Measuring Diversity in Benchmark NIDS Datasets
WP5 - PO5.1

Benoit Nougnanke, Gregory Blanc, Thomas Robert

Plénière SuperviZ - 11/03/2025

Context

NIDS Dataset Quality Matters

- NIDS Construction and Evaluation
  - Effective ML-based NIDS depends on high-quality datasets [1, 2]
- Existing datasets suffer from label errors, bias, **low diversity** [3, 4]
- **Overfitting risk**: Flawed datasets reduce model generalization to real-world scenarios

Motivation

Why Measure Diversity?

- **Unmeasured dataset diversity**: Its effect on performance, generalization, and robustness remains unclear/unmeasured.
- **"Bad Design Smells"**: Existing studies highlight poor data diversity as a key issue in NIDS datasets [3]
- **Quantification gap**: Lack of structured measurement method of diversity tailored for NIDS datasets

Approach

Research Questions:

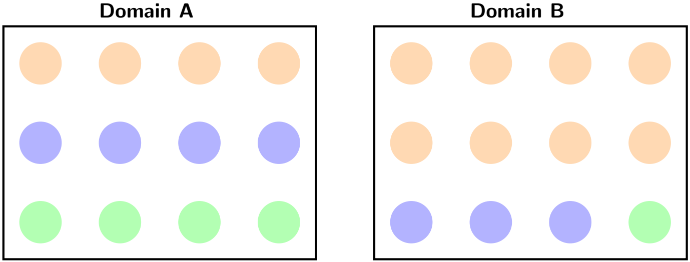
1. *Systematic Measurement Framework*: How to **systematically measure** NIDS data diversity?
2. *Bio-diversity Inspired NIDS Data Diversity*: How to leverage **ecological diversity measures** for NIDS data diversity quantification?
3. *NIDS Diversity and ML-based NIDS Performance*: How does NIDS **data diversity impact ML-based NIDS** performance and generalization?

Approach

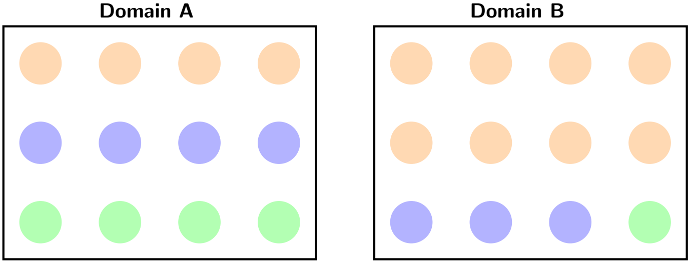
Research Questions:

1. *Systematic Measurement Framework*: How to **systematically measure** NIDS data diversity?
2. *Bio-diversity Inspired NIDS Data Diversity*: How to leverage **ecological diversity measures** for NIDS data diversity quantification?
3. *NIDS Diversity and ML-based NIDS Performance*: How does NIDS **data diversity impact ML-based NIDS** performance and generalization?

Diversity Concept: Ecology Inspirations [5]



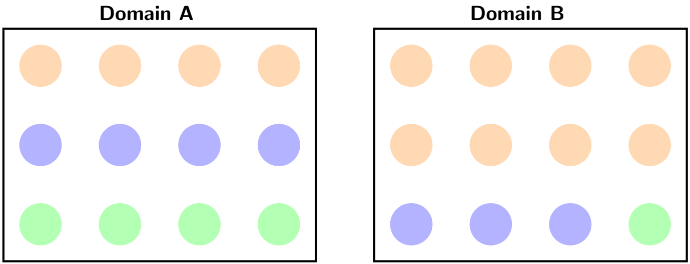
Diversity Concept: Ecology Inspirations [5]



Main Diversity Characteristics:

- Richness: number of 'species'
  - D(A) = 3.0 vs. D(B) = 3.0

Diversity Concept: Ecology Inspirations [5]

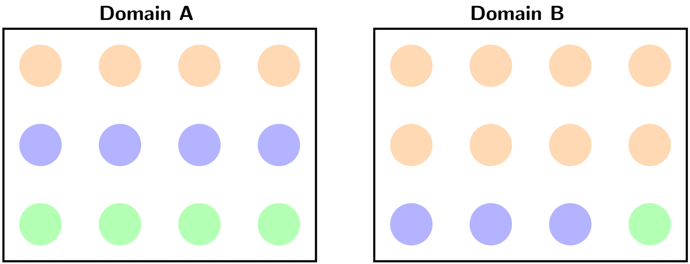


Main Diversity Characteristics:

