Case Study: Understanding Internet Anomalies

Hardeep Kaur Takhar, Luiz Felipe Oliveira, Ljiljana Trajković

ljilja@cs.sfu.ca

Communication Networks Laboratory

http://www.sfu.ca/~ljilja/cnl

School of Engineering Science

Simon Fraser University, Vancouver
British Columbia, Canada

Roadmap

- Introduction
- Description of datasets
- Intrusion detection systems
- Machine learning for anomaly detection
- Methodology and performance evaluation
- Conclusion and references

Roadmap

- Introduction
- Description of datasets
- Intrusion detection systems
- Machine learning for anomaly detection
- Methodology and performance evaluation
- Conclusion and references

Introduction

- The Internet has been highly susceptible to malicious attacks:
  - worms
  - viruses
  - denial of service (DoS) and distributed denial of service (DDoS)
  - power and other outages
  - ransomware attacks
  - router misconfigurations
  - IP hijacks
- Attacks compromise the availability of resources to legitimate users by flooding the network, excessively consuming network resources, and overwhelming servers with a large number of requests.

Roadmap

- Introduction
- Description of datasets
- Intrusion detection systems
- Machine learning for anomaly detection
- Methodology and performance evaluation
- Conclusion and references

Description of Datasets: BGP

- Well-known BGP anomalies include:
  - Worm attacks (Code Red, Nimda, Slammer)
  - Power link failures (Moscow, Pakistan)
  - Ransomware attacks (WannaCrypt, WestRock)
- BGP RIPE and Route Views datasets consist of 37 features extracted from BGP update messages collected during periods of Internet anomalies.

Data Collections: Réseaux IP Européens (RIPE)

