Computer Networks 170 (2020) 107103

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Longitudinal analysis of a campus Wi-Fi network

José Camacho [a,*], Chris McDonald [b], Ron Peterson [c], Xia Zhou [c], David Kotz [c]

[a] Department of Signal Theory, Telematics and Communications, School of Computer Science and Telecommunications - CITIC, University of Granada (Spain)
[b] Department of Computer Science and Software Engineering, The University of Western Australia, Perth 6009, Australia
[c] Department of Computer Science, Dartmouth College, Hanover, NH 03755, United States

ARTICLE INFO

Article history:
Received 27 June 2019
Revised 3 January 2020
Accepted 10 January 2020
Available online 15 January 2020

Keywords:
Wireless network
Wi-Fi
802.11
SNMP
Network analysis
Network modeling

ABSTRACT

In this paper we describe and characterize the largest Wi-Fi network trace ever published: spanning seven years, approximately 3000 distinct access points, 40,000 authenticated users, and 600,000 distinct Wi-Fi stations. The 7TB of raw data are pre-processed into connection sessions, which are made available for the research community. We describe the methods used to capture and process the traces, and characterize the most prominent trends and changes during the seven-year span of the trace. Furthermore, this Wi-Fi network covers the campus of Dartmouth College, the same campus detailed a decade earlier in seminal papers about that network and its users' network behavior. We thus are able to comment on changes in patterns of usage, connection, and mobility in Wi-Fi deployments.

1. Introduction

Now a mature technology, Wi-Fi (IEEE 802.11) plays an essential role at the edge of the Internet. Despite countless studies about its capabilities, behavior under various network conditions and use cases, much remains unknown about how large enterprise Wi-Fi networks behave when faced with the traffic demands of thousands of daily users. Although researchers have characterized early Wi-Fi networks [1–4], the technology and usage has evolved dramatically in the two decades since 802.11 was first introduced.

In this paper, therefore, we publish and describe a massive trace of a live, production Wi-Fi network, on the same campus that was first documented in the Dartmouth traces from 2001 [1,2] and 2003 [3,4], to provide new insights into the changes resulting from the evolution of mobile client technology, Wi-Fi network technology, and user behavior. We describe a foundational characterization of a *seven-year* capture of the Dartmouth campus, providing a platform on which others can explore deeply in a variety of directions. Understanding and forecasting user behavior and connectivity patterns in Wi-Fi is important for network managers, protocol designers and software developers, to improve current practices of network design and improve network performance [5,6].

Specifically, this paper makes four primary contributions:

- We analyze a trace with a 7-year span of time that includes 5 billion records regarding the activity of approximately 3000 distinct access points (APs), 40,000 authenticated users, and 600,000 distinct Wi-Fi stations. Unfortunately, we found a gap of approximately two years where a significant amount of information was missing. Therefore, the actual period useful for most analyses is 5 years.
- We identify active device-connection sessions across time, a process fundamental to the proper interpretation of connection patterns.
- We discuss in detail current usage patterns and trends in the campus Wi-Fi, both in terms of users and devices, including user mobility and device manufacturer distribution. We derive updated models for the realistic simulation of Wi-Fi environments.
- We release a data set with anonymized information about connection sessions in 2018, for which the amount of traffic and duration of connection sessions is available, and another data set with anonymized information about association traps in 2012–2018.

The next section introduces the network and our methods for data capture and analysis. Section 3 describes our method to identify connection patterns, and Section 4 analyzes the network infrastructure and number of connections over time. Section 5 focuses on usage patterns and Section 6 analyzes mobility patterns. Section 7 contrasts results with those in previous papers on the Dartmouth Wi-Fi and present models of usage of the network,

* Corresponding author.
E-mail address: josecamacho@ugr.es (J. Camacho).

https://doi.org/10.1016/j.comnet.2020.107103
1389-1286/

Table 1
General statistics across papers reporting on the Dartmouth Wi-Fi network.

Concept	Kotz & Essien '05 [2]	Henderson et al. '08 [4]	Camacho et al. '19
Capture Features			
Year	2001	2003–2004	2012–2018
Length	11 weeks	17 weeks	7 years
Measurement technologies	Syslog + SNMP polling + tcpdump sniffers	Syslog + SNMP polling + tcpdump sniffers + IP telephony records	SNMP traps
# Entities			
Stations (/week)	1706 (155)	7134 (420)	624,903 (1716)
Users	< 1706 (estimated)	< 7134 (estimated)	38,096 (authenticated)
APs	476	566	3330
Buildings	161	188	200
SSIDs	1	1	20
Average Density			
APs/Building	3.0	3.0	16.7

which can be used in simulation. Section 8 discusses related work, and Section 9 summarizes our conclusions. The Appendices discuss technical details on the analysis of the capture and the data sets released.

2. Environment and data capture

Dartmouth College is a small liberal-arts university in a rural part of northeastern United States, on a campus comprising over 200 buildings on 200 acres. At the end of 2018 there were approximately 6500 students, 3300 staff, and 1000 academic faculty affiliated. The preceding seven-year period saw over 12,000 Dartmouth-affiliated users connecting to the Wi-Fi network. Dartmouth installed the first-ever campus-wide Wi-Fi network in 2001 by deploying 476 APs, and researchers reported on its usage in a 2002 paper [1,2]. In late 2003 there were 566 APs in the network, and researchers reported on the changing usage patterns [3,4]. These early networks were relatively simple, with no authentication required and only a single Service Set Identifier (SSID). By the end of 2018 the Dartmouth network included over 3000 APs, sophisticated authentication, and up to 20 different SSIDs. In our seven-year capture, the dominant SSIDs include *Dartmouth Secure* (the WPA2-Enterprise authenticated college network), *Dartmouth Public* (an unsecured public-access network), and *eduroam* [7] (for students and faculty visiting from other universities). *Dartmouth Secure* was entirely replaced by *eduroam* at the end of the capture, allowing us to analyze this evolution. We chose to limit our study to the recent seven-year period because the network infrastructure (comprising Cisco network controllers and access points) was reasonably consistent throughout those seven years.

Table 1 compares the main features of our capture with those in the previous papers. This comparison is limited, though, by the fact that the papers used different measurement technologies and analysis strategies. For example, the recent capture has limited information about traffic volumes (bytes in/out per station or per AP), whereas the early papers had detailed data about traffic volume. The recent capture has no information about the mix of application types, whereas the early papers used tcpdump to capture packet headers on a representative subset of the campus. The recent capture has information about user identity (username) whereas the early papers had only network identity (MAC and IP addresses). Overall, the recent trace is much larger: longer time duration, greater number of APs and client stations, and there are even more buildings covered by the network.

To collect the recent trace, the Dartmouth network operators configured the Cisco network controllers to forward a record of network activity to the research team's servers in the form of Simple Network Management Protocol (SNMP) traps [8]. Unfortunately, a significant number of traps were not captured from the fourth

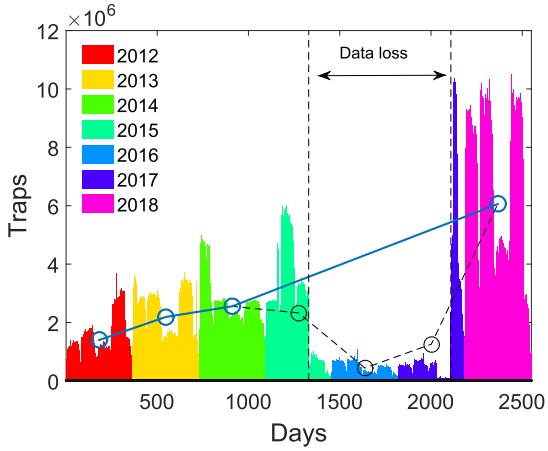


Fig. 1. Number of traps per day, over time. In this and following figures, the vertical dashed lines identify the period where data loss was experienced, and circles represent yearly averages.

quarter of 2015 to the third quarter of 2017. Dartmouth's network staff confirmed that, during this period, most of the Wi-Fi controllers were misconfigured and were not sending traps. Although we must omit this period for most of our analysis, we present the available data for completeness; in relevant figures, the affected period is marked with two vertical dashed lines. The same figures also present trends in yearly averages using both a dashed curve (connecting all years) and a solid curve (omitting this period of data loss). Note that figures have varying scales on their x- and y-axes, some with large magnitudes indicated with exponential notation.

Fig. 1 depicts the evolution of the number of traps in the capture; over seven years the network created an average of approximately three million traps per day, although the number grew steadily from 2012 through 2018. Within each year it is also possible to see coarse fluctuations corresponding to the academic calendar: periods with a high number of traps corresponding to winter, spring, and fall terms, and a quieter period corresponding to summer term. Deeper dips are sometimes visible around the late-December holidays.

The seven-year capture is thus a trace comprising a sequence of records ("traps"); each record includes a trap type (TT) and a set of fields labeled with object identifiers (OIDs). (For details, see Appendix A.) We use this data to identify who associated to the network, when the association took place, the device and APs involved in the connection and, thus, the approximate location and movement of each device and user throughout the capture. We fur-

ther use the data to extract a sequence of network *sessions*, as described below; for a subset of sessions we can identify the amount of traffic generated. No information about the traffic content or protocols is available.

3. Connection sessions

To identify usage patterns and understand the data, we require a rigorous definition of a connection *session*, that is, some notion of the period when a user's station is connected to the network. When using WPA2, stations are first *authenticated* to the network and then *associated* to an AP. In this process, there is an exchange of protocol messages, as defined in the 802-11i and 802-11r standards [9,10]. When the session is finished, a station may disassociate and deauthenticate from the network. In the event that it does not, the AP (or its controller) terminates the session after a configurable period of inactivity.

While the association process is clear, detecting sessions in a data capture can be challenging, particularly in a multi-SSID environment. Further, not all steps may trigger a SNMP trap, or the trap may not reach the collecting server, and it is difficult to distinguish the establishment of new sessions from roaming events within the same session.

To overcome this problem, previous work estimated that a session starts every time an association message is received, provided at least 30s has elapsed since the previous such message [1–4]. However, there was no sound justification of this specific time threshold.

In this work, we leverage an OID called *SessionID*, which appears in certain types of traps. According to Cisco's definition [11] the "*Session ID feature allows a single session identifier to be used for all ... authenticated sessions*". Thus, the SessionID is expected to unequivocally identify an authenticated session. Unfortunately, SessionID traps were not regularly produced prior to 2018 and, even then, there were many missing traps. Instead, we use the *Association* trap type, available throughout the seven-year capture, and track sessions by noting the station MAC in each such trap.

