Flutes vs. *Cellos*: Analyzing Mobility-Traffic Correlations in Large WLAN Traces

Babak Alipour* babak.ap@ufl.edu
Leonardo Tonetto† tonetto@in.tum.de
Aaron Yi Ding† ding@in.tum.de
Roozbeh Ketabi* roozbeh@ufl.edu
Jörg Ott† ott@in.tum.de
Ahmed Helmy* helmy@ufl.edu

*Computer and Information Science and Engineering
University of Florida, Gainesville, USA

†Department of Informatics
Technical University of Munich, Munich, Germany

Abstract—**Two major factors affecting mobile network performance are *mobility* and *traffic* patterns. Simulations and analytical-based performance evaluations rely on models to approximate factors affecting the network. Hence, the understanding of mobility and traffic is imperative to the effective evaluation and efficient design of future mobile networks. Current models target either mobility or traffic, but do not capture their interplay. Many trace-based mobility models have largely used pre-smartphone datasets (e.g., AP-logs), or much coarser granularity (e.g., cell-towers) traces. This raises questions regarding the relevance of existing models, and motivates our study to revisit this area. In this study, we conduct a multi-dimensional analysis, to *quantitatively* characterize mobility and traffic spatio-temporal patterns, for laptops and smartphones, leading to a detailed integrated mobility-traffic analysis. Our study is *data-driven*, as we collect and mine capacious datasets (with 30TB, 300k devices) that capture all of these dimensions. The investigation is performed using our systematic (*FLAMeS*) framework. Overall, dozens of mobility and traffic features have been analyzed. The insights and lessons learnt serve as guidelines and a first step towards future *integrated mobility-traffic models*. In addition, our work acts as a stepping-stone towards a richer, more-realistic suite of *mobile test scenarios* and *benchmarks*.**

I. Introduction

Human mobility has been studied extensively and many models have been derived. The spectrum ranges from simple synthetic mobility models to complex trace-based models, capturing different properties with varying degrees of accuracy [1], [2]. Similarly, network traffic has been studied increasingly for wireless networks: for rather stationary users (as in WLANs) (e.g., [3], [4]) and potentially mobile users as for cellular networks (e.g., [5], [6]). Such analyses range from metrics such as flow count, sizes, and traffic volume to service usage (e.g., visited web sites, backend services).

Both mobility and network usage, characterize different aspects of human behavior. In this sense, we have a *mobility plane* and a *(network) traffic plane*. In reality, these two planes are likely interdependent. Human mobility may be influenced by network activity; for example, a person slowing down to read incoming messages. Also, network activity may be influenced by mobility and location; stationary users may produce/consume more data than those walking, and people may use different services in different places [7].

In earlier studies, this interdependence has not been widely considered, and models for both mobility and network traffic planes have been developed and evaluated largely in isolation. For example, when evaluating mobile systems' performance, traffic generation generally follows regular patterns, drawn from common simple distributions (e.g., exponential or uniform), while assuming neither transmission nor reception of data impacts mobility. Simply observing people walking while staring at (or reacting to) their smartphones suggests, however, that such interdependencies need to be captured properly. Understanding the mobility-traffic interplay is imperative to the effective evaluation and efficient design of future mobile algorithms ranging from user behavior prediction and caching, to network load estimation and resource allocation.

In this paper, we take a stab at understanding the interconnection of the mobility and traffic planes. To do this properly, we need to consider the nature of mobile devices people use: one class of devices is merely intended for stationary use, typically while the user is seated—this primarily holds for laptop computers, dubbed *cellos*. In contrast, another class—smartphones, which we refer to as *flutes*—lend themselves to truly mobile use[1]. We focus our analysis on these two classes because they have been around long enough to have extensive datasets to build upon. We stipulate that the interconnection of the mobility and traffic is modulated by the device(s) a mobile user is carrying. Therefore, we follow two main lines of investigation: we develop a framework to differentiate between cellos and flutes, and study both the mobility and traffic patterns for each of those types.

Specifically, the main goal of this paper is to quantitatively investigate the following questions in-depth: *(I) How different are mobility and traffic characteristics across device types, time and space? (II) What are the relationships between these characteristics? (III) Should new models be devised to capture these differences? And, if so, how?*

To answer these questions, a multi-dimensional (comparative) analysis approach is adopted to investigate mobility and traffic spatio-temporal patterns for flutes and cellos. We drive our study with capacious datasets (30TB+) that capture all the above dimensions in a campus society, including over 300k devices (Sec. IV). A systematic *F*ramework for *L*arge-scale

[1] Throughout this paper we shall use the terms *flutes* and *smartphones* interchangeably, and the terms *cellos* and *laptops* interchangeably.

*A*nalysis of *M*obil*e* *S*ocieties (*FLAMeS*) is devised for this study, that can also be used to analyze other multi-sourced data in future studies. Our main contributions include:

1) *Integrated mobility-traffic analyses* (Sec. VII): this study is the first to quantify the correlations of numerous features of mobility and traffic simultaneously. This can identify gaps in existing mobile networking models, and reopen the door for future impactful work in this area.
2) *Flutes vs. Cellos analysis* (Sec. V–VI): the device-type classification presented here, facilitates another important dimension to understand. This is particularly important as new generations of portable devices are introduced, that are different than the laptops, traditionally considered in earlier studies.
3) *Systematic multi-dimensional investigation framework* (Sec. III): *FLAMeS* provides the scaffolding needed to process, in multiple dimensions, many features of large sets of measurements from wireless networks, including AP-logs and NetFlow traces. This systematic method can apply to other datasets in future studies.

II. Related Work

To characterize mobility and network usage, existing studies have covered various aspects, including human mobility, device variation, and dataset analysis.

Human Mobility: Given its importance in various research areas, human mobility has received significant attention. We refer the reader to [1], [2] for surveys of mobility modeling and analysis. For spatial-temporal patterns, [8] and [9] reveal the regularity and bounds for predicting human mobility using cellular logs. A recent study has highlighted the importance of combining different datasets to study various features simultaneously [10]. Our observations are similar to [8], [9] which reassure the intrinsic properties of human mobility, despite the differences in granularity and population characteristics across datasets. To advance the understanding of human mobility, we integrated different datasets to correlate mobility and network traffic.

Device Variation: Usage and traffic patterns of different device types have been studied from various perspectives ([11], [12], [13], [14], [15], [16]). However, those findings are based on classifications that rely on either MAC addresses or HTTP headers solely. The former is rather limited and the latter may have serious privacy implications and are often unavailable. In [17], authors use packet-level traces from 10 phones and application-level monitoring from 33 Android devices to analyze smartphone traffic. Although this allowed fine-grained measurements, the approach is invasive and limited in scalability, leading to small sample sizes and restricted conclusions. They also do not compare the traffic of smartphones with that of "stop-to-use" wireless devices (i.e. cellos) nor do they measure spatial metrics. To characterize usage pattern for users with multiple wireless devices, Das et al. carried out wireless trace analysis on a university campus, covering more than 32k users [18]. Their study revealed usage difference between laptops/tablets and smartphones in terms of time, packets, content, intermittent and overlapping usage. Compared with our study to correlate mobility and traffic, their work targets at device usage patterns and security aspects. In our method, the combination of MAC and NetFlow allowed us to classify majority of observed devices while preserving users' privacy.

Dataset Analysis: The most recent work on WLAN traces [19] revealed surprising patterns on increases of long-term mobility entropy by age, and the impact of academic majors on students' long-term mobility entropy. The authors of [7] investigated correlations and characteristics of web domains accessed by users and their locations of users based on NetFlow and DHCP logs from a university campus in 2004. They propose a simulation paradigm whose parameters are extracted, producing realistic scenarios for simulations. However, the study uses data from pre-smartphone era and does not distinguish between device types. It also does not analyze the relationship between mobility and traffic. On both WiFi and cellular networks, the authors of [20] performed an in-depth study on smartphones traffic, highlighting the benefits and caveats/limitations of using MPTCP. Distributions of flow, Inter-arrival time (IAT) and arrival rate at APs of "static" flows has been analyzed against popular distributions (e.g., Exp, Weibull, Pareto, Lognormal) [21]. Lognormal was found to best fit the flow sizes, while at small time scales (i.e. hourly), IAT was best described by Weibull but parameters vary from hour to hour. We analyze flows on a much larger scale, newer dataset including smartphones, and identify Lognormal distribution as the best fit for flow sizes, and beta as best for IAT, regardless of device type. Xu et al. conducted a large scale ISP trace analysis that covers over 9600 cellular towers and 150k subscribers of city Shanghai in August 2014 [22]. Their study identified the mapping between time-domain traffic patterns and five types of urban functional regions, yielding several insights on mobile traffic patterns across time, location and frequency domains. This work is complementary to our study on campus WLAN traces as they focus on cellular networks in urban areas.

