December 2019

AUTOMATIC NETWORK DEVICE IDENTIFICATION USING IDENTIFIERS OF CONTACTED SERVERS

Jan Kohout

AUTOMATIC NETWORK DEVICE IDENTIFICATION USING IDENTIFIERS OF CONTACTED SERVERS

AUTHORS:
Jan Kohout

ABSTRACT

This proposal provides techniques for a system to identify network devices (e.g., classify their types, operating system families, etc.) based on information associated with contacted servers. In some implementations, the system may be designed to include an auto-update mechanism that enables the system to adapt to changing behavior of devices automatically. Thus, only initial labeled seeds may be needed to initialize the system, which may then automatically adapt itself without manual retraining.

DETAILED DESCRIPTION

With increasing number and types of devices in corporate networks (e.g., due to Internet of Things (IoT), Bring Your Own Device (BYOD), etc. trends), it is difficult to manage records about the type of each device connecting to a network. Moreover, not all types of logs of network communication may provide enough information to simply identify devices (e.g., by Media Access Control (MAC) addresses) connecting to a network.

This proposal provides techniques that may utilize analysis of contacted servers' hostnames to identify devices connecting to a network. Unlike other information, records containing information about contacted servers can be found in most network traffic logs. Tokenization and text mining techniques may be utilized to identify hostnames, parts of hostnames (e.g., specific substrings, etc.), and/or special tokens in Uniform Resource Locators (URLs) that may be indicative of certain types of devices or Operating System (OS) families that can be used to identify types of devices. Thus, techniques presented herein may enable a system to automatically learn and identify devices connecting to a network.

Certain types of devices may contact specific types of servers (e.g., identified by their hostnames). For example, a hostname for a web page optimized for mobile devices may have format corresponding to 'm.something.com'. Similarly, devices running certain operating systems developed by Apple® may contact hostnames having a format such as 'something.apple.com' (or other top level domain (TLD)) to perform updates, download new apps, etc. Similarly, devices running Microsoft® Windows® may visit hostnames having a format such as 'something.microsoft.com' and/or may contact servers that have a 'microsoft' substring or similar in their hostname. These are only a few examples; for other operating systems, developers, devices, etc. hostname/substring formatting may be analogical.

Thus, individual parts of hostnames (e.g., substrings that are parts of a complete hostname) of contacted servers or entire hostnames, either of which can be referred to herein as 'tokens', can be used to identify types of devices based on device activity that may be observed for a certain period of time. Behavior of each device in a network can be represented by capturing frequencies with which one or more given tokens may be observed in device communications. Frequency vectors representing the devices can be passed to classifiers or clustering algorithms to either directly classify a device to a device type or to classify a device to a group of devices having similar behavioral patterns (which are likely of the same type).

In order to perform the representations and classification effectively, a dictionary of indicative tokens may be created. An important aspect this proposal is that the dictionary can be built and maintained automatically using text data analysis techniques.

Consider, for example, that initial seeds in the form of representations of devices for which their device type is known (e.g., there can be previously known/identified devices in the network) may first be created. Next, representations of other devices can be built and groups of similarly behaving devices can be created. A similarity of two representations can be computed through the use of similarity measures applicable to frequency vectors that capture frequencies of tokens observed in device communications. For example, methods such as Jaccard index, cosine similarity, or Euclidean distance can be used to determine similarity between two representations.

Created groups of similarly behaving devices can be labeled using the labels from the seed devices by propagating these labels to all devices in a same group. The group labels can then be treated as topics and tokens observed in communication of devices from the group can be treated as words observed for the given topic. By treating the labels as topics and the tokens as words observed for a given tope, topic analysis methods (e.g., Latent Dirichlet Allocation, etc.) can be used to determine tokens that are most relevant and specific for a given topic. These techniques may provide for the ability to identify tokens that are often observed in communications of devices from a given group while they may not be observed in communications of devices from other groups. Therefore, such tokens can be considered as highly indicative for the type of devices that a given group represents and can be added to the dictionary for the given group.

As the behavior of devices and entire groups of devices evolves, new tokens can be automatically identified as indicative from each group and added to the dictionary. Additionally, old tokens from the dictionary that may no longer be observed in communications for a given group can be excluded from the dictionary for the group in order to maintain the dictionary size.

Thus, a key feature of the techniques of this proposal is the ability to use traffic from known devices to automatically infer tokens that are indicative for various types of devices. These tokens can be then used to classify unknown devices by observing hostnames and/or URLs visited from these devices without active probing of the devices.