We thus use the SessionID traps to analyze the duration and amount of traffic and number of connections in 2018, and then the station MAC in association traps to generalize the analysis to the prior years. This are also the two data sets that we put available for the community.

3.1. Traffic and duration of SessionIDs in 2018

Fig. 2 shows the number of SessionIDs in terms of duration and volume of traffic. The analysis is limited to SessionIDs in 2018 with available starting and end times in the capture. Interestingly, a majority of SessionIDs have a duration between 5 and 7 min and consume tens of KBytes of traffic. We can also see very short connections of between 10 and 30 s and with no reported traffic, and longer SessionIDs, with a duration of several hours and high traffic volumes.

We can achieve a better understanding of the previous figure if we distinguish according to the main values of "Reason Code" in the disassociation trap: 64% of SessionIDs finish with Reason Code 4 ("Disassociation due to inactivity") and 24% with Reason Code 2 ("Previous Authentication no longer Valid"). Fig. 3 shows the percentages of SessionIDs separately in terms of duration and volume of traffic, including SessionIDs with Reason Code equal to 2 and 4. If we focus on the SessionIDs terminated due to inactivity in Fig. 3(a), there is a clear boost at 5 min, suggesting that the inactivity threshold for most SSIDs was configured to be 5 min. We can also see in Fig. 3(b) that the vast majority of connections with no traffic have Reason Code 2, and thus are attributed to an authentication problem. We conclude that these 'sessions' do not represent true, active connections. In a separate analysis (not shown), we also observed that some devices generate several overlapping SessionIDs with no traffic, particularly if the device roams between two APs during association. This effect may be related to the aggressive association behavior observed in earlier work [2]. Since unsuccessful connection attempts generate traps, and sessions generate multiple traps, the number of traps is much higher than the number of active sessions.

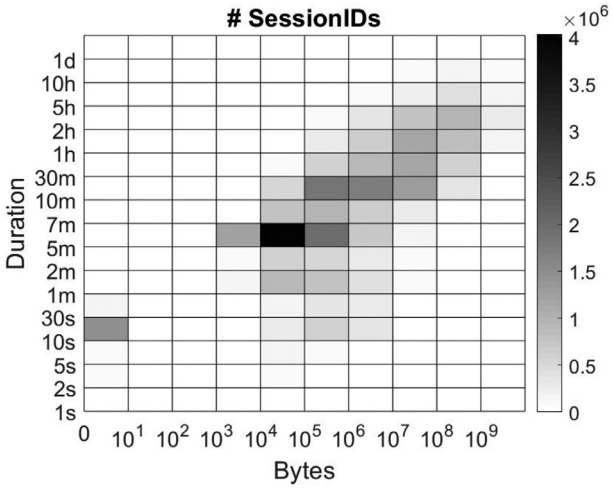


Fig. 2. Heatmap of SessionIDs in terms of duration and bytes of traffic.

The large majority of connections were terminated due to inactivity and were, therefore, shorter than their calculated duration because the station had to be inactive for at least five minutes before the disassociation occurred. In Fig. 4 we present a close-up view of the distribution of SessionIDs with durations between 5 and 10 min. We can see that the number of SessionIDs are more evenly distributed than in Fig. 2, as a result of the use of shorter bins. This even distribution is more realistic considering the disparate ways of using the network, and leads us to the conclusion that the true duration of most user sessions is actually 5 min shorter than the corresponding duration of the SessionID, given that 5 min has been determined as the inactivity threshold. If we subtract 5 min from the SessionIDs with longer duration and terminated due to inactivity, we observe that about 29% of these are shorter than one minute. Furthermore, some longer SessionIDs may be masking the appearance of several shorter user sessions. This would happen when the time between sessions is below the 5 min threshold, so that the device is not disassociated.

We hypothesize that these short sessions are caused by smartphone applications (or systems) that need to make quick connections to Internet services, e.g., to check for new messages or to update their status. Although our traces do not contain the information needed to verify this hypothesis, there is ample evidence in the research literature that smartphones and other mobile devices often make short Wi-Fi network connections. For example, the 2012 DozyAP paper found long inter-packet arrival times in a study that led them to optimize Wi-Fi energy consumption [12]. A 2011 study of handheld (and non-handheld) Wi-Fi devices found small flow sizes were common [13]. This was also confirmed in a recent 2018 study [14].

3.2. Characterizing connection sessions in the entire capture

Inspired by the previous analysis, we analyze the sessions in the entire capture using the following definition: a *session* is a period of time during which the station appears frequently in the trace, such that the temporal gap between any two adjacent occurrences

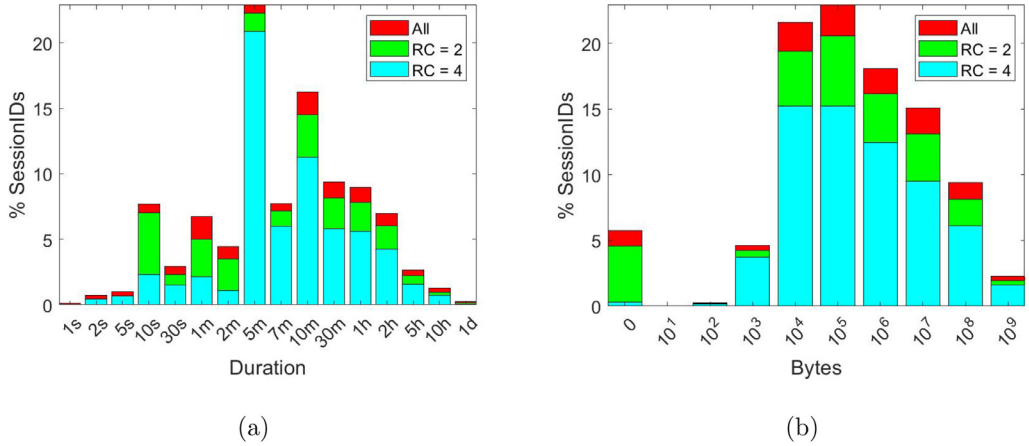


Fig. 3. Histogram of percentages of SessionIDs in terms of duration (a) and bytes of traffic (b). Plots include all SessionIDs and those with Reason Code 2 ("Previous Authentication no longer Valid") and 4 ("Disassociation due to inactivity"). Each bin contains the number of SessionIDs with duration or volume of traffic between the marked one and the one in the next bin.

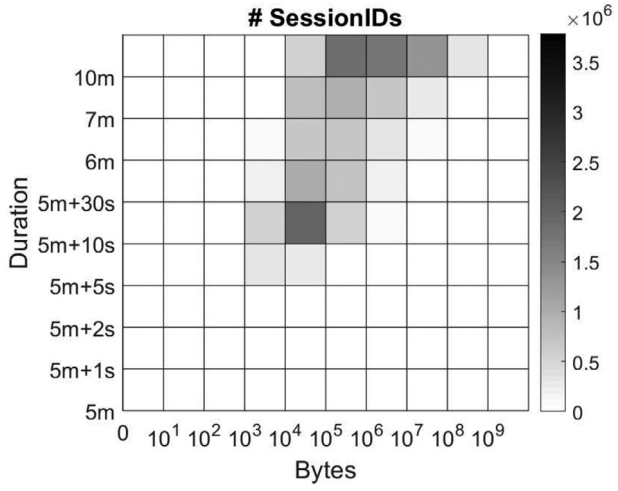


Fig. 4. Heatmap of SessionIDs in terms of duration and bytes of traffic: zoom between 5 and 10 min.

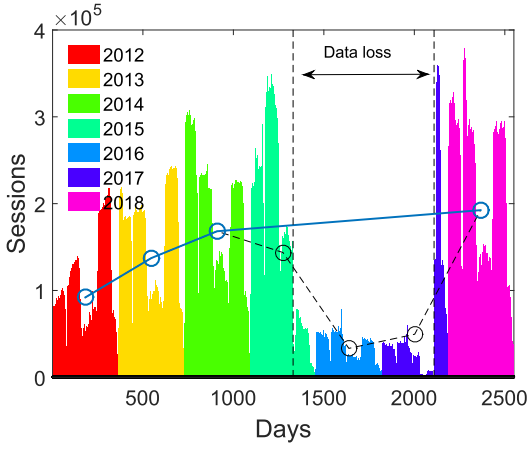


Fig. 5. Number of sessions per day, over time.

is no more than some constant d; we set d to 5 min. Here we limit our analysis to association traps, which were consistently produced across the seven years.

In Fig. 5 we show the evolution of the number of sessions in time. The plot shows an almost steady number of sessions since 2014, despite a jump in the number of traps (as noted in Fig. 1). Indeed, the network staff confirmed that there was a 2017 change in trap configuration, which caused the increase in the number of traps per session. This variability in trap configurations emphasizes why it was essential to consider *sessions*, rather than *traps*, to properly interpret connection patterns.

Fig. 6 compares the daily evolution of active sessions between weekdays and weekends for two representative years. Contrast this plot with the results from 2003–04, which show a similar daily profile on weekdays and weekends for active cards (stations) [3,4]. Active cards per hour are not exactly the same as active sessions per hour, but both numbers are expected to be similar, given that the hourly median of sessions per station was below 1. Fig. 6 shows that the daily activity approached a minimum during the night, consistently across the week. It also reached a weekday maximum, with around 15,000 sessions on average, at midday. In 2001 and 2003, respectively, there were only 500 and 1400 active cards in the busiest hour, and 200 and 400 in the least active hour [4]. This 10-fold increase in the last 15 years reflects an increase in the number of users and number of devices per user, but there are other reasons. Modern mobile devices, such as smartphones, transparently connect to Wi-Fi even when not in use – e.g., to download new mail or to receive inbound Messages – whereas 2001–04 devices (laptops) only connected to Wi-Fi when their user was seated with laptop opened for work.

4. Infrastructure analysis

Fig. 7 presents the number of active APs across the capture, and provides more information on the data loss in 2015–17. During that interval, the collected traps correspond to a reduced set of APs. If we discount that interval, we can observe the increased incorporation of APs to the network.