- Richness: number of 'species'
  - D(A) = 3.0 vs. D(B) = 3.0

- Relative abundance (evenness)
  - D(A) = 3.0 vs. D(B) = 2.28

Diversity Concept: Ecology Inspirations [5]



Main Diversity Characteristics:

- Richness: number of 'species'
  - D(A) = 3.0 vs. D(B) = 3.0
- Relative abundance (evenness)
  - D(A) = 3.0 vs. D(B) = 2.28
- Similarity between species

Diversity Measure: True Diversity vs. Diversity Index [8, 9]

Generalized Dq Framework: Diversity of Order (q)

$$D_q = \left(\sum_{i=1}^{S} p_i^q\right)^{\frac{1}{1-q}} = \left(\sum_{i=1}^{S} p_i * p_i^{q-1}\right)^{\frac{1}{1-q}}$$

where q controls sensitivity to rare vs. dominant species

Diversity Measure: True Diversity vs. Diversity Index [8, 9]

Generalized Dq Framework: Diversity of Order (q)

$$D_q = \left(\sum_{i=1}^{S} p_i^q \right)^{\frac{1}{1-q}} = \left(\sum_{i=1}^{S} p_i * p_i^{q-1} \right)^{\frac{1}{1-q}}$$

where q controls sensitivity to rare vs. dominant species

Key Diversity Indices:

- **Species Richness (D0)**: Counts distinct entities (e.g., *species*)
- **Shannon Entropy (D1)**: Measures uncertainty in species distribution
- **Gini-Simpson Index (D2)**: Probability that two randomly selected samples belong to different species
- **Rao's Quadratic Entropy (D2)**: Captures similarity between species

Diversity Measure: True Diversity vs. Diversity Index [8, 9]

Generalized Dq Framework: Diversity of Order (q)

$$D_q = \left(\sum_{i=1}^{S} p_i^q \right)^{\frac{1}{1-q}} = \left(\sum_{i=1}^{S} p_i * p_i^{q-1} \right)^{\frac{1}{1-q}}$$

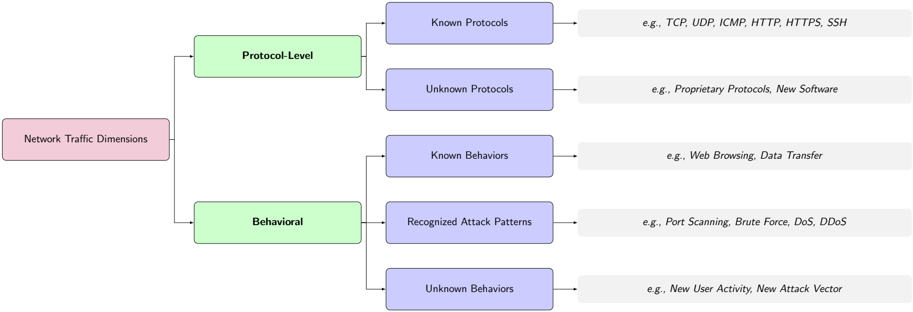
where q controls sensitivity to rare vs. dominant species

Key Diversity Indices:

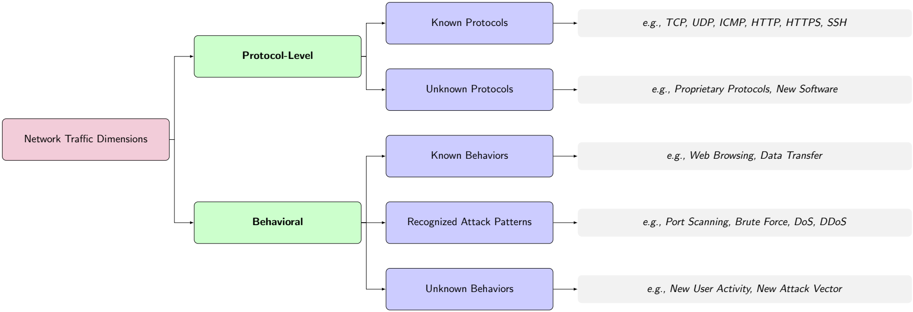
- **Species Richness (D0)**: Counts distinct entities (e.g., *species*)
- **Shannon Entropy (D1)**: Measures uncertainty in species distribution
- **Gini-Simpson Index (D2)**: Probability that two randomly selected samples belong to different species
- **Rao's Quadratic Entropy (D2)**: Captures similarity between species