- Regional Internet Registry for Europe, Middle East, and Central Asia

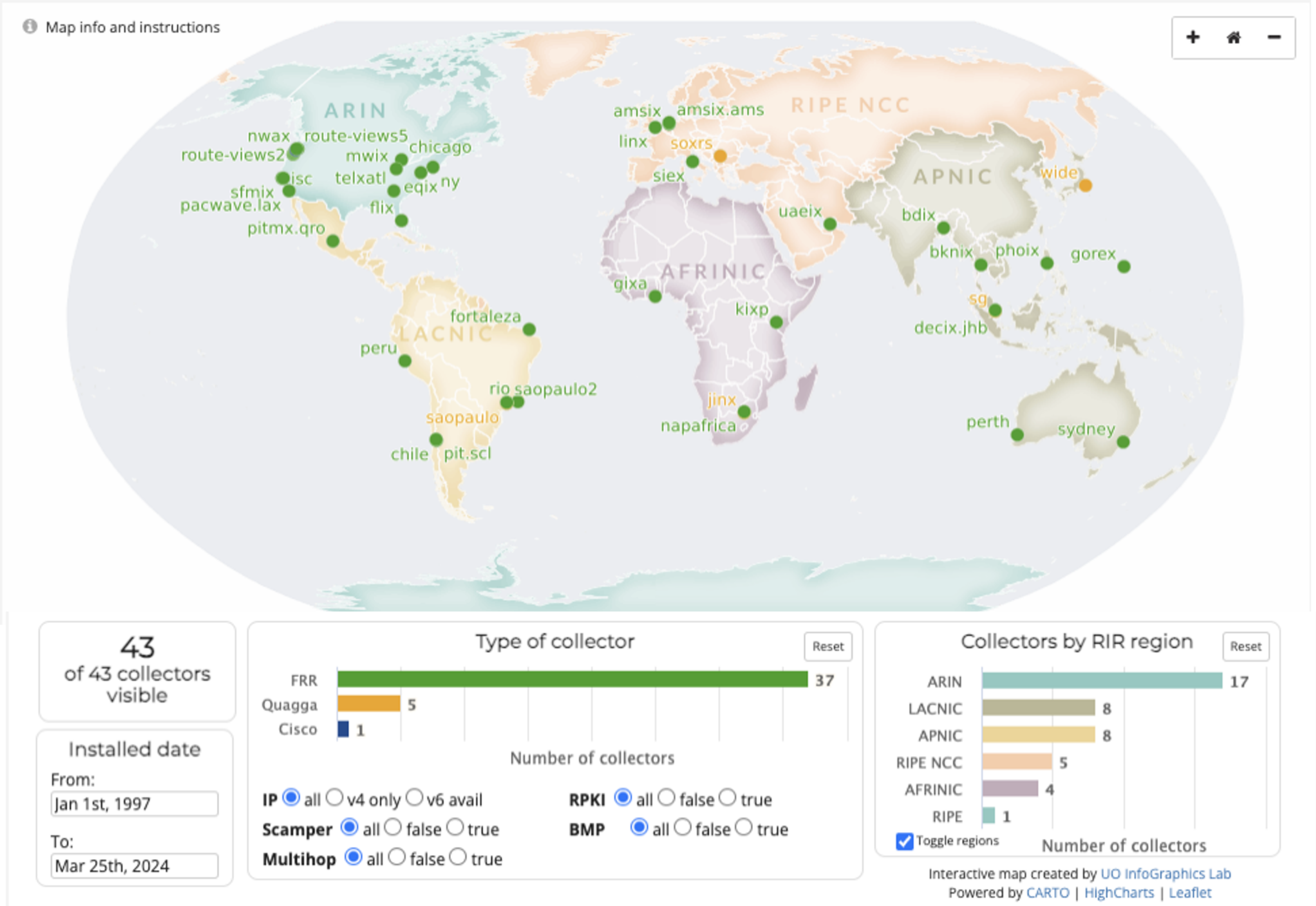
Source: https://www.ripe.net/about-us/

Data Collections: Routing Information Service (RIS)

Name	City
RRC00	Amsterdam, NL
RRC01	London, GB
RRC03	Amsterdam, NL
RRC04	Geneva, CH
RRC05	Vienna, AT
RRC06	Otemachi, JP
RRC07	Stockholm, SE
RRC10	Milan, IT
RRC11	New York, NY, US
RRC12	Frankfurt, DE
RRC13	Moscow, RU
RRC14	Palo Alto, CA, US
RRC15	Sao Paolo, BR
RRC16	Miami, FL, US
RRC18	Barcelona, ES
RRC19	Johannesburg, ZA
RRC20	Zurich, CH
RRC21	Paris, FR
RRC22	Bucharest, RO
RRC23	Singapore, SG
RRC24	Montevideo, UY
RRC25	Amsterdam, NL
RRC26	Dubai, AE

27 Remote Route Collectors (RRCs)
Source: https://ris.ripe.net/docs/route-collectors/
Map created using https://www.zeemaps.com

Data Collections: University of Oregon Route Views Project



Source: https://www.routeviews.org/routeviews/index.php/map/

Description of Datasets: NSL-KDD

- An improved version of the KDD’99 intrusion dataset based on the DARPA 1998 testbed.
- It contains 9 weeks of collected traffic when various intrusions were introduced in a simulated US Air Force base network.
- The *tcpdump* utility was used to collect traffic from:
  - Transport Control Protocol (TCP)
  - User Datagram Protocol (UDP)
  - Internet Control Message Protocol (ICMP)
- Each network connection is represented by 41 features:
  - 38 numerical and 3 categorical features.

Description of Datasets: CIC Testbed

- CICIDS2017, CSE-CIC-IDS2018, CICDDoS2019 datasets include intrusions that exploited various network vulnerabilities executed using tools for malicious attacks.
- Features include duration, size of packets, number of packets, and number of bytes.
- CICIDS2017: collected between 03.07.2017 and 07.07.2017 including 84 features.
- CSECIC-IDS2018: collected between 14.02.2018 and 02.03.2018 including 83 features.
- CICDDoS2019: collected between 03.11.2018 and 01.12.2018 extracting 87 features.

Roadmap

- Introduction
- Description of datasets
- Intrusion detection systems
- Machine learning for anomaly detection
- Methodology and performance evaluation
- Conclusion and references

Intrusion Detection Systems

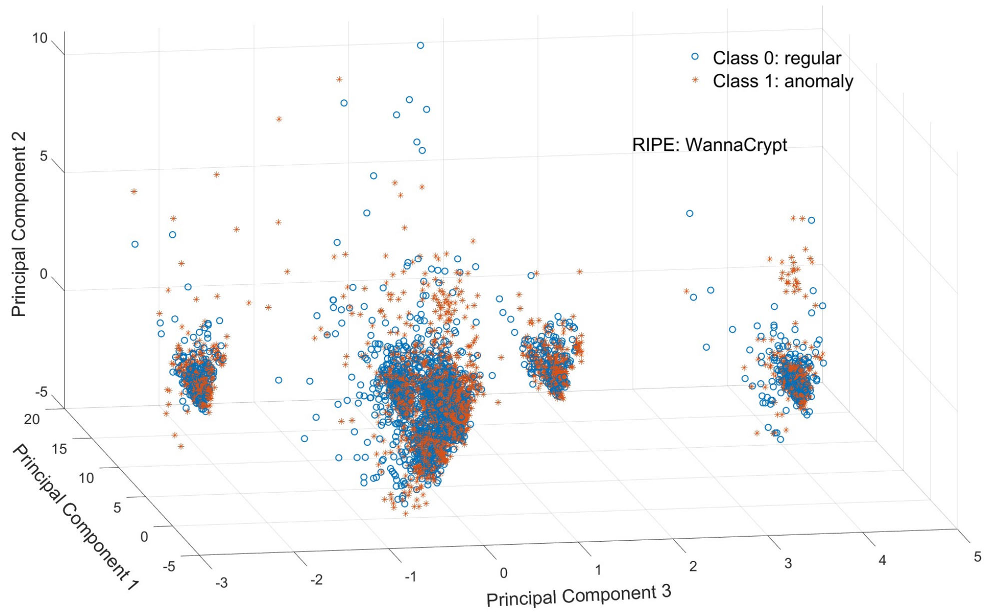
- The Internet lacks security infrastructure and, as such, it is exposed to viruses, worms, power outages, ransomware attacks, IP hijacks, and misconfigurations.
- Various methods and tools to detect network intrusions have been reported.
- Anomalies: The generated unusual patterns in routing traffic data.
- Categorized as:
  - point anomalies: individual data points that significantly deviate from the expected behavior.
  - contextual anomalies: depend on specific conditions.
  - collective anomalies: multiple instances of joint anomalous behavior.

