# A Novel Approach to IP ID Classification

Shujie Zhao and Haya Shulman
Fraunhofer Institute for Secure Information Technology SIT
{firstname.lastname}@sit.fraunhofer.de

December 18, 2020

The IP identifier (IP ID) in the IP header plays a vital role in packet fragmentation and reassembly. This field value can be assigned in diverse ways depending on Operating Systems, e.g., a global counter on Windows, a per-destination incrementor on Linux, or a pseudo-random generator under iOS. In recent decades, IP ID has drawn a lot of security concerns in the research community as a side-channel that can be exploited to implement various attacks. The IP ID abuse ranges from covert channels via global ID fields, DNS cache poisoning using sequentially incremental IP ID mechanisms (e.g., per-target or global), to user tracking through hash-based random number generation. As a consequence, identification disclosure makes the host involved susceptible to IP ID-based attacks. Therefore, understanding IP ID behaviors in the wild is essential to scrutinize the security threats induced by vulnerable identifiers.

However, realistic IP ID implementations have not been clarified in the literature. To the best of our knowledge, there is only one previous work, Salutari, Cicalese & Rossi (2018), aiming to classify the real-world IP ID behaviors through a machine learning (ML) algorithm based on six statistical features (e.g., expectation and entropy) of an ID sequence. Thereby we are motivated to characterize ID observations to discriminate different behaviors. Unlike Salutari *et al.* (2018)'s work, we assume that IP ID behavior is a function of time because, in most cases, the ID value will change temporally or dynamically. We employ two unique IP addresses hosted by a single physical machine to alternately send probes at a constant sampling rate in a specific protocol to collect a list of $N-$length ID time series (each series is assembled by ID values and their receiving time) from 99 randomly selected and BGP routable autonomous systems (ASes). We will then train a support vector machine (SVM) model on the ID series collected and apply the training model to distinguish IP ID behaviors automatically. The major challenge of this work is to determine appropriate features for achieving a high classification accuracy. We will conduct feature extraction in the time and frequency domains. Combined with our preliminary research, we consider the following characteristics of $N$ ID time samples: 1. $L$ **dominant frequencies** 2. **Autocorrelation** 3. **Cross correlation** 4. **ID velocity** 5. **ID acceleration** 6. **ID standard deviation**.

## References

Flavia Salutari, Danilo Cicalese & Dario J Rossi (2018). A closer look at ip-id behavior in the wild. In *International Conference on Passive and Active Network Measurement*, 243–254. Springer.

**crypto day matters 32** (2021)
Beiträge vom 32. Kryptotag, 15.01.2021, online.
doi:10.18420/cdm-2021-32-32