III. Systematic multi-dimensional analysis

To methodically analyze statistical characteristics and correlations in multiple dimensions, we introduce the *FLAMeS* framework (Fig. 1). The main components include: I. Data collection and pre-processing, II. Flutes vs. cellos mobility and traffic analysis, and III. Integrated mobility-traffic analysis.

The two main purposes of this work are to understand and **quantify** the *gaps between flutes and cellos*, and the *interaction between the mobility and traffic dimensions*. Individual mobility and traffic analyses for flutes and cellos are conducted in Sections V and VI, respectively, with detailed reporting for spatio-temporal features showing significant gaps. In Sec. VII, the most important mobility and traffic features are identified and their correlation quantified.

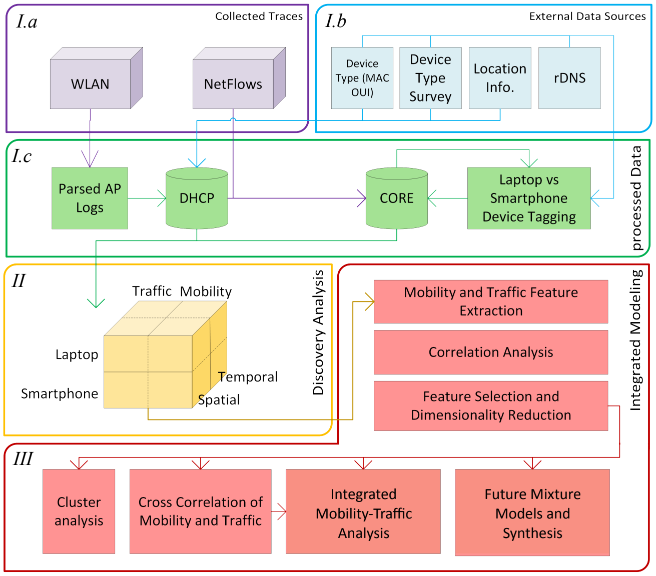


Fig. 1: *FLAMeS* system overview.

IV. Experimental Setup and Datasets

We drive our framework with large-scale datasets from multiple sources, capturing the mobility and traffic features in different dimensions. In this section, we introduce the two data sets and their preprocessing, and present the device type classification into flutes and cellos.

The *input datasets* in this study are specifically chosen to capture: 1. location, mobility and network traffic information, 2. smartphone and laptop devices, 3. spatio-temporal features, and 4. scale in number of devices and records. The total size is >**30TB**, consisting of two main parts: WLAN Access Point (AP) logs, and Netflow records (details in Tables I, II).

A. WLAN AP Logs

These logs are collected from 1760 APs in 138 buildings over 479 days on a university campus, and contain association and authentication events from 316k devices in 2011-2012. It contains over 555M records, with each record including the device's MAC and assigned IP addresses, the associated AP and a timestamp. Locations of the APs are approximated by the building locations where they are installed, i.e., (longitude, latitude) of Google Maps API. To validate this, we fetched 8000 mapped APs around the campus area from a crowd-sourced service, *wigle.net*. For the 130 matched APs in 42% of buildings (i.e., 58 bldgs), all were less than 200m from their mapped location; an error of less than 1.5% of the campus area. This is very reasonable for our study purposes.

B. NetFlow Logs

Over **76 billion** records of NetFlow traces were collected from the same network, over 25 days in April 2012. A *flow* is defined as a consecutive sequence of packets with the same transport protocol, source/destination IP and port number, as identified by the collecting gateway router. An example of major Netflow data fields is presented in Table II.

The NetFlow records are matched with the wireless associations (from the AP logs) using the dynamic MAC-to-IP address mapping from the DHCP logs. We refer to the result as *CORE* dataset (Table I). They are also augmented with location and website information using reverse DNS (rDNS)[2].

TABLE I: Summary of datasets. B=billion.

	# Records		Traffic Vol. (TB)		# MAC	
	DHCP	CORE	TCP	UDP	WLAN	CORE
Flutes	412.0 M	2.13 B	56.18	4.50	186.0 K	50.3 K
Cellos	101.0 M	4.20 B	73.85	12.90	93.2 K	27.1 K
Total	557.5 M	6.53 B	134.39	17.61	316.0 K	80.0 K

C. Device Type Classification

To classify devices into flutes and cellos, we utilize several observations and heuristics. To start, note that a device manufacturer (with OUI) can be identified based on the first 3 octets of the MAC address[3]. Most manufacturers produce one type of device (either laptop or phone), but some produce both (e.g., Apple). In the latter case, OUI used for one device type is not used for another. We conducted a survey to help classify 30 MAC prefixes accurately. Using OUI and survey information, we identify and label 46% of the total devices (90k cellos and 56k flutes). Then, from the NetFlow logs of these labeled devices, we observe over 3k devices (92% of which are flutes) contacting *admob.com*; an ad platform serving mainly smartphones and tablets (i.e. flutes). This enables further classification of the remaining MAC addresses. Finally, we apply the following heuristic to the dataset: (1) obtain all OUIs (MAC prefix) that contacted *admob.com*; (2) if it is unlabeled, mark it as a flute. Overall, over 270k devices were labeled (180k as flutes), covering 86% of the devices in AP logs and 97% in NetFlow traces, a reasonable coverage for our purposes. Out of $\approx 80k$ devices in the NetFlow logs, $\approx$ 50K are flutes and $\approx$ 27K cellos.

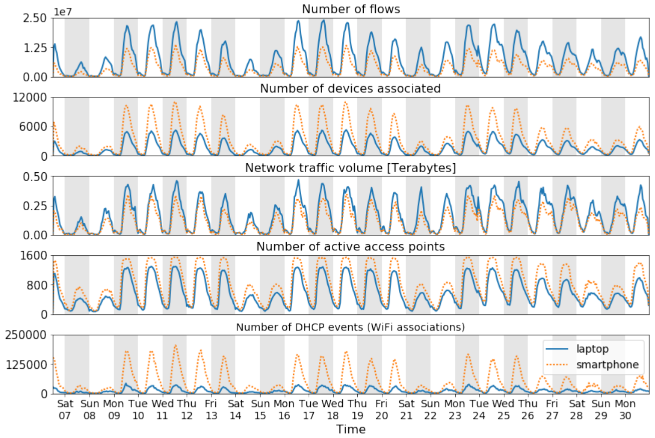


Fig. 2: Time series for 25 days of combined AP-NetFlow Core traces

Fig. 2 shows the temporal plot for the combined traces over 25 days, after device classification. Throughout, the number

[2] Detailed dataset merging and the query computing system are available in the Appendix II of our technical report [23].

[3] MAC address randomization does not affect our association trace.

of flows and total traffic volume is clearly higher for cellos, even with an overall higher number of flutes connected. Also note the device activities in a *diurnal* and *weekly* cycles, with the peaks occurring during weekdays, as expected. Wed, 25th, was the last day of classes, explaining the decline in network activity afterwards. The plot motivates our analyses for flute vs cellos, over *weekends vs weekdays*, for the rest of the study.

V. Mobility analysis

This section covers the *temporal* and *spatial* mobility analyses. For all metrics, unless otherwise noted, we investigate 479 days. A summary of studied metrics and their most significant statistical values are presented in Tab. III along with mean and median ratios for comparison. From that list, we further investigate in this section those metrics that show the most interesting or non-trivial differences between *flutes* and *cellos*.