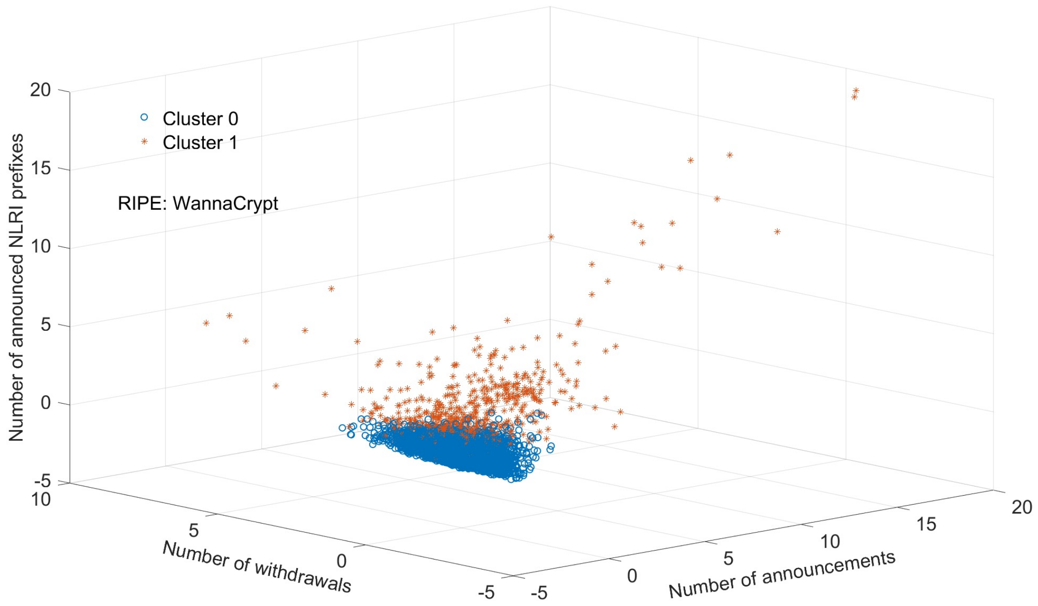
Roadmap

- Introduction
- Description of datasets
- Intrusion detection systems
- Machine learning for anomaly detection
- Methodology and performance evaluation
- Conclusion and references

Regular and Anomalous Classes

- Feature selection
- Label refinement





WannaCrypt RIPE training dataset: Regular and anomalous clusters based on principal components (left) and *k*-means clustering (right).

Machine Learning for Anomaly Detection

- Supervised, unsupervised, and semi-supervised machine learning techniques are used to detect network anomalies.
- Supervised machine learning algorithms:
  - Support vector machine (SVM) and naïve Bayes (NB) may achieve desired performance using smaller datasets but require longer training time
- Deep learning algorithms:
  - CNNs, RNNs, autoencoders, transformers (attention mechanism), generative adversarial networks
  - Rely on backpropagation and may use variable number of hidden layers

Machine Learning for Anomaly Detection

- Fast machine learning algorithms have been successful in generating models for large datasets and have shorter training times:
  - Broad Learning System (BLS
  - Gradient Boosted Decision Tree (GBDT)
- BLS: updates weights using pseudo-inverse
- GBDT: relies on decision trees