On average, there were fewer than 100 sessions per AP and day (not shown), and this number dropped in 2018, as expected, due to the deployment of a large number of APs in 2016–17. We ranked the APs according to their average number of sessions per day in Fig. 8(a). The ranking was performed for each year, so all years show a monotonic decrease. The figure shows the 500 highest-ranked APs. Because each AP is assigned to a campus building, we repeated the same computation for the buildings in Fig. 8(b), where the 100 highest ranked-buildings are shown. The number of observed sessions depends on the location and role of the building where the AP is deployed, with some APs and buildings much busier than others (notably, libraries and student centers). The load of

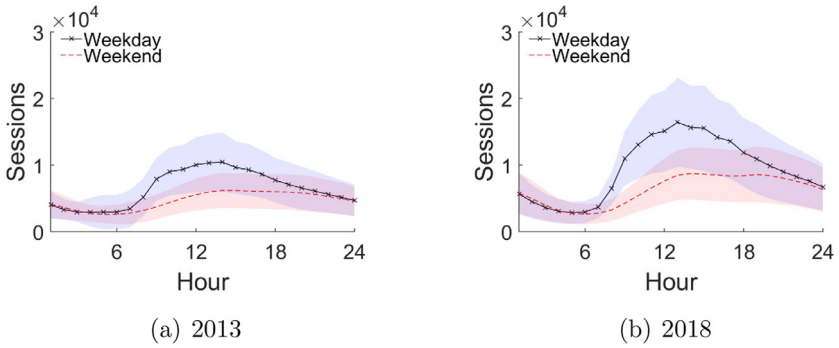


Fig. 6. Average number and standard deviation of sessions per hour during the day.

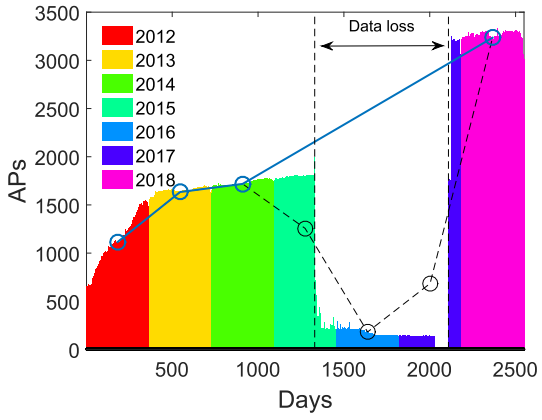


Fig. 7. Number of APs per day, over time.

Table 2
Average number of sessions per day of the 10 highest ranked APs in 2018.

AP	Building Type	2012	2013	2014	2018
Collis-2-4-AP	Student Center	0	642	2527	4733
Baker-3-4-AP	Library	1301	964	2113	3,356
Carson-3-2-AP	Academic/library	126	490	1613	2,723
Thayer-113-10	Academic	0	0	0	2,648
Baker-2-11-AP	Library	0	0	0	2,483
Baker-2-2-AP	Library	2092	1135	867	2,281
53commons-2-6-AP	Dining Hall	716	375	991	2,075
53commons-2-4-AP	Dining Hall	753	424	1082	2,051
BerryLib-3-7-AP	Library	1488	959	1643	2,041
BerryLib-3-13-AP	Library	0	0	0	2,009

the busiest APs (in terms of number of sessions per day) has doubled since 2014.

Combining Figs. 5, 7 and 8 we can see that the number of sessionIDs increased over time – first from 2012 to 2014 and then more steadily. At the same time, more APs were deployed. This additional deployment served to reduce the load in most locations, but not for the busiest APs, which continued to increase their load. Therefore, we can conclude that the deployment may not have been effective in managing load in those locations.

The busiest APs and buildings, according to their 2018 ranking, are listed in Tables 2 and 3, which also show the change for those APs and buildings across the capture. Libraries and the student center top both lists, as expected. Several of the busiest APs have been recently deployed, as indicated by their lack of sessions in earlier years. In Fig. 9, we show the cumulative density function (CDF) of the number of sessions, distributed across APs, for each of the busiest buildings. We can see that the distribution of the load can be heavily skewed in some buildings, and that the distribution can be very different from building to building. In some buildings, the deployment appears to have been focused on coverage rather than the demand.

In Fig. 10 we show the ranking of SSIDs according to the average number of sessions. Basically, traffic is dominated by three SSIDs: Dartmouth Secure, Dartmouth Public and eduroam. The latter was in its infancy in the campus in 2014, but by the end of the capture it had completely replaced Dartmouth Secure. We can see a clear dominance of the two authenticated networks in the last four years (74% of sessions) in comparison to the main public networks (below 19%, including Dartmouth Public and Dartmouth Library Public).

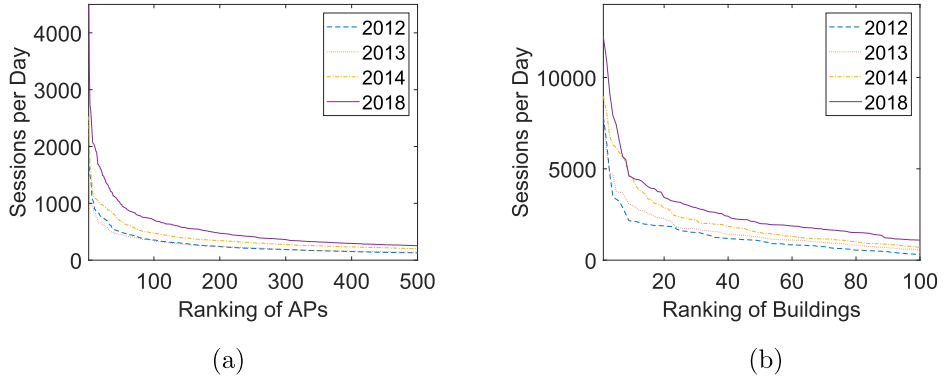


Fig. 8. Busiest (a) 500 APs and (b) 100 buildings, according to their average number of sessions per day. The ranking is performed for each year, with the maximum of 4733 sessions in a single AP and 12,200 sessions in a single building, both seen in 2018.

Table 3
Average number of sessions per day of the 10 highest ranked buildings in 2018. *In 2012 it appears that the logging infrastructure may not have yet included Collis, which certainly had Wi-Fi coverage.

Building	Type	2012	2013	2014	2018
BerryLib	Library	7844	7875	8967	12,204
Baker	Library	6519	6008	8124	10,976
Collis	Student Center	0*	3754	6315	9,132
53commons	Dining Hall	3481	3690	5868	7910
Hopkins	Arts/Dining/Student Center	4715	4746	5734	7,430
Webster	Library	1540	4694	6763	6,521
Topliff	Residence	2522	2916	4126	5,668
Kemeny	Academic	2075	3080	3341	5,514
Alumni	Gymnasium	433	2304	3097	4,614
Cummings	Academic	2895	2535	3767	4,527

Table 4
Type of authenticated users in the network from 2012 to 2018.

Type	Number
Dartmouth NetIDs	25,112
Other eduroam identifiers	12,984
Estimated Authenticated Users	38,096

Table 5
Type of users as in 2018 (User ID Dartmouth).

Type	Number
Total GR Student	675
Total UG Student	4,272
Total Student	4947
Total Staff	2,429
Total Faculty	1365
Total Other	2,738
Total	11,479

5. User and station analysis

Many of the SNMP traps include information about the network user (person) and network station (mobile device) connected to the network. This section presents statistics about those users and their stations.

5.1. User statistics

The SNMP trace uses several means for identifying users. Regarding free-access networks (SSID *Dartmouth Public* or *Dartmouth Library Public*), user names in the trace can be either a name freely selected by the user during log-in (in the first years) or the reserved word 'Guest'. The access-controlled networks (SSID *Dartmouth Secure* and, later, *eduroam*) use Dartmouth identification numbers ('NetIDs'). The *eduroam* SSID also allowed guests from other eduroam universities to authenticate with their email address [7]. Table 4 depicts the major types of user identifiers we found in the authenticated networks, including almost 13,000 email addresses from an institution other than Dartmouth[1]

The Dartmouth IT staff provided a 2018 database that mapped NetID to user type. Table 5 presents a summary of that database. A total of nearly 11,500 users are listed, including around 5000 students, 2400 staff, and 1400 academic faculty.

Fig. 11 presents the number of active users per day, over seven years. There is a clear yearly pattern, with three large periods of usage corresponding to the three terms in the academic year. It also shows an abnormally high peak at the end of 2017 and the first half of 2018, where the number of users far exceeded the number specified in Table 5, and then an unexpectedly low number in the latter half of 2018. This particular pattern is an artifact resulting of the process of replacing *Dartmouth Secure* by *eduroam*, as explained in detail in Appendix A, Users.

More broadly, we found a steady average of 25 sessions per user per day in the network (not shown). Considering the daily average of 100 sessions per AP, we observe that the deployment roughly includes 1 AP for every 4 users.

5.2. Station statistics

We counted 600,000 distinct MAC addresses in the seven year capture. We wondered whether these represented actual stations, or whether some of the MAC addresses were dynamically chosen, as discussed in a recent paper [15]. Typically, MAC addresses are fixed and unequivocal. Manufacturers purchase contiguous blocks of MAC addresses with the same three-byte prefix, referred to as an *Organizationally Unique Identifier (OUI)*. Then, they assign a single MAC in the block for each station they manufacture. Therefore, in principle, a MAC address unequivocally and permanently identifies station and manufacturer. The list of OUIs is publicly available [16].

A mobile station may actively search for APs by sending *probe request messages* that include the station MAC. This means that an interested observer could track the movements of a station (and, so, of its user) by collecting the probe requests in the wireless medium. Notice this is possible without requiring access to network services, or the knowledge or permission of users. One use case could be a shopping mall, which may profit from personalized marketing by identifying customer shopping habits. To avoid this form of tracking, some devices employ MAC address randomization (MAR).

Martin et al. notes that "*Nearly all randomization schemes utilize locally assigned MAC addresses to perform randomization*" and "*99.12% of all locally assigned MAC addresses are randomized addresses*" [15]. More importantly, "*When a mobile station attempts to connect to an AP, however, it reverts to using its globally unique MAC address. As such, tracking smartphones becomes trivial while they are operating in an associated state.... If we find matching locally assigned MAC addresses in authentication, association, or data frames,...the randomization scheme is likely Windows 10 or Linux.*"

Table 6 shows the number of station MACs found in the capture, arranged by the type of MAC. As expected, the number of global (unique and unicast) MACs is dominant, with only 7000 local MACs out of the total 632,000. Indeed, when we matched unique MACs against the public list of OUIs [16], we found that Dartmouth's campus is dominated by Apple laptops, smartphones,

[1] A negligible number of other "User" identifiers, including errors and PC identifiers, was also found.

