
Emergent Research Forum (ERF)

Martin (Dae Youp) Kang
University of Memphis
dkang3@memphis.edu

Sungyong Um
National University of Memphis
umsy@comp.nus.edu.sg

Ted (Euntae) Lee
University of Memphis
elee@memphis.edu

Horim Kim
Hallym University
xhr011027@hallym.ac.kr

Abstract

Benchmarking methodology can provide organizations with a way of choosing an appropriate ISSP by referring to others in the industry. However, selecting a proper benchmarking target in establishing of information system security policy is a challenge. This paper proposes an artifact to select a target organization by quantitatively measuring the similarity of organizations’ systems. Our proposed artifact includes required formulas and an algorithm which compute the similarities. The artifact is based on moment generating function with probabilistic distribution through moment method and generalized method of moment. We expect the artifact may resolve the challenge of selecting a target organization by increasing the accuracy of similarity identification.

Keywords

Information security, information security policy, benchmarking, data analytics, design science.

Introduction

Information Security (InfoSec) breaches and risk can lead to severe damage in an organization’s system and processes, with negative influences on organizational performance such as financial losses and the loss of organizations’ trustworthiness (Hovav and Gray 2014). Establishing an appropriate Information Security Systems Policy (ISSP) is a critical issue for organizations to prevent breaches and manage risk effectively. Benchmarking methodology can provide organizations with a way to choose an appropriate ISSP by referring to others in the industry while offering an ISSP that is generally accepted in the industry (Whitman and Mattord 2013). However, organizations do not have a robust, standardized way of finding a target organization with which to quantitatively measure and compare their diverse security issues. As such, they lose their chances to update ISSPs effectively in light of continuously changing security threats. This paper suggests an effective way to select a target organization by quantitatively measuring the similarity of organizations’ systems.

Common ways to select benchmarking targets are to use technical and non-technical measures to estimate the appropriateness of benchmarking. Technical measures include the frequency of virus checks and the number of detected virus types extracted from the log-file of a server computer. Using these measures, organizations can compare the average, frequency, and rate of attacks in terms of the similarity of their system with others. Non-technical measures include legal concerns and policy design (Knapp et al. 2009), education (Tsohou et al. 2015), organizational culture (e.g., employee, manager and users (D’Arcy et al. 2014) and the costs and benefits of compliance (Vance and Siponen 2012). Organizations can use these to complement the technical measures to estimate the simultaneous impact on the effectiveness of adopting ISSP.
However, the technical and non-technical measures may have limited empirical robustness because of the non-linear nature of InfoSec, which seriously violates the assumption of estimations. For instance, suppose two different firms in the IT industry produce similar software products. Firm A uses a cloud service and its asset is frequently managed by a cloud server, while Firm B manages the data on individual computers. The virus infection rates, as determined by the frequency of virus check of both firms, are similar in terms of variance and average. However, the number of detected viruses in Firm A is higher when the cloud service is highly active but distributed rather uniformly in Firm B. In this case, Firm A and Firm B should not pursue benchmarking due to the differences in their technological environments. To prevent such mismatches, there is a compelling need for an advanced benchmarking methodology.

To address the compelling need, we propose a new benchmarking methodology based on a design science approach. The methodology quantitatively and robustly compares the ISSP of each organization by reflecting the nonlinearity issue. In particular, we use the theorem of the moment that uniquely generates a measure to identify the similarity of ISSP across various organizations. The methodology is based on the algorithm created by multiple statistical approaches such as the generalized method of moment, exponential moving average, and detection of inflection point. This paper offers a step-by-step process to capture the impact of ISSP implementation on organizational performance.

**Design Science Approach**

This research follows a design science paradigm proposed by Peffers et al. (2007): problem identification, objectives of a solution, design & development, demonstration, evaluation, and communication. Following Peffers et al. (2007), we first discuss the problems and solutions for benchmarking methodology and the required steps to implement the artifact for benchmarking. Then, we propose an evaluation method to test the artifact, which will be part of our future research. Communication takes place in academic conference proceedings.

**Problem Identification**

Selecting a proper benchmarking target is the essential challenge in the benchmarking process. The organization should first identify other organizations with a similar environment to pick an appropriate target. We assume that if organizations share similar system environments such as their goals of InfoSec management, InfoSec technology, and resources, then their usages in information security management systems (ISMS) will be similar. For instance, suppose that there is an ISSP measurement for a firewall (an example of ISMS) which measures the performance of firewall policy. The ISSP measurement is a random variable which is the number of detected viruses corresponded to the strictness of firewall policy. If two different organizations share similar probabilistic distributions of the ISSP measurement, then they may be considered to have the same environment of virus attacks in terms of technology, organizational culture and required system resources.