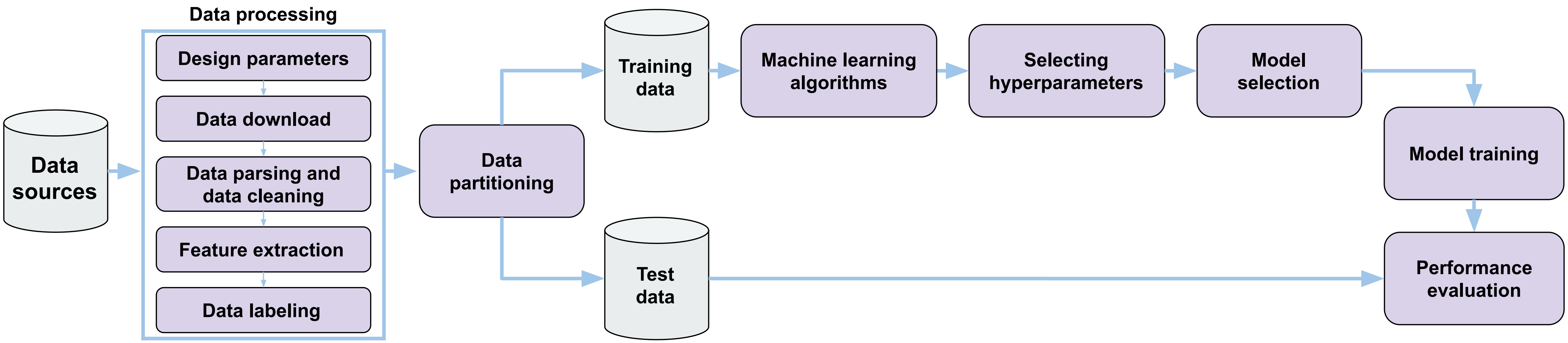
Supervised Learning Algorithms

- Support vector machine
- Deep learning
- Broad learning system
- Gradient boosting decision tree

Roadmap

- Introduction
- Description of datasets
- Intrusion detection systems
- Machine learning for anomaly detection
- Methodology and performance evaluation
- Conclusion and references

Designing Machine Learning Models

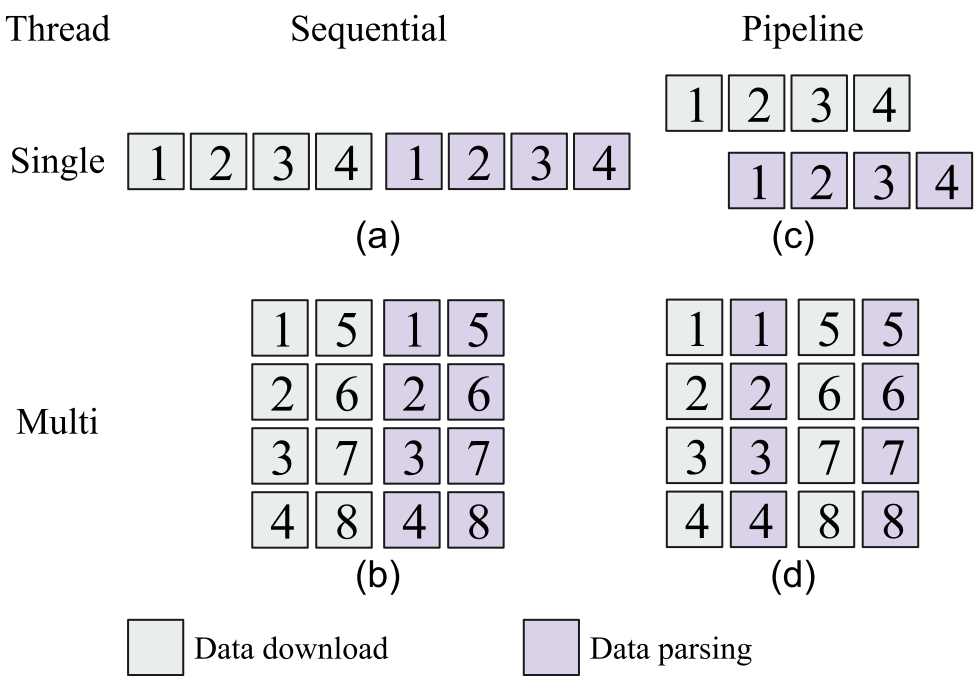


Designing machine learning models to classify network anomalies.
The steps include data processing, data partitioning, cross-validation to calculate hyperparameters, model selection and training, and performance evaluation.

