Recognition of Similar NetFlow Data in Decentralised Monitoring Environments

Georg Eisenhart, Simon Volpert, Jan Braitinger, Jörg Domaschka
Institute of Information Resource Management
Ulm University, Germany
{firstname.lastname}@uni-ulm.de

***Abstract*—One of the main challenges in the analysis of NetFlow data in decentralised monitoring environments comes from merging datasets from different independent sites. One problem is to identify similar data points which can impact derived metrics from such data directly. This article provides a proof of concept how similarity measurements based on distance metrics can be used to identify similar or related flows from different datasets. For this, several domains are outlined which can benefit from this approach to support validation of research scenarios and data analysis.**

***Index Terms*—netflow, monitoring, distance/similarity measures**

I. Introduction

In large ISP networks, NetFlow data is collected at multiple points of interest to obtain an insight into the network traffic composition. Flows traversing a network might be exported on multiple devices on the path for hosts communicating with each other. Though, similar or related data can lead to multiple captures of the same flow which leads to misinterpretation of the traffic while analysing the data. This introduces the need for mechanisms to identify or merging flows traversing network devices. To deal with this problem, the BelWü[1] actually monitors NetFlow traffic only on selected border interfaces to minimize the impact of this effect. This also comes into play, whilst merging collected flow data from multiple monitoring instances, or different sites and networks.

The main questions in this ongoing work are *(i)* how similar or related NetFlow data affects monitoring data analysis, *(ii)* how can these similar or related data points be treated in decentralised monitoring environments and *(iii)* how can the identification of similar NetFlow data be used for classification.

II. Related Work

Several related work uses distance metrics for analysing NetFlow data. Some selected will be presented in the following which indicates distance measurement as a viable tool in NetFlow data analysis.

Tayal et al. [1] provides an approach for identifying recurring network flows by generating a communication graph between hosts over distinct time intervals. Their approach utilises the manhattan distance between selected features of the flows for matching edges in the graph. However the approach focuses on detecting botnet traffic by calculating the distance between incoming and outgoing flows for each distinct connection.

Terzi et al. [2] uses similarity measurements on NetFlow data for anomaly and outlier detection. Data is preprocessed and clustered with k-means for a subsequently distance measurement of the cluster. While this work focuses on identifying anomalies by high distances, our approach is to find most similar NetFlow data for our domains as described in Section III.

III. NetFlow Data Collection

The BelWü monitors 28 border interfaces on 9 routers which are exporting NetFlow v9 with a 1:32 sampling rate to collect flow data leaving or entering the BelWü network. This leads to 20000 flows per second in average with peaks up to 160000 flows per second. The flows are collected with *goflow* as a NetFlow collector to re-encode them with *protobuf* and sending the data to an *Apache Kafka* cluster. For further consumption, the flows are getting enriched with additional metadata e.g. *snmp* data from the interface which the flow originates, protocol names or geolocation data. Currently these data are used for DDoS mitigation and the detection of high-volume networks [3].

Identifying similar NetFlow data enables network operations centres to monitor more points of interest in their network. To circumvent distortion of live data analysis due to multiple data points of the same flow, such identification is necessary.

Our approach can also be used to detect the presence of the same flow at specific points in the network to support the validation of research scenarios in disciplines like traffic engineering, service function chaining, traffic routing and flow-based validation for proof of transit. This enables NetFlow based analysis of datasets over a long range to provide an overview of the routing in a network to a given timestamp.

Another domain is to find similar data points by a given blueprint. Simplified, it shall be able to find all similar flows in datasets, which matches a predefined artificial flow as input. This can be useful for classification scenarios, where predefined patterns are to be analysed.

IV. NetFlow Identification

The related work in Section II shows that distance measurements can be applied to NetFlow data. In our current work we are examining different distance metrics like euclidean, manhattan and cosine to show which fits best for our approach.

