Recognition of Similar Netflow Data in Decentralized Monitoring Environments

Georg Eisenhart, Simon Volpert, Jan Braitinger, Jörg Domaschka
Institute of Information Resource Management
Ulm University

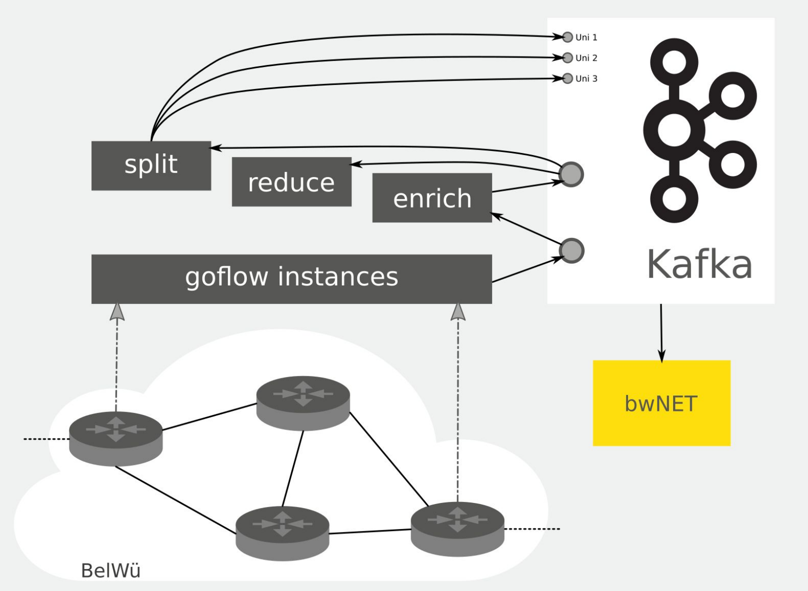
Motivation

- NetFlow Monitoring
  - Extend data acquisition
  - Need for tightly monitored networks
- Data Analysis
  - Aggregating NetFlow data from multiple Sites
  - Deep insight in network traffic composition
  - Provide the ability for Proof of Transit in SFC szenarios
  - Classification of NetFlow Data can support threat detection

- Need to identify similar data points

NetFlow Data Acquisition

- Monitoring Netflow
  - Only on selected interfaces
  - Enriched data
  - Provide for further consumption
- Extension to decentralized monitoring environments
  - Multiple instances
  - Merge data from different sites and networks
  - More tightly meshed monitoring



[1] bwNetFlow: A Customizable Multi Tenant Flow processing Platform for Transit Providers
[2] https://github.com/bwNetFlow/flowpipeline

Use Cases

- Circumvent distortion of data analysis
  - Multiple but different data points of the same flow
  - This can impact derived metrics and analysis results
- Detect presence of same or similar flows at specific points in the network
  - Validation of research scenarios like traffic engineering, traffic routing, service function chaining
- Classification of NetFlows
  - Detection of similar data points by a given artificial blueprint

Research Questions

- How do similar or related NetFlow data affect data analysis?
- How can similar data points be treated in decentralized monitoring environments?
- How can we use the identification of similar NetFlow data for classification?

Related Work

- Botnet traffic detection by calculating distances between incoming and outgoing flows

[3] A. Tayal, N. Hubballi, and N. Tripathi, "Communication recurrence and similarity detection in network flows," in 2017 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Dec. 2017, pp. 1–6. doi: 10.1109/ANTS.2017.8384174.

- Anomaly and outlier identification on NetFlow data based on similarity measurements

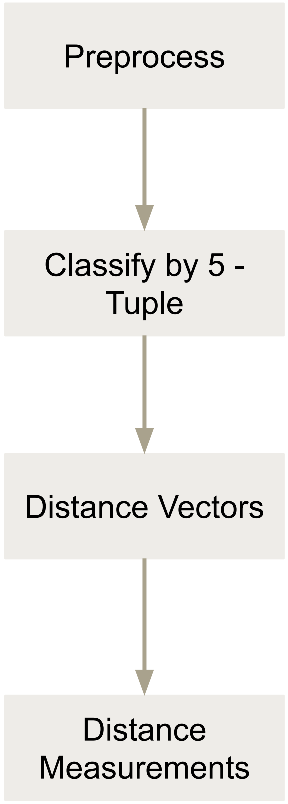
[4] D. S. Terzi, R. Terzi, and S. Sagiroglu, "Big data analytics for network anomaly detection from netflow data," in 2017 International Conference on Computer Science and Engineering (UBMK), Oct. 2017, pp. 592–597. doi: 10.1109/UBMK.2017.8093473.