ML SoA: Vendi Score (Dq) - Uses sample similarity and eigenvalues to quantify diversity [6, 7]

Network Traffic Categorization



Network Traffic Categorization

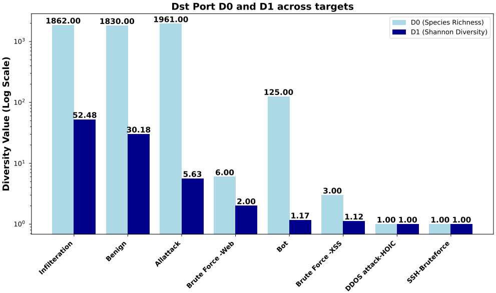


Key Features for Traffic Categorization:

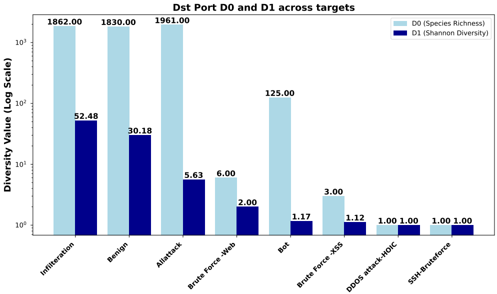
- Dst Port, Protocol
- Flow Duration, Flow IAT Mean, TotLen Fwd Pkts, Pkt Size Avg, Flow Byts/s, Flow Pkts/s, Down/Up Ratio
- Clustering to identify traffic species $\rightarrow$ profiling

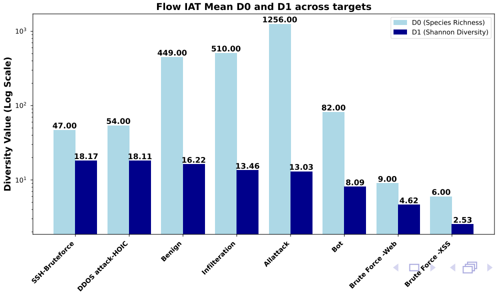
Related works: [10] [11] [12]

Initial Experiments on CICIDS 2018 subset (100k samples)

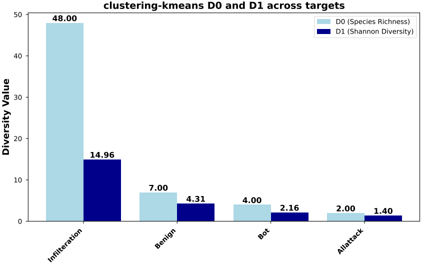


Initial Experiments on CICIDS 2018 subset (100k samples)



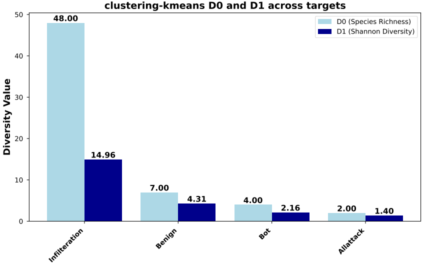


Initial Experiments on CICIDS 2018 subset (100k samples)

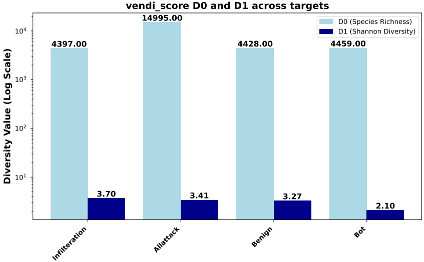


(a) D0 and D1 - Clustering (K-Means)

Initial Experiments on CICIDS 2018 subset (100k samples)

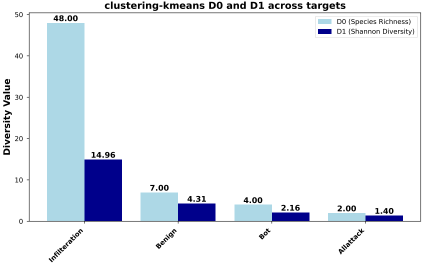


(a) D0 and D1 - Clustering (K-Means)

