Link-Layer Identification of Device Signatures: Wi-Fi Sensing for Crowd Analytics

Jennifer A. Finley

ABSTRACT

The Johns Hopkins University Applied Physics Laboratory (APL) Link-Layer Identification of Device Signatures (LLIDS) research effort uses machine learning techniques to identify unique wireless device signatures from patterns in link-layer data. Identifying signatures can increase situational awareness, assist in estimating crowd sizes, provide pattern of life, and protect facilities and infrastructure through activity surveillance. Link-layer Wi-Fi data are unique because they can be collected without access to a network and with devices that have low size, weight, and power (SWaP) requirements. The LLIDS multilayer system design breaks down link-layer data into unique device signatures using a combination of pattern recognition and state-of-the-art algorithms.

INTRODUCTION

The link layer is the second layer of the Open Systems Interconnection (OSI) model.[1] It consists of the physical addressing of devices to networks and resides between the physical layer, which consists of the physical signals, and the network layer. Link-layer data are easy to collect because of their location in the OSI model. Since the data are in the form of frames and not raw signals, specialized hardware is not needed, unlike in the physical layer. Since the data are not connected to a network, they can be collected right out of the air, unlike in the network layer. Any device with Wi-Fi enabled is transmitting link-layer data, and it is very easy to collect the data with devices with low size, weight, and power (SWaP) requirements. In addition, the number of wireless devices is rapidly increasing, so the amount of data available is also increasing.[2] Most Americans (85%) own a smartphone and carry it regularly, so the number of people present in a location is strongly correlated to the number of devices.[3]

The research described in this article primarily focuses on wireless link-layer signatures, specifically Wi-Fi (IEEE 802.11)[4] signatures. Most Wi-Fi link-layer analyses use probe request frames as the data source. These frames are broadcast messages that devices use to look for wireless networks in the area to connect to. Devices broadcast these messages continuously when Wi-Fi is enabled. The frames are unencrypted plain-text messages that can easily be processed into identifying features for further analysis.

Link-layer data have historically been used for a variety of analytics. For example, retailers in shopping malls previously used link-layer data for retail analytics to better understand the trends and behaviors of customers.[5] These historical analyses typically utilized the

device MAC address, a globally unique identifier, which could be used to identify and track individual devices. However, in 2014 manufacturers started implementing MAC address randomization on their devices to increase privacy,[6] making these previous analyses impossible because the globally unique identifier was removed.

The research described in this article focuses on extracting other identifying information to differentiate device signatures from probe requests since the broadcasted MAC address is no longer a unique identifier. The proposed identification system enables some of the previously described analyses, such as crowd analytics, to be conducted again. However, since the broadcasted MAC address is not the globally unique address specific to that device, the identified devices cannot be tied to an individual or the device itself without some additional intelligence (camera feeds, etc.). By identifying link-layer device signatures, a variety of analyses are possible to aid various missions, such as crowd analysis, threat detection, and anomaly detection. This work was completed under APL's Asymmetric Operations Sector Independent Research and Development program.

BACKGROUND

Since the onset of MAC address randomization, there have been many studies on techniques to circumvent MAC address randomization for identification and characterization purposes. Some studies have focused on reverse engineering randomization from probe requests to better understand the methods used.[7,8] While these techniques provide valuable insight into randomization techniques used by manufacturers, they are not generalizable since randomization is not standardized. Other studies have looked at active methods to differentiate devices, but these methods require interaction with the devices and a response, rather than using the passive frames already sent by the device.[9]

A variety of methods successfully analyze probe requests across different manufacturers. The regularity of interarrival time of frames is a popular method to differentiate devices by binning the times to recognize patterns for each device.[10,11] Cunche, Ali Kaafar, and Boreli look at the Service Set Identifiers (SSIDs), or network names, that devices are probing for to identify devices.[12] All these techniques have been successful with a small set of devices but have not proved successful for large, dense environments (50+ devices). Probe request analysis has also been used for purposes other than identification. Sequence numbers from probe requests have been used to identify MAC address spoofing by looking for multiple sequences occurring at the same time, which is indicative of multiple devices with the same address.[13] Vanhoef et al. track devices with probe requests using information element fingerprints, clustering, and sequence number techniques.[14]