- Clustering algorithms with euclidean distance metric as input for network intrusion detection systems

[5] L. Dias, S. Valente, and M. Correia, "Go With the Flow: Clustering Dynamically-Defined NetFlow Features for Network Intrusion Detection with DynIDS," in 2020 IEEE 19th International Symposium on Network Computing and Applications (NCA), Nov. 2020, pp. 1–10. doi: 10.1109/NCA51143.2020.9306732.

Methodology

- Preprocess
  - Filter for relevant traffic to be analyzed
  - TCP and UDP traffic
  - Calculate identifier for source port and destination port pair (due to unidirectional flow data)
  - Calculate identifier for address pair
- Classify
  - Classify each flow by the Netflow 5 - Tuple (SrcAddress, DstAddress, SrcPort, DstPort, Protocol)
  - Usage of above mentioned Identifiers

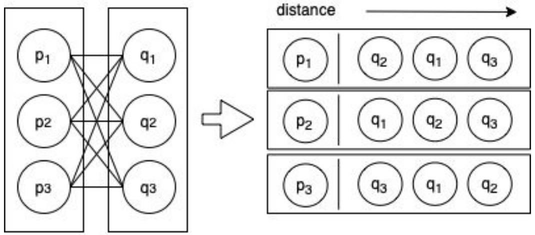


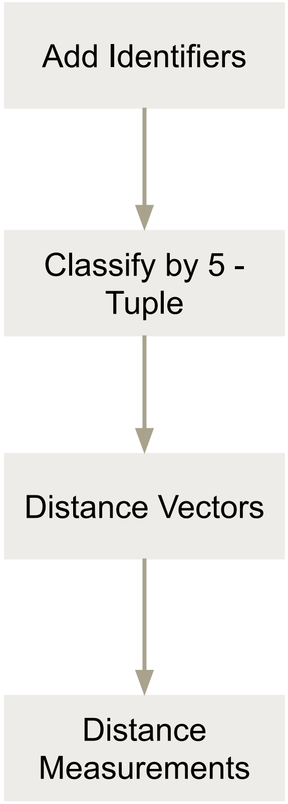
Methodology

- Create Distance Vectors
  - Transmitted Bytes
  - Transmitted Packets
  - Timestamp of flow

$$\vec{v} = \begin{pmatrix} Bytes \\ Packets \\ Timestamp \end{pmatrix}$$

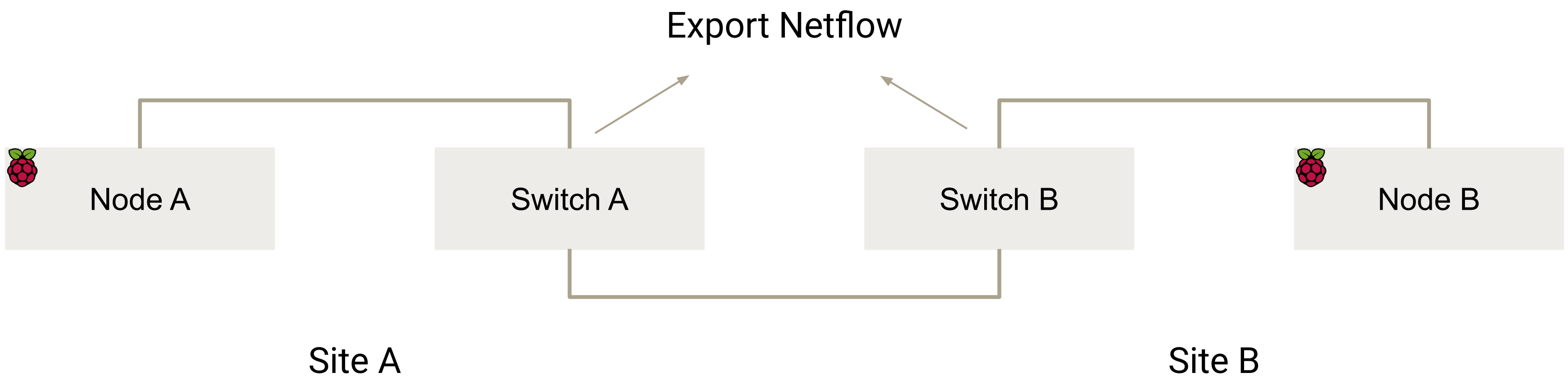
- Distance Measurements
  - Use euclidean or manhattan distance over vectors v
  - Calculated between all flows from different routers within the same class
  - Minimal distance represents most similar flows