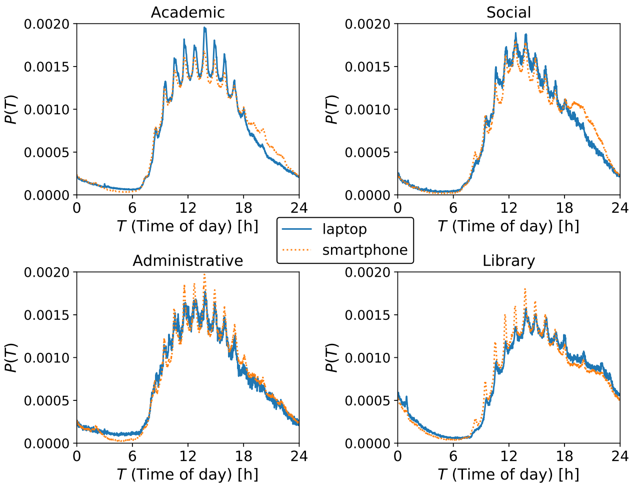


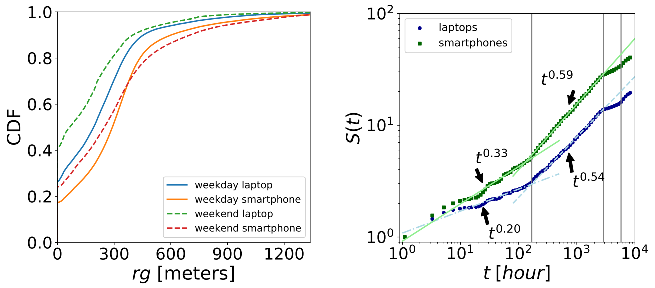
Fig. 3: PDF Session start over time of the day.

A. Session start probability

We define a session as the period between WLAN associations. The distributions of session start times across the day for four building categories are depicted in Fig. 3. aspects of analysis for this metric. The start times of the Sessions match the periodic beginning of classes, but mainly in *Academic* buildings, where users move mostly at the start and end of classes. In these places, activity drops sharply for *cellos* at 5pm, with considerable *flutes* activity until 8pm. For *Social* and *Library* buildings, *the probability of new sessions remains higher for a few more hours into the evening, and the times users tend to leave are more spread out*. We do not make similar observation during weekends, which is expected when the day is, unlike weekdays, not governed by a class schedule. For most visitors, the session start distributions show a smooth shape and no significant differences between device types (omitted for brevity).

B. Radius of Gyration

This metric, GYR, captures the size of the geospatial dispersion of a device's movements, denoted by r_g and computed as $r_g = \frac{1}{N}\sum_{k=1}^{N}(\vec{r_k} - \vec{r_s})^2$, where $\vec{r_1},...,\vec{r_N}$ are positional vectors of a device and $\vec{r_s}$ is its center of gravity.



(a)

(b)

Fig. 4: (a) Radius of gyration (rg for the device types). (b) Visited locations $S(t)$. Vertical lines at 7, 120 and 240 days.

Grouping devices by their r_g after six months of observation, we look at its evolution since the first time they are observed. Unsurprisingly (cf. [8]), after an initial transient period of about one week, this value stabilizes even across different semesters (not shown).

We split the traces into weekdays and weekends, presenting the distributions in Fig. 4a. Comparing these groups for *cellos*, we notice a substantial reduction in their overall mobility while, for *flutes*, this difference is not so pronounced. This might be due to students having fewer activities on weekends, a tendency to study at a single building like a library, or just not carry their cellos; we will revisit this aspect in Sec. VII. *Flutes*, being "always-on" devices, are able to capture movements at pass-by locations, dining areas, and bus stops and thus are better suited to capture the fine-granular mobility of their users than cellos.

Despite the 8.1km^2 area of the campus (approximate radius of 1.42km), buildings with related fields of study (e.g. Fine Arts) are somewhat clustered. Computing the distance between the k-nearest neighboring buildings, for $k = 22$ and $k = 9$ (average number of visited buildings for *flutes* and *cellos*) the median distances are 295m and 172m, respectively. Due to their focus on classes, students attending have limited area of activity on weekdays, which explains the observed *radius of gyration*.

We also evaluated: (1) *diameter* DIA, the longest distance between any pair of $\vec{r_k}$ points; (2) *max jump* LJM, the longest distance between a pair of consecutive $\vec{r_k}$ points; and (3) *total trajectory length* TJM, the sum of all trips made by a device. The distributions of these metrics are similar to *Radius of Gyration* and therefore not shown. Table III summarizes the most significant statistical values for these metrics.

C. Visitation preferences and interests

We count the number of unique buildings visited by a user, BLD, and define a *preferred building* as the location where a device has spent most of its time in a given day, which we measure in minutes and refer to as PDT. We approximate the latter by the formula $t_b = \sum_{k=1}^{N_b} S_k$, where t_b is the time spent, N_b the total number of sessions and $S_1...S_N$ the time

TABLE II: NetFlow (top) and AP logs/DHCP (bottom) sample data

Start time	Finish time	Duration	Source IP	Destination IP	Protocol	Source port	Destination port	Packet count	Flow size
1334332274.912	1334332276.576	1.664	173.194.37.7	10.15.225.126	TCP	80	60482	157	217708

User IP	User MAC	AP name	AP MAC	Lease begin time	Lease end time
10.130.90.3	00:11:22:33:44:55	b422r143-win-1	00:1d:e5:8f:1b:30	1333238737	1333238741

TABLE III: General results for mobility. Upper values are for weekdays and lower ones for weekends (in red color). **LJM**: maximum jump [m]; **DIA**: diameter [m]; **TJM**: total trajectory length [m]; **GYR**: radius of gyration [m]; **BLD**: no. uniq. buildings; **APC**: access point count; **PDT**: time spent at preferred building [minutes]; **DLT**: total session time at each building.

	Flutes (F)			**Cellos (C)**			**Ratio (C/F)**	
	μ	*mdn*	σ	μ	*mdn*	σ	μ	*mdn*
LJM	435 350	296 168	813 683	178 97	1 1	624 312	0.409 0.277	**0.003** 0.006
DIA	549 425	411 179	874 739	195 107	1 1	642 338	0.355 0.252	**0.002** 0.006
TJM	1582 1036	707 279	2336 1793	378 252	1 1	1444 1766	0.239 0.243	**0.001** 0.004
GYR	396 330	290 248	2725 1368	321 178	191 65.1	3265 1800	1.102 1.247	1.019 1.4
BLD	5.4 2.8	3 2	5.6 4.1	1.8 1.5	1 1	2.1 1.8	0.811 0.539	0.659 0.262
APC	11.8 7.2	6 4	13.3 8.8	3.7 3	2 2	4.8 3.8	0.333 0.536	0.333 0.5
PDT	225 223	161 135	219 272	248 278	164 189	254 292	0.314 0.417	0.333 0.5
DTL	316 326	235 247	302 308	316 316	217 221	305 309	1 0.97	0.92 0.89

duration of *each session* at a building b, here referred as DLT. Interestingly, cellos have slightly longer stays but both have medians around 2:40 hours. The similarity of the distributions, combined with a lower number of visited locations indicate that cellos are used mostly when users remain longer periods at places.

Fig. 4b highlights the differences between *flutes* and *cellos* on the required time t to visit $S(t)$ locations. *After an initial exploration period of one week the rates of new visits change similarly for both device types, and new exploration rates show up at 120 and 240 days.* These could be explained by the weekly schedules of the university as well as the usual length of a lecture term (≈ 4 months).

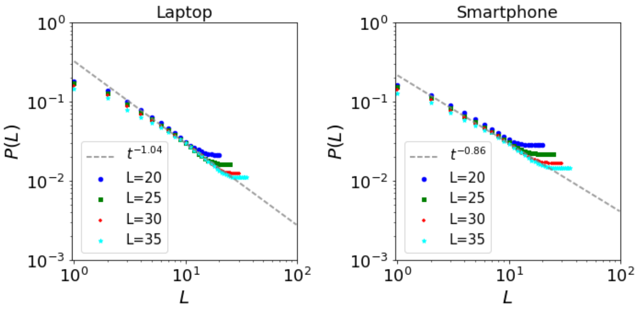


Fig. 5: Zipf's plot on L visited access points.