LINK-LAYER IDENTIFICATION OF DEVICE SIGNATURES

Overview

As mentioned, link-layer data are unique in that they are very easy to collect, and with the explosion of smartphones in today's markets, there is an abundance of Wi-Fi data available for collection. By collecting the unencrypted link-layer Wi-Fi data broadcast from smart devices, signatures for those devices can be formed to differentiate the devices and estimate the number of smart devices present. Since there is no standardization for probe requests, the data can vary significantly depending on the operating system, manufacturer, device model, and user configuration and usage. With the many dimensions to these data, the Link-Layer Identification of Device Signatures (LLIDS) system provides a dynamic solution that is independent of device and manufacturer and does not require knowledge of the environment. This research uses a combination of machine learning techniques that are designed to handle large data sets that are more representative of public settings in today's world.

Data Set

The identification system was created primarily using the Sapienza data set from Community Resource for Archiving Wireless Data at Dartmouth (CRAWDAD) for development and testing.[15] The Politics 1 data set includes a packet capture collected during a political meeting in Rome, Italy, on February 22, 2013. The collection occurred before MAC address randomization, so the addresses in the capture can be used as ground truth of devices. MAC address randomization was then added to the data set for testing. For all the probe requests from a given device, the one true MAC address was replaced with anywhere from three to ten total addresses. The data set includes roughly 8 h of data and 1.7 million probe requests. It was split into 2-h blocks and then split into separate development and testing data sets.

Feature Extraction

As with any machine learning algorithm, the output is only as good as the input data it is provided. To assign each frame to a final device signature, identifying features needed to be extracted from each probe request. Each frame serves as the exemplar for all algorithms. To start, a large number of potential features were extracted from probe request frames, and the final list was chosen based on features with the highest average merit across the data set (i.e., features that provided the most identifying information to differentiate devices). Ultimately the features extracted as inputs to the different algorithms include frame length, radio duration, frame length subtracted by SSID length, channel, first byte

of address, signal strength, sequence number, time, data rate, resolved Organizational Unique Identifier (OUI), and SSID. If the OUI or SSID field is empty, the feature is set to an empty string. This research only uses fields required in the 802.11 probe request specification as features, rather than optional tagged fields, so the system will be compatible with future versions of 802.11 Wi-Fi.[4]

Identification System Design

The identification system concept is designed for scalability and robustness. The multilayer approach, shown in Figure 1, uses machine learning to break down large data sets into manageable sample sizes. Initial research using only clustering techniques found that devices of the same manufacturer were frequently grouped together, so additional layers were added to the system design. The system has three layers, as shown in Figure 1: (1) the manufacturer group classification layer, (2) the clustering layer, and (3) the device decision logic layer. The second layer uses unsupervised learning, also known as clustering techniques, to differentiate devices. The first layer was added to separate the frames by manufacturer group so the clustering algorithm can detect device-specific differences rather than manufacturer-specific differences. The third and final layer was added to ensure that the final device signature assignments are unique. This last layer uses a combination of state-of-the-art algorithms to break down the clusters into unique device identifications. The identification system uses the following steps with the assumption that there is no MAC address spoofing (multiple devices with the same MAC address):

1. Extract features from each probe request exemplar.
2. Classify the exemplar to a manufacturer group (layer 1).
3. Cluster all exemplars in a manufacturer group into clusters (layer 2).
4. Regroup exemplars with the same MAC address to the same cluster assignment.
5. Break down each cluster into unique device signatures by using a combination of state-of-the-art algorithms and decision logic (layer 3).