Methodology and Performance Evaluation

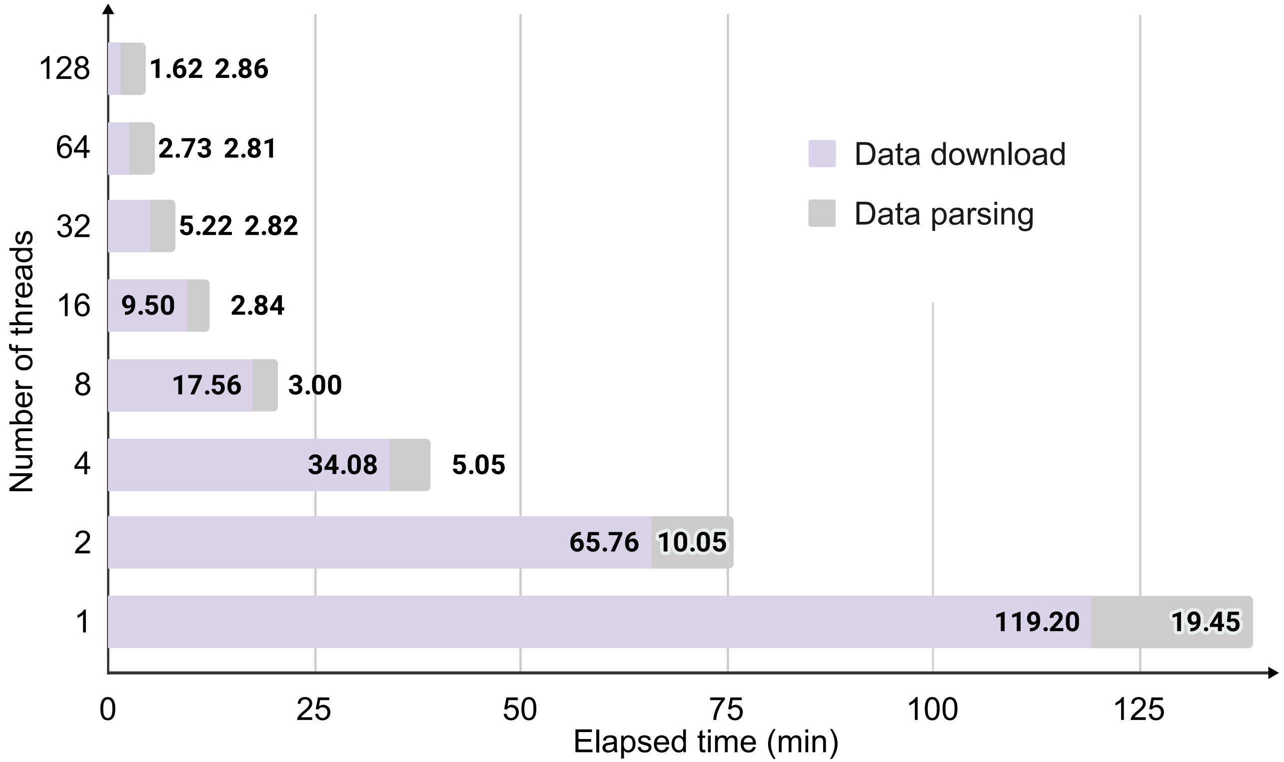
- Data sources
  - Design parameters
  - Data download
  - Data parsing and cleaning
  - Feature extraction
- Data labeling
- Data partitioning
- Machine learning algorithms
- Calculating hyperparameters
- Model selection
- Model training
- Performance evaluation

Data Download and Parsing



Data download and parsing: (a) data download is completed before parsing; (b) multi-core CPUs with data download completed before parsing; (c) data parsing begins before the completion of downloads; (d) simultaneous use of the pipeline and multi-threading.

