

# Peer-group Behaviour Analytics of Windows Authentications Events Using Hierarchical Bayesian Modelling

Iwona Hawryluk

Henrique Hoeltgebaum, Cole Sodja, Tyler Lalicker, Joshua Neil

# Introduction

- ◆ Majority of systems rely on signature- or rule-based methods
  - ◇ Vulnerable to zero-day attacks
  - ◇ Low precision □ high FPs
- ◆ UEBA for FP reduction using peer-groups
  - ◇ Challenge in how to create the clusters

Goal

Detect anomalous spikes in the  
**hourly numbers of unique target entities**

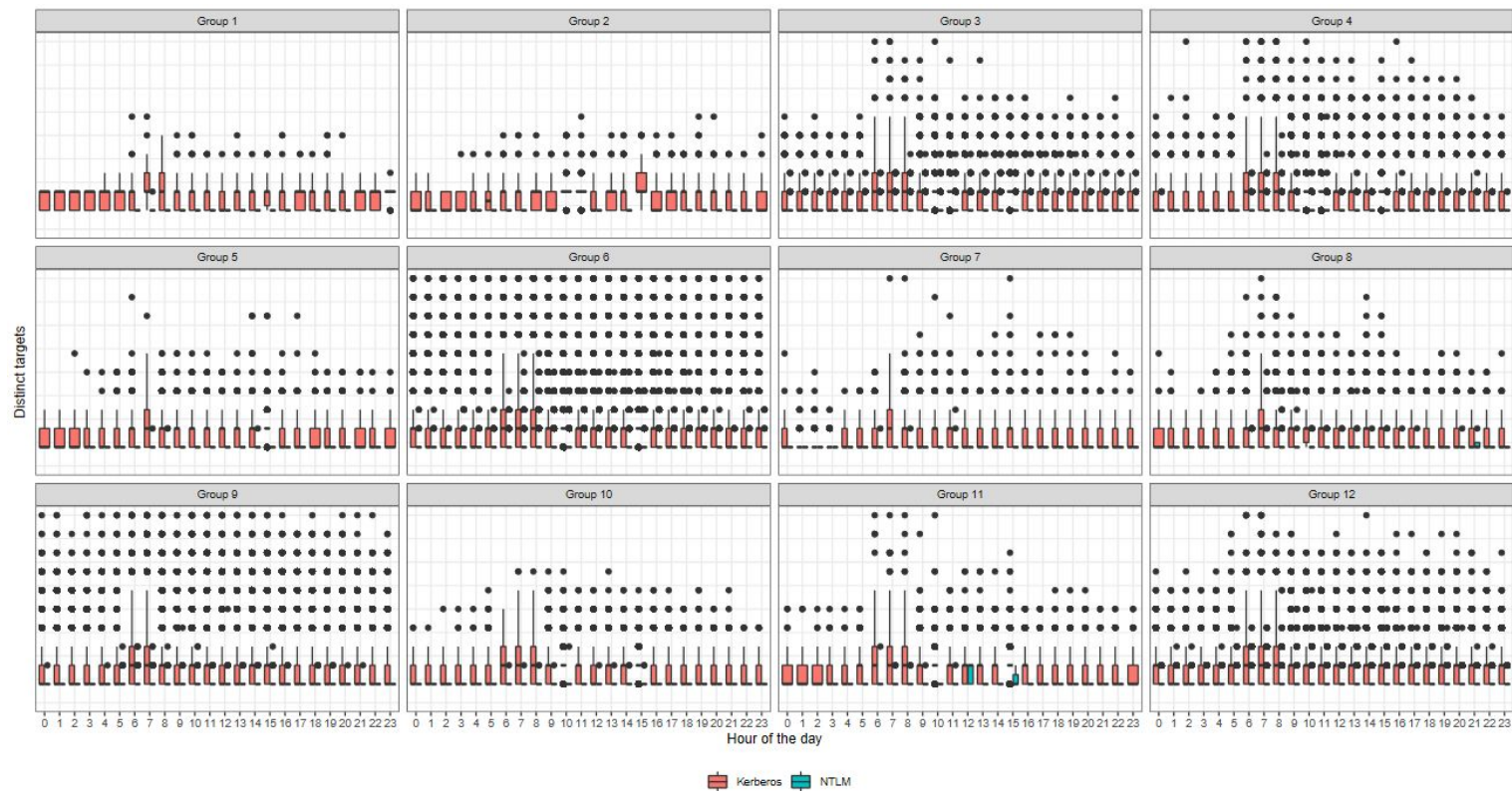
## Contributions

- ◆ Fully Bayesian hierarchical framework for conducting anomaly detection
  - ◇ 4 methods of clustering users based on their behaviour
  - ◇ 6 statistical models which we fit to authentications logs
  - ◇ Full Bayesian framework for conducting anomaly detection
  - ◇ Real-life data example

## Data

- ◆ Individual networked computers running the MS Windows
- ◆ Authentications from a single enterprise, single country
- ◆ Only successful, method 'Kerberos' or 'NTLM'
- ◆ No system accounts, no privileged accounts
- ◆ Joined with HR records (job title, division, etc)
- ◆ 27 days of data, aggregated at hourly level

# Data



# 1. Grouping

HR data

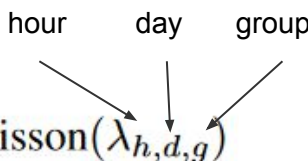
1. HR records
2. Time series – univariate ARIMA + Gaussian Mixture Model (GMM)
3. K-means – Singular Value Decomposition (SVD) + k-means
4. GMM – SVD + GMM
5. Spectral bi-cluster [1, 2]

Data-driven

## 2. Modelling

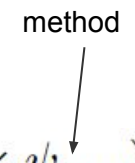
### 6 Bayesian models

#### Model 4

$$\begin{aligned} y_{h,d,g} &\sim \text{Poisson}(\lambda_{h,d,g}) \\ \log(\lambda_{h,d,g}) &\sim N(0, 5). \end{aligned}$$


A diagram for Model 4 showing three variables: 'hour', 'day', and 'group'. Arrows from each of these variables point to the  $\lambda_{h,d,g}$  parameter in the Poisson distribution equation above.