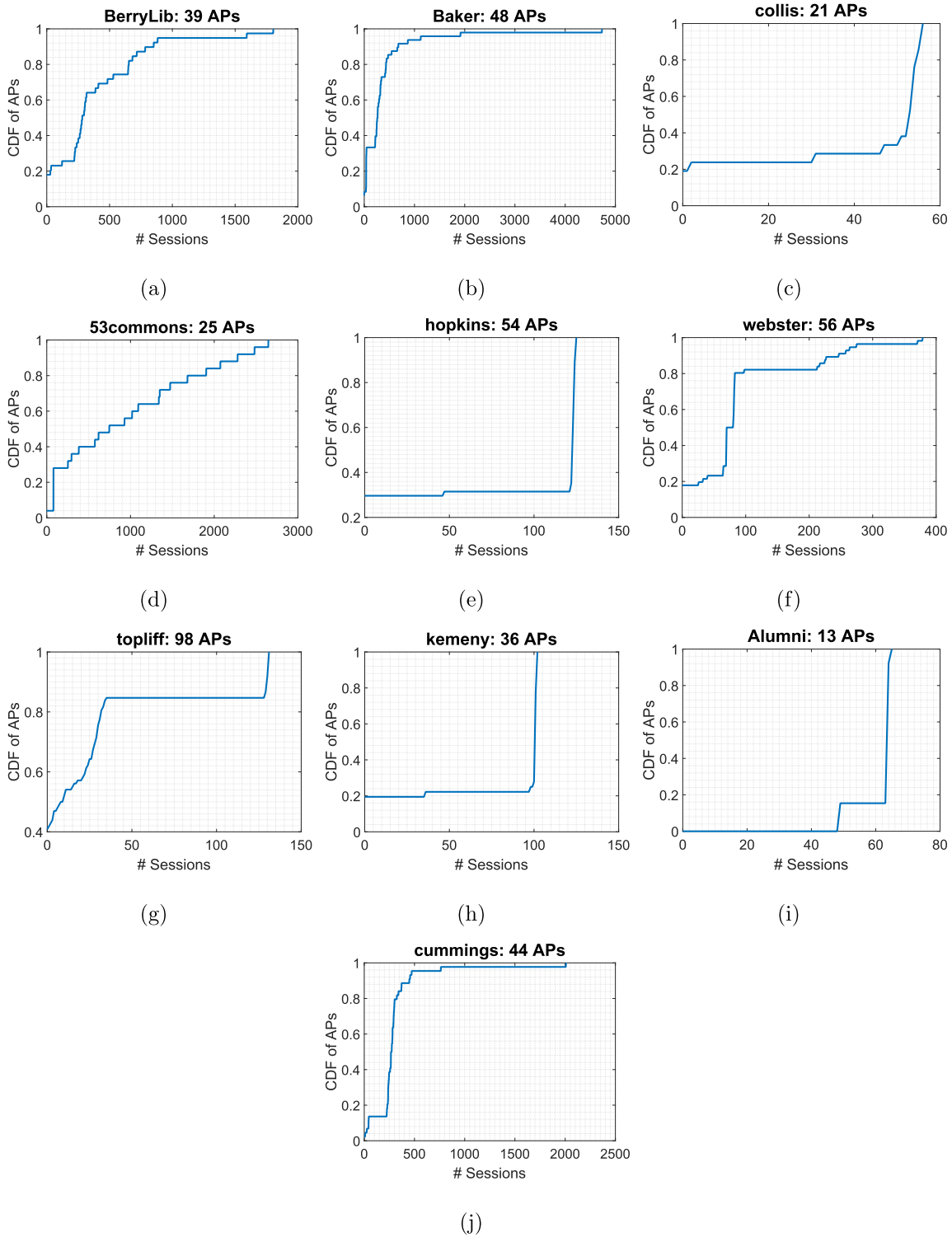


Fig. 9. CDF of APs per building in terms of the number of sessions for the busiest buildings.

Table 6
Type of station MACs in the network.

Type	Number
Global MACs	624,903
Local MACs	7062
Unicast MACs	631,965
Multicast MACs	5
Total	631,970

and tablets; see Table 7. No other vendor OUI represented more than 7% of stations.[2]

[2] It is sometimes difficult to map from OUI to vendor name, because some vendors use multiple brands of Wi-Fi chips, and some brands of Wi-Fi chips are used in multiple vendors' products. Apple purchases chips from multiple vendors (including Murata) but assigns the chips Apple OUIs – at least, most of the time. Table 7 is necessarily an approximation of the population of specific brands of hardware on campus.

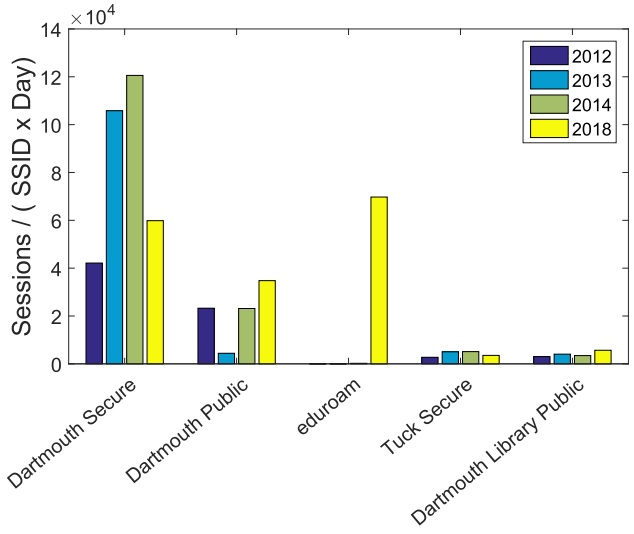


Fig. 10. The highest-ranked SSIDs, according to their average number of sessions per day.

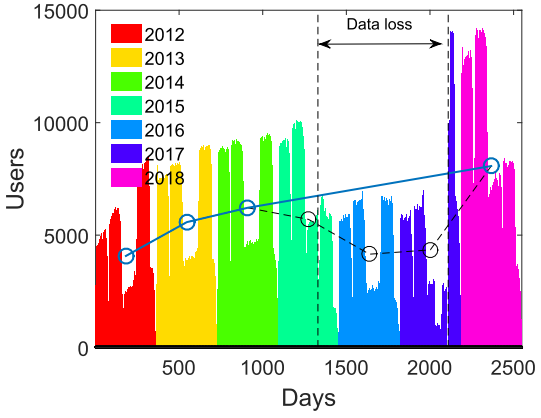


Fig. 11. Number of users per day, over time.

Table 7
Stations according to OUI.

Manufacturer	Percentage
Apple, Inc.	72.07
Intel Corporate	6.65
Samsung Electronics Co.,Ltd	4.96
Murata Manufacturing Co., Ltd.	2.24
Hon Hai Precision Ind. Co.,Ltd.	1.69
LG Electronics (Mobile Communications)	1.58
Motorola Mobility LLC, a Lenovo Company	1.45
HTC Corporation	1.33

Fig. 12 presents the number of stations per day, showing a moderate and steady increase throughout the capture. Although (not shown) we found a steady average of 12 sessions per day, per station, the number of sessions per day varied widely across stations. We ranked the station MACs according to their average number of sessions per day in Fig. 13. Stations, unlike APs, may be inactive for long periods. For this reason, to compute the averages, we only consider the days on which a station was active. The figure shows medians between 1 and 5 sessions per day, but some stations average well above 100 sessions per day.

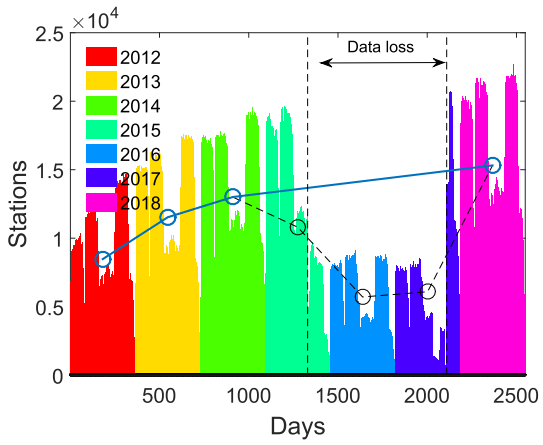


Fig. 12. Number of stations per day, over time.

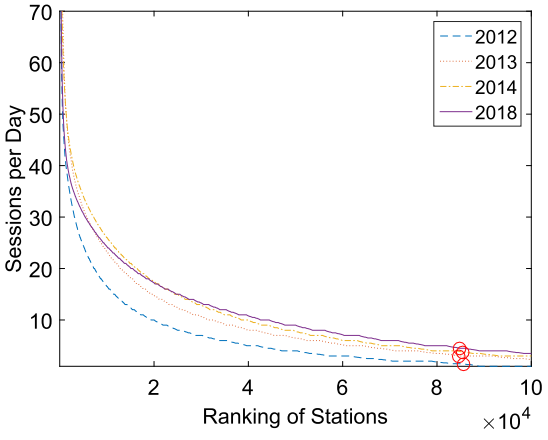
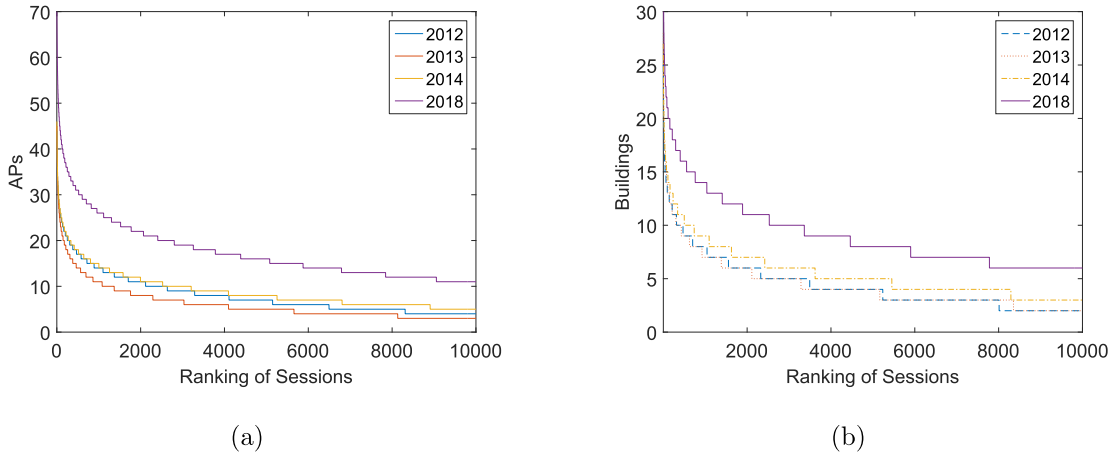


Fig. 13. Number of sessions per day of the 10,000 highest ranked stations according to their average number of sessions per day. The ranking is performed for each year, with the maximum of 230.5 (average sessions per day) exhibited by one station in 2018. Median values are highlighted with circles.