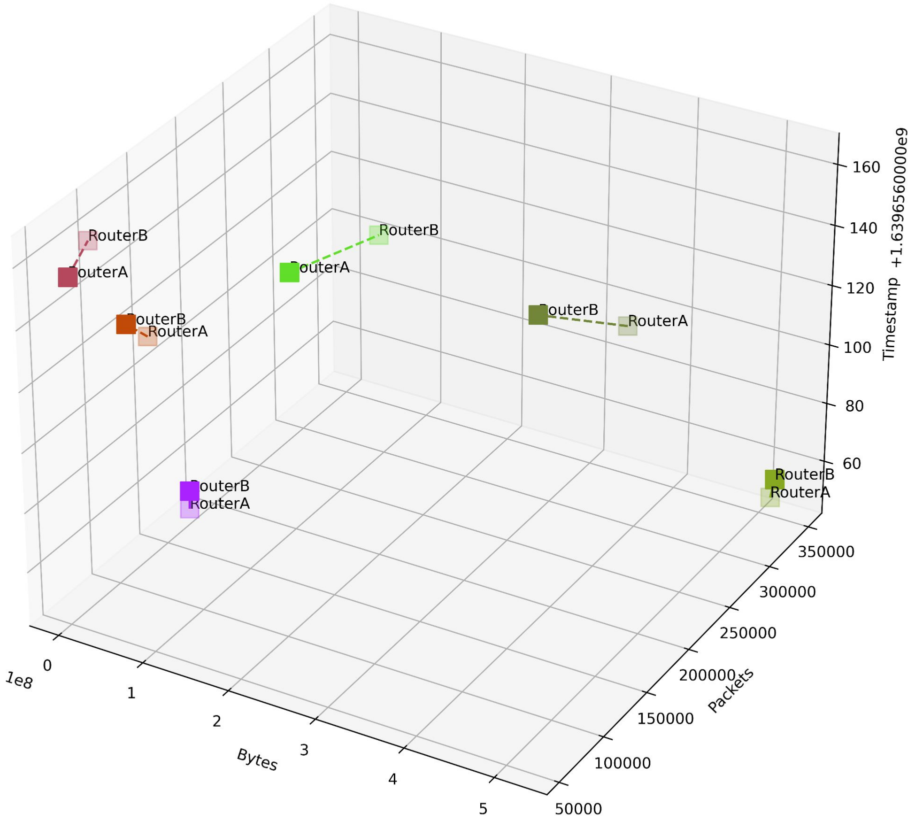
Testing Environment

- Two Layer-3 Switches
  - Export Netflow for each device
- Two Nodes (Raspberry PI 3B)
  - Multiple file transfers between Node A and Node B for traffic generation



Preliminary Results

- Small sample size for validation
  - 20 flow records per router
  - Recognition of 6 pairs of similar flows
  - Other flows treated as background traffic
  - Manhattan and euclidean distance gives comparable results
  - Large datasets have to be evaluated automatic in future



Euclidean distance of similar flows (TCP)

Summary and Outlook

- We provide a proof of concept to identify similar Netflow data from multiple data sources
- Support for use cases in our research field
  - Distortion of data analysis
  - Detecting presence of similar flows at different devices, …

- Refine approach
  - Make software scalable for more data sources (n > 2)
- Compare results of different distance measures for their suitability
  - Target euclidean, manhattan and cosine similarity
- Evaluate more techniques and technologies
  - e.g. bloom filter and AI technologies like General Adversarial Networks

References

[1] D. Nägele, C. B. Hauser, L. Bradatsch, and S. Wesner, “bwNetFlow: A Customizable Multi-Tenant Flow Processing Platform for Transit Providers,” in 2019 IEEE/ACM Innovating the Network for Data-Intensive Science (INDIS), Nov. 2019, pp. 9–16. doi: 10.1109/INDIS49552.2019.00007.

[2] https://github.com/bwNetFlow/flowpipeline

[3] A. Tayal, N. Hubballi, and N. Tripathi, “Communication recurrence and similarity detection in network flows,” in 2017 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Dec. 2017, pp. 1–6. doi: 10.1109/ANTS.2017.8384174.

[4] D. S. Terzi, R. Terzi, and S. Sagiroglu, “Big data analytics for network anomaly detection from netflow data,” in 2017 International Conference on Computer Science and Engineering (UBMK), Oct. 2017, pp. 592–597. doi: 10.1109/UBMK.2017.8093473.

[5] L. Dias, S. Valente, and M. Correia, “Go With the Flow: Clustering Dynamically-Defined NetFlow Features for Network Intrusion Detection with DynIDS,” in 2020 IEEE 19th International Symposium on Network Computing and Applications (NCA), Nov. 2020, pp. 1–10. doi: 10.1109/NCA51143.2020.9306732.

Thank you

Any Questions?