A Dictionary Learning Based Anomaly Detection Method for Network Traffic Data

Taha Yusuf Ceritli[1], Barış Kurt[2], Çağatay Yıldız[2], Bülent Sankur[3]
Ali Taylan Cemgil[2]

[1]Bogazici University, Dept. of Computational Science and Engineering,
[2]Bogazici University, Dept. of Computer Engineering,
[3]Bogazici University, Dept. of Electrical and Electronics Engineering

Abstract

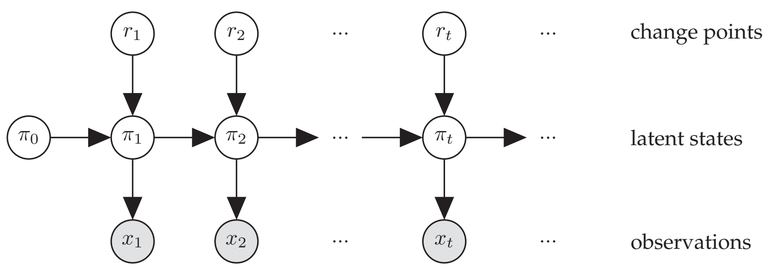
We propose a **dictionary learning** scheme to extract **network traffic** pattern templates for different types of anomalies together with the normal traffic via **Nonnegative Matrix Factorization (NMF)**. We employ **Bayesian Change Point Models** on the representation of the running network traffic in terms of those templates to detect network anomalies. We test and evaluate our methods on a simulated **Session Initiation Protocol (SIP)** network, alongside attacks generated by a commercial network vulnerability scanning tool.

Motivation

- **Problem**: Detecting Distributed Denial of Service (DDoS) attacks in SIP networks.
- **Contribution**: Combination of Bayesian Change Point models with NMF.

Methodology

Bayesian Change Point Model (BCPM):



$$\pi_0 \sim \Omega(\pi_0)$$
$$r_t \sim p^{r_t}(1-p)^{(1-r_t)}$$
$$\pi_t | r_t, \pi_{t-1} \sim [r_t = 0]\delta(\pi_t - \pi_{t-1}) + [r_t = 1]\Omega(\pi_t)$$
$$x_t | \pi_t \sim \Theta(x_t; \pi_t)$$

Nonnegative Matrix Factorization (NMF):

