FlexHash: Hybrid Locality Sensitive Hashing for IoT Device Identification

Nathan Thom*, Jay Thom*, Batyr Charyyev, Emily Hand, Shamik Sengupta
Department of Computer Science and Engineering, University of Nevada, Reno, USA
1664 N. Virginia street m/s 0171 Reno, NV 89557 775-784-6905
Email: nathanthom@nevada.unr.edu, jthom@unr.edu, bcharyyev@unr.edu, emhand@unr.edu, ssengupta@unr.edu

The utility and ease-of-use offered by Internet of Things (IoT) devices has lead to a rapid increase in popularity for private, commercial, industrial, and healthcare applications. It is estimated that as many as 30 billion devices are connected globally with continued rapid growth projected for the near future. While these devices offer many benefits, they also bring increased risk to the private data of their users, as well as increased vulnerability for the networks they are connected to. New methods are required for reliably tracking device behavior and network membership as MAC addresses are easily spoofed and subverted or rogue devices can be difficult to detect.

Various techniques for IoT device identification through network traffic fingerprinting with machine learning have been proposed in the literature. While effective, these methods are not without their weaknesses; many requiring complex feature extraction and engineering which can introduce a high degree of overhead, and require extensive domain knowledge (both in networking and machine learning techniques) to select appropriate features for a given classifier. The classification of identical devices has not been adequately addressed, an issue that is important when monitoring multiple similar devices on the same network. Many previous studies classify heterogeneous devices in lab environments limited to only traffic generated by known devices, while in realistic environments there always exists a variety of background noise from unknown or unrelated devices. For this reason, we also include background noise in the form of random network traffic, as well as traffic generated from unknown IoT devices to show our system performs well in a realistic setting.

To address these challenges we introduce FlexHash, a hybrid locality sensitive hashing (LSH) method. While a cryptographic hash will produce an entirely different output if even a small change is made to the input, LSH will produce similar hash values for similar inputs. Because IoT devices are typically simple in their functionality, similarity hashes are useful for fingerprinting network traffic. Our method extends this utility in two ways; first, it combines the benefits of LSH with the power of machine learning, allowing for highly accurate single packet device identification. Second, it optimizes the similarity hashing function by providing the ability to adjust various parameters in the hashing process, allowing the output to be tuned for use with specific sets of devices. This is done by converting a hash digest generated from network traffic in the form of *.pcap files* to strings of base-10 values which become input feature vectors for a machine learning model. Since we use the hash of the traffic data we avoid the need for feature selection and extraction, a computationally expensive process.

To demonstrate the effectiveness of FlexHash, and our identification system as a whole, we perform the following experiments. First, we analyze the system's ability to classify devices as belonging to one of three categories. Smart Plug, Smart Lightbulb, and Web Camera. Next, we task our method with identifying different devices of identical make and model. Both of these experiments are performed with and without background network noise to ensure that results are consistent with a realistic network environment. Finally, we demonstrate the importance of hash parameter tuning on data representation and model performance.

Contributions of this work are as follows:

- We develop FlexHash, a novel locality sensitive hashing algorithm that enables adjustments to the hashing parameters (accumulator length, window size, and n-grams) providing improved optimization and flexibility.
- We implement a network traffic fingerprinting method combining FlexHash with machine learning, and perform accurate IoT device identification requiring only a *single packet* of data.
- We evaluate this system by classifying device genre and *identical devices* in the presence of similar peers while also including realistic *background noise*.
- We collect traffic data from three categories of 8 identical IoT devices, which we share with the research community.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under award number 2019164. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

* Equal contributing authors