However, the probabilistic distribution patterns of ISSP measurements can be nonlinear. The patterns are not consistent across diverse samples of organizations due to the impacts of internal and external factors on the InfoSec system. For instance, the public disclosure of system vulnerability motivates hackers to attack in diverse ways (Arora et al. 2006), and InfoSec damage and fixing efforts are different depending on the maturity of users and information systems due to InfoSec investment (Nazareth and Choi 2015). The literature offers empirical evidence that any ISSP measurements may fluctuate depending on the organization’s environment. This means that the ISSP measurements may have multiple probabilistic distributions in the sample. To detect this multiple probabilistic distributions, we introduce a moving window perspective. We use a moving window to identify the multiple moments attribute to multiple probabilistic distributions in sample data.

**Design and Development**

We systematically create an artifact to address the problems in the previous subsection. First, we define the required formulas to identify probabilistic distributions. In particular, the equations compute the similarities of organizations using a moment generating function (MGF). Then, we provide an implementation strategy based on the algorithm approach (Algorithm 1).

The distribution of a random variable (i.e., ISSP measurement) is characterized in terms of MGF. The probabilistic distribution is uniquely determined by its MGF. To get MGF, we define the required terms...
(Definition 1). The expected value of \( E(x^n) \), where \( x \) is a random continuous variable\(^1\) to the positive number \( n \), is determined by \( f(x) \) probabilistic density function (PDF).

\[
E(x^n) = \int_{-\infty}^{\infty} x^n f(x)dx
\]

**Definition 1**

We can derive the MGF \( M_x(t) \) based on definition 1:

\[
M_x(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx
\]

**Definition 2**

Based on the theorem of “MGF uniquely determines probability distributions \(^2\)”, the similarities of organizations are determined according to the similarities of moment. To elicit the formula, let \( s = e^t \) and \( c_i \) be the subtraction functions of \( fX(i) \) and \( fY(i) \) where \( i \) increases from zero to \( n \):

\[
\int_{-\infty}^{\infty} e^{tx} \cdot fX(x)dx = \int_{-\infty}^{\infty} e^{ty} \cdot fY(y)dx \rightarrow \int_{-\infty}^{\infty} s^x[fX(x) - fY(x)]dx = 0
\]

**Definition 3**

Realistically, it is difficult to derive moment in samples. To derive the moment, we adopt the estimation of moment displays indivisible and inconsistent probability. As such, we employ the generalized method of moment (GMM) to estimate the moment in a consistent way and find the patterns of PDF in the sample space. Let \( \theta \) be the parameter of the sample space and \( E[f(x, \theta)] \) be a set of population moments (i.e., subset of sample space). \( f_n(\theta) \) is the corresponding sample counterpart. The criterion function is \( Q_n(\theta) = f_n(\theta)'W_n f_n(\theta) \) where \( W_n \) is the weighting matrix that depends on the strictness of ISSP (i.e., vector of sample space). Then, the GMM estimator is given by:

\[
\hat{\theta} = \arg\min Q_n(\theta)
\]

**Definition 4.1**

If the sample sets do not satisfy the conditions of the GMM, the artifact employs a method of moment (MM) estimator instead.

\[
f_n(\theta) = \frac{\sum_{i=1}^{n} f(x_i, \theta)}{n}
\]

**Definition 4.2**

To detect the multiple probabilistic distributions, we suggest to use moving window. Our proposed artifact divides the probabilistic distributions in a sample by moving window which consists of the measurement interval between \( p \) and \( q \) in sample space. We propose to use the exponential moving average (EMA) to get the moving averages of the probabilistic distribution for each moving window. EMA may reduce the possible errors (e.g., outlier) in the moving window by standardizing the values. EMA is computed from the each window in the sample, where \( \alpha \) is the coefficient degree of weighting decrease over the set of previous windows:

\[
EMA_p = \begin{cases} 
\alpha \cdot x_p + (1 - \alpha) \cdot EMA_q, & p = 1 \\
\alpha \cdot x_p + (1 - \alpha) \cdot EMA_q, & p > 1 
\end{cases}
\]

**Definition 5**

Finally, we propose a formula to merge two similar samples of ISSP:

\[
M_{x+y}(t) = M_x(t) \cdot M_y(t)
\]

**Definition 6**

Based on these definitions, we propose the following algorithm. First, the algorithm elicits the process to set the required variables to find MGF in a sample. Second, the algorithm proposes a generalized method to seek multiple MGFs, which attribute to multiple probabilistic distributions, in a sample. Then, the algorithm suggests an implementation strategy to compare the different samples. The algorithm also includes a strategy to merge similar samples as the outcome of benchmarking.

---

\(^1\) In this paper, the artifact only describes a continuous perspective for the purpose of simplicity. However, the artifact can be applied to discrete variables using the expected value of \( E(x^n) \) determined by \( \sum_{x} x^n f(x) \) with the same notation in the body of the text.