We also consider the number of unique access points a device associates with, APC, which provides a finer spatial resolution than the building level. Furthermore, the probability of finding a device at its *L-th* most visited access point is shown in Fig. 5. When taking buildings as aggregating points for location, the values become $L^{-1.36}$ for *cellos* and $L^{-1.16}$ for *flutes*. These approximations validate previous work on human mobility [8], yet highlighting differences between device types.

D. Sessions per building

To study AP utilization over time, we look at the session duration distribution, or session duration dispersal kernel P(t), depicted in Fig. 6. The smaller inner plots represent the same metric, limited to four types of buildings.

We noted that the five-minute spikes correspond to default idle-timeout for the used WiFi routers. On the other hand, the *knees* at 1 and 2 hours could be explained by the typical duration of classes. They are only noticeable at Academic buildings (shown inside inner plots) and during weekdays (not shown). This leads us to conclude that despite the differences in distributions of device types, *flutes* and *cellos* present *certain similarities in their usage, such as during classes.* To differentiate *pass-by* access points, we examine all sequences of three unique APs where all session durations are lower than 5 minutes (typical idle-timeout). We observed these APs clustered at buildings that also had major bus stops nearby.

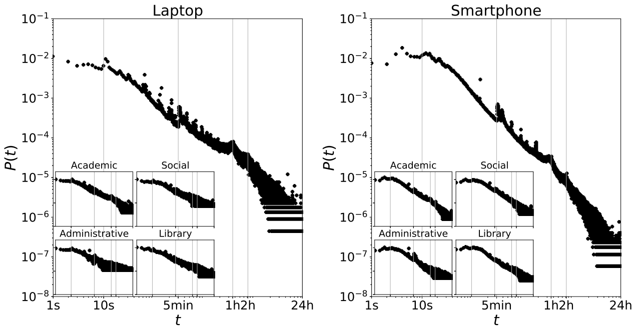


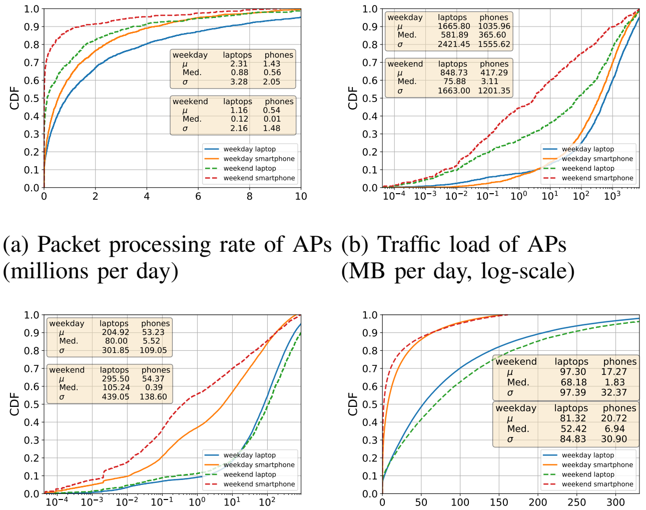
Fig. 6: Probability $P(t)$ of session duration t.

VI. Traffic Analysis

In this section, we compare different *traffic* characteristics, across *device types*, *time* and *space*. For this purpose, we start with statistical characterization of *individual* flute and cello flows. Next, we measure how these flows, *put together*, affect the network patterns across APs and buildings. Finally, *user behavior* is analyzed by monitoring weekly cycles, data rates, and active durations. By quantifying *temporal* and *spatial* variations of traffic across device types, we make a case for new models to capture such variations based on the most relevant attributes. Table IV summarizes the results.

A. Flow-level statistical characterization

We compare the following distributions using maximum likelihood estimation (MLE) and maximizing goodness-of-fit



(a) Packet processing rate of APs (millions per day)

(b) Traffic load of APs (MB per day, log-scale)

(c) User data consumption (MB per day, log-scale)

(d) User active time (minutes per day)

Fig. 7: Distribution plots

estimation: Gaussian, Exponential, Gamma, Weibull, Logistic, Beta and Lognormal[4].

1) Size: Flow size is the sum of bytes for all packets within a single flow. On weekdays, average *size of individual flute flows* is $> 2x$ **larger** *than cello flows* (2070 vs. 822 bytes), while median is $> 4x$ **larger** (678 vs. 142 bytes). There are no significant changes on weekends (See Appendix IV in [23] for details).

The average packet size within a single flow also provides insight into packet-level behavior of services on mobile devices. We notice that the average *packet size of flute flows* is ≈**50% larger** *than that of cellos* (212 vs 144 bytes on weekdays, 205 vs 142 on weekends). Comparing weekdays and weekends, median size of flute packets drops on weekends whereas it remains *the same* for cellos. In fact, comparing cello flows on weekdays and weekends shows *no significant difference* in terms of average packet size (p-value$>$ .05). In spite of smaller flows, the average cello generates *2.7 times traffic* as an average flute because the average cello is responsible for *3.7 as many flows* as a flute. Analyzing distributions of flow size and average packet size in our datasets shows that *Lognormal* distribution is the best fit, with varying parameters for each device type (See Appendix IV in [23] for details).

2) Packets: This metric is the count of packets within each flow. The mean and median packet counts per flow are 7.06 and 5 in flutes and 3.64 and 2 in cellos, during weekdays. The means drop slightly on weekends. Packet counts per flow match the *Lognormal* distribution well for flows of both device types. The average flute flow is bigger in size and has *more packets* (with higher variance) but there are *fewer* flows coming from these devices. This is analyzed further for TCP/UDP flows (Sec. VI-A5).

3) Runtime: Flow runtime is the period of time the flow was active (equal to a flow's $finishtime - starttime$). Flute flows have a mean and median of 1868ms and 128ms respectively on weekdays, while these numbers are 1639ms and 64ms for cellos. Both device types show increase in means during weekends (flutes by 204 and cellos by 164), indicating that although there are fewer devices online during weekends, they are more active. The low medians in either group corresponds to many *short-lived* flows with few packets, showing little variation across device type, time or space.

4) Inter-arrival times (IAT): Median of the flow *IAT*[5] at access points is 6ms for cello flows and 4ms in case of flutes, on weekdays (similar on weekends), which suggests that the majority of access points handle flows from either device type at nearly the same rate. However, average *IAT* is ≈ 143ms for flute and ≈ 78ms for cello flows, as there are more cellos with very high rate of flows. Flow *IAT* in our datasets matches a **beta** distribution well (See Appendix IV in [23]) with a *very high estimated kurtosis* and *skewness* (estimated at 58 & 6.9 respectively). The high estimated kurtosis illustrates that there are *infrequent extreme values*, which explains the observed highly elevated standard deviation of *IAT*. Higher average IAT of flutes, combined with the higher standard deviation compared to cellos (596 vs 284), shows that *flutes face more extreme periods of inactivity, which can be caused by higher mobility and packet loss.*

5) Protocols: TCP accounts for *78.5%* of cello flows (**84.6% of bytes**) and *98.2%* of flute flows (**91.6% of bytes**). The higher presence of UDP in cellos is reasonable, considering that UDP applications (e.g., multi-player games, video conferencing and file sharing) are more likely to be used with cellos. Comparing the number of packets in flows, in case of TCP, the average number of packets in cello flows is almost half that of a flute flow (4.6 vs 8.8), and the average packet size of flutes is 22% higher than that of cellos. This supports our earlier observation regarding the bigger flows sizes of flutes. However, for UDP, the two device types are similar in terms of average packet count per flow (2.5 & 2.87 for cellos and flutes respectively) and average packet size (119 for both). This conforms to low latency requirements of many UDP applications.

Given these differences, traffic classification using machine learning [26] could benefit from considering device types to train models. We investigate this in VII-B.

