnerabilities within the field of IA, including cyber attacks (CA), the development of comprehensive and standardized approach should define many concepts. Methods, including those which measure the impact, outcome and potential risk of an attack are additional needs. As IA is an emerging field, it would be prudent to borrow validated methodologies from existing fields. One field with common similarities is public health. National defense forces have active programs in place that use public health methods to increase the security of the US. [3] In IT, previously adopted aspects from the health fields include the concept and operationalization of

viruses and bacteria. Simply stated, common modalities between health systems and IA systems are not limited to the IT systems used in public health and medicine.

II. Epidemiologic Prediction Models

Two concepts of potentially significant benefit are public health surveillance systems (PHSS) and estimated risk calculation. Using an adapted PHSS would allow the creation of statistical models to determine the estimated risk of a negative event, thus creating opportunities for intervention and reduction of risk. Risk reduction reduces the impact of the negative event. Public health focuses on methods of prevention and care. While each approach strives to positively impact health, the approach, methods, point of implementation and cost vary. Prevention generally is more cost beneficial and is implemented prior to the onset of a negative event.

III. Public Health Surveillance Systems

PHSSs collect and warehouse data that enable the understanding of public health issues, such as disease pathology and the calculation of risk, identification of vectors and treatment interventions. [4-6] Surveillance systems also include data from health registries, such as births, chronic diseases, infectious diseases and deaths. The World Health Organization (WHO) has released guidelines for the development and implementation of surveillance systems. [4, 5] Using WHO guidelines, many nations have developed and implemented successful surveillance programs designed to import existing data from various systems into one unified system that allows the estimation of the distribution of disease burden throughout a population and therefore allows public health workers to track shifting patterns on a local, state, regional or national level. These systems include laboratory values, hospitalization information, behavioral data and other data sources that impact disease acquisition and progression. Prevention programs are based on the outcomes and evaluation of these data.

IV. Susceptible, Exposed, Infectious and Recovered Model

The SEIR (susceptible, exposed, infectious, recovered) Model can be used to design the PHSS framework. This model accounts for various person types in a population in respect to disease nosology and pathology. In IA, this could be measured through a similar marker, such as the exposure of PCs to a virus (TABLE 1). To determine the rates of disease and estimate the risk, using the SEIR model, it is important to identify the frequency of the population in each category.

Susceptible persons are those at risk of contracting a disease. An applied example is the risk of exposure to a blood borne pathogen (or a PC at risk of contracting a virus). Persons sharing injection drug equipment are at increased risk of exposure to blood borne pathogens compared to persons who do not share injection equipment. [4, 5] File sharing PCs are at increased risk for contracting a virus compared to non-file exchanging PCs. Infectious persons are susceptible persons who were exposed and contracted the pathogen and who may transmit the pathogen. Not all susceptible persons will be exposed or become infectious. Infectious PCs are susceptible and exposed PCs with the virus who can transmit to other PCs. Recovered persons are those who successfully received treatment. Recovered PCs contracted the virus and were successfully treated.

The SEIR model allows the calculation of the percent of the population affected at each stage thereby providing the ability to track changes over time. In the example, the impact of injection drug equipment sharing (or file sharing) on the total population who tested positive for the pathogen (virus) without another high risk factor.

Using time series analyses, interventions can be measured for their impact on the distribution, frequency and intensity of event. Identifying the level of impact on each area and by risk factor allows an understanding of the overall risk. Multivariable modeling allows the investigation and identification of risk from multiple factors. PHSSs using the SEIR model could be adapted to IA to measure data from multiple sources thereby increasing the probability of developing a robust statistical model.

Table 1. Translation Example of the Susceptible, Exposed, Infectious and Recovered (SEIR) Model from Infectious Diseases (ID) to Information Assurance (IA)

	Infectious Diseases Modeling	**Information Assurance Modeling**
Negative	Blood Borne	Virus

Event	Pathogen	
Susceptible – at risk population	Persons at risk of contracting the pathogen	PC at risk of contracting a virus
Exposed – population in contact with event	Persons sharing injection drug use equipment	PCs engaging in file sharing
Infectious – population exposed and impacted by the event who can further transmit the event	Persons with the blood borne pathogen who continued to engage in sharing injection equipment	Infected PCs that continue to engage in file sharing
Recovered – at risk population which was exposed and may or may not have been infectious that have received successful treatment (may be further susceptible)	Persons with blood borne pathogen who were successfully treated.	Infected PCs that were restored to a non-infected status and are operational.