6. Mobility analysis

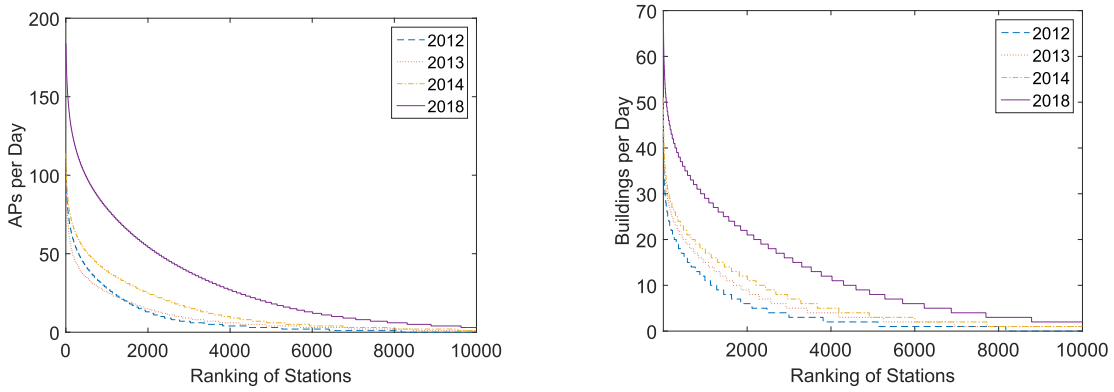
There are two types of mobility that are of interest in this paper. On the one hand, the mobility within sessions is interesting from the network-management perspective: stations that roam within a session require the network infrastructure to perform fast, reliable hand-offs as a station roams. On the other hand, the geographical mobility of a user, estimated from the mobility of a station over time, is interesting to researchers who study patterns in the ebb and flow of people around campus. An understanding of patterns of user location can also be helpful in planning network infrastructure. In this section we study the mobility across APs and buildings, the latter in order to avoid being distracted by fine-grain mobility or the 'ping-pong' effect (when a station rapidly roams to and from adjacent APs) observed in previous papers [1,2].

In Fig. 14, we rank the sessions by the number of APs (Fig. 14(a)) and buildings (Fig. 14(b)) they traversed, and average the result across all days, representing a distribution of mobility of the set of sessions occurring during a day. Medians (not shown) were 1 AP and building per session, which means most stations remained in the same area during a session. However, there was a large number of sessions roaming to more than one AP and building, with a maximum daily average of 70 APs and 36 buildings in 2018. We can also see that the mobility has significantly increased from 2014 to 2018 for the sessions with higher mobility within a



(a) (b)

Fig. 14. Mobility in number of APs (a) and buildings (b) per session. The top 10,000 sessions are ranked by mobility, with the maximum of 70/36 seen in 2018. The median number of APs and buildings visited by a session was 1.



(a) (b)

Fig. 15. Station mobility, ranked by average number of APs (a) and buildings (b) per day. The top 10,000 stations are ranked by mobility, with the maximum daily average of 184 APs and 70 buildings seen in 2018.

day, likely reflecting an increase in smartphones and wearable Wi-Fi devices.

The mobility of stations during the day is shown in Fig. 15. Station mobility increased significantly from 2014 to 2018 for top-ranked stations. The yearly median (not shown) rose from 4 APs and 1–2 buildings in the period 2012–14, to 8 APs and 4 buildings in 2018. This change is likely due to an increased prevalence of smartphones and similar devices that users carry while they move.

We were surprised to find that some stations associated with *more than 70 buildings in a single day*; a study of the detailed track for some of these station-days indicates these numbers are plausible, if exceptional. We speculate the tracks were caused by a smartphone carried by a security guard or mail-delivery person making the rounds of buildings on campus, passing by (if not entering) as many as 70 buildings.

7. Deriving updated connectivity models in Wi-Fi

Table 8 compares the connection statistics reported by papers about the Dartmouth Wi-Fi network. We can see that main traffic characteristics are consistent throughout the entire life of the campus network: diurnal dominance in traffic, daily and weekly cyclostationarity, and skewed distribution of the load (in terms of traffic or number of sessions) in APs and buildings. The present capture also shows the yearly pattern of traffic, which could not be observed in previous works due to the limited time span. We have also seen an increase of mobility.

We have seen a maintained session duration.[3] In 2018, almost one third of the sessions were shorter than one minute, when computed only for sessions terminated due to inactivity and by subtracting the 5-min estimated inactivity threshold. Fewer than 15% of sessions were longer than one hour, but an indeterminate number of them may have been formed by shorter sessions separated by less than the inactivity threshold. One potential explanation for the low session length is that modern devices generate a huge number of short sessions. Smartphones, in particular, will periodically ping a server to determine whether any new messages or mail has arrived, and that requires them to be online, initiating a quick session. Devices may be terminating these sessions to save energy, leading to very short sessions.

As discussed above, the number of sessions (or active cards) per hour has increased by an order of magnitude, but daily patterns remain very similar: minimum at night, maximum shortly after noon. The number of active stations in a single day has also increased by an order of magnitude. Finally, even after increasing

[3] It is difficult to draw precise conclusions because of changes in the network technology and configuration. In this paper we were able to derive a session-ending threshold (five minutes) from evidence related to the network configuration – whereas the earlier papers used a threshold (30 min) that was less clearly related to network configuration.)

Table 8
Connection statistics reported in papers about the Dartmouth Wi-Fi network. Notes: *These intervals were described in [4] for the captures published in 2005 and 2008. The interval for this paper corresponds to the yearly averages in Fig. 12. **Computed for SessionIDs with Reason Code 4 and corrected with the inactivity threshold.

Concept	Kotz & Essien '05 [2]	Henderson et al. '08 [4]	Camacho et al. '19
Traffic/Connection Patterns			
Diurnal	✓	✓	✓
Daily/Weekly	✓	✓	✓
Yearly			✓
Load skewed across APs	✓	✓	✓
Load skewed across Buildings	✓	✓	✓
Increase in mobility			✓
Session duration			(only 2018)**
< 1m	27%	–	29%
< 1h	71%	–	88%
#Sessions			(only 2018)
Minimum (hours)	~ 200 (4–7 a.m.) cards/h	~ 500 (5–6 a.m.) cards/h	2738 (4–6 a.m.) sess
Maximum (hours)	~ 500 (1–4 p.m.) cards/h	~ 1400 (1–2 p.m.) cards/h	16,410 (12-2 p.m.) sess
Busiest AP	71 cards/h	–	> 200 sess
Busiest Building	76 cards/h	~ 340 cards/h	> 500 sess
Busiest Station			> 70 sess
Max. buildings per Session			> 30 sess
Stations			
Stations per day*	800–1,000	3,000–3,500	8,474–15,310
Max. buildings in a Day			> 70

AP deployment by 700% (from 476 to 3330 APs), the load on the busiest APs and buildings has increased.

7.1. Models

The development (or validation) of detailed new models (network models, traffic models, or mobility models) is out-of-scope for this characterization paper. We encourage researchers to use our findings, in particular those summarized in Table 8, to derive simulation models useful in research. As input to that endeavor, we extract the following approximate distributions observed in 2018.

- Session length (S_L): from Fig. 4 the distribution of session length observed in 2018, after correcting for the inactivity threshold, follows a log-normal distribution: $S_L \sim$ Log-N(5.4,3.6). The quality of the approximation is shown in Fig. 16(a).
- Session data (S_D): from Fig. 4 the distribution of session data content observed in 2018 also follows a log-normal distribution, which depends on the session length: $S_D \sim$ Log-N$(1.1\log(S_L)+6.7, 0.3\log(S_L)+4.4)$. Examples of the approximation for 5-min and 1-h sessions are shown in Fig. 16(b)–(c).
- Number of sessions per AP: the distribution observed for 2018 in Fig. 8(a) follows a piece-wise decaying logarithm:
  - The top [1.10] most-loaded APs: $y = 4800 \cdot (-0.6)^x + 2070$
  - The next [11.100] most-loaded APs: $y = 1870 \cdot (-0.035)^x + 670$
  - The next [101.1000] most-loaded APs: $y = 800 \cdot (-0.005)^x + 160$

  The quality of the approximation is shown in Fig. 16(d), and a detail for the 10 most-loaded APs in Fig. 16(e).
- Number of sessions per station: the distribution observed for 2018 in Fig. 13 follows a piece-wise decaying logarithm:
  - The top [1.100] most-loaded stations: $y = 121 \cdot (-0.04)^x + 107$
  - The next [101.1000] most-loaded stations: $y = 86 \cdot (-0.004)^x + 45$
  - The rest of stations: $y = 33 \cdot (-0.00003)^x$

  The quality of the approximation is shown in Fig. 16(f), and a detail for the 100 most-loaded stations in Fig. 16(g).
- Number of buildings per day: the distribution observed for 2018 in Fig. 15 follows a piece-wise decaying logarithm:
  - The top [1.100] most-mobile stations: $y = 21 \cdot (-0.03)^x + 48$
  - The next [101.1000] most-mobile stations: $y = 25 \cdot (-0.002)^x + 26$
  - The rest of stations: $y = 40 \cdot (-0.0003)^x$

  The quality of the approximation is shown in Fig. 16(h), and a detail for the 100 most-mobile stations in Fig. 16(i).

Note these specific models may not generalize to other campuses, with the exception of the models for the session length and data, which can be seen to agree well with findings in other works [17,18].

8. Related work

As already discussed, the evolution of the Dartmouth Wi-Fi network has been previously studied in a series of two papers. First, Kotz and Essien [1,2] analyzed 11 weeks of traffic starting September 2001. The capture was composed of syslog records with authentication and roaming events, SNMP logs and payload captures. They found high variance in the traffic load, and a large number of sessions with an excess of roaming between close APs and an indication of reduced mobility. In a second paper [3,4], Henderson, Kotz and Abyzov assessed the evolution of the network by comparing the first capture with a second capture of 17 weeks starting November 2003, with the same type of data plus VoIP connections. The study found a large increase of usage, with a significant load of P2P traffic, and a limited increase of mobility for specific device types, mainly VoIP devices. The present paper represents the third of the series to assess the evolution in the usage of the Dartmouth Wi-Fi, and analyze a mature technology that can be contrasted to those initial deployments. A major contribution of the present paper is the analysis of a much larger capture, corresponding to 7 years of Wi-Fi connections. Very relevant is the fact that the current analysis includes the emergence of smartphones, which now dominate network connections and have altered usage, traffic and mobility patterns to a large extent. As a result, the increase of traffic (17-fold) and mobility represents the major difference with previous analysis.