After establishing the similarities and differences of flows, the next step is to evaluate whether the individual variations in flows lead to different *aggregate traffic behaviors* from viewpoint of the network.

[4] For distribution comparison, significance threshold $p-value$ is set at .05.

[5] IAT is important in simulation and modeling of networking protocols, traffic classification [24], congestion control and traffic performance [25]. Our flow-level IAT analysis can also be used for measuring delay and jitter effects.

B. *Network-centric (spatial) analysis*

We examined the load of APs in all buildings on a daily basis to provide insight into differences from the viewpoint of the network. For each AP, we calculate flow metrics for every weekday and weekend.We focus our analysis on the first three weeks of NetFlow traces to avoid significant user behavior change during exams period, as already shown in Fig. 2.

First, we measure the daily packet and flow arrival rates at APs. The median flow rates are $42k$ and $20k$ per weekday for cellos and flutes respectively (7.5k and 0.5k on weekends). The average number of cello packets processed daily by APs is **1.6 times higher** than flute packets (Fig. 7a). Each AP handles, on average during weekdays, ≈ 27 *cello packets per second* and ≈ 17 *flute packets per second*, dropping to ≈ 13.5 and ≈ 6.25 on weekends. This indicates that, during the weekends, a high percentage of access points are not utilized, with *60% of APs seeing no flute flows* and *70% receiving no cello flows*. However, *at least one AP in >80% of buildings sees traffic*, supporting observations of less mobility during weekends.

Next, we look at traffic volume. On average weekdays, 90% of APs handle < $5GB$ of cello traffic ($2.5GB$ on weekends), whereas the same percentage handles < $3GB$ of flute traffic ($1GB$ on weekends) (Fig. 7b). Flutes are more mobile, visit a higher number of unique APs and have bigger flow sizes but they are still responsible for *less overall network load.*

Thus, the individual differences of flute and cello flows result in *heterogeneous aggregate traffic patterns* in time (different days) and space (APs at different buildings)[6]. With that established, in order to take steps towards modeling and simulation, we also need to analyze the behavior of users.

C. *User behavior (temporal) analysis*

Here, we measure traffic patterns from a user-centric perspective. We identified gaps in diurnal and weekly cycles (Fig. 2) as well as traffic flow features of individual *users* including data consumption, packet rates, and network activity duration.

1) Data consumption: Fig. 7c shows daily data consumption, with 90% of cellos consuming **< 700MB** and 90% of flutes using **< 200M** on weekdays. Surprisingly, for cellos on campus during weekends, average data consumption is even higher whereas data consumption of flutes drops sharply.

2) Packet rate: On weekdays, cellos on average generate $\approx$**318K packets**, while flutes only average $\approx$**84K packets** per day. On weekends, the few on-campus cellos see greatly increased number of packets, with an average daily packet rate of ≈ 495K. Weekend flutes also have a modestly *increased* packet count, with an average of ≈ 96K flows.

3) Active duration: Total active time of devices serves well to demonstrate the differences between time spent online by users of different device types. We rely on NetFlow to measure **'active'** time instead of AP association time. This allows us to distinguish user's *idle* presence in the network from its *activity* periods. Cellos have **4x** average active time compared to flutes in our traces (≈ 81 vs ≈ 21 min on weekdays, ≈ 97 vs ≈ 17 min on weekends). Overall, 90% of cellos are active for <**3.5h** and 90% of flutes are active for <**1h** (Fig. 7d). As evident in various metrics, the cellos appearing on weekends are more active than the average cello on weekdays.

Overall, the data consumption of flutes seems to be *more bursty* in nature, with **bigger** flows and **lower active duration**. This could be due to more intermittent usage of flutes and also bundling of network requests to save battery on these devices. In addition, there are fewer devices on campus during weekends, but those remaining devices are more active and consume more data than average.

VII. Integrated Mobility-traffic analysis

By studying the relationship between features from mobility and network traffic, we examine whether the *fusion* of these dimensions provides a case for the necessity of an *integrated mobility-traffic model* and introduce steps towards a combined model (Sec. VII-B).

A. *Feature engineering*

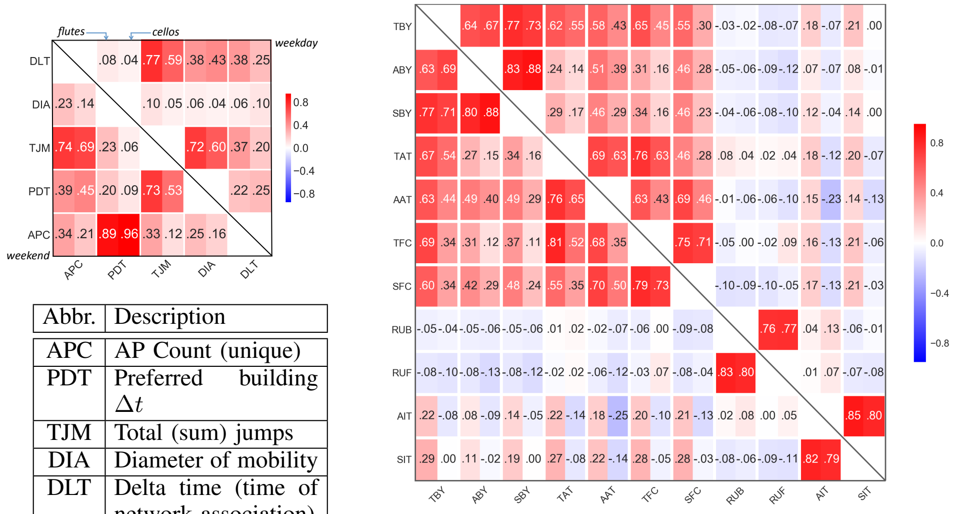
To simplify analysis and interpretation, and reduce dimensionality, we identify the most important features. First, we study the relationships among variables from *mobility* and *traffic* dimensions separately. Then, from this subset of combined features, we investigate whether clusters of user devices appear in the dataset. For this, we use correlation feature selection (*CFS* [27]), to obtain uncorrelated features[7], but highly correlated to the classification. Finally, we quantify correlations between mobility and traffic metrics (See abbreviations in Fig. 8).

1) Mobility: The *CFS* algorithm was run on 8 features (in Sec. V), and kept only *5* (to be used in the cross-dimension analysis). Fig. 8a visualizes the linear dependence between mobility features, comparing flutes and cellos on weekdays and weekends. Close inspection reveals temporal correlation relationships. For example, for cellos on weekdays, there is a **strong** correlation (0.96) between preferred building time (*PDT*) and time of network association (*DLT*), but weak correlation (0.1) on weekends, suggesting that most of weekend online time is spent at preferred buildings (e.g., libraries).

2) Traffic: We extract statistical measures for traffic metrics (Sec. VI) per device per day. The *CFS* algorithm was run on 19 features, reducing them to 11. A summary of these metrics is provided in Table IV. The correlations are depicted in Fig. 8b. The analysis shows us that average number of packets and bytes are positively correlated, but negatively correlated with variance of bytes and uncorrelated with IAT. Average IAT (*AIT*) seems to be mostly independent from other traffic features, but as *AIT* increases, its standard deviation (*SIT*) also greatly increases which could be due to device mobility; bearing further investigation on traffic-mobility interactions. Interestingly, active time is *weakly correlated* with number of flows and packets, which shows that users who remain online longer are *not* necessarily consuming traffic at a high rate.

[6] A more in-depth analysis is presented in the Appendix IV in [23].

[7] We show Pearson correlation for simplicity, and omit non-linear correlation for brevity.



