Estimating Probability of a Cyber-Security Breach
Meghan Anthony, Maria Ishmael, Erik Santa, Arkady Shemyakin, Gary Stanull and Natalie Vandeweghe
University of St. Thomas

Introduction

Information technology is the engine that drives the U.S. economy, giving it a competitive advantage in global markets by providing better services and facilitating greater productivity. Unfortunately, information systems are subject to a variety of threats and the threat landscape is constantly changing. To determine the risk these threats pose we need to evaluate the likelihood of their success in exploiting known and unknown vulnerabilities. This involves an accurate assessment of both impact and probability of breaches.

The purpose of this paper is to define a predictive model, based on known system attributes, for assessing risk associated with information systems. The goal is to provide decision-makers with the best possible information about the probability of a security breach so they can make informed decisions on how to best address the risk.

Most statistical papers dealing with cyber-security are dedicated to analysis of the breach data and development of distribution models for frequency of the breaches and severity of the associated losses. We will deal with a different problem: How to estimate probability of a breach for a specific database application. This estimate would make it possible to assign “risk ratings” and help to identify the most vulnerable applications based on some of their observable characteristics. However, obtaining such estimates will require us to build distribution models of these characteristics not only for “broken” applications, but also for “unbroken” ones.

To illustrate this point, let us consider an attribute $A$ of an application (such as size, type of data, or the industry). Using the Bayes formula let us assess the probability of an application with this attribute to be broken:

$$P(B | A) = \frac{P(B)P(A | B)}{P(B)P(A | B) + P(U)P(A | U)}, \quad (1)$$

where $P(B)$ is overall prior probability of an application to be broken, $P(U) = 1 - P(B)$; while $P(A | B)$ and $P(A | U)$ are the probabilities to observe attribute $A$ among broken and unbroken applications respectively. The latter two can be estimated from the historical data when such data are available. This estimation requires certain knowledge regarding the population of unbroken applications, which is not characteristic for a typical analysis of the breach data.
Probability of Breach Given Application Size

For an illustration let us concentrate on health-related data and single out such application attribute as record size measured as its total number of records. We will be considering the following grouping of application record sizes:

- $S_1$: Below 10,000
- $S_2$: 10,000 – 30,000
- $S_3$: 30,000 – 100,000
- $S_4$: 100,000 – 1 million
- $S_5$: Above 1 million

Suppose that independently of the application size, the probability of that to be breached (or broken) is estimated as $P(B)$, the prior probability of a breach. Then the posterior probability of a breach conditional on the size of the class $S_k$, where $k = 1, 2, 3, 4, 5$, can be evaluated as

$$P(B / S_k) = \frac{P(B)P(S_k / B)}{P(B)P(S_k / B) + (1 - P(B))P(S_k / U)},$$

where $P(S_k / B)$ and $P(S_k / U)$ will be estimated from two different parametric distribution models for the random size $X_B$ of a broken application $F_B(x) = P(X_B \leq x)$ and the size $X_U$ of an unbroken application $F_U(x) = P(X_U \leq x)$:

$$P(S_k / B) = \int_{s_k} x dF_B(x), P(S_k / U) = \int_{s_k} x dF_U(x).$$

Distribution Models for Application Size

For the data on breaches we used 1,572 data points recorded in the HHS database [2]. For the data on the unbroken applications we use a sample size of 81 obtained from industry experience of application analysis from 2014 to 2015.

Application size is measured as the total number of records $R$ log-transformed and shifted with natural thresholds 500 (minimum record size for broken data) and 10 (minimum record size for unbroken data) so that $X_B = \ln(R) - \ln(499)$ and $X_U = \ln(R) - \ln(9)$.

For variables $X_B$ and $X_U$ three two-parameter distribution models are considered: normal (Gaussian), suggested in [1]; gamma; and Weibull. Maximum likelihood estimates (MLE) were obtained for all three models. In Figure 1 and Figure 2 the boundary of the shaded area
corresponds to empirical CDF, the best normal fits are depicted by dashed lines, fitted gamma in red and fitted Weibull in green.

Figure 1. Distribution \( F_B(x) = P(X_B \leq x) \).