Layer 1: Manufacturer Group Classifier

Layer 1 is a multiclass supervised learning classifier designed to classify each probe request frame to a manufacturer group. The purpose of this classifier is to break data sets into manufacturer groups since the manufacturer is a high-level identifier for devices. By separating a large data set into smaller manufacturer-specific sets, the clustering algorithms in layer 2 can better detect the device-specific differences rather than the manufacturer-specific differences. The classifier was developed with 11 phone manufactures as the output classes. The manufacturers include Apple, Google, Hon Hai, HTC, Huawei, LG, Murata, Nokia, Samsung, Sony, and TCL. These manufacturers were chosen based on prevalence in data sets and market share.[16] While the results of this classifier provide some indication of the true manufacturer of the devices, the classifier is not optimized for profiling the manufacturer of devices since it is only designed with 11 classes.

The following features are used in the classification: frame length, radio duration, frame length subtracted by SSID length, channel, and first byte of address. These features are similar to those used in the clustering algorithms but optimized for classification accuracy. Several classification model types were tested, and a subset of the results is shown in Table 1.[17] As shown in the table, the ensemble – bagged trees model was the most accurate, with 89.7% accuracy, and was ultimately used in the identification system. The decision tree model was close in

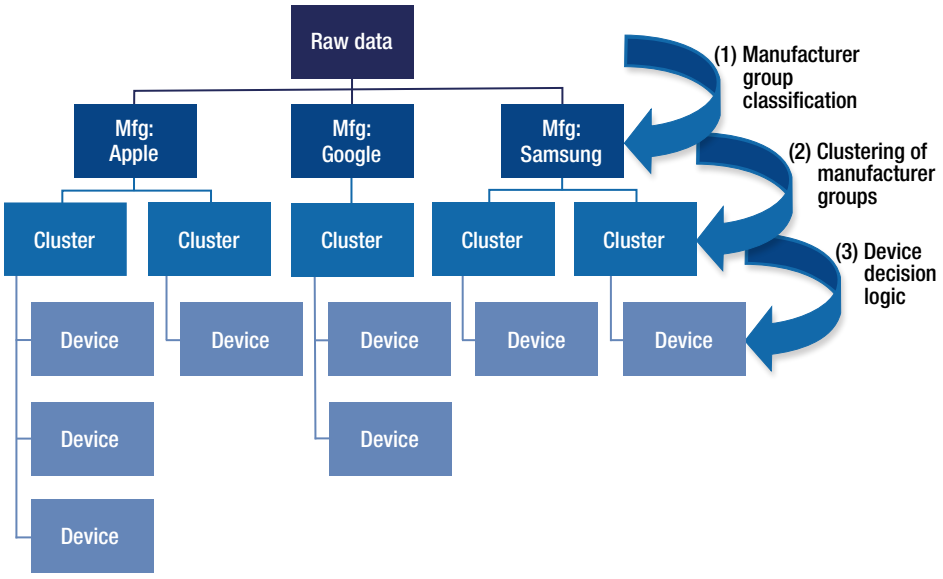

Figure 1. Identification system design. The multilayer approach uses machine learning to break down large data sets into manageable sample sizes. The system has three layers: (1) the manufacturer (Mfg) group classification layer, (2) the clustering layer, and (3) the device decision logic layer.

Table 1. Classification accuracy by model type

Classifier Model Type	Accuracy (%)
Ensemble – bagged trees	89.7
Fine decision tree	82.7
Gaussian naive Bayes	62.4
Cubic support-vector machine (SVM)	72.0

The ensemble – bagged trees model was the most accurate and was ultimately used in the identification system.

accuracy to the ensemble model, but the Bayesian and support-vector machine (SVM) models performed notably worse. This is likely because the features chosen are mostly non-Gaussian.

Layer 2: Clustering

The clustering layer is the foundation of this research, which aims to differentiate devices on a frame-by-frame basis. The clustering algorithm cannot require a number of clusters as an input since the number of devices is unknown and handling non-Gaussian features. The features used include frame length, radio duration, frame length subtracted by SSID length, first byte of address, and signal strength. Several algorithms were tested, but the density-based OPTICS (ordering points to identify the clustering structure) and agglomerative algorithms were ultimately chosen since they were the best fits for the algorithm requirements.[18,19] The OPTICS algorithm is the primary algorithm. However, the agglomerative algorithm requires all requests be assigned to a cluster, while OPTICS uses a "–1" assignment for outliers. This difference can be important for various use cases, so both were kept in the system for future modularity. Initial testing showed that an average of 75% of frames were regrouped to the same cluster compared with the worst-case scenario of 35% where the devices were assigned using their random MAC address. While these algorithms perform well, multiple devices are often still found in each cluster in dense environments. Thus, for scalability, another layer was added to separate the unique devices.