V. Defining Prevention

Prevention is the attempt to mitigate the impact of negative health events, control the spread of disease to uninfected persons, and improve the overall health of persons in a population (Table 2). [7] Primary prevention attempts to change circumstances so that an event does not occur or is significantly delayed. Translating this concept to IA, a basic definition is that primary prevention includes the steps necessary to delay or arrest a negative IT event from occurring in system, such as installing anti-virus software. Secondary prevention in health focuses on infected persons to reduce the pathogen transmission. Secondary prevention in IA would address the negatively impacted systems to stop cascading effects of an attack across multiple systems, such as removing infected machines from a network until the virus is quarantined or removed. Tertiary prevention focuses on infected persons and attempts to halt or reduce disease progression. IA tertiary prevention would work within impacted systems to restore full access and operations, such as stopping a network collapse through addressing individual PC issues and restoring the individual PCs for overall health.

Table 2. Interpreting Primary, Secondary and Tertiary Prevention from Public Health to Information Assurance.

	Public Health	**Information Assurance**
Prevention	The methods used to reduce the occurrence, severity and/or probability of a negative health event on an individual or a population	The methods used to reduce the occurrence, severity and/or probability of a successful cyber attack on a predefined unit (entity or geographic).
Primary Prevention	Attempt to reduce the probability of a negative health event to as close to zero as possible for as long as possible.	Attempt to reduce the probability of a successful cyber attack to as close to zero as possible for as long as possible.
Example of Primary Prevention	Providing clean injection equipment for injecting drug users.	Provide free anti-virus software to PCs on a system network.
Secondary Prevention	Focuses treatment or other methods on currently infected persons to reduce the pathogen	Intervention on compromised machines to reduce the possibility of cascading or

	transmission.	amplified effects.
Example of Secondary Prevention	Educate infected persons on how to reduce transmission risk to other persons.	Stop cascading effects of an attack across multiple systems by removing infected machines from a network until the virus is quarantined or removed.
Tertiary Prevention	Tertiary prevention focuses on infected persons and attempts to halt or reduce disease progression.	Work within impacted systems to restore full access and operations, such as stopping a network collapse through addressing individual PC issues and restoring the individual PCs for overall health.
Example of Tertiary Prevention	Provide treatment for the blood borne pathogen to the infected person.	Work within impacted systems to restore full operations of individual machines and thereby the overall system.

VI. Estimating Risk

Risk estimation is a priority in many fields including health, information systems, environmental sciences and financial industries. Understanding risk allows the identification of factors and the development of interventions to reduce vulnerabilities. Additional benefits to understanding risk include performance improvement; enhanced security and understanding the impact of negative events on system participants. In health, estimating the risk of the outcome of disease allows public health providers to design interventions to mitigate the risks for persons who are potentially susceptible.

Multiple factors are related to risk and its estimation, therefore a statistical model needs to be able to include several potential factors and determine one risk statistic.

VII. Multivariable Logistic Regression Models

In public health, multivariable models are used to calculate an estimated risk for an outcome. These methods are being exported to other fields for prediction, such as the US Geological Service predicting wild fire debris fields and in the prediction of mobile malware attacks and defense in IT. [8, 9] Multivariable logistic regressions (MLR) calculate inferential statistics by determining the effects of a set of predictor variables on a dichotomous outcome. [10] In disease modeling, MLRs are selected because of the dichotomy of the outcome and the ability to include variables with different data types. The common outcome levels are non-diseased, assigned a value of 0; and diseased, assigned a value of 1. The model calculates the estimated risk for the outcome of diseased.

A primary strength of the MLR is flexibility. These models allow variables with different types of data and provide a measurement of each on the impact of the population at risk. The data types in the predictor variables may be dichotomous, nominal, ordinal, interval or ratio.

Predictive statistical models estimate risk for the population at risk by measuring the influence of each variable in the presence of the each other variables. [11] The output of the model measures the effect of each variable in the environment of all variables. Interactions between the variables can also be included. The model is then compared to the data through the use of a goodness of fit test statistic. This analysis indicates the level of variance in the data accounted for by the model. Better models explain more variance leaving less to chance. [10] The full model includes all variables of interest, based on scientific criteria. Removal of statistically non-significant variables is commonly completed in a stepwise progression creating a series of reduced models. The level of statistical significance, or alpha (α), is determined a priori to the investigation. Simply, alpha refers to the number of times out of 100 that an investigator accepts as the probability of rejecting a true statement even though it is true or in the context of modeling, that a variable may be retained in the model when it should be removed.