	APC	PDT	TJM	DIA	DLT
DLT		.08 .04	.77 .59	.38 .43	.38 .25
DIA	.23 .14		.10 .05	.06 .04	.06 .10
TJM	.74 .69	.23 .06		.72 .60	.37 .20
PDT	.39 .45	.20 .09	.73 .53		.22 .25
APC	.34 .21	.89 .96	.33 .12	.25 .16	

Abbr.	Description
APC	AP Count (unique)
PDT	Preferred building Δt
TJM	Total (sum) jumps
DIA	Diameter of mobility
DLT	Delta time (time of network association)

(a) Mobility

	TBY	ABY	SBY	TAT	AAT	TFC	SFC	RUB	RUF	AIT	SIT
TBY		.64 .67	.77 .73	.62 .55	.58 .43	.65 .45	.55 .30	-.03 -.02	-.08 -.07	.18 -.07	.21 .00
ABY	.63 .69		.83 .88	.24 .14	.51 .39	.31 .16	.46 .28	-.05 -.06	-.09 -.12	.07 -.07	.08 -.01
SBY	.77 .71	.80 .88		.29 .17	.46 .29	.34 .16	.46 .23	-.04 -.06	-.08 -.10	.12 -.04	.14 .00
TAT	.67 .54	.27 .15	.34 .16		.69 .63	.76 .63	.46 .28	.08 .04	.02 .04	.18 -.12	.20 -.07
AAT	.63 .44	.49 .40	.49 .29	.76 .65		.63 .43	.69 .46	-.01 -.06	-.06 -.10	.15 -.23	.14 -.13
TFC	.69 .34	.31 .12	.37 .11	.81 .52	.68 .35		.75 .71	-.05 .00	-.02 .09	.16 -.13	.21 -.06
SFC	.60 .34	.42 .29	.48 .24	.55 .35	.70 .50	.79 .73		-.10 -.09	-.10 -.05	.17 -.13	.21 -.03
RUB	-.05 -.04	-.05 -.06	-.05 -.06	.01 .02	-.02 -.07	-.06 .00	-.09 -.08		.76 .77	.04 .13	-.06 -.01
RUF	-.08 -.10	-.08 -.13	-.08 -.12	-.02 .02	-.06 -.12	-.03 .07	-.08 -.04	.83 .80		.01 .07	-.07 -.08
AIT	.22 -.08	.08 -.09	.14 -.05	.22 -.14	.18 -.25	.20 -.10	.21 -.13	.02 .08	.00 .05		.85 .80
SIT	.29 .00	.11 -.02	.19 .00	.27 -.08	.22 -.14	.28 -.05	.28 -.03	-.08 -.06	-.09 -.11	.82 .79	

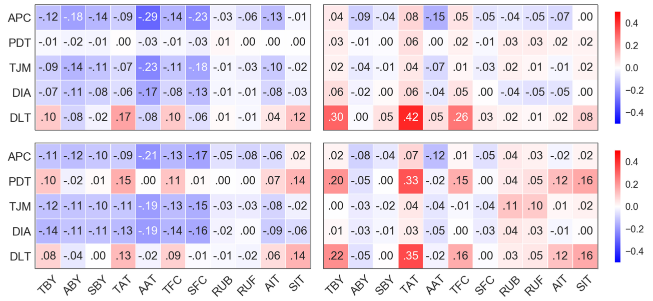
Abbr.	Description
TBY	Total flow bytes
ABY	Avg. flow bytes
SBY	Std. flow bytes
TAT	Total active time
AAT	Avg. active time
TFC	Total flow count
SFC	Std. flow counts
RUB	UDP bytes / total bytes
RUF	UDP flows / total flows
AIT	Avg. IAT
SIT	Std. IAT

(b) Traffic

Fig. 8: Correlation plots for (a) *mobility* and (b) *traffic* features. Each cell's left half is for flutes and right half is for cellos, the upper right triangle is for weekdays and the lower left for weekends.

TABLE IV: Summary of traffic features used for integrated mobility-traffic analysis (per device, averaged over all days; see Fig. 8 for abbreviations). Upper values are for weekdays and lower ones for weekends (in red color).

	Flutes (F)			Cellos (C)			Ratio (C/F)	
	μ	***mdn***	σ	μ	***mdn***	σ	μ	***mdn***
TBY [MB]	96.77 80.96	11.47 0.86	194.52 195.15	373.08 448.87	144.68 180.23	554.54 623.86	3.85 5.54	**12.61** 209.56
ABY	5.48 4.54	0.74 0.15	14.02 14.16	15.67 18.06	7.34 8.34	25.81 28.71	2.85 3.97	**9.91** 55.6
SBY	10.56 8.09	1.57 0.13	23.76 21.48	30.59 33.21	13.77 15.42	49.82 53.39	2.89 4.10	**8.77** 118.61
TAT	1,330 1,059	388.6 90.89	2,517 2,497	5,123 5,883	3,003 3,861	6,444 6,934	3.85 5.55	7.73 42.48
AAT	63.14 50.60	27.97 12.98	86.69 85.27	188.26 206.89	166.93 184.17	138.70 156.53	2.98 4.08	5.96 14.18
TFC [K]	7.2 5.7	1.7 0.3	15.61 15.01	33.5 38.5	17.1 20.6	60.10 88.52	4.65 6.75	**10.05** 68.66
SFC	515.6 361.05	177.3 30.18	907.7 796.6	1,640 1,673	1,181 1,215	2,081 2,098	3.18 4.63	6.66 40.27
RUB	0.05 0.06	0.00 0.00	0.19 0.22	0.07 0.08	0.00 0.00	0.22 0.23	1.4 1.33	N/A N/A
RUF	0.07 0.09	0.00 0.00	0.18 0.22	0.12 0.13	0.02 0.02	0.22 0.24	1.71 1.44	N/A N/A
AIT	3.36 2.95	2.24 1.74	3.59 3.60	3.40 3.18	2.45 2.27	3.51 3.39	1.01 1.07	1.09 1.3
SIT	5.22 4.09	3.44 1.98	5.50 5.06	5.14 4.72	3.18 2.79	5.28 4.96	0.98 1.15	0.92 1.41



	TBY	ABY	SBY	TAT	AAT	TFC	SFC	RUB	RUF	AIT	SIT
APC	-.12	-.18	-.14	-.09	-.29	-.14	-.23	-.03	-.06	-.13	-.01
PDT	-.01	-.02	-.01	.00	-.03	-.01	-.03	.01	.00	.00	.00
TJM	-.09	-.14	-.11	-.07	-.23	-.11	-.18	-.01	-.03	-.10	-.02
DIA	-.07	-.11	-.08	-.06	-.17	-.08	-.13	-.01	-.01	-.08	-.03
DLT	.10	-.08	-.02	.17	-.08	.10	-.06	.01	-.01	.04	.12

	TBY	ABY	SBY	TAT	AAT	TFC	SFC	RUB	RUF	AIT	SIT
APC	.04	-.09	-.04	.08	-.15	.05	-.05	-.04	-.05	-.07	.00
PDT	.03	-.01	.00	.06	.00	.02	-.01	.03	.03	.02	.02
TJM	.02	-.04	-.01	.04	-.07	.01	-.03	.02	.01	-.01	.02
DIA	.06	-.02	.00	.06	-.04	.05	.00	-.04	-.05	-.05	.00
DLT	.30	.00	.05	.42	.05	.26	.03	.02	.01	.02	.08

	TBY	ABY	SBY	TAT	AAT	TFC	SFC	RUB	RUF	AIT	SIT
APC	-.11	-.12	-.10	-.09	-.21	-.13	-.17	-.05	-.08	-.06	.02
PDT	.10	-.02	.01	.15	.00	.11	.01	.00	.00	.07	.14
TJM	-.12	-.11	-.10	-.11	-.19	-.13	-.15	-.03	-.03	-.08	-.02
DIA	-.14	-.11	-.11	-.13	-.19	-.14	-.16	-.02	.00	-.09	-.06
DLT	.08	-.04	.00	.13	-.02	.09	-.01	-.01	-.02	.06	.14

	TBY	ABY	SBY	TAT	AAT	TFC	SFC	RUB	RUF	AIT	SIT
APC	.02	-.08	-.04	.07	-.12	.01	-.05	.04	.03	-.02	.02
PDT	.20	-.05	.00	.33	-.02	.15	.00	.04	.05	.12	.16
TJM	.00	-.03	-.02	.04	-.04	-.01	-.04	.11	.10	.01	.02
DIA	.01	-.03	-.01	.03	-.05	.00	-.03	.04	.03	-.01	.00
DLT	.22	-.05	.00	.35	-.02	.16	.00	.03	.05	.12	.16

Fig. 9: Correlation plots of mobility vs. traffic on weekdays (top) vs. weekends (bottom) for flutes (left) and cellos (right).