Other prior works have analyzed Wi-Fi traces to mine user activity and behavioral patterns and to conduct mobility and us-

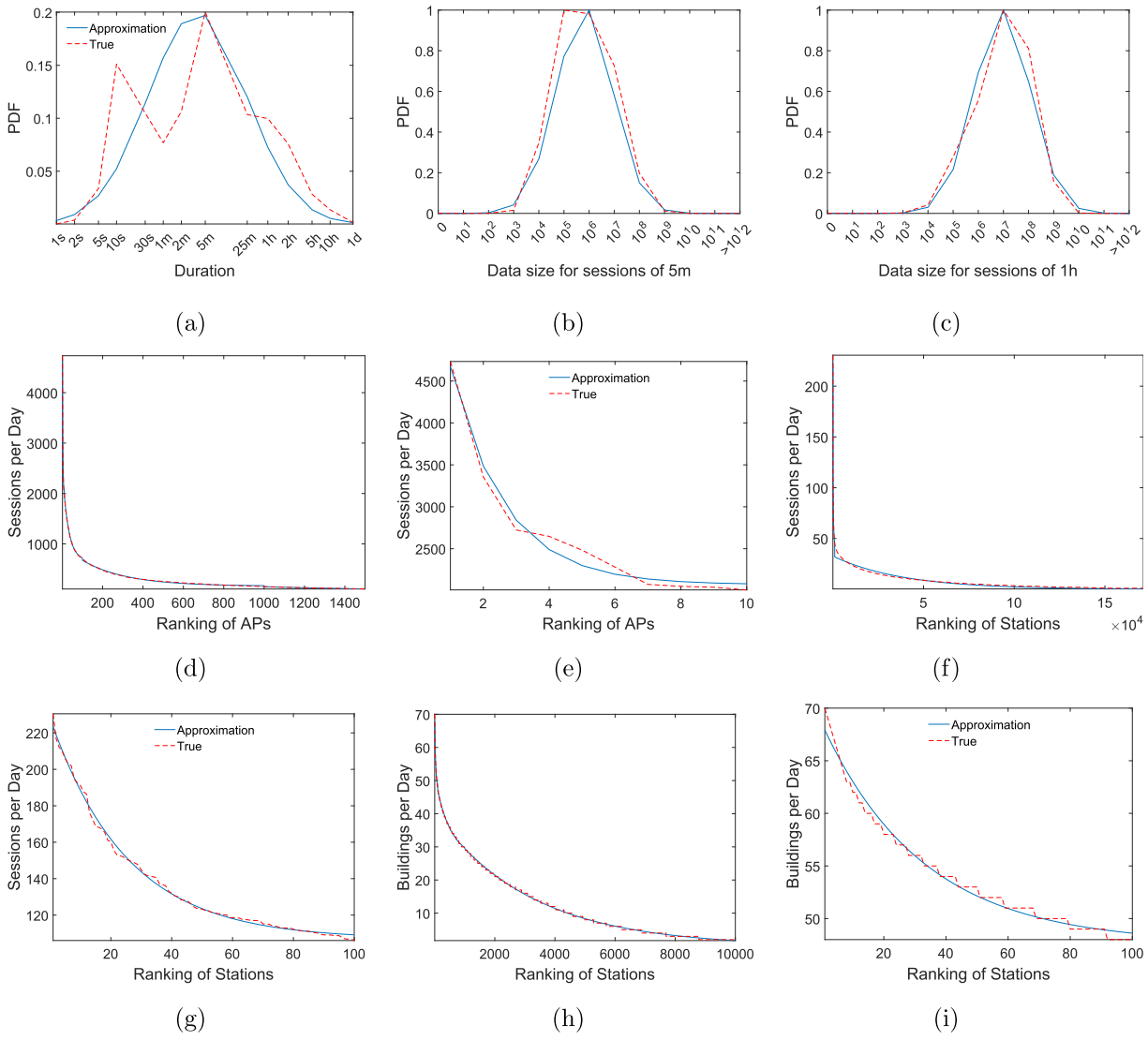


Fig. 16. Approximations vs measured.

age analysis. These works, however, all rely on traces over a time span shorter than the dataset used in this paper. As one example, Poucin, Farooq and Patterson [19] used one-week Wi-Fi connection logs from 2015 to mine users' activity patterns in a university campus with unsupervised machine learning, in particular PCA and k-means clustering. Another, based on connection logs of several months in 2005, Meneses and Moreira [20] studied mobility and network usage in a university campus with more than 550 APs. Afanasyev et al. [17] conducted similar analysis with a capture of connection logs over 28 days in 2008 from the Google Wi-Fi network in Mountain View, California, with around 500 APs. Ruiz-Ruiz et al. [21] analyzed and visualized large Wi-Fi facilities using two-week trace data. Lyu et al. [18] leveraged four-month SNMP data in 2018 to analyze a Wi-Fi deployment with more than 8K APs and 40K active users. Wei et al. [22] reviewed behavioral patterns using one-year connection logs plus a month of traffic data logs. Alipour et al. [14] combined Netflow (25 days) and connection records (479 days) from 2011–2012. Cao et al. [23] investigated the predictability of human mobility through four-month Wi-Fi logs from 2015. Shi et al. [6] studied Wi-Fi scans on smartphones through around 1000 days ending in 2015. Serrano et al. [24] reviewed the relevant literature related to Wi-Fi, including traffic and usage analysis of real traces; all cited work studied traces from 2000 to 2008 and presented traces of fewer than 1200 APs and capture periods no longer than one month. By comparison, the dataset analyzed in this paper is far larger than any of the previously cited references, and presents the most up-to-date analysis, making our observations and derived models of special value.

Two relevant dimensions of interest in the study of large Wi-Fi deployments are privacy and usage patterns. Regarding privacy, Martin et al. [15] performed a study on MAC Address Randomization (MAR), a privacy-preserving methodology applied by some manufacturers to prevent user location and movements from being tracked. The study showed that MAR is only applied when a device is actively probing for APs, but not when associating to a network. The authors argued that when associated, using the actual, unequivocal, MAC address of the device is not a risk for privacy, since associated devices show reduced mobility. In our capture, all devices are in the process of association or already associated, and therefore they use the actual MAC of the device. However, being a large campus network, with thousands of APs covering 200 acres, and considering the high degree of mobility of smartphones, MAC addresses in connection logs do actually pose a significant risk for privacy. Recent similar experiences [25] support this idea.

Regarding usage patterns, the analysis of the Google Wi-Fi Network in 2008 [17] identified three user populations with different

traffic, mobility, and usage patterns: modem users, hotspot users, and smartphone users. The latter show high mobility and low traffic demand, with most sessions below 2-h length and 10-MB usage. This finding aligns with the general pattern of session length and traffic volume that we found in our capture, which we actually expect to be dominated by smartphone users. However, mobility has increased: in the Google trace, only a handful of users in [17] connected to more than 16 APs per day, while in our capture we found several thousands of stations exceeding that number, with a maximum daily average of 114 APs and 50 buildings in 2014 and 184 APs and 70 buildings in 2018 (see Fig 15(a)). Lyu et al., in their contemporary work [18] to ours, have also shown conclusions consistent with the ones presented here in terms of the length of the sessions (median of 100s per session) and mobility (median of association to 20 APs per day). Alipour et al. [14] also explored differences between laptops and smartphones, describing a methodology to classify devices according to the MAC addresses. Wei et al. [22] found patterns of load within the day and differences between regular and non-regular users of the network, showing conclusions consistent with those presented here.

9. Conclusion

In this paper we present a detailed characterization and analysis of the most-recent seven years of data (2012–18) collected about the Dartmouth Wi-Fi network, and highlight some of the changes over time. The analysis shows the evolution of the infrastructure and adoption of Wi-Fi technology by users and leads to the following conclusions.

- The number of connection sessions has increased 10-fold in the last 15 years, but remained reasonably stable in the last 5 years.
- Much as in the previous analyses of the same network, the number of active sessions shows marked daily, weekly and yearly patterns.
- Usage patterns in 2018 reflect that 29% of sessions were very short: under 1 min. These short sessions may reflect the background activity of idle devices that check-in periodically with a remote server.
- We found a gross mean of 100 daily sessions per access point (AP), 25 daily sessions per user and 12 daily sessions per device. These numbers reflect a provisioning of one AP per 4 users and 8 devices, on average. However, we also found that the distribution of sessions per AP, per user, and per station were highly uneven, with maximums in the thousands per AP and hundreds per station.
- The network has multiple SSIDs, and a majority of sessions (74%) occurred on the primary authenticated networks. While there is a public-access network across the entire campus, this network carried fewer than 19% of sessions.
- Most mobile sessions have doubled their average mobility in the last 5 years, likely as a result of the emergence of smartphones.
- Device mobility during a day has also increased in the last 5 years, likely because smartphones are always “on” even when the screen is locked and they are not in use – unlike the laptops of 2004, which had no network activity when closed and being carried – and even a pocketed smartphone remains periodically active to check for new messages.
- Indeed, we found examples of high mobility in which the device is connecting to the network even when the user may not be explicitly trying to use the network. This behavior may impinge on user privacy, because the network’s connection logs can be used to track user movements. MAC Address Randomization (MAR) does not address this concern, since MAR is seldom used when a device is associated to the network.

From these and other observations throughout the paper, we derived models (Section 7.1) that could be used to drive realistic simulations of modern Wi-Fi networks, and we provide the research community with two anonymized data sets with preprocessed information about connection sessions, so that other groups can extend our research.

Declaration of Competing Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

CRediT authorship contribution statement

José Camacho: Conceptualization, Data curation, Funding acquisition, Software, Visualization, Writing - original draft. **Chris McDonald:** Software, Data curation, Writing - original draft. **Ron Peterson:** Software, Data curation. **Xia Zhou:** Writing - review & editing. **David Kotz:** Conceptualization, Funding acquisition, Methodology, Resources, Writing - original draft.

Acknowledgment

This work was supported by Dartmouth College, and in particular by the many network and IT staff who assisted us in configuring the Wi-Fi network infrastructure to collect data, and who patiently answered our many questions about the network and its operation. We furthermore appreciate the support of research colleagues and staff who have contributed to our data-collection and data-analytics infrastructure over the years: most notably Wayne Cripps, Tristan Henderson, Patrick Proctor, Anna Shubina, and Jihwang Yeo. Some of the Dartmouth effort was funded through support from ACM SIGMOBILE and by an early grant from the US National Science Foundation under award number 0454062. The first author was partly supported by the Ministerio de Educación, Cultura y Deporte under the Programa Estatal de Promoción de Talento y su Empleabilidad en I+D+i, Subprograma Estatal de Movilidad, del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013–2016 and the Fulbright program.