Figure 2. Distribution \( F_U(x) = P(X_U \leq x) \).

Model selection using Akaike information criterion (AIC) brings about the results in Table 1.

<table>
<thead>
<tr>
<th>Model</th>
<th>Normal</th>
<th>Gamma</th>
<th>Weibull</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_B )</td>
<td>5964.4</td>
<td>5129.9</td>
<td>5117.2</td>
</tr>
<tr>
<td>( X_U )</td>
<td>435.5</td>
<td>479.3</td>
<td>447.6</td>
</tr>
</tbody>
</table>

Table 1. AIC Values (the lower the better).
Based on AIC values from Table 1 and graphs of $F_B(x)$ in Figure 1 and $F_U(x)$ in Figure 2, we recommend the choice of Weibull distribution for both cases: $F_M(x) = 1 - e^{-(x/\theta_M)^M}$ with scale $\theta_M$ and shape $\tau_M$ estimated separately for models $M = B, U$ via MLE:

$$\tau_B = 1.089, \theta_B = 1.937; \tau_U = 3.156, \theta_U = 11.036.$$ 

**Bayes Factors**

Based on the results from the previous section, we can estimate $P(S_k/B)$ and $P(S_k/U)$. Obtaining posterior probabilities of a breach directly from (2) would require an additional specification of the prior probability $P(B)$, generally unknown. Therefore we will use Bayes factors defined for each $k = 1, \ldots, 5$ as

$$BF_k = \frac{P(S_k/B)}{P(S_k/U)}.$$ 

It follows from (2) that for sufficiently small $P(B)$ the Bayes factors approximate the ratios of probabilities $BF_k \approx \frac{P(B/S_k)}{P(B)}$ and thus represent appropriate “adjustments” to prior probabilities of breach, which take into account the application size. The results are summarized in Table 2.

<table>
<thead>
<tr>
<th>$S_k$</th>
<th>$P(S_k/B)$</th>
<th>$P(S_k/U)$</th>
<th>$BF_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq 10,000$</td>
<td>0.800</td>
<td>0.213</td>
<td>3.756</td>
</tr>
<tr>
<td>$10,000 \leq 30,000$</td>
<td>0.096</td>
<td>0.102</td>
<td>0.941</td>
</tr>
<tr>
<td>$30,000 \leq 100,000$</td>
<td>0.054</td>
<td>0.128</td>
<td>0.422</td>
</tr>
<tr>
<td>$100,000 \leq 1,000,000$</td>
<td>0.038</td>
<td>0.248</td>
<td>0.153</td>
</tr>
<tr>
<td>$\geq 1,000,000$</td>
<td>0.011</td>
<td>0.308</td>
<td>0.036</td>
</tr>
</tbody>
</table>

**Table 2. Conditional Probabilities and Bayes Factors.**

These results can be interpreted as follows: when we start with a small prior probability of a breach $P(B)$, additional knowledge that the application size belongs to a certain class $S_k$, makes
it possible to estimate the posterior probability as \( P(B / S_k) \approx P(B) \times BF_k \). According to the first and the last rows of Table 2, small size of an application makes it more likely to be broken, while very large size makes it much less vulnerable. This may be characteristic for the breaches of healthcare applications being often caused by thefts of laptops and storage devices, which are unlikely to contain large size applications.

**Conclusions:**

- Weibull distribution provides the best overall fit for \( B \) and \( U \)
  - Data points in the tail of the dataset (large record sizes) do not deviate from the best-fit trend-line as was noticed in [1] for Gaussian model.
- Bayes factors can be used to evaluate posteriors for small prior probabilities \( P(B) \).
- Probability of data breach can be effectively adjusted based on the record size.

**Next Steps:**

- Perform similar analysis and calculate Bayes factors based on the record size for different breach types (e.g., theft of a laptop versus a large-scale intentional hacking).
- Identify entity attributes such as revenue that correlate with the record size.
  - Use these attributes as proxies to model distribution of record sizes for “unbroken” applications.
- Analyze the Privacy Rights Clearinghouse dataset [6] to expand the scope of analysis to the industries beyond healthcare.
- Estimate risk to information systems based on the record size.

**References**