Examining weekdays and weekends, correlation trends among traffic features remain similar for either device type.

3) Cross-dimension: Studying correlations across mobility and traffic dimensions, based on subsets of features selected by *CFS*, is a solid step towards an integrated mobility-traffic model. Results are presented in Fig. 9. We find that as the numbers of unique APs/buildings visited (*APC*, *BLD*) *increase*, the average active time (*AAT*), and total and std. of flow counts (*TFC* and *SFC*) *decrease* markedly (significant negative correlation). Surprisingly, there is no noticeable change in total traffic consumed with change in *APC* (negligible correlation), suggesting bundling of more packets in flute flows. (Similar correlation between mobility diameter and the above traffic features) Average IAT (*AIT*) of flutes also rises slightly as mobility metrics *decrease*; for cellos this correlation is almost *nonexistent*. This reinforces our *"stop-to-use"* categorization; cellos are movable but are not active in transit. To sum, *flutes score high on mobility metrics*, have an overall lower flow count and network traffic but produce bigger flows on average. For cellos, on weekends the more time spent at preferred buildings the higher the total active time (*TAT*) and flow counts; this effect exists to a lesser degree for flutes. On weekdays, such correlation does not exist.

B. Steps towards modeling

Here we present our steps towards an integrated mobility-traffic model, with various applications in simulation and protocol design. We utilize daily mobility and traffic features of users during a week[8]. First, we examine how different mobility and traffic features are for flutes and cellos using machine learning. Second, we investigate whether natural convex clusters of users appear in the dataset. These steps verify that the differences of mobility and traffic characteristics across device types are *significant*. We also find that *combining*

[8] Introducing a new detailed model is part of future research.

features from *mobility* ***and*** *traffic* makes this distinction even more *clear*. Finally, mixture models are used to model and synthesize simulated data points of each device type, finding that the accuracy the mixture model *increases* when trained on *combined* features.

1) Supervised classification: Having shown significant differences throughout this study, we used support vector machines (SVM) on different subsets of features to examine the feasibility of device type inference as well as the relationship between mobility and traffic characteristics. These sets include mobility and traffic features *separately*, then *combined*, and then combined with *weekend/weekday labels*. Using *solely mobility features* achieves ≈65% accuracy, while *traffic features alone*, obtains ≈79% accuracy. Using all mobility and traffic variables *combined*, the trained model achieves ≈81% accuracy. Then, as the **combined** feature set is extended to include *weekdays and weekends* independently, the trained SVM yields an accuracy of ≈86%. This suggests that users' behavior (both flutes and cellos) is *more distinguishable* when looking at **combined** mobility and traffic features; especially when *temporal* features such as weekdays are considered separately from weekends. We note that such behavior gaps are *not* the same for both device types and a model should to take that into account.

2) Unsupervised clustering: To investigate natural convex clusters, we used K-means algorithm. Using *mobility features only*, the best mean silhouette coefficient is achieved on k=2 and 4. However, cluster sizes are highly skewed and at k=2, $\approx$ 60% of devices are correctly clustered. *Traffic features alone*, at k=2, results in $\approx$ 81.2% accuracy. **Combining** mobility and traffic features, *increases* the accuracy to $\approx$ 81.5%. While some flutes and cellos are similar in terms of mobility and traffic, the clusters of the combined features clearly illustrate **two distinct modes** (especially in *traffic*) and the *high homogeneity* of the clusters hints at *disjoint sets of behaviors* in mobility and traffic dimensions, governed by the device type.

3) Mixture model: To take a step towards synthesis of traces based on our datasets, we trained Gaussian mixture models (GMM) on *combined mobility and traffic features*. From the combined model (CM), we acquired simulated samples. We used Kolmogorov-Smirnov (KS) statistic to compare the simulated samples with the real data and found that CM is able to capture the behaviors of each device type. (Average KS statistic of features is ≈ 0.15 **for flutes** and ≈ 0.14 **for cellos**. See Appendix V in [23] for details.) Importantly, we noted, the combined model produces samples whose *traffic* features match the original data **better**, compared with training a GMM *on traffic features alone* (based on KS statistic), hinting at a key relationship between mobility and traffic. However, comparing mobility features of CM with a GMM trained on mobility features alone shows no improvement to slightly worse results.

Overall, this suggests that there is significant potential for an ***integrated mobility-traffic model*** that captures the differences and ***relationships*** of features, across ***device types, time*** *and* ***space***. We leave detailed comparison of combined modeling with separate modeling of mobility and traffic for future work.

VIII. Conclusion

In this study, we mine large-scale WLAN and NetFlow logs from a campus WiFi network to answer three questions: *(I) How different are mobility and traffic characteristics across device types, time and space? (II) What are the relationships between these characteristics? (III) Should new models be devised to capture these differences? And, if so, how?* We build the *FLAMeS* framework for systematic processing and analysis of the datasets. Using MAC address survey, OUI matching and web domain analysis, we put devices into two categories: flutes (*"on-the-go"*) and cellos (*"stop-to-use"*). We then study a multitude of mobility and traffic metrics, comparing flutes and cellos across time and space. On average, flutes visit twice as many APs as cellos, while cellos generate ≈2x more flows than flutes. However, flutes flows are 2.5x larger in size, with ≈2x the number of packets. The best fit for location preference is **Zipfian**, for flow/packet sizes is **Lognormal**, and for flow IAT at APs is **beta** distribution. Furthermore, flute traffic drops sharply on weekends whereas many cellos remain active. Across mobility and traffic dimensions, we spot a negative correlation for flutes between mobility and flow duration but negligible correlation with traffic size; for cellos, this effect is less pronounced. We find a negative correlation with APs visited and the active time, particularly for flutes. However, no correlation exists APs visited and traffic for cellos. We *quantified* correlations *across both mobility and traffic dimensions*. Finally, we applied machine learning and trained a mixture model to synthesize data points and verified that the **combined** mobility-traffic features capture the *differences* in metrics ***better*** than *either mobility or traffic separately*. Many of our findings are not captured by today's models, and they provide insightful guidelines for the design of evaluation frameworks and simulations models. Hence, this study answered the questions posed, introduced a strong case for newer models, and provided our first step towards a future integrated mobility-traffic model.

IX. Acknowledgements

We would like to thank Prof. Alin Dobra for help in setting up part of the computing cluster, and the anonymous reviewers for their useful feedback. The term *'cello mobility'* was suggested by Prof. Mostafa Ammar and used here with permission. Partial funding for this effort was provided by NSF Award 1320694, August-Wilhelm Sheer fellowship at Technical Univeristy-Munich, and Aalto University.

References

[1] J. Treurniet, "A Taxonomy and Survey of Microscopic Mobility Models from the Mobile Networking Domain," *ACM CSUR*, 2014.

[2] A. Hess, K. A. Hummel, W. N. Gansterer, and G. Haring, "Data-driven Human Mobility Modeling: A Survey and Engineering Guidance for Mobile Networking," *ACM CSUR*, 2016.

[3] D. Kotz and K. Essien, "Analysis of a Campus-Wide Wireless Network," *Springer Wireless Networks*, vol. 11, no. 2, January 2005.

[4] T. Henderson, D. Kotz, and I. Abyzov, “The changing usage of a mature campus-wide wireless network,” *Elsevier Computer Networks*, vol. 52, no. 14, October 2008.

[5] G. Maier, F. Schneider, and A. Feldmann, “A First Look at Mobile Hand-held Device Traffic,” in *Proc. of IEEE PAM*, 2010.

[6] Y. Zhand and A. Arvidsson, “Understanding the Characteristics of Cellular Data Traffic,” in *ACM SIGCOMM CellNet workshop*, 2012.

[7] S. Moghaddam and A. Helmy, “SPIRIT: A simulation paradigm for realistic design of mature mobile societies,” in *IWCMC ’11*, 2011.

