

# Statistical Causality Analysis of Alert Data

Xinzhou Qin and Wenke Lee

Presented by Annie Chen

# Overview

- Introduce Information Security
- Overview of proposed solution
- Discussion on each stage of the process
- Experiments

# Information Security

- Protect computers from malicious attacks
- Mechanism includes authentication systems, firewalls, intrusion detection systems, antivirus... etc
- Lots and lots of data overwhelms security administrators

# Alert Analysis

- Reduce redundancy of alarms
- Intelligently integrate and correlate alerts
- Construct attack scenarios
- Present high-level aggregated information

# Contribution of this Paper

- Existing solutions use “signatures” of known attacks to perform pattern matching which cannot detect new types of attack
- Use time series causality analysis to construct attack scenario
- This approach can discover new patterns of alert relationships without depending on prior knowledge of attack scenarios

O  
V  
E  
R  
V  
I  
E  
W

**Raw Alert** (timestamp, source IP, dest IP, ports, user, process, attack class, sensor ID)



**Aggregate & cluster**

**Hyper Alert**



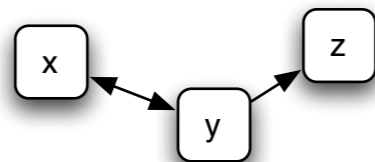
**Prioritize using Naive Bays**

**Target Alert**



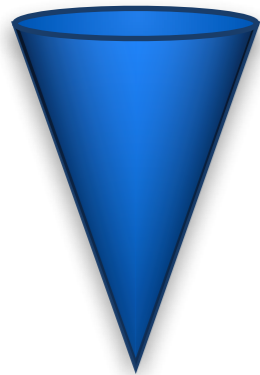
**Remove background alerts & apply pair-wise Granger Causality Test**

**Attack Scenario**



# Alert Aggregation & Clustering

**Raw Alert** (timestamp, source IP, dest IP, ports, user, process, attack class, sensor ID)

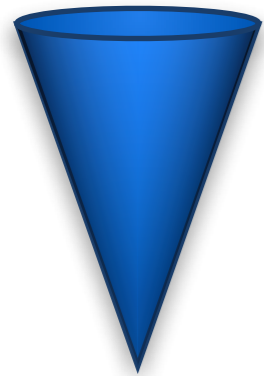


1. aggregate on everything except timestamp
2. fuse same alert from different sensors if timestamp close
3. cluster everything except timestamp

**Hyper Alert** - a time ordered sequence of alerts in the same cluster

# Alert Prioritization

Hyper Alert - a time ordered sequence of alerts in the same cluster

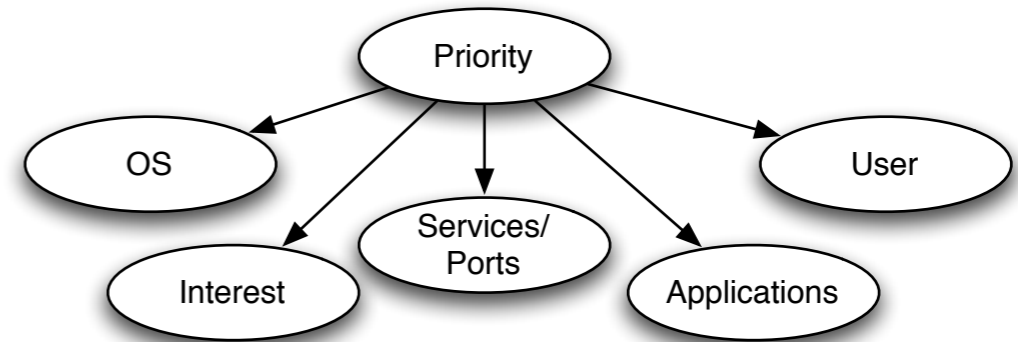


- Prioritize each hyper alert based on its relevance to mission goal using Naive Bayes
- Construct a priority computation model for Hyper Alert using predefined CPT from author. For variables OS, Interest, Services, Applications, User
- Using predefined CPT from author
- Match it against the configurations of the target networks and hosts

Target Alerts

# Alert Prioritization

Priority (High|Low) as dependent binary class variable  $H$ , conditional on variables  $e^i$



$$P(H_i | e^1, e^2, \dots, e^N) = \frac{P(H_i) P(e^1, \dots, e^N | H_i)}{P(e^1, \dots, e^N)}$$

Assume conditional independence of variables  $e^i$

$$P(H_i | e^1, e^2, \dots, e^N) = \gamma P(H_i) \prod_{k=1}^N P(e^k | H_i)$$

# Granger Causality

1. The cause occurs before the effect; and
2. The cause contains information about the effect that is unique, and is in no other variable

A consequence is that the causal variable can help forecast the effect variable

# Granger causality not real causality?

- Granger causality is more a definition of the *incremental predictability* between two time series variable
- If A Granger-causes B, we cannot say that controlling A we influence B