Threads for WestRock Ransomware Attack Data

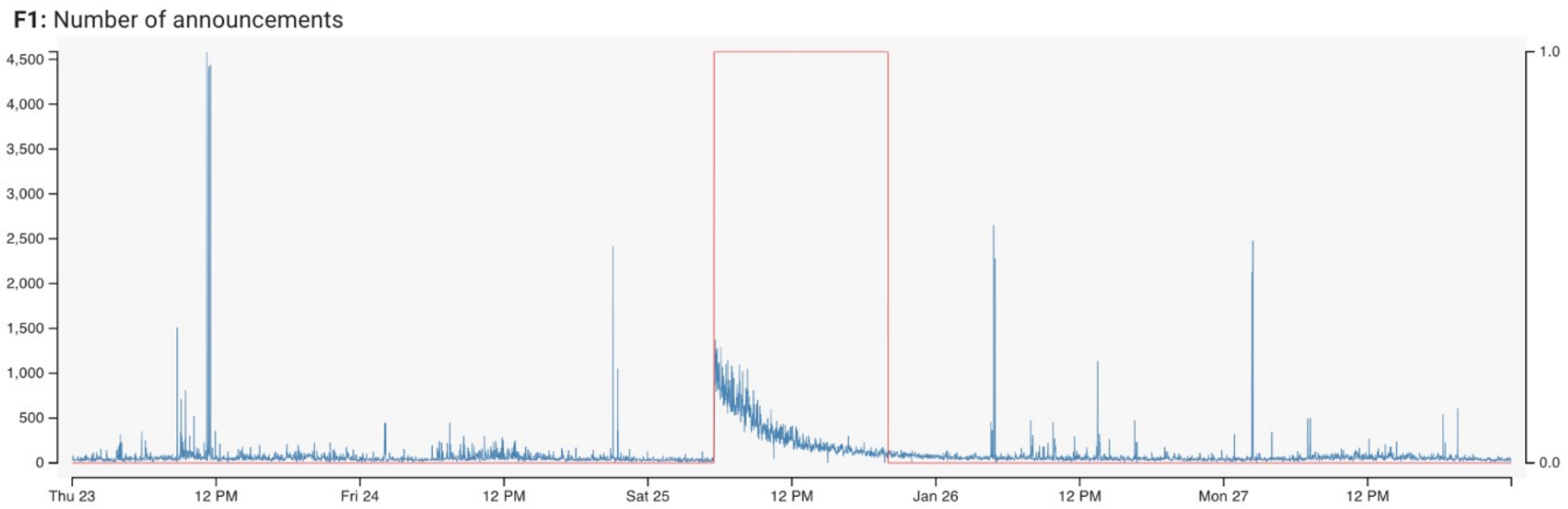


Data download and data parsing steps as functions of the number of threads for WestRock ransomware attack data collected from the RIPE RIS (rrc04) between 21.01.2021 and 31.01.2021.

Time Series Data Partition

- Partitions based on the entire data:

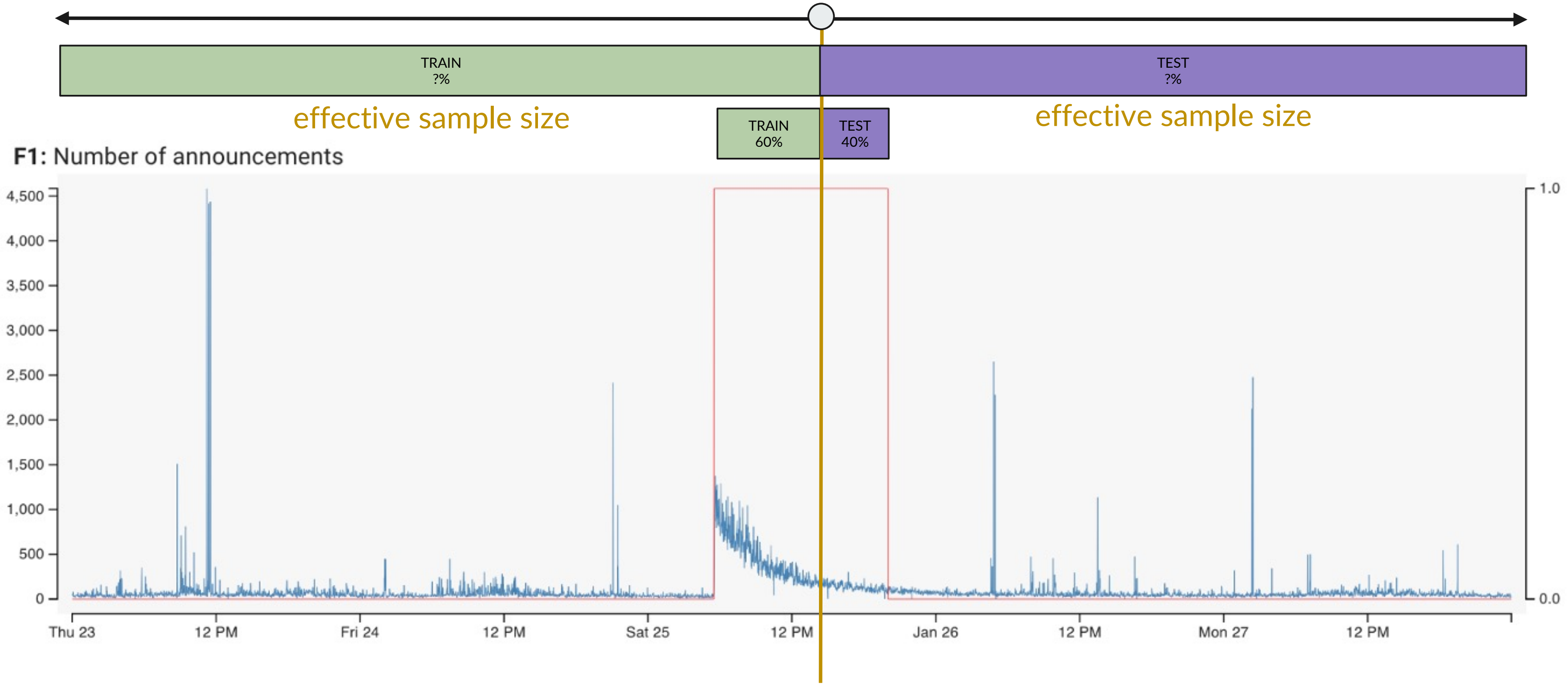
TRAINING: 60%	TEST: 40%



Source: RIPE, Slammer 2003

Time Series Data Partition

- Partitions based on the anomalous data:



Source: RIPE, Slammer 2003

Machine Learning Models: Best Performance

Dataset	Algorithm	Training time (s)	Accuracy (%)	F-Score (%)
Code Red	LightGBM	0.04	92.41	0.00
Nimda	LightGBM	0.46	81.67	40.94
Slammer	LightGBM	0.38	93.06	46.67
Moscow	LSTM4	9.16	97.12	44.88
Pakistan	LSTM4	10.64	73.78	21.10
WannaCrypt	VCFBLS	3.97	54.92	46.70
WestRock	VCFBLS	4.14	55.33	70.31
NSL-KDD	VCFBLS	31.32	83.58	83.70
CS-CIC-2018	VCFBLS	21.38	98.84	92.47

Models based on worm, blackout, ransomware, NSL-KDD, and CIC datasets

Roadmap

- Introduction
- Description of datasets
- Intrusion detection systems
- Machine learning for anomaly detection
- Methodology and performance evaluation
- Conclusion and references

Conclusion

- We described steps for generating various machine learning models using supervised learning algorithms.
- Datasets collected during reported anomalies that included worms, viruses, denial service attacks, blackouts, and ransomware attacks.
- While LightGBM models offered shorter training time than models generated using the LSTM and BLS algorithms, results indicated that model performance greatly depends on the used dataset.

References: Data Sources

- RIPE NCC:
  https://www.ripe.net
- University of Oregon Route Views project:
  http://www.routeviews.org
- IODA: Internet Outage Detection and Analysis:
  https://ioda.inetintel.cc.gatech.edu/
- NSL-KDD dataset:
  https://www.unb.ca/cic/datasets/nsl.html
- CIC-IDS2017, CSE-CIC-IDS2018, CIC-DDoS2019 datasets:
  https://www.unb.ca/cic/datasets/
- CAIDA: Center for Applied Internet Data Analysis:
  https://www.caida.org/projects/ioda/
  http://www.caida.org/home/

References: Tools

- Python: https://pypi.org

  Pandas: https://pandas.pydata.org/
- PyTorch
  https://pytorch.org/docs/stable/nn.html
- zebra-dump-parser:
  https://github.com/rfc1036/zebra-dump-parser
- BGP C# tool:
  http://www.sfu.ca/~ljilja/cnl/projects/BGP_datasets/index.html
- IEEE DataPort
  Border Gateway Protocol (BGP) datasets:
  - https://ieee-dataport.org/open-access/border-gateway-protocol-bgp-routing-records-reseaux-ip-europeens-ripe-and-bcnet
  - https://ieee-dataport.org/open-access/border-gateway-protocol-bgp-routing-records-route-views

References: Intrusion Detection

- J. P. A. Maranhão, J. P. C. L. da Costa, E. P. de Freitas, E. Javidi, and R. T. de Sousa, Jr., "Noise-robust multilayer perceptron architecture for distributed denial of service attack detection," *IEEE Commun. Lett*., vol. 25, no. 2, pp. 402–406, Feb. 2021.
- P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "A detailed investigation and analysis of using machine learning techniques for intrusion detection," *IEEE Commun. Surveys Tut*., vol. 21, no. 1, pp. 686–728, First quarter 2019.
- A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surv. Tut*., vol. 18, no. 2, pp. 1153–1176, 2016.
- M. C. Libicki, L. Ablon, and T. Webb, The Defenders Dilemma: Charting a Course Toward Cybersecurity. Santa Monica, CA, USA: RAND Corporation, 2015.
- V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Comput. Surv.,* vol. 41, no. 3, pp. 15:1–15:58, July 2009.

References: Deep Learning

- S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Oct. 1997.
- G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *Computing Research Repository (CoRR)*, abs/1207.0580, pp. 1–18, Jul. 2012.
- K. Cho, B. van Merriënboer, C. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder–decoder for statistical machine translations,” in *Proc. 2014 Conf. Empirical Methods Natural Lang. Process*., Doha, Qatar, Oct. 2014, pp. 1724–1734.
- I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning. Cambridge*, MA, USA: The MIT Press, 2016.
- K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, “LSTM: a search space odyssey,” *IEEE Trans. Neural Netw. Learn. Syst*., vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

References: BLS and GBDT

- C. L. P. Chen and Z. Liu, “Broad learning system: an effective and efficient incremental learning system without the need for deep architecture,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 10–24, Jan. 2018.
- C. L. P. Chen, Z. Liu, and S. Feng, “Universal approximation capability of broad learning system and its structural variations,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1191–1204, Apr. 2019.
- Broad Learning System: http://www.broadlearning.ai/
- T. Chen and C. Guestrin, “XGBoost: a scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.
- G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, “LightGBM: a highly efficient gradient boosting decision tree,” in *Proc. Int. Conf. Neural Inform. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, 3146–3154.
- L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” in *Proc. Int. Conf. Neural Inform. Process. Syst.*, Montreal, Quebec, Canada, Dec. 2018, 6639–6649.