Appendix A. Traps, OIDs, and sessions.

Trap structure Fig. A.17 shows an example of an SNMP trap as received. Each trap comprises a header, with timestamp and sender and collector information, followed by a variable number of triplets related to object identifiers (OIDs) following the format ‘< OID > = < type >: < value > ’ and separated by hashes (#). OIDs are partly represented in the ASN.1 notation, which can be translated into more meaningful OID names using the relevant Management Information Base (MIB). An important OID is the trap type (TT), which is the second triplet in the figure, highlighted by a rectangle. In the TT, the value is also an OID: ‘< TT ;>= OID: < OID >’.

Distribution of traps Table A.9 presents the most common OIDs in the capture and Table A.10 the most common TTs. The description of OIDs can be easily found in OID information repositories available on-line [26,27]. Inspecting the table we can see that the capture included information about APs (their name, MAC address, communication slot), users (name/ID) and stations/devices (MAC address). It also contained TTs related to the process of associating user devices to the network. Some of these TTs report the volume of traffic generated per connection, but this information was incomplete and could only be found for a subset of connections. The description of TTs, as defined by Cisco Networks, and their period of capture, is in Table A.11.

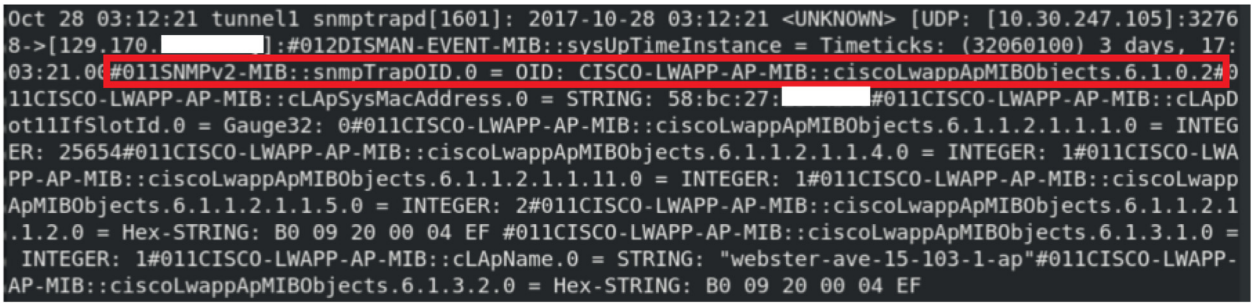


Fig. A.17. Example of an SNMP trap in the data capture. The second OID, highlighted by a rectangle, represents the trap type. Parts of an IP and a MAC address have been hidden.

Table A.9
SNMP OIDs found in more than 20% of traps. CLAM refers to CISCO-LWAPP-AP-MIB, AWM to AIRESPACE-WIRELESS-MIB and CLDCM to CISCO-LWAPP-DOT11-CLIENT-MIB.

Label	% of traps
CLAM::cLApDot11IfSlotId	46.83
AWM::bsnAPName	35.94
AWM::bsnStationMacAddress	34.32
AWM::bsnStationAPIfSlotId	34.32
AWM::bsnStationAPMacAddr	34.32
AWM::bsnStationUserName	34.32
AWM::bsnUserIpAddress	34.29
CLAM::cLApName	29.76
CLDCM::cldcClientByIpAddress	20.77
CLDCM::cldcClientByIpAddressType	20.77
CLDCM::cldcClientMacAddress	20.55
CLDCM::cldcApMacAddress	20.49

Table A.10
Trap Types (TTs) found in more than 1% of traps. AWM refers to AIRESPACE-WIRELESS-MIB and CLDCM to CISCO-LWAPP-DOT11-CLIENT-MIB.

Label	% of traps
AWM::bsnDot11StationAssociate	19.43
CLDCM::ciscoLwappDot11ClientMovedToRunState	17.73
CLDCM::ciscoLwappDot11ClientSessionTrap	12.66
AWM::bsnDot11StationDeauthenticate	6.93
CLDCM::ciscoLwappDot11ClientAssocDataStatsTrap	5.57
AWM::bsnAuthenticationFailure	3.55
CLDCM::ciscoLwappDot11ClientDisassocDataStatsTrap	2.13
AWM::bsnDot11StationAssociateFail	1.98

Traps content Our capture included TTs related to the association process. Unfortunately, due to configuration changes during the seven years, not all TTs were captured throughout. It is useful to consider the content of some trap types, summarized for some fields of interest in Table A.12. Connection traps consistently contained the User Name and the AP MAC, relating user with AP, and thus that user's approximate location. Station MAC and AP Name were not included in all cases, and the SSID only appeared in a limited number of traps. Session ID and connection packets/bytes appeared in a couple of traps each, enabling the identification of sessions and their associated volume of traffic. The "SessionID" is a concatenation of three values: a monotonically increasing unique 32-bit integer, the station MAC address, and the session start timestamp (also 32-bit integer). One anonymized example is "5b5e6b70/d2:63:—:—:e3:06/18477800".

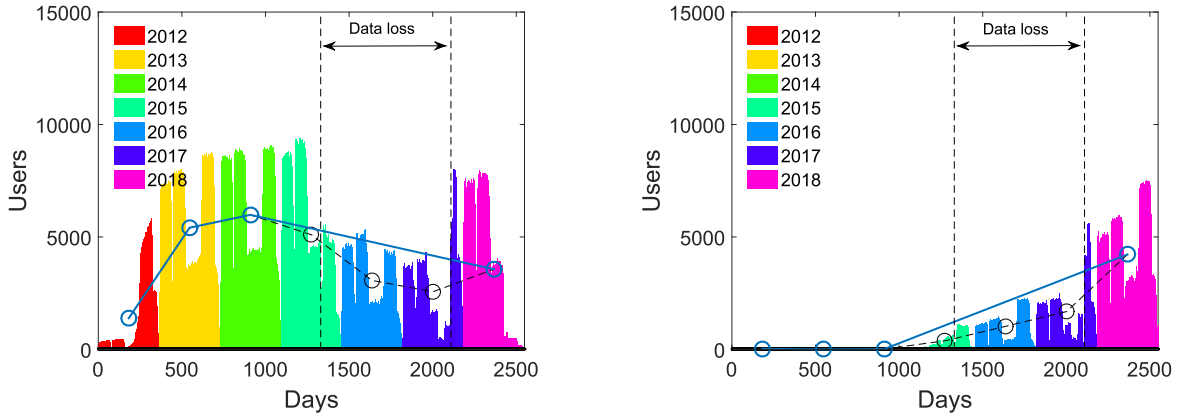
Sessions Identification Looking at the TTs available, we have (at least) two potential ways to identify connection sessions. The SessionID was one possibility. However, a main drawback is that SessionIDs were only present in two TTs, ClientSessionTrap and ClientDisassocDataStatsTrap which, in turn, were only captured since

Table A.11
Description of traps in the connection process. For the sake of readability, we use short trap names. Corresponding complete names can be found in Table A.9.

Trap Type (period of capture)	Description [27]
StationAssociate (2012–2018)	The associate notification shall be sent when a client associates with an AP.
ClientAssocDataStatsTrap (2018)	The associate notification shall be sent when the Station sends a association frame.
ClientMovedToRunState (2012–2018)	This notification is generated when the client completes the PEM state (authentication-association) and moves to the RUN state (associated).
ClientSessionTrap (2014–2018)	Issued when the client completes the PEM state and moves to the RUN state.
StationDeauthenticate (2012–2018)	The deauthenticate notification shall be sent when the Station sends a Deauthentication frame.
StationDisassociate (2017–2018)	The disassociate notification shall be sent when the Station sends a Disassociation frame.
ClientDisassocDataStatsTrap (2018)	The disassociate notification shall be sent when the Station sends a Disassociation frame.
StationAssociateFail (2012–2018)	The associate failure notification shall be sent when the Station sends an Association frame with a status code other than 'successful'.
StationAuthenticateFail (2012–2018)	The authenticate failure notification shall be sent when the Station sends an Authentication frame with a status code other than 'successful'.

Table A.12
Partial content of traps in the connection process. For the sake of readability, we use short trap names. Corresponding complete names can be found in Table A.9. X*: this trap includes the roaming AP MAC.

Trap Type	Station MAC	AP MAC	User Name	AP Name	SSID	Session ID	Packets /Bytes
StationAssociate	X	X	X	X			
ClientAssocDataStatsTrap		X*	X		X		X
ClientMovedToRunState	X	X	X	X	X		
ClientSessionTrap		X	X	X	X	X	
StationDeauthenticate	X	X	X	X			
StationDisassociate	X	X	X	X			
ClientDisassocDataStatsTrap		X	X	X	X	X	X
StationAssociateFail	X	X	X	X			
StationAuthenticateFail	X	X	X	X			



(a) Dartmouth Secure (b) eduroam

Fig. A.18. Number of Dartmouth users on *Dartmouth Secure* (a) and *eduroam* (b), each day, over time.

2014 and 2018, respectively. This means that we could only identify the complete timespan of a SessionID in 2018, where we had both associations and disassociations, and even in this year there were missing traps. Alternatively, we could use the station MAC to identify connection sessions in other traps available during the entire capture, like StationAssociate, ClientMovedToRunState and StationDeauthenticate. In the paper, we used the SessionID traps to analyze the duration and amount of traffic and number of connections in 2018, and then the station MAC in association traps to generalize the analysis to the complete capture.

Users The number of user names shows abnormally high peak at the end of 2017 and the first half of 2018, and then an unexpectedly low number in the latter half of 2018. This particular pattern is explained in Fig. A.17, where we show the evolution of the users in the primary authenticated SSIDs, *Dartmouth Secure* and *eduroam*. Dartmouth pushed all users to use the *eduroam* SSID in late 2018, and shut down *Dartmouth Secure* near the end of 2018. If we skip the misconfiguration period (2015–17) and focus on the following one, we can see that at the end of 2017 and the first half of 2018 both SSIDs coexist. During this period, many users connected through both SSIDs but with different user names: their user ID and the corresponding email address, respectively. This double-counting explains the anomalously high and unrealistic peak in Fig. 11.

Appendix B. Data sets

We provide two anonymized data sets for the research community: a) the SessionIDs in 2018, where we found starting and ending traps, and b) the association traps found from 2012 to 2018. Both data sets are presented in JSON format and are released for use by researchers.