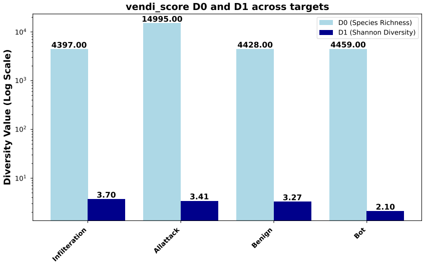


(b) D0 and D1 - Vendi Score

Initial Experiments on CICIDS 2018 subset (100k samples)



(a) D0 and D1 - Clustering (K-Means)



(b) D0 and D1 - Vendi Score

Diversity measurement depends strongly on how we define clusters (species)

Key Insights & Future Directions

Main Takeaways:

- **NIDS Dataset Quality Matters**:
  How does dataset diversity influence ML-based NIDS performance and generalization?
- **Bio-inspired Diversity Metrics**:
  Adapting ecological diversity measures (Dq and Vendi Score) for evaluating NIDS dataset diversity

Key Insights & Future Directions

Main Takeaways:

- **NIDS Dataset Quality Matters**:
  How does dataset diversity influence ML-based NIDS performance and generalization?
- **Bio-inspired Diversity Metrics**:
  Adapting ecological diversity measures (Dq and Vendi Score) for evaluating NIDS dataset diversity

Challenges & Future Work:

- **Reliability** and **validity** of the proposed diversity metrics
- Bridging diversity measures to ML performance: Evaluating how diversity influences model generalization
- Integrating diversity into dataset lifecycle: creation, curation, evaluation

Key Insights & Future Directions

Main Takeaways:

- **NIDS Dataset Quality Matters**:
  How does dataset diversity influence ML-based NIDS performance and generalization?
- **Bio-inspired Diversity Metrics**:
  Adapting ecological diversity measures (Dq and Vendi Score) for evaluating NIDS dataset diversity

Challenges & Future Work:

- **Reliability** and **validity** of the proposed diversity metrics
- Bridging diversity measures to ML performance: Evaluating how diversity influences model generalization
- Integrating diversity into dataset lifecycle: creation, curation, evaluation

Thank you for your attention!

References I

[1] J. Halvorsen, C. Izurieta, H. Cai, and A. Gebremedhin, "Applying generative machine learning to intrusion detection: A systematic mapping study and review," *ACM Computing Surveys*, vol. 56, no. 10, pp. 1–33, 2024.

[2] S. Layeghy, M. Gallagher, and M. Portmann, "Benchmarking the benchmark—comparing synthetic and real-world network ids datasets," *Journal of Information Security and Applications*, vol. 80, p. 103689, 2024.

[3] R. Flood, G. Engelen, D. Aspinall, and L. Desmet, "Bad design smells in benchmark nids datasets," in *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*, pp. 658–675, IEEE, 2024.

[4] L. Liu, G. Engelen, T. Lynar, D. Essam, and W. Joosen, "Error prevalence in nids datasets: A case study on cic-ids-2017 and cse-cic-ids-2018," in *2022 IEEE Conference on Communications and Network Security (CNS)*, pp. 254–262, IEEE, 2022.

[5] T. Leinster, *Entropy and diversity: the axiomatic approach*. Cambridge university press, 2021.

[6] D. Friedman and A. B. Dieng, "The vendi score: A diversity evaluation metric for machine learning," *arXiv preprint arXiv:2210.02410*, 2022.

[7] A. P. Pasarkar and A. B. Dieng, "Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning," *arXiv preprint arXiv:2310.12952*, 2023.

[8] L. Jost, "Entropy and diversity," *Oikos*, vol. 113, no. 2, pp. 363–375, 2006.

[9] H. Tuomisto, "A consistent terminology for quantifying species diversity? yes, it does exist," *Oecologia*, vol. 164, no. 4, pp. 853–860, 2010.

References II

[10] G. Nychis, V. Sekar, D. G. Andersen, H. Kim, and H. Zhang, “An empirical evaluation of entropy-based traffic anomaly detection,” in *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, pp. 151–156, 2008.

[11] S. Landau-Feibish, Z. Liu, and J. Rexford, “Compact data structures for network telemetry,” *ACM Computing Surveys*, 2025.

[12] T. Bühler, R. Schmid, S. Lutz, and L. Vanbever, “Generating representative, live network traffic out of millions of code repositories,” in *Proceedings of the 21st ACM Workshop on Hot Topics in Networks*, pp. 1–7, 2022.