Publications: http://www.sfu.ca/~ljilja

Journal publication:

- Z. Li, A. L. Gonzalez Rios, and Lj. Trajkovic, “Machine learning for detecting the WestRock ransomware attack using BGP routing records,” *IEEE Communications* Magazine, vol. 61, no. 3, pp. 20–26, Mar. 2023.
- Z. Li, A. L. Gonzalez Rios, and Lj. Trajkovic, “Machine learning for detecting anomalies and intrusions in communication networks,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 2254-2264, July 2021.

Book chapters:

- Q. Ding, Z. Li, S. Haeri, and Lj. Trajković, “Application of machine learning techniques to detecting anomalies in communication networks: datasets and feature selection algorithms” in *Cyber Threat Intelligence,* M. Conti, A. Dehghantanha, and T. Dargahi, Eds., Berlin: Springer, pp. 47–70, 2018.
- Z. Li, Q. Ding, S. Haeri, and Lj. Trajković, “Application of machine learning techniques to detecting anomalies in communication networks: classification algorithms” in *Cyber Threat Intelligence,* M. Conti, A. Dehghantanha, and T. Dargahi, Eds., Berlin: Springer, pp. 71–92, 2018.

Publications: http://www.sfu.ca/~ljilja

Conference publications:

- Z. Li and Lj. Trajković, “CyberDefense: tool for detecting network anomalies and intrusions,” *IEEE Int. Conf. Syst., Man, Cybern.*, Honolulu, HI, USA, Oct. 2023.
- H. Takhar and Lj. Trajković, “BGP feature properties and classification of Internet worms and ransomware attacks,” *IEEE Int. Conf. Syst., Man, Cybern.*, Honolulu, HI, USA, Oct. 2023.
- T. Sharma, K. Patni, Z. Li, and Lj. Trajković, “Deep echo state networks for detecting Internet worm and ransomware attacks” *IEEE Int. Symp. Circuits and Systems*, Monterey, CA, USA, May 2023.
- Z. Li, A. L. Gonzalez Rios, and Lj. Trajković, “Classifying denial of service attacks using fast machine learning algorithms,” in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Melbourne, Australia, Oct. 2021, pp. 1221-1226 (virtual).
- K. Bekshentayeva and Lj. Trajkovic, "Detection of denial of service attacks using echo state networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Melbourne, Australia, Oct. 2021, pp. 1227-1232 (virtual).
- Z. Li, A. L. Gonzalez Rios, and Lj. Trajković, “Detecting Internet worms, ransomware, and blackouts using recurrent neural networks,” in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Toronto, Canada, Oct. 2020, pp. 2165-2172 (virtual).

Publications: http://www.sfu.ca/~ljilja

Conference publications:

- A. L. Gonzalez Rios, Z. Li, K. Bekshentayeva, and Lj. Trajković, "Detection of denial of service attacks in communication networks," in *Proc. IEEE Int. Symp. Circuits and Systems*, Seville, Spain, Oct. 2020 (virtual).
- Z. Li, A. L. Gonzalez Rios, G. Xu, and Lj. Trajković, "Machine learning techniques for classifying network anomalies and intrusions," in *Proc. IEEE Int. Symp. Circuits and Systems*, Sapporo, Japan, May 2019 (virtual).
- A. L. Gonzalez Rios, Z. Li, G. Xu, A. Dias Alonso, and Lj. Trajković, "Detecting network anomalies and intrusions in communication networks," in *Proc. 23rd IEEE International Conference on Intelligent Engineering Systems 2019*, Gödöllő, Hungary, Apr. 2019, pp. 29–34.
- Z. Li, P. Batta, and Lj. Trajković, "Comparison of machine learning algorithms for detection of network intrusions," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Miyazaki, Japan, Oct. 2018, pp. 4248–4253.
- P. Batta, M. Singh, Z. Li, Q. Ding, and Lj. Trajković, "Evaluation of support vector machine kernels for detecting network anomalies," in *Proc. IEEE Int. Symp. Circuits and Systems*, Florence, Italy, May 2018, pp. 1-4.
- Q. Ding, Z. Li, P. Batta, and Lj. Trajković, "Detecting BGP anomalies using machine learning techniques," in *Proc. IEEE International Conference on Systems, Man, and Cybernetics,* Budapest, Hungary, Oct. 2016, pp. 3352–3355.