# Granger Causality Test (GCT)

Model variable  $y$  by two auto-regression models, namely for two time series variables  $y$  and  $x$  with size  $N$

$$y(k) = \sum_{i=1}^p \theta_i y(k-i) + e_0(k)$$

$$y(k) = \sum_{i=1}^p \alpha_i y(k-i) + \sum_{i=1}^p \beta_i x(k-i) + e_1(k)$$

where  $p$  is lag length

$\alpha_i, \beta_i, \theta_i$  are computed in the process of solving OLS which finds the parameters in order to have the minimum estimation error

# Granger Causality Index (GCI)

GCT compares the residuals of the two models

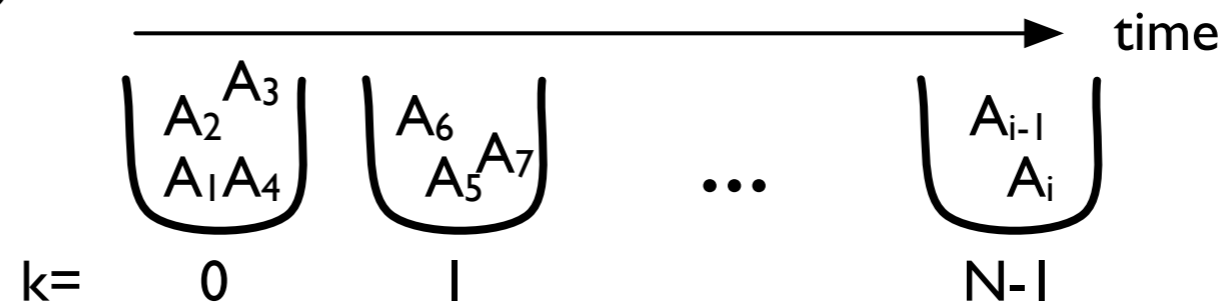
$$R_0 = \sum_{k=1}^T e_0^2(k) \quad R_1 = \sum_{k=1}^T e_1^2(k)$$

$$GCI = \frac{(R_0 - R_1)/p}{R_1/(T - 2p - 1)} \sim F(p, T - 2p - 1)$$

If GCI value larger than critical value in Fisher's F-test,  
then x Granger-causes y

# GCT Alert Correlation

1. Formulate each Hyper Alert into univariate time series

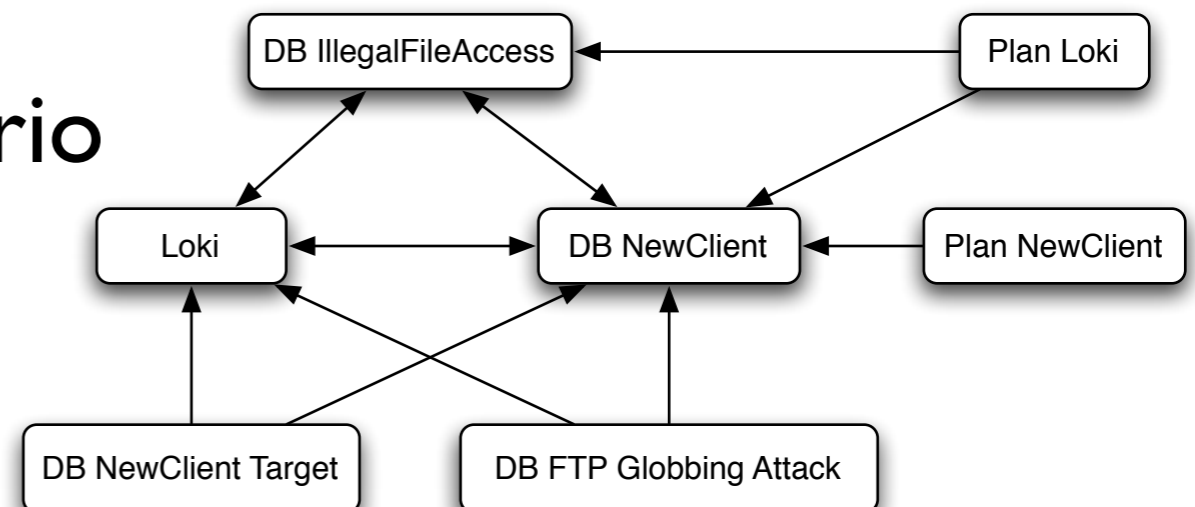


time series  $a(k)$  = size of bucket  $k$

2. Remove background Alerts using Ljung-Box test and expert/student

# GCT Alert Correlation

1. Apply GTC for pair-wise alert correlation for each target alert with all other alerts
2. Rank GCI of pairs that passed the F-test and select top  $m$  candidate alerts as being causally related to the target alert
3. Inspect/filter relations
4. Construct attack scenario



# Experiments

# Discussion