$$X \approx BW$$
$$x_{i,t} \approx [BW]_{i,t} = \sum_j b_{i,j} w_{j,t}$$

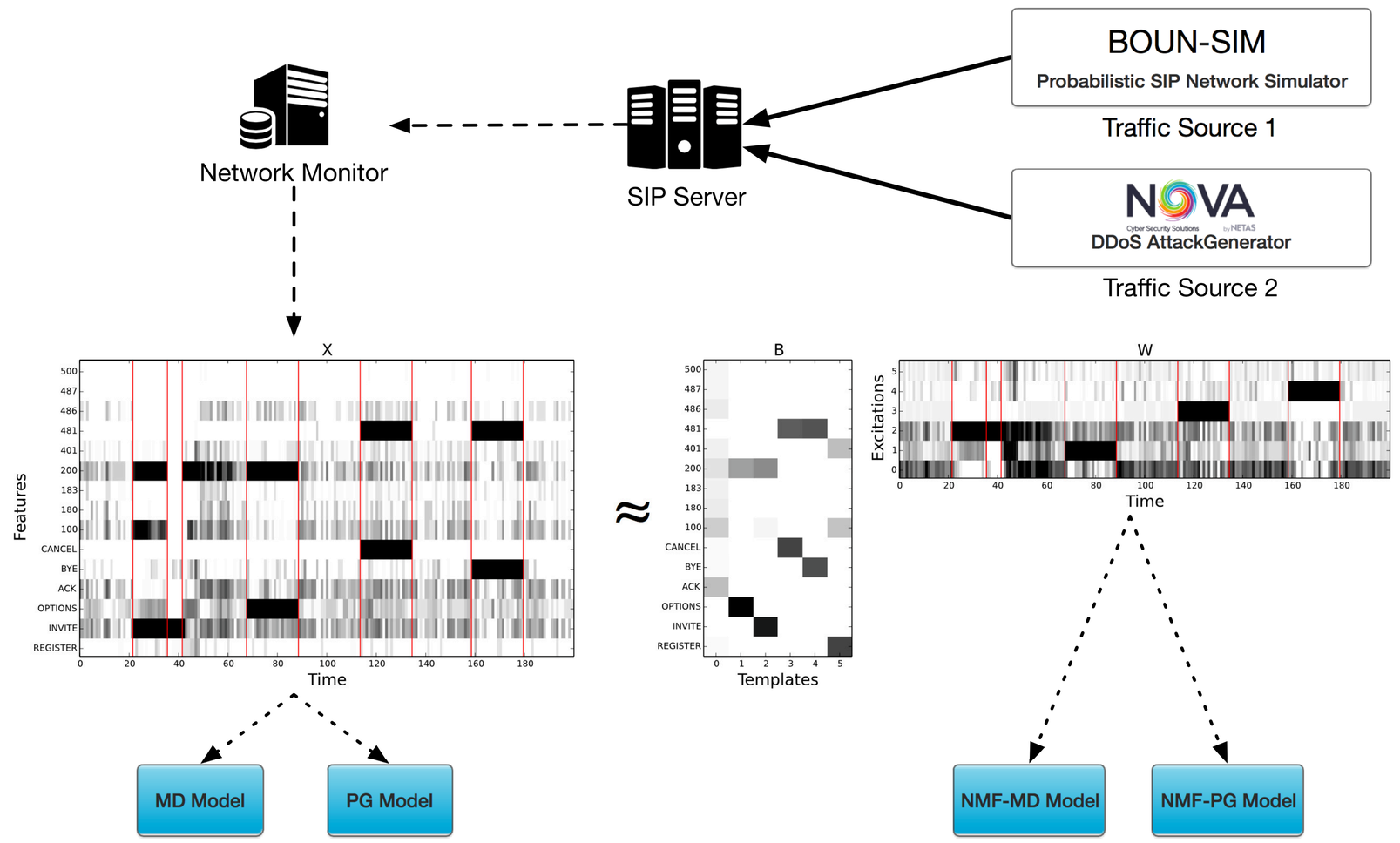
References

BOUN SIP Simulator. `https://github.com/cagatayyildiz/boun-sim`, 2016. [Online; accessed 17-June-2016].

Fearnhead, P. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and computing*, 16(2):203–213, 2006.

Xiaohong, G., Wang, W., and Zhang, X. Fast intrusion detection based on a non-negative matrix factorization model. *Journal of Network and Computer Applications*, 32(1):31 – 44, 2009.

Experimental Setup



Models

1. **Multinomial-Dirichlet (MD) Model**:

$$\Theta(x_t; \pi_t) \equiv \mathcal{M}ult(x_t; \pi_t)$$
$$\Omega(\pi_t; \alpha) \equiv \mathcal{D}ir(\pi_t; \alpha)$$

2. **Poisson-Gamma (PG) Model**:

$$\Theta(x_{i,t}; \pi_{i,t}) \equiv \mathcal{PO}(x_{i,t}; \pi_{i,t}), \quad \forall i \in [1, I]$$
$$\Omega(\pi_{i,t}; \alpha, \beta) \equiv \mathcal{G}(\pi_{i,t}; \alpha, \beta), \quad \forall i \in [1, I]$$

3. **NMF[J]-MD Model**:

$$x_t \approx B w_t$$
$$\Theta(w_t; \pi_t) \equiv \mathcal{M}ult(w_t; \pi_t)$$
$$\Omega(\pi_t; \alpha) \equiv \mathcal{D}ir(\pi_t; \alpha)$$

4. **NMF[J]-PG Model**:

$$x_t \approx B w_t$$
$$\Theta(w_{j,t}; \pi_{j,t}) \equiv \mathcal{PO}(w_{j,t}; \pi_{j,t}), \quad \forall j \in [1, J]$$
$$\Omega(\pi_{j,t}; \alpha, \beta) \equiv \mathcal{G}(\pi_{j,t}; \alpha, \beta), \quad \forall j \in [1, J]$$

Results

Experiment Settings:

1. Traffic Intensity: **Low** or **High**.
2. Number of SIP clients: **50** or **250**.
3. Attack Scenario: A stream of **5** types of DDoS attacks replicated with different **flood rates** and **fluctuation levels**.

	Low		High	
	50	250	50	250
MD	0.77	0.68	0.81	0.60
NMF4-MD	0.93	0.66	**0.94**	**0.83**
NMF5-MD	**0.95**	**0.70**	0.83	0.67
NMF6-MD	0.84	0.63	0.82	0.60
NMF7-MD	0.83	0.62	0.85	0.58
PG	0.30	0.29	0.31	0.28
NMF4-PG	0.52	**0.56**	**0.50**	**0.49**
NMF5-PG	**0.53**	0.51	0.43	0.45
NMF6-PG	0.49	0.46	0.42	0.43
NMF7-PG	0.47	0.46	0.43	0.41

Table 1: F-scores of the proposed models under different experiment settings.

Future Work

- Classification of network attacks.
- Separating normal and malicious traffic by classifying per-user traffic.
- Non-parametric and online learning of template vectors.

Conclusions

- Dictionary based network representation helps increasing anomaly detection success.
- MD models yield much better results than PG models.
- As the NMF rank increases, F-scores tend to decrease.