Automated feature engineering VS manual feature engineering in MalPaCA for network flow

Sung kyung Park (S.Park-4@student.tudelft.nl)

CSE3000 – Technical University of Delft
Supervised by Azqa Nadeem and Sicco Verwer

1. Background

MalPaCA = Malware Packet-sequence Clustering and Analysis

- Discovers the distinct behaviors of each uni-directional connection from the network flow.
- Accuracy and explainability

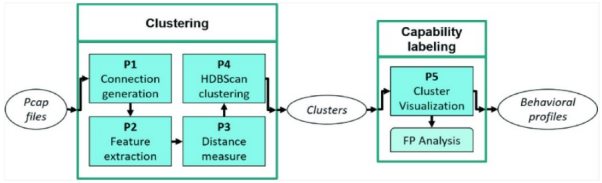

Figure 1: Baseline MalPaCA pipeline

Focus

- P2 – Feature Extraction:
  - B: {packet size, packet interval, src port, dst. port}
  - AE: {size, flags, dest. IP address, fragment offset, protocol, src IP address, header checksum, TOS, TTL, src port, dst. port and TCP/UDP checksum}

2. AE integrated MalPaCA

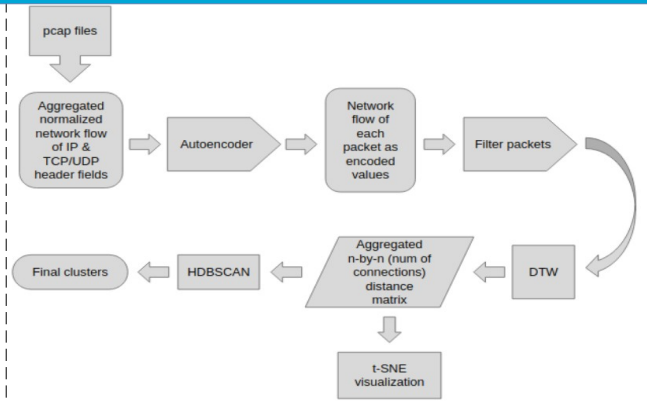

Figure 2: AE integrated MalPaCA pipeline

3. Methodology

Autoencoder design:

- AE variant = Undercomplete autoencoder → literature study
- Number of hidden layers = 7 → literature study
- Number of neurons = [12, 65, 35, 20, 5] → literature study + grid search

4. Experimental setup

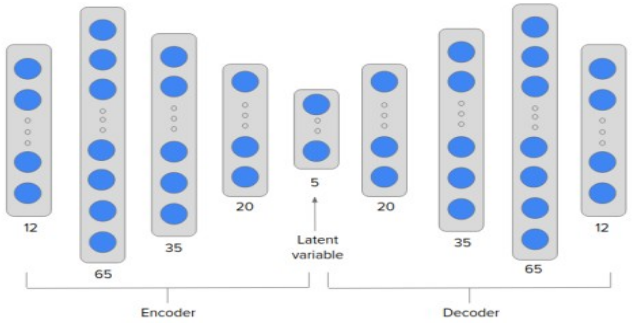

Figure 3: MalPaCA AE design

Accuracy evaluation metrics:

- Malware cluster purity (MCP)
- Noise percentage
- Silhouette score

5. Result

Baseline MalPaCA

- MCP = 60%
- Noise = 9.3%
- Silhouette = 0.48

Interpretabrility is comprehensive and values are meaningful

AE integrated MalPaCA

- MCP = 44%
- Noise = 20.3%
- Silhouette = 0.21

Interpretability is limited due to encoded values. Although, helps further Wireshark investigation

6. Conclusion

- Baseline MalPaCA outperforms AE integrated MalPaCA.
- Feature set change resulting cluster
- Difference: IP address (+), flags (+), offset (+), protocol (+), checksum (+), TOS (+), TTL (+), time interval (-)

7. Future Work

- Initial feature selection for AE input (e.g. time interval)
- Search for clustering algorithms / hyper parameters

AE integrated MalPaCA heatmaps

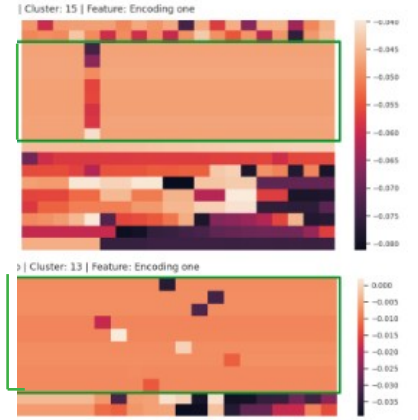

Figure 4: AE MalPaCA - C&C-Torii heatmap

Figure 7: AE MalPaCA - DDoS heatmap

Baseline MalPaCA heatmaps

Figure 5: Baseline MalPaCA – C&C gap heatmap

Figure 6: Baseline MalPaCA – C&C-Torii gap heatmap

Figure 8: AE MalPaCA – Benign NTP heatmap