Layer 3: Device Decision Logic

The purpose of the third layer of the identification system is to ensure the uniqueness of the device signatures. The layer is designed to look at each cluster and determine whether there are multiple devices in the cluster and, if there are, separate them. This is done by calculating a similarity score between the data of all MAC addresses in the cluster and using a threshold to combine data above a certain similarity where the data are assumed to be from the same device. The similarity score is calculated using a weighted combination of the result of two algorithms: interarrival time similarity and SSID set similarity.[10,12] Both algorithms were derived from techniques found in published papers and implemented based on their documentation, but the implementation may not be identical to that from the paper. The result of this layer is that each probe request frame is given a device signature assignment that is unique to that run of the identification system. This final assignment can be used to determine the number of devices present for crowd analytics and other applications.

RESULTS

Identification System Performance

Since there are several layers and algorithms in this system, it is important to evaluate the system using several metrics. The main metrics of this system are the number of devices identified, homogeneity, and completeness. It is critical to look at all three metrics because it is possible to improve performance of one metric at the cost of another. Thus, all must be considered to ensure there are no performance trade-offs. The following results compare the LLIDS identification system with three algorithms found in the literature. The interarrival time similarity and SSID set similarity algorithms are discussed earlier. The sequence number correlation algorithm looks at the sequence number over time to look for discrepancies in the sequences (i.e., multiple sequences happening at the same time) to differentiate devices.

Figure 2 shows the average number of identified devices for the different techniques compared with the true number of devices in the data set, the goal. The data sets ranged from 10 to 1,000 devices, which is represented logarithmically on the *x* axis. LLIDS, interarrival time similarity, and SSID set similarity all perform well by closely estimating the number of true devices, shown by the bottom black line. The sequence number correlation algorithm performs poorly by estimating the worst-case result, the total number of MAC addresses present, which includes randomization and is significantly more than the number of devices. This plot demonstrates LLIDS's ability to accurately identify the number of devices present, which is important for analyses such as crowd analytics among others.

The two plots in Figure 3 show the homogeneity and completeness scores for the identification techniques for the differently sized data sets.[20] Each score is measured from 0 to 1, where a score of 1 is ideal. These scores are commonly used to evaluate the performance of unsupervised learning algorithms. Completeness is a measure of whether all frames of a device are in the same signature (i.e., ability to regroup all frames from a device). Homogeneity is a measure of whether each signature contains only frames of a single device (i.e., ability to identify unique signatures). Figure 3, left, shows that the

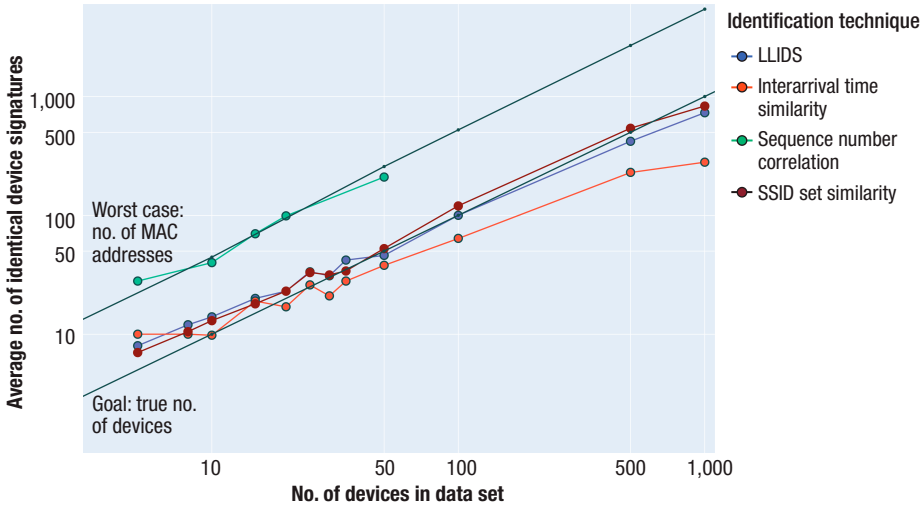

Figure 2. Number of identified devices vs. true devices. LLIDS, interarrival time similarity, and SSID set similarity all perform well by closely estimating the number of true devices. This demonstrates LLIDS's ability to accurately identify the number of devices present.

completeness score for all the algorithms is comparable. However, Figure 3, right, shows significant performance differences among the algorithms' homogeneity scores. The sequence number correlation algorithm performs the best but at the cost of accurately estimating the number of devices, as shown in Figure 2. This indicates that the algorithm is doing well at differentiating the devices but is doing so by separating devices into many different signatures. On the other end of the spectrum, the interarrival time and SSID similarity algorithms accurately estimate the number of devices but at the cost of homogeneity. This indicates that these algorithms are mixing probe request frames across different signatures, so the results are not unique. LLIDS performed relatively well, with an average homogeneity score of 0.8 and a completeness score of 0.7, demonstrating the system's ability to accurately identify individual device signatures. The LLIDS system's good performance across all three metrics indicates that there are no clear performance trade-offs and that the identification system is viable for many different use cases.

Potential bias in the system was also studied. First, the breakdown of performance defined by the homogeneity and completeness for each manufacturer group was analyzed to ensure that the identification system is not biased toward certain manufacturers. For example, while Apple devices are much more prevalent (roughly 50% of devices in the CRAWDAD data set), their performance was similar to that of other manufacturer groups. Thus, there was no significant difference in performance between manufacturer groups despite significant differences in frequency in the data set. Second, the amount of data available for each device was analyzed to ensure that the system is not biased toward devices with more data or more frequent probe requests. The system performed well for devices with as little as five samples, with an average score of 0.8 for completeness and 0.98 for homogeneity. In fact, the system performed worse with large amounts of data (100+ samples) for a given device, where completeness decreased while homogeneity increased. This is likely caused by the clustering algorithms. Since the algorithms are density based, the device is split into multiple clusters when there are large amounts of data available, as observed in the completeness and homogeneity scores. Because of this finding, it is recommended that the system be run on mid-sized batches of data where fewer than 100 samples are available for a device. Depending on the density of the environment, this could be between 15 and 90 min.

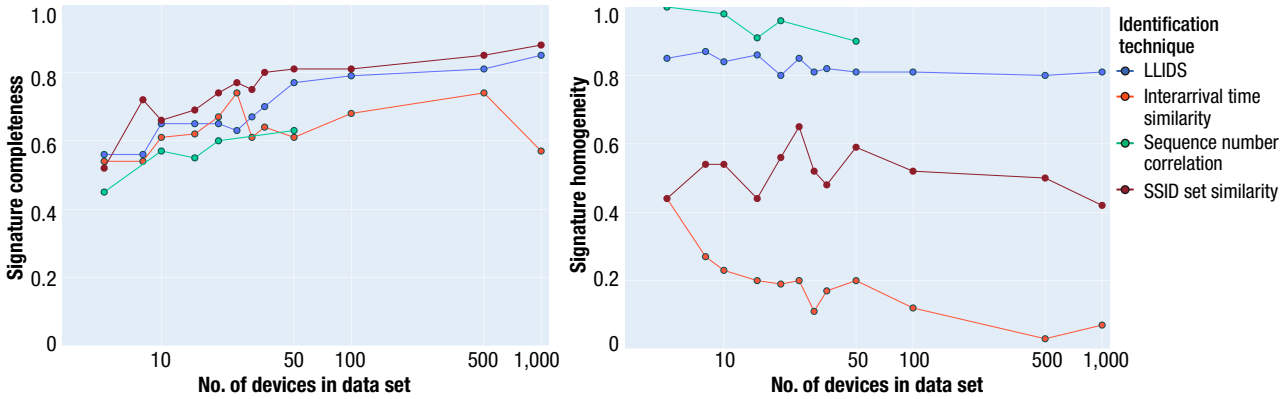

Figure 3. Signature completeness (left) and homogeneity (right). LLIDS performed relatively well, with an average homogeneity score of 0.8 and a completeness score of 0.7, demonstrating the system's ability to accurately identify individual device signatures.

Crowd Estimation Demonstration

The results using the Sapienza data set proved promising, but since the data were from 2013, additional testing was required to ensure that LLIDS can also identify modern devices. A collection was conducted on APL's campus in a building lobby where staff members are required to swipe their badges to gain entry. APL has a substantial interest in advanced technologies to maintain the security of its property, and this building lobby is a prime location for assessing the potential benefits of the LLIDS identification system in an environment that contains modern devices deploying MAC address randomization. The results of the collection were provided to APL's head of facilities security.

To protect the privacy of the device owners, the MAC addresses and SSIDs were hashed to obscure any potential identifier that could tie the identified device signatures to the owners. In addition, the packet captures collected were deleted after being processed into feature tables, so the raw data cannot be used for other purposes. Probe requests were collected for a week in 2-h blocks using a small, single-board computer in monitoring mode. Each 2-h block was processed through the LLIDS identification system and interarrival time identification algorithm. These results were compared with the unique staff count in the area collected from APL's badge reader data, which serves as the ground truth, and the number of MAC addresses, which serves as the worst case during each 2-h block.

Table 2 outlines the comparison of the target staff count with the different identification methods given by the total count from the identification method divided by the target staff count. The LLIDS output is very close to the target, with an average 1.2 times the staff count. This result would likely be improved by removing the devices found in the space at all times, such as laptops and other smart devices such as TVs. The interarrival time algorithm overestimates the target staff count by 14 times. The number of unique MAC addresses observed through the week severely overestimates the number of staff members, with an average 43 times more addresses than staff members. This demonstrates that the LLIDS system closely estimates the number of true devices (estimated here by the staff count), especially compared with the MAC address count, which grossly overestimates the number of devices and is the main identifier available today.

Table 2. Comparison of identification methods

Identification Methods	Comparison to Staff Count
LLIDS signatures	1.2×
Interarrival time signatures	14×
MAC addresses	43×

The LLIDS system closely estimates the staff count, especially compared with the MAC address count, which overestimates the number of devices.

The total numbers of identified LLIDS signatures and unique staff count for each 2-h collection are shown in Figure 4. The LLIDS identification system results closely correlate to the number of staff members present, which indicates that the system can accurately estimate the number of devices present. Provided with just the LLIDS results, one can see clear crowd trends from the week of the collection, including the high peaks, which indicate the workdays compared with the weekend days, and the large increases and decreases, which indicate the arrival and departure times of staff members.

DISCUSSION

Wi-Fi Sensing for Public Safety

This research has many potential applications, including Wi-Fi sensing for human movement feedback. Understanding human movement after emergency alerts are issued is important to determine the effectiveness of alerts and strategies for future alerts.[21] Wi-Fi sensing with LLIDS can be used for crowd estimation to provide feedback about human movement and the efficacy of alerts. This information would be delivered

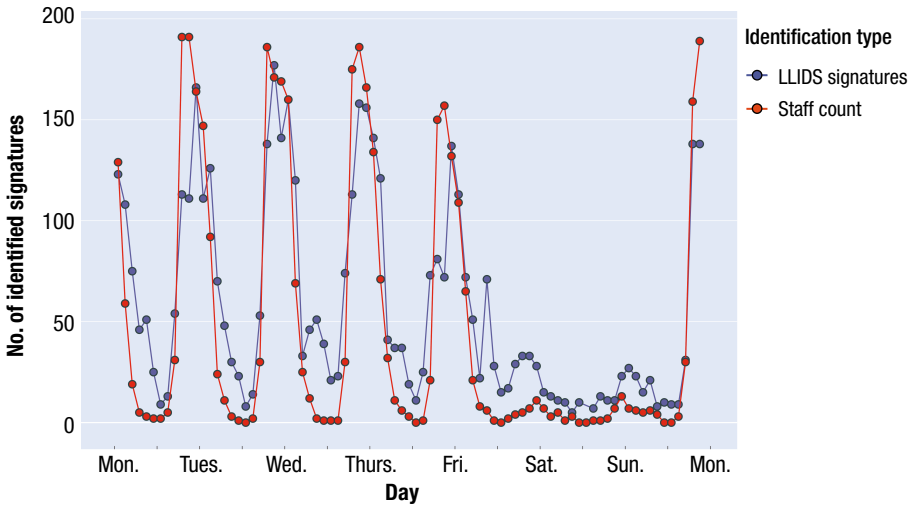

Figure 4. LLIDS signatures vs. unique staff count. The LLIDS identification system results closely correlate to the staff members present, and provided with just the LLIDS results, one can see clear crowd trends from the week of the collection.

in near real time and could provide situational awareness of movement to assist decision-making related to future alert implementation. Since this technique uses Wi-Fi signals, which have a range of roughly 60 m, this method is not appropriate for analyzing large-scale areas such as states or interstate highways. Instead, focusing on high-impact, high-density areas such as city hot spots, public transit locations, bridges, toll stations, and even gas stations would provide important indicators of the public's movement. Observing these high-impact areas would provide decision-makers with near-real-time feedback on the movement of crowds and the effectiveness of alerts. In addition to identifying movement after alerts, this observation system could also notify officials of other emergencies, such as terrorists threats or accidents, by discerning anomalies in human movement at these locations that are consistent with public emergencies.

Next Steps

The team is working with various collaborators to continue development and to apply the identification system to specific mission use cases such as the public safety case discussed. The next steps for development depend on the mission application and needs, but several recommendations are listed here. For crowd analysis use cases such as the public safety example, more testing with larger data sets representative of dense, high-impact environments with modern devices is recommended. Other missions may require building out specific algorithms for analysis and deploying units for collection. A near-real-time system will be of interest to groups with a need for rapid feedback. To collect data in near real time, the LLIDS system would need to be set up to run batch processing using a rolling time window to get frequent updates and enough data for accurate results. Many missions will also need to correlate results from different identification batches since the signatures found in the current LLIDS system are unique to that run. This correlation would allow for more accurate identifications and understanding of long-term analytics. The correlation algorithms would require a similar multilayer design but would compare device signature objects instead of the individual probe request frames. For data collection use cases, a hardware study is encouraged to survey various potential collection devices and determine the best devices and methods. The network interface of the small, single-board computer used for this research timed out after a few hours and required a reboot, which resulted in lost data during the downtime. The team will explore extensions of the frame/packet-based machine learning techniques of the research, such as Bluetooth or cellular device identifications. These applications require different features and optimized algorithms but would use the fundamentals of the research.

SUMMARY

Unsupervised learning methods, such as clustering, in combination with additional algorithms proved to be a successful technique for differentiating devices of various manufacturers in dense environments. The multilayer identification system successfully breaks down large data sets and more accurately identifies device signatures. More development and testing is necessary to create confidence in the identification system's use for specific missions. However, the system's initial results for crowd analysis are very promising. With the increasing prevalence of wireless devices in people's lives, the LLIDS system provides an important technique for crowd analytics for use cases such as public safety and disaster relief, among others.

REFERENCES

[1]Practical Networking, "OSI model," https://www.practicalnetworking.net/series/packet-traveling/osi-model/ (accessed Sep. 1, 2021).

[2]Transforma Insights, "Number of Internet of Things (IoT) connected devices from 2019 to 2030 (in millions), by region," statista, 2020.

[3]Pew Research Center, "Mobile technology and home broadband 2021," 2021.

[4]IEEE, "IEEE 802.11 wireless local area networks," https://www.ieee802.org/11/ (accessed Sep. 1, 2021).

[5]L. Sweeney, "My phone at your service," Feb. 2014, https://www.ftc.gov/news-events/blogs/techftc/2014/02/my-phone-your-service.

[6]L. Mirani, "A tiny technical change in iOS 8 could stop marketers spying on you," Jun. 9, 2014, https://qz.com/218437/a-tiny-technical-change-in-ios-8-could-stop-marketers-spying-on-you/.

[7]J. Freudiger, "Short: How talkative is your mobile device? An experimental study of Wi-Fi probe requests," in *WiSec'15*, New York City, 2015.

[8]B. Misra, "MAC randomization—Analyzed!," AirTight Networks, http://blog.mojonetworks.com/ios8-mac-randomization-analyzed/ (accessed Sep. 1, 2021).

[9]S. Bratus, C. Cornelius, D. Kotz, and D. Peebles, "Active behavioral fingerprinting of wireless devices," in *WiSec'08*, Alexandria, VA, 2008.

[10]J. Franklin, D. McCoy, P. Tabriz, V. Neagoe, J. Van Randwyk, and D. Sicker, "Passive data link layer 802.11 wireless device driver fingerprinting," in *Secur. '06: 15th USENIX Secur. Symp.*, 2006.

[11]C. Matte, M. Cunche, F. Rousseau, and M. Vanhoef, "Defeating MAC address randomization through timing attacks," in *ACM WiSec 2016*, Darmstadt, Germany, 2016.

[12]M. Cunche, M. Ali Kaafar, and R. Boreli, "Linking wireless devices using information contained in Wi-Fi probe requests," in *Pervasive and Mobile Computing*, Elsevier, 2013.

[13]J. Wright, "Detecting wireless LAN MAC address spoofing," Johnson & Wales University, 2003.

[14]M. Vanhoef, C. Matte, M. Cunche, L. S. Cardoso, and F. Piessens, "Why MAC address randomization is not enough: An analysis of Wi-Fi network discovery mechanisms," in *ASIA CCS '16*, Xi'an, China, 2016.

[15]M. Barbera, A. Epasto, A. Mei, S. Kosta, V. Perta, and J. Stefa, "The sapienza/probe-requests dataset (v. 2013-09-10)," A Community Resource for Archiving Wireless Data at Darmouth (CRAWDAD), Sep. 10, 2013, https://crawdad.org/sapienza/probe-requests/20130910/Politics1/index.html.

[16]Counter Point Weekly, "Global smartphone market share: By Quarter," Aug. 5, 2021, https://www.counterpointresearch.com/global-smartphone-share/.

[17]MathWorks, "Classification," https://www.mathworks.com/help/stats/classification.html?s_tid=CRUX_lftnav (accessed Sep. 1, 2021).

[18]scikit learn, "sklearn.cluster.OPTICS," https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS.html (accessed Sep. 1, 2021).

[19]scikit learn, "sklearn.cluster.AgglomerativeClustering," https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html (accessed Sep. 1, 2021).
[20]scikit learn, "Clustering—Homogeneity, completeness and V-measure," https://scikit-learn.org/stable/modules/clustering.html#homogeneity-completeness-and-v-measure (accessed Sep. 1, 2021).
[21]IPAWS Program Management Office, "Integrated Public Alert and Warning System (IPAWS): Strategic plan: Fiscal year 2014-2018," Federal Emergency Management Agency, 2010.

Jennifer A. Finley, Asymmetric Operations Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Jennifer A. Finley works as a networking and communications engineer in the Mission Critical Communications Group of APL's Asymmetric Operations Sector. She has a bachelor's in electrical engineering from Lehigh University and a master's in electrical engineering with a concentration in telecommunications from Johns Hopkins University. Jennifer has experience in test and evaluation, network and protocol analytics, link-layer analytics, and 4G and 5G systems. Her email address is jennifer.finley@jhuapl.edu.