A logo of a company

Description automatically generatedFaculty of Engineering   
Cairo university   
Computer department

# Statistical anomaly-based intrusion detection systems

BY CMPlexity  
SUPERVISED BY PROF/MAHA AMIN

# Introduction:

In today’s digital landscape, enterprises are increasingly vulnerable to cyber threats, which have grown in sophistication and subtlety. Traditional signature-based Intrusion Detection Systems (IDS) often fail to detect novel or emerging attacks, as these systems rely on predefined patterns that do not adapt to new, unknown threats. Additionally, the high rate of false positives from conventional IDS creates a noisy environment that diverts critical attention from genuine threats.

# Background & Problem Statement:

On January 24, 2003, the [W32.SQLExp.Worm](http://securityresponse.symantec.com/avcenter/venc/data/w32.sqlexp.worm.html) (later named Slammer/Sapphire) was released into the wild. This worm exploited a stack-based buffer overflow vulnerability in Microsoft's SQL Server 2000 software (including MSDE 2000). While vulnerabilities affecting Microsoft products are nothing new, the speed at which this worm propagated was extremely novel - scary in fact. The worm was released and within ten minutes it compromised 90% of all vulnerable systems worldwide. Before this incident, worms of this type were merely theoretical, given serious consideration primarily in the academia. It takes even the fastest vendors hours or days to produce a signature for rule-based intrusion detection (RBID) systems. In the case of this worm however, a vulnerable network would be compromised in a matter of seconds, much too quickly for even the most diligently updated RBID system. So, what is the solution to a worm that doubles its infection rate every 8.5 seconds? This question is what our research aims to answer, resulting in a statistical-based (also referred to as behavior-based) anomaly detection.

Intrusion Detection the Statistical Way

Statistical-based systems (SBIDs) take a different approach to signature-based IDS. The concept of the SBID system is simple: it determines "normal" network activity and then all traffic that falls outside the scope of normal is flagged as anomalous (not normal). SBID systems attempt to learn network traffic patterns on a particular network. This process of traffic analysis continues as long as the SBID system is active, so, assuming network traffic patterns remain constant, the longer the system is on the network, the more accurate it becomes. By analyzing network traffic and processing the information with complex statistical algorithms, SBID systems look for anomalies in the established normal network traffic patterns. All packets are given an anomaly score (indicating the degree of irregularity for the specific event) and if the anomaly score is higher than a certain threshold, the IDS will generate an alert. The key to any SBID system is its ability to learn and distinguish normal from anomalous network activity. The model will flag low-probability events as potential intrusions by monitoring deviations in metrics such as mean, median, and standard deviation.

# Research Objectives:

* This project seeks to develop a statistics-based IDS that uses anomaly detection methods to identify deviations from normal network and host behavior patterns.
* Anomaly Detection: Implementing univariate, multivariate, and time-series models to detect deviations in real-time, focusing on abnormal event frequency, unusual packet flows, and atypical user activities.
* Evaluating Effectiveness: Assessing the IDS model’s accuracy in identifying genuine threats by measuring false positive and false negative rates on real-world data.

# Statistical Questions:

1. **How can descriptive statistics be used to establish baseline network traffic patterns for an IDS?**
   * We’ll explore how measures like mean, median, and variance can define "normal" behavior, creating a basis for detecting anomalies.
2. **What statistical thresholds are most effective for anomaly detection in network traffic?**
   * We’ll examine the effectiveness of thresholds like standard deviations and percentiles for flagging unusual events in network traffic data.
3. **How does Z-score analysis compare to percentile-based thresholds in detecting network intrusions?**
   * We’ll investigate the comparative strengths and weaknesses of Z-score and percentile thresholds for identifying deviations from normal traffic.
4. **Can moving average and exponential smoothing techniques effectively identify time-based anomalies in network traffic?**
   * We’ll assess these time-series methods for detecting sudden spikes or drops in traffic, which may indicate intrusions.
5. **How can statistical hypothesis testing, such as the Chi-square test, be applied to detect deviations in network traffic patterns?**
   * We’ll discuss using hypothesis tests to check if observed traffic patterns deviate significantly from expected distributions.
6. **What is the effectiveness of the t-Test and ANOVA in comparing network traffic features under normal vs. attack conditions?**
   * We’ll explore the role of these tests in detecting significant differences in traffic features between typical and suspicious periods.

# Methodology:

* Gaussian distribution: The Gaussian distribution, also known as the normal distribution, is frequently used in statistical-based anomaly detection. It assumes that the normal behavior of the data follows a bell-shaped curve. Anomalies are identified as instances that fall outside a specified range or threshold based on the estimated mean and standard deviation of the data. Instances that lie in the tails of the distribution, beyond a certain number of standard deviations from the mean, are considered anomalies.
* Mahalanobis distance: The Mahalanobis distance measures the distance between a data point and the center of a distribution, considering the correlation between variables. It accounts for the covariance structure of the data and is particularly useful when the variables are correlated. The Mahalanobis distance can be used to detect anomalies by comparing the distance of each data point to a threshold value. Points with a large Mahalanobis distance are considered anomalies.
* Z-score method: The Z-score method is a simple statistical technique for anomaly detection. It calculates the standard deviation from the mean for each data point and expresses it as a Z-score. The Z-score represents the number of standard deviations a data point is away from the mean. Anomalies are identified as data points with a Z-score exceeding a specified threshold. This method is particularly useful when the data is normally distributed.
* Hypothesis testing: Hypothesis testing is a statistical technique used to determine the likelihood that an observed deviation from the expected behavior is due to chance or represents an anomaly. Commonly used hypothesis tests include the t-test, chi-square test, or Kolmogorov-Smirnov test. These tests compare the observed data to a reference distribution or expected behavior and calculate a p-value. If the p-value is below a predefined significance level, the deviation is considered significant, and the instance is flagged as an anomaly.

# Contribution to Knowledge

* Empirical evidence on the comparative performance of Z-score analysis, percentile-based thresholds, and hypothesis testing approaches in differentiating malicious network activities from benign traffic. This knowledge can inform the design of next-generation IDS solutions.
* Evaluation of the effectiveness of various statistical techniques, including univariate, multivariate, and time-series models, in detecting deviations from normal network traffic patterns. This will provide insights into the most suitable anomaly detection methods for different types of intrusions.