[8] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, “Understanding individual human mobility patterns,” *Nature*, vol. 453, no. 7196, 2008.

[9] C. Song, T. Koren, P. Wang, and A.-L. Barabási, “Modelling the scaling properties of human mobility,” *Nature Physics*, vol. 6, no. 10, 2010.

[10] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, “Exploring human mobility with multi-source data at extremely large metropolitan scales,” *MobiCom ’14*, 2014.

[11] G. Maier, F. Schneider, and A. Feldmann, “A first look at mobile hand-held device traffic,” in *PAM ’10*. Springer, 2010.

[12] U. Kumar, J. Kim, and A. Helmy, “Changing patterns of mobile network (WLAN) usage: Smart-phones vs. laptops,” *IWCMC ’13*, 2013.

[13] X. Chen, R. Jin, K. Suh, B. Wang, and W. Wei, “Network performance of smart mobile handhelds in a university campus wifi network,” in *IMC ’12*. ACM, 2012.

[14] A. Gember, A. Anand, and A. Akella, “A comparative study of handheld and non-handheld traffic in campus wi-fi networks,” in *PAM ’11*.

[15] M. Afanasyev, T. Chen, G. M. Voelker, and A. C. Snoeren, “Analysis of a mixed-use urban wifi network: when metropolitan becomes neapolitan,” in *SIGCOMM ’08*. ACM, 2008.

[16] I. Papapanagiotou, E. M. Nahum, and V. Pappas, “Smartphones vs. laptops: comparing web browsing behavior and the implications for caching,” *SIGMETRICS ’12*, vol. 40, no. 1, 2012.

[17] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin, “A first look at traffic on smartphones,” *IMC ’10*, 2010.

[18] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra, “Characterization of wireless multidevice users,” *ACM Transactions on Internet Technology (TOIT)*, vol. 16, no. 4, pp. 29:1–29:25, Dec. 2016. [Online]. Available: http://doi.acm.org/10.1145/2955096

[19] P. Cao, G. Li, A. Champion, D. Xuan, S. Romig, and W. Zhao, “On human mobility predictability via wlan logs,” in *INFOCOM ’17*.

[20] A. Nikravesh, Y. Guo, F. Qian, Z. M. Mao, and S. Sen, “An in-depth understanding of multipath TCP on mobile devices,” in *MobiCom ’16*.

[21] X. G. Meng, S. H. Y. Wong, Y. Yuan, and S. Lu, “Characterizing flows in large wireless data networks,” *MobiCom ’04*, 2004.

[22] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, “Understanding mobile traffic patterns of large scale cellular towers in urban environment,” *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 1147–1161, April 2017.

[23] “Appendix - Tech. Report.” [Online]. Available: https://s3.amazonaws.com/infocom2018/infocom18_flutes_cellos.pdf

[24] A. W. Moore and D. Zuev, “Internet traffic classification using bayesian analysis techniques,” *SIGMETRICS ’05*, vol. 33, no. 1, 2005.

[25] V. Paxson and S. Floyd, “Wide area traffic: the failure of Poisson modeling,” *IEEE/ACM ToN*, vol. 3, no. 3, jun 1995.

[26] T. Nguyen and G. Armitage, “A survey of techniques for internet traffic classification using machine learning,” *IEEE Communications Surveys & Tutorials*, vol. 10, no. 4, 2008.

[27] M. A. Hall, “Correlation-based feature selection for machine learning,” Ph.D. dissertation, The University of Waikato, 1999.

[28] W.-J. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, “Modeling time-variant user mobility in wireless mobile networks.”

[29] P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. C. González, “Discovering urban activity patterns in cell phone data,” *Transportation*, vol. 42, no. 4, pp. 597–623, 2015.

[30] C. Boldrini and A. Passarella, “Hcmm: Modelling spatial and temporal properties of human mobility driven by users’ social relationships,” *Computer Communications*, vol. 33, no. 9, 2010.

Appendices

Here we further describe various aspects of our submitted work to Infocom 2018, which were not included in the original document for brevity.

I. Merging Datasets

In order to study network traffic across devices and APs, it is necessary to match the NetFlow records with wireless associations (from WLAN dataset). This task requires the MAC-IP mapping. The IP addresses are dynamically assigned using DHCP but DHCP session logs were not directly available and had to be derived. We define the duration of a DHCP lease as the time between two consecutive associations of a device with any AP; i.e. when a device connects to AP_1, a session starts and once the user device connects to AP_2, the first session ends and a new one starts. Fig. 10 illustrates the associations of a sample device with different APs at different times. The first session would have the IP given by AP_1 and a lease time $t_2 - t_1$, and so on. (total of 5 sessions in this example) The last association is discarded as we do not know the duration of that IP assignment. Combining these derived-DHCP records with the *Location Information* and *Device Type Classification* we create the **DHCP** table.

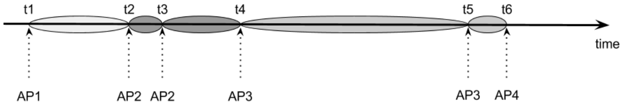


Fig. 10: Wireless association for a device at different times.

The derived DHCP and NetFlow datasets were then merged to form what we refer to as the **CORE** dataset for our study. The unique identifiers between the two are the clients’ IPs in addition to start and end time of flows, hence the need for a DHCP-like set. For a DHCP lease session *LS*, all flows whose IP address is the same as the lease *and* whose entire lifetime falls within the lease duration, are associated with *LS*.

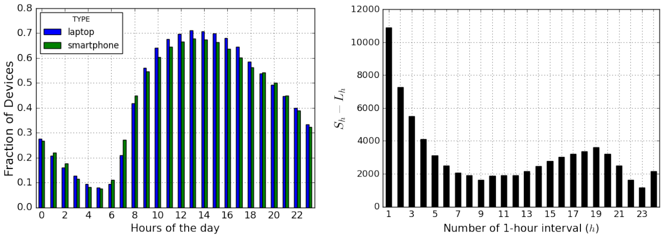
Given these traces, cellular usage cannot be analyzed. However, this does not significantly impact analysis for two reasons: 1) The traces already capture a very large user-base, with tens of thousands of active devices. This raises confidence in our analysis of a real world WLAN. 2) The WiFi campus coverage is ubiquitous, with 1760 APs installed in the vast majority of populated areas. Also, most laptops on campus lack cellular connectivity, and many smartphones use WiFi for their data to avoid cellular data costs.

II. Computing System

The size of the datasets is ≈30TB in raw text format, mostly consisting of NetFlow data and ≈0.5TB for AP logs. There were several challenges in managing and mining the largescale datasets that required a thorough preparation, to run on a fast machine with plenty of resources/memory. We explored several techniques and pipelines for extraction, transformation, loading (ETL) and querying of big data and chose tools from Apache Hadoop ecosystem. We use Hive as our data warehouse (tables stored in Parquet format). Apache Spark is the compute engine for data processing and analysis tasks. Computation runs on two nodes, each with 64 cores and ≈0.5TB of memory. Further discussion of the system and comparison to others is out of scope of this document.

III. Mobility Analysis

For completeness, we include further analysis of the mobility aspects of our dataset, discussed in Section V of the conference paper.



(a) (b)

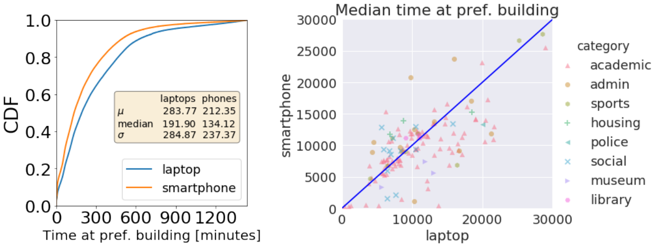
Fig. 11: Hourly associations.

A. Hourly associations

Measuring device associations every hour, Fig. 11a shows the percentage of devices with at least one event as a function of hours of the day. The majority of devices appear online between 9am and 8pm, with the hours between 2am and 6am having less than 20% of devices associating. We find no major differences between flutes' and cellos' distributions, as many users potentially own