[1] https://www.belwue.de/

The NetFlow data of each source are classified and preprocessed for getting similarity based on distance measures. The raw data has to be filtered and classified by the NetFlow 5-Tuple (*SrcAddr, DstAddr, SrcPort, DstPort, Protocol*). This is enough for identifying a unique flow in a small time range, but doesn't give the accuracy for long term analysis or our outlined domains from Section III. In addition there shall be taken more features into account to identify a similar flow, like the timestamp, transmitted bytes and amount of packets. These are used to create the distance vector $\vec{v}$ in Equation (1) which is used to calculate the different distance measures. The vectors can be calculated for each preprocessed and reduced dataset (p and q) and then be applied to distance measures.

$$\vec{v} = \begin{pmatrix} Bytes \\ Packets \\ Timestamp \end{pmatrix} \tag{1}$$

As shown in Figure 1, each vector p has to be checked which vector q if its most similar vector. If multiple vectors of p point to q, the vector with the least distance is the most similar. Regarding the classification scenario to identify similar flows by a blueprint, it should be possible to define a range, which allows a certain amount of deviation in the vector space to be able to identify flows which inherits similar patterns as the given blueprint.

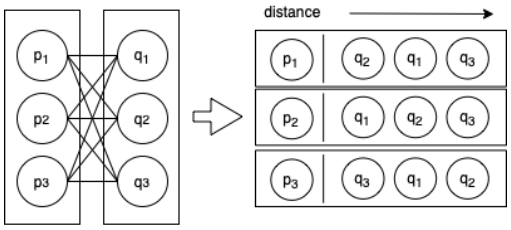

Figure 1. Get lowest distances of $\vec{p_i}$ to $\vec{q_j}$

V. Testing Environment

In the test setup, based on two nodes (Raspberry Pi 3B) and two layer-3 switches, NetFlows are examined. Each node is connected to a switch to represent an independent site, which are connected via WAN. Both switches are configured identically to have the same caching interval for NetFlow export. Thus, the probability for a 1:1 flow mapping is higher to validate the proof of concept. For traffic generation, larger files are exchanged between these two nodes, so that traffic traverses both switches. For each switch NetFlow data is exported to a distinct sink to have different datasets for each. After filtering, preprocessing the data and generating the input vectors for the distance measurements, first results are analysed and evaluated.

The results of the euclidean distance analysis are shown for example in Figure 2. In the test scenario, preliminary results show, that the euclidean distance and the manhattan distance gave promising results for our testing dataset.

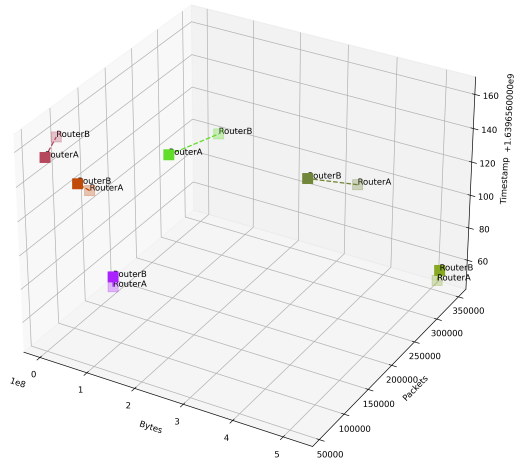

Figure 2. Euclidean distance analysis with vector $\vec{v}$

VI. Conclusion and Outlook

With this work we provide a proof on concept to identify similar NetFlow data in decentralised monitoring environments. This approach can support straight forward use cases like finding similar data points from data gathered and stored on multiple sites. But it may also be suitable for more sophisticated use cases like analyse traffic routing to a given timestamp, supporting proof of transit in tightly monitored computer networks or identifying periodic occurrences of certain traffic. In further work we want to refine our approach and compare the results of different distance measurements for its suitability to NetFlow datasets. There is also the question for appropriate threshold values for the distance metrics to indicate at which point a shortest distance is no longer representing similar data points. Another research field for the future is to evaluate other techniques like bloom filter or the usage of AI technologies like Generative Adversarial Networks to identify similar flows.

Acknowledgment

This work was supported by the bwNET2020+ project which is funded by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK).

References

[1] 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.

[2] 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.

[3] 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.