In order to demonstrate the ability to identify devices using the above described techniques, consider example results of classifying three most prevalent operating system families—Apple®-based devices, Android®-based devices, and Windows®-based devices. Figure 1, below, provides various graphs illustrating precision and recall of the classification techniques described herein in which the classifications may evolve over time to provide improved classification quality. As shown in Figure 1, rising trends illustrate that, as devices are observed for a longer time, the quality of classifications improves as more indicative representations of device behavior are obtained. It should be emphasized that the behavior of devices was modeled only using hostnames of contacted servers and no other information for the example illustrated in Figure 1. Overall, the classification performance can be assessed as very good; for example, in the case classifying Windows-based devices, the classification performance utilizing techniques presented herein may be almost perfect.

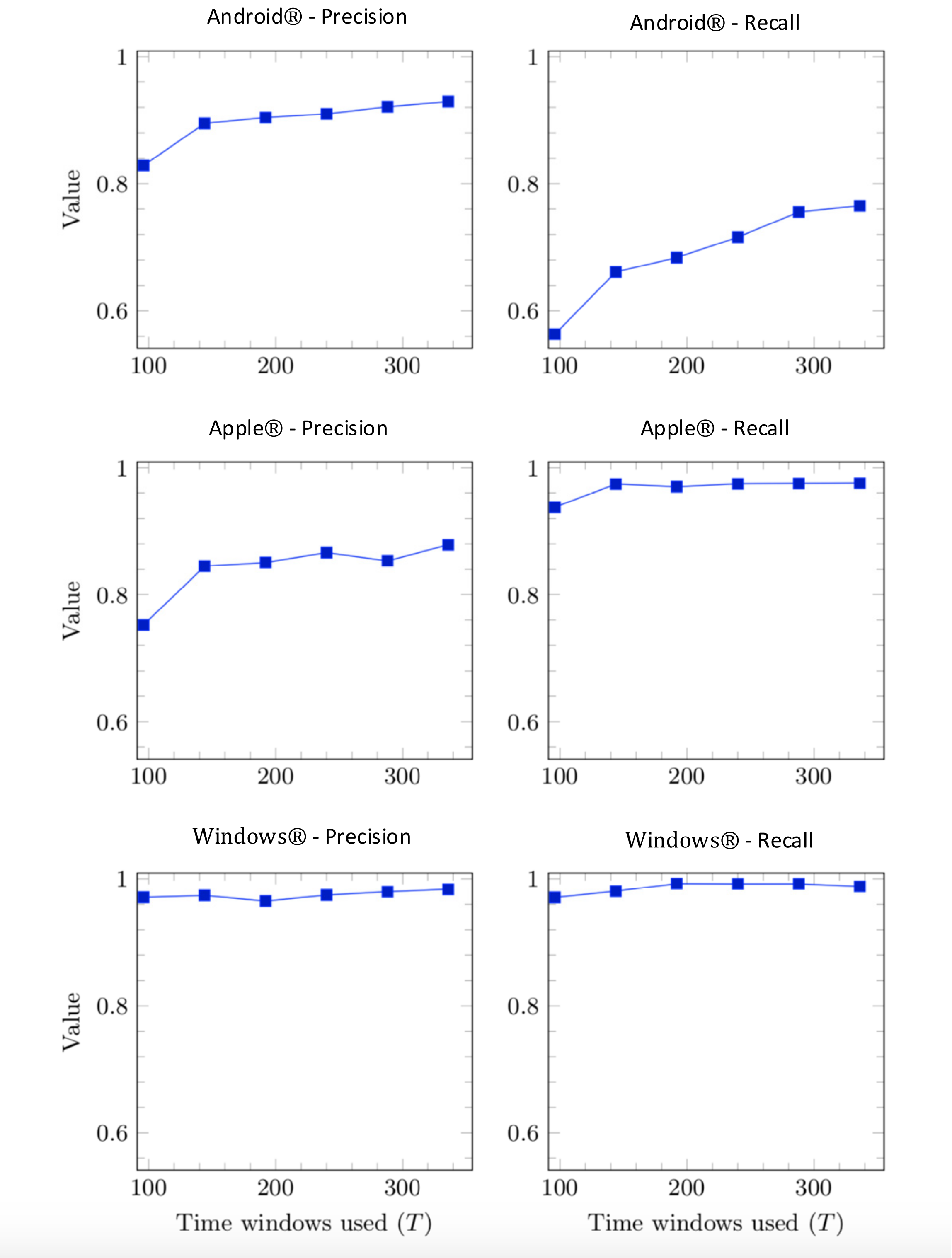
Android® - Precision
Android® - Recall
Apple® - Precision
Apple® - Recall
Windows® - Precision
Windows® - Recall
Value
1
0.8
0.6
100
200
300
Time windows used (T)

Figure 1

In summary, this proposal provides techniques for a system to identify network devices (e.g., classify their types, operating system families, etc.) based on information associated with contacted servers. In some implementations, the system may be designed to include an auto-update mechanism that enables the system to adapt to changing behavior of devices automatically. Thus, only initial labeled seeds may be needed to initialize the system, which may then automatically adapt itself without manual retraining.