At each step of variable removal, the reduced model is compared to the full model using a -2 log Likelihood test. This test compares the reduced model's 'fit' to the data to determine if it fits at least as well as the full model. The goal is to identify the simplest model with

the best fit. The law of parsimony states that the simplest model, meaning the model with the fewest variables, which explains the most variance, is the best model.

VIII. Conclusions

Development and validation of an operational method that identifies the predicted risk introduced into the overall organizational risk would accomplish several goals. As this modeling technique could be employed on various levels, such as an organization, state or national level, the overall costs savings could be extraordinary. The development of a Cyber Security Surveillance System (CS^3) framework that pulls data from multiple and varied sources would allow the estimation and prediction of risk in several domains. Based on the predicted level of risk, the organization could assign resources in a more efficient manner.

Predictive risk would also allow for the quantification of prevention interventions and allow decision makers to see which aspect of security has the least risk, therefore presenting a low threat – and those aspects with the highest risk which present the highest threat. Tailored interventions would save time, increase system security and allow a better commitment of resources. Similar models are used the world over in many areas of health and defense. Cyber security with its basis in information assurance a field with tremendous data resources should be an earlier adopted of this method.

Acknowledgement

Special appreciation goes to James McCoy and Aaron Chestnut for assistance with this manuscript.

References

[1] Major, J.A. *Advanced Techniques for Modeling Terrorism Risk*. 2002; Available from: http://www.gravitascapital.com/Research/Risk/Risk%20Measurement%20and%20Modelling/Modelling%20Terrorism%20Risk%20feb03.pdf.

[2] Marks, P., *Dot-dash-diss: The gentleman hacker's 1903 lulz*, in *NewScientist*2011.

[3] Parker, R.D., *Increasing Security through Public Health: A Practical Model.* J Spec Oper Med, 2011. **11**(4): p. 4-8.

[4] *Guidelines for Second Generation HIV Surveillance*, 2000, World Health Organization and Joint United Nations Programme on HIV/AIDS: Geneva.

[5] *Guidelines on surveillance among populations most at risk for HIV*, 2011, UNAIDS / World Health Organization: Geneva.

[6] Meynard, J.-B., et al., *Proposal of a framework for evaluating military surveillance systems for early detection of outbreaks on duty areas.* BMC Public Health, 2008. **8**(1): p. 146.

[7] Labonte, R. and M. Gagnon, *Framing health and foreign policy: lessons for global health diplomacy.* Globalization and Health, 2010. **6**(1): p. 14.

[8] Rupert, M.G., Cannon, S.H., Gartner, J.E., Michael, J.A., and Helsel, D.R., *Using logistic regression to predict the probability of debris flows in areas burned by wildfires, southern California, 2003–2006.* , U.S.G. Survey, Editor 2008.

[9] Dunham, K., *Mobile Malware Attacks and Defense, 1st Edition*2008: Syngress Publishing.

[10] Freedman, D.A., *Statistical Models: Theory and Practice*2005, New York: Cambridge University Press.

[11] Parker, R. and K. Rüütel, *A Surveillance Report of HIV Status and High Risk Behaviors Among Rapid Testing Participants in Tallinn, Estonia.* AIDS and Behavior, 2011. **15**(4): p. 761-766.

Author Biographies

R. David Parker, PhD is a research assistant professor of medicine at the University of South Carolina, Department of Medicine (US). He holds a doctor of philosophy in epidemiology from the University of South Carolina in Columbia, SC, US (2008), a master of arts in sociology and a bachelor of science in psychology both from Georgia Southern University in Statesboro, Georgia, US (2001, 1994). He is currently working on a graduate certificate in Information Assurance at the University of South Carolina, in Columbia, South Carolina, US.

Csilla Farkas, PhD is an associate professor at the University of South Carolina, Department of Computer Science and Engineering. She holds a doctor of philosophy in information technology and a master of science in computer science both from George Mason University in Fairfax, Virginia, US (2000, 1993). Additional degrees include of bachelor of science in computer science from the Institution of Computer Science in Budapest, Hungary (1989) and a bachelor of science in geology from Eötvös Loránd University also in Budapest, Hungary.