

# Multivariate Statistical Profiling for Anomaly Detection

Nong Ye, Qiang Chen, Syed Masum Emran, and Sean Vilbert  
Arizona State University

## Purpose

In this paper we present a multivariate statistical technique based on Hotelling's  $T^2$  to detect coordinated actions of intrusions.

## Method

In our study, we consider the 284 event types in the UNIX systems as variables and apply Hotelling's  $T^2$  to analyze them. Each activity on a host machine is logged in an audit file and is characterized by a correspondent event type. We study the characteristics of each event type and the relationship between the different event types by comparing the recent-past behaviors to the norm profile. We define the recent-past probability from time  $t-k$  to  $t$  for event type  $i$  based on the Exponentially Weighted Moving Average (EWMA) technique. It is given by:

$X_i(t) = I \times 1 + (1 - I) \times X_i(t - 1)$  if the observed event at time  $t$  falls into the  $i$ th event type

$X_i(t) = I \times 0 + (1 - I) \times X_i(t - 1)$  if the observed event at time  $t$  does not fall into the  $i$ th event type.

In the training, we obtain the sample mean  $\bar{X}$  and sample covariance matrix  $S$  as the estimators to actual mean  $\mu$  and covariance matrix  $S$  for a long-term  $T^2$  profile of normal activities. Combined with other techniques, we build up a norm profile for the testing, a signal should be produced if an observed event deviates from the norm profile. Hotelling's  $T^2$  is defined as the generalized distance from a  $p$ -dimensional sample point  $X = (x_1, x_2, \dots, x_p)'$  to its sample mean  $\bar{X}$ . If we know the mean vector  $\mu$  and covariance matrix  $S$ , we compute the distribution of the  $T^2$  statistics for a  $p$ -dimensional observation vector  $X$  as follows:

$T^2 = (X - \mu)' \Sigma^{-1} (X - \mu) \sim \mathbf{c}_{(p)}^2$ ; where  $\mathbf{c}_{(p)}^2$  represents a central chi-square distribution.

When the true parameters are unknown,  $T^2$  is given by:  $T^2 = (X - \bar{X})' S^{-1} (X - \bar{X})$ ;

where  $S$  is an estimate of  $S$  for  $p$  variables and  $m$  observation:  $S = \frac{1}{(m-1)} \sum_{i=1}^m (X_i - \bar{X})(X_i - \bar{X})'$ .

The larger a computed value of  $T^2$  is, the more likely the observed activities deviate from the norm profile, and thus the more likely the observed activities come from intrusions.

## Results

We compute the false alarm rate and the detection rate of the Hotelling's  $T^2$  technique on the testing data. The results show a low false alarm rate and a high detection rate.

## New or Breakthrough Aspect of Work

We apply the EWMA technique to represent the observed behavior in the recent past changes in the profiles. Hotelling's  $T^2$  is used to account for the correlation of multiple variables for detecting coordinated actions of intrusions.

## Conclusions

The results of Hotelling's  $T^2$  shows its power in intrusion detection.