\(^2\) We use the proof of the theorem in Chung (2001).
Step | Action | Highlights of implementation
--- | --- | ---
1 | Create a pair set in a progression which consists of key and value. The key is the strictness of ISSP (i.e., vector value of sample space) and the value is the domain of probability density function. | The progression can be implemented using a hashmap.
2 | Get MGF of the progression based on Definition 2. | The scope MGF (i.e., interval) is the entire dataset. 
3 | Identify the distribution of the sample (i.e., paired value in progression). Returns true value when it is identifiable and false when it is vice versa. | The identification can be manually computed based on the relevant theorems in a specific distribution. In this case, the outlier should be controlled. For other options, use computational testing such as Anderson-Darling and Shapiro-Wilk (Razali and Wah 2011). True means the sample has one identical distribution over the sample, meaning the algorithm is ready to compare with other companies.
4 | If step 3 returns false, check whether there are multiple points of inflection in the sample distribution. | We suggest using simple linear regression to find the knee\(^3\) in the distributions of the sample based on Salvador and Chan (2004). The inflection is detected when the square error sums (SES) in the boundary between the pair of straight lines in PDF most closely fit the curve. If the square error sums drastically change (e.g., greater than absolute 0.45) then the inflection is detected in the window. \(S_p\) is the SES of the moving window which has initial points of \(p\) to \(q\). The formula is \(S_p = \sum_{i=p}^{q} (x_i-\mu)^2\). The moving window points, \(p\) and \(q\), are the freedom to control the SES because of the size of the window. If the \(S_p\) satisfies certain thresholds, then it returns true and false if vice versa.
5 | If step 4 returns true, set the variables for the moving window that is changing (i.e., moving forward) over the SES searching process. If step 4 returns false, the sample dataset is limited to extract moment. Terminate the algorithm. | Assign a set of moving window variables. Each element in the set (i.e., each moving window) has an initial point and end point. \(p_\text{end}\) is the notation for the initial point and \(q_\text{end}\) is the end point in the \(n^\text{th}\) moving window.
6 | Repeat steps 1 to 5 to derive MGF of each moving window by changing its variables over the iterations. Save the information of MGF by creating a new set of progressions for each moving window. Compare the MGF of one to the other samples and derive similarity. | To implement MGF, use definition 4.1 or 4.2 to estimate momentum. Exercise definition 3 for the comparison of samples. In detail, divide the value in the window by the PDF of each sample and take the natural logarithm. If the value is similar to a certain threshold (e.g., 95%), the ISSPs of the two organizations are similar in a certain point of the moving window (step 4) or the entire sample set (step 2).
7 | If similarity is detected, merge the variables using MGF and report the suggestion. | Implement Definition 6. Let two samples, \(x\) and \(y\), follow the normal distribution. The average of \(x\) is 2 and \(y\) is 0. The standard deviations are 3 and 4, respectively. The MGF of each sample \(M_x(t) = e^{zt}\) and \(M_y(t) = e^{zt}\). Thus, the merged MGF is \(M_{xy}(t) = e^{zt}\). The average is 2 and the standard deviation is 5 according to the merged MGF. Inverse the merged MGF to get a specific benchmarking value.

**Algorithm 1**

**Demonstration and Evaluation**

For the step of demonstration, we are in the process of developing a prototype program in the Java platform using ND4J (N-Dimensional Array for Java) to implement the proposed algorithm. Java provides a fast iteration process for hashing data, and ND4J delivers a complex convolutional process with an efficient programming method. The prototype includes a visualization tool that shows the moving window and curves graphs to compare similarities for each frame of the moving window.

Based on the prototype, we will conduct a two-stage evaluation. First, we will evaluate variable validity (i.e., the validity of the similarity). The proposed definitions are theoretically rigorous based on the literature and mathematical constructs that have been tested in academia. However, the applicability of similarities to benchmarking ISSP regarding the efficiency, similarity detection performance, and generalizability has not been validated in the previous studies. To secure the validity of this variable, we conduct cross-validation methods using other similarities methods such as vague soft set and fuzzy set, which compute the similarities using entropies (Wang and Qu 2013). Additional implementation strategies for other methods will be provided in our future research.

---

\(^3\) The point of maximum curvature of a magnetization curve where saturation occurs.
Second, we will conduct an empirical study using the proposed artifact with log-data of InfoSec systems of multiple organizations. The main goal is to derive practical implications by demonstrating the applicability of the artifact in real benchmarking practice. The case study will also provide academic implications through the analysis of various ISSPs in heterogeneous groups of organizations in the industry where the organizations reside.

Conclusion

Effective ISSP is vital in an organization. The benchmarking of ISSP is a methodology that fulfills the needs of the business. Organizations may achieve their goals in benchmarking ISSP by using the artifact proposed in this research. The artifact can be used in academia to develop research ideas through generating the patterns of ISSP in an industry. For instance, research may focus on the dynamics of heterogeneous organizations that share similar characteristics of ISSP (e.g., ISSP related to availability, confidentiality, and accessibility). Then, a research argument regarding the relationships of the shared characteristics and the economic perspective of organizations can be established. These types of post analysis and research will be continued in our future studies.

REFERENCES


Wang, C., and Qu, A. 2013. "Entropy, Similarity Measure and Distance Measure of Vague Soft Sets and Their Relations,” Information Sciences (244), pp. 92-106.