#### Model 6

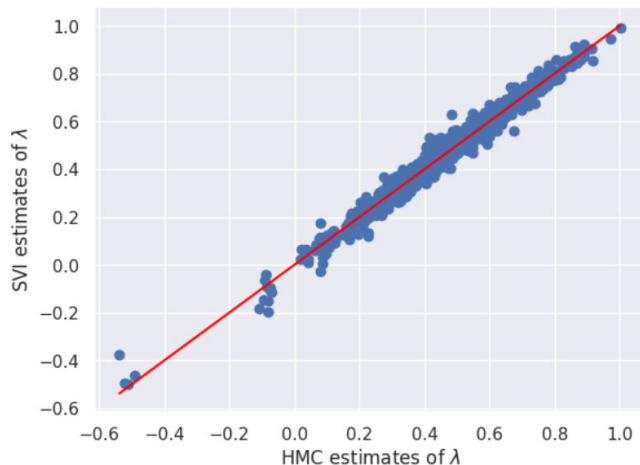
$$\begin{aligned} y_{h,d,g,m} &\sim \text{Poisson}(\lambda_{h,d,g} \times \psi_{m,g}) \\ \log(\lambda_{h,d,g}) &\sim N(0, 5) \\ \log(\psi_{m,g}) &\sim N(\mu_m, 5) \\ \mu_m &\sim N(0, 5) \end{aligned}$$


A diagram for Model 6 showing the variable 'method'. An arrow points from 'method' to the  $\psi_{m,g}$  parameter in the Poisson distribution equation above.



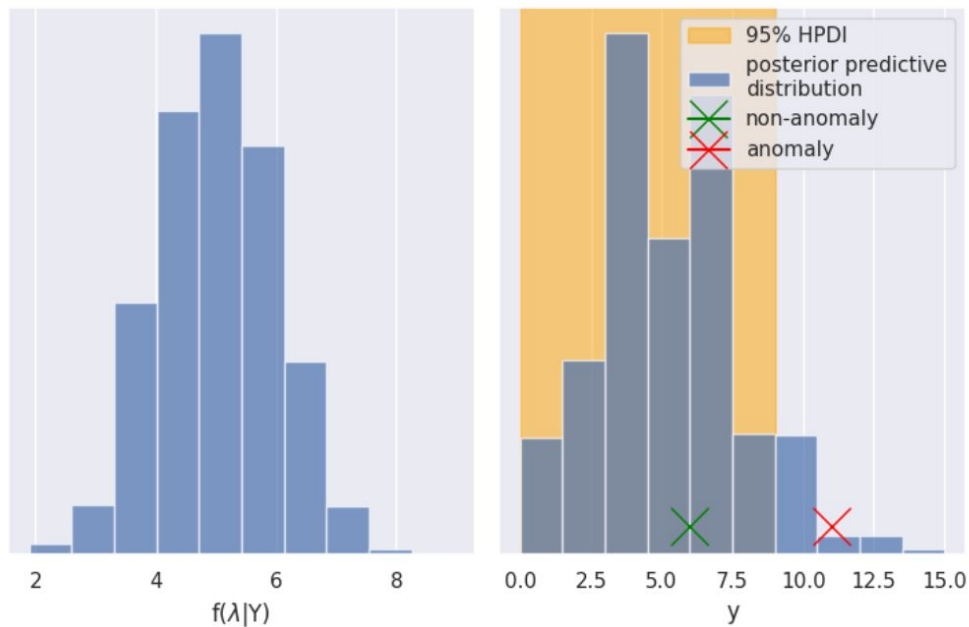
## 2a. Model fitting - technical details

- ◆ Whole analysis run in Python on AWS
- ◆ Parameters estimation done in NumPyro (<https://num.pyro.ai>)
- ◆ HMC (MCMC) vs Stochastic Variational Inference (SVI):
- ◆ fitting time from **51 mins to 2.5 mins** (reduced cost)



### 3. Anomaly detection

1. Get posterior predictive distribution from the model fitting
2. Evaluate Highest Posterior Density Interval (HPDI)
3. If an observed count is higher than HPDI  $\rightarrow$  anomaly



# Results

		HR data	Data-driven grouping method			
		HR	TS	k-means	GMM	Bi-cluster
No grouping	M1	1.10%	1.10%	1.10%	1.10%	1.10%
	M2	0.70%	0.70%	0.70%	0.70%	0.70%
seasonality	M3	0.70%	0.70%	0.71%	0.69%	0.61%
	M4	0.63%	0.68%	0.66%	0.63%	0.63%
auth. method	M5	0.60%	0.63%	0.60%	0.57%	0.60%
	M6	0.61%	0.62%	0.59%	0.57%	0.61%
N groups		12	3	15	8	3

Total counts in the test set = 400k, so reduction from 1.1% to 0.57% means ~2120 fewer alerts

# Conclusions

- ◆ Summary: New UEBA method proposed which can perform anomaly detection in authentication counts based on data-driven users grouping (no need for HR records);
- ◆ Limitations: Unaware of real threats within the considered data set;
- ◆ Future work: shrinkage methods for proposing a better-suited hierarchy

securonix

Thank you!