Sessions from 2018 The first data set contains the following information per SessionID:

InitTS	Init timestamp
FinalTS	End timestamp
Duration	Session length
PacketsSent	#Packets uploaded
BytesSent	#Bytes uploaded
PacketsRecv	#Packets downloaded
BytesRecv	#Bytes downloaded
UserName	User ID
UserMAC	Station MAC address
SSID	SSID
APName	AP Name
ReasonCode	Reason for the end of the session

Note regarding the computation of SessionIDs: In the case of several starting traps (ClientSessionTrap) for the same SessionID code, we chose the first one. In the case of several ending traps (ClientDisassocDataStatsTrap) for the same SessionID code, we chose the last one. We found that fewer than 0.01% of SessionIDs had duration exceeding 2 days. For this reason, and to avoid a large computational burden, we only considered SessionIDs of duration less than two days.

Association traps found from 2012 to 2018 The second data set is built from the ClientMovedToRunState traps. We chose these traps because its number (Table A.10) and availability (Table A.11: from 2012 to 2018) is similar to that of StationAssociate traps, but with the additional advantage that the former contain the SSID (Table A.12). The following information is stored per association trap:

TS	timestamp
UserName	User ID
UserMAC	Station MAC address
SSID	SSID
APName	AP Name

Anonymization In each dataset, we replaced each identifier (UserName, UserMAC, APName) with a consistent, unique pseudonym of the same format. The same approach was used in the public release of the early Dartmouth traces [28].

References

[1] D. Kotz, K. Essien, Analysis of a campus-wide wireless network, in: Proceedings of the ACM International Conference on Mobile Computing and Networking (MobiCom), 2002, pp. 107–118, doi:10.1145/570645.570659. Revised and corrected as Dartmouth CS Technical Report TR2002-432.

[2] D. Kotz, K. Essien, Analysis of a campus-wide wireless network, Wirel. Netw. 11 (1–2) (2005) 115–133, doi:10.1007/s11276-004-4750-0.

[3] T. Henderson, D. Kotz, I. Abyzov, The changing usage of a mature campus-wide wireless network, in: Proceedings of the ACM International Conference on Mobile Computing and Networking (MobiCom), ACM Press, 2004, pp. 187–201, doi:10.1145/1023720.1023739.

[4] T. Henderson, D. Kotz, I. Abyzov, The changing usage of a mature campus-wide wireless network, Comput. Netw. 52 (14) (2008) 2690–2712, doi:10.1016/j.comnet.2008.05.003.

[5] H.A. Omar, K. Abboud, N. Cheng, K.R. Malekshan, A.T. Gamage, W. Zhuang, A survey on high efficiency wireless local area networks: next generation wifi, IEEE Commun. Surv. Tutor. 18 (4) (2016) 2315–2344.

[6] J. Shi, L. Meng, A. Striegel, C. Qiao, D. Koutsonikolas, G. Challen, A walk on the client side: monitoring enterprise wifi networks using smartphone channel scans, in: 35th Annual IEEE International Conference on Computer Communications, INFOCOM 2016, San Francisco, CA, USA, April 10–14, 2016, 2016, pp. 1–9, doi:10.1109/INFOCOM.2016.7524453.

[7] Eduroam, Eduroam: World wide education roaming for research & education, Accessed: 2018-09-30. (https://www.eduroam.org/).

[8] J. Case, M. Fedor, M. Schoffstall, J. Davin, A Simple Network Management Protocol (SNMP), RFC 1157, RFC Editor, 1990. https://www.rfc-editor.org/rfc/rfc1157.txt.

[9] IEEE Standard 802.11i, Accessed: 2019-04-30. (https://standards.ieee.org/standard/802_11i-2004.html).

[10] IEEE Standard 802.11r, Accessed: 2019-04-30. (https://standards.ieee.org/standard/802_11r-2008.html).

[11] IEEE 802.1X Common Session ID, Cisco Systems (2012).

[12] H. Han, Y. Liu, G. Shen, Y. Zhang, Q. Li, Dozyap: Power-efficient Wi-Fi tethering, in: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, in: MobiSys '12, ACM, New York, NY, USA, 2012, pp. 421–434, doi:10.1145/2307636.2307675.

[13] A. Gember, A. Anand, A. Akella, A comparative study of handheld and non-handheld traffic in campus wi-fi networks, in: N. Spring, G.F. Riley (Eds.), Passive and Active Measurement, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 173–183.

[14] B. Alipour, L. Tonetto, A.Y. Ding, R. Ketabi, J. Ott, A. Helmy, Analyzing mobility-traffic correlations in large WLAN traces: Flutes vs. cellos (2018). CoRR arXiv: 1801.02705.

[15] J. Martin, T. Mayberry, C. Donahue, L. Foppe, L. Brown, C. Riggins, E.C. Rye, D. Brown, A study of MAC address randomization in mobile devices and when it fails (2017). CoRR arXiv: 1703.02874.

[16] OUI listing, (Mar. 2019). http://standards-oui.ieee.org/oui.txt.

[17] M. Afanasyev, T. Chen, G.M. Voelker, A.C. Snoeren, Usage patterns in an urban wifi network, IEEE/ACM Trans. Netw. 18 (5) (2010) 1359–1372, doi:10.1109/TNET.2010.2040087.

[18] F. Lyu, J. Ren, N. Cheng, P. Yang, M. Li, Y. Zhang, X.S. Shen, Big data analytics for user association characterization in large-scale wifi system, in: 2019 IEEE International Conference on Communications, ICC 2019, Shanghai, China, May 20–24, 2019, 2019, pp. 1–6, doi:10.1109/ICC.2019.8761511.

[19] G. Poucin, B. Farooq, Z. Patterson, Activity patterns mining in Wi-Fi access point logs, Comput Environ Urban Syst 67 (2018) 55–67, doi:10.1016/j.compenvurbsys.2017.09.004.

[20] F. Meneses, A.J.C. Moreira, Large scale movement analysis from wifi based location data, in: 2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 2012, pp. 1–9.

[21] A.J. Ruiz-Ruiz, H. Blunck, T.S. Prentow, A. Stisen, M.B. Kjaergaard, Analysis methods for extracting knowledge from large-scale WiFi monitoring to inform building facility planning, in: IEEE International Conference on Pervasive Computing and Communications (PerCom)(PERCOM), 00, 2014, pp. 130–138, doi:10.1109/PerCom.2014.6813953.

[22] X. Wei, N. Valler, H.V. Madhyastha, I. Neamtiu, M. Faloutsos, Characterizing the behavior of handheld devices and its implications, Comput. Netw. 114 (2017) 1–12.

[23] P.Y. Cao, G. Li, A.C. Champion, D. Xuan, S. Romig, W. Zhao, On human mobility predictability via WLAN logs, in: 2017 IEEE Conference on Computer Communications, INFOCOM 2017, Atlanta, GA, USA, May 1–4, 2017, 2017, pp. 1–9, doi:10.1109/INFOCOM.2017.8057234.

[24] P. Serrano, P. Salvador, V. Mancuso, Y. Grunenberger, Experimenting with commodity 802.11 hardware: overview and future directions, IEEE Commun. Surv. Tutor. 17 (2) (2015) 671–699, doi:10.1109/COMST.2015.2417493.

[25] M. Hang, I. Pytlarz, J. Neville, Exploring student check-in behavior for improved point-of-interest prediction, in: International Conference on Knowledge Discovery and Data Mining, in: KDD'18, ACM, 2018.

[26] Object Identifier (OID) repository, Accessed: 2018-09-30. (http://oid-info.com).

[27] OID Cisco, Accessed: 2018-09-30. (http://snmp.cloudapps.cisco.com/Support/SNMP/do/BrowseOID.do).

[28] D. Kotz, T. Henderson, I. Abyzov, J. Yeo, CRAWDAD dataset dartmouth/campus (v. 2009-09-09), 2009, (Downloaded from http://crawdad.org/dartmouth/campus/20090909).

José Camacho is Associate Professor in the Department of Signal Theory, Telematics and Communication and researcher in the Information and Communication Technologies Research Centre, at the University of Granada, Spain. He holds a degree in Computer Science from the University of Granada (2003) and a Ph.D. from the Technical University of Valencia (2007). His Ph.D. was awarded with the second Rosina Ribalta Prize to the best Ph.D. projects in the field of Information and Communication Technologies (ICT) from the EPSON Foundation, and with the D.L. Massart Award in Chemometrics from the Belgian Chemometrics Society. His research interests include exploratory data analysis, anomaly detection and optimization with multivariate techniques applied to data of very different nature, including manufacturing processes, chemometrics and communication networks. He is especially interested in the use of exploratory data analysis to Big Data.

Chris McDonald received the BSc degree in computer science and mathematics and the PhD degree in computer science both from the University of Western Australia. He currently holds the appointments of associate professor in the School of Computer Science and Software Engineering at the University of Western Australia (UWA) and adjunct associate professor in the Department of Computer Science at Dartmouth College, New Hampshire. He has recently taught in the areas of computer networking, security and privacy, mobile and wireless computing, software design and implementation, C programming, and operating systems at UWA and Dartmouth. Together with these areas, his research interests include wireless, ad hoc, and mobile networking; network simulation; and computer science education. He is a member of the ACM.

Ronald Peterson has been a Senior Programmer in the Dartmouth Computer Science Department for 20 years, providing software engineering support for mobile computing, mobile health, and wireless sensor network projects.

Xia Zhou is an Associate Professor in the Department of Computer Science at Dartmouth College. She co-directs the DartNets (Dartmouth Networking and Ubiquitous Systems) Lab, and the Dartmouth Reality and Robotics Lab (RLab). She received her PhD in Computer Science at UC Santa Barbara in June, 2013, working under the supervision of Prof. Heather Zheng. She was a visiting faculty in National Taiwan University from December 2016 to February 2017, and in University of Cambridge from April 2017 to June 2017. Her research interest lies broadly in mobile computing and its intersection with other disciplines. Most of her current projects center on light - an ubiquitous medium around us.

David Kotz is the Champion International Professor in the Department of Computer Science. He previously served as Interim Provost, as Associate Dean of the Faculty for the Sciences, as the Executive Director of the Institute for Security Technology Studies, and on the US Healthcare IT Policy Committee. His research interests include security and privacy, pervasive computing for healthcare, and wireless networks. He has published over 200 refereed papers, obtained over $67m in grant funding, and mentored nearly 100 research students. He is a Fellow of the IEEE, a Distinguished Member of the ACM, a 2008 Fulbright Fellow to India, and an elected member of Phi Beta Kappa. After receiving his A.B. in Computer Science and Physics from Dartmouth in 1986, he completed his Ph.D in Computer Science from Duke University in 1991 and returned to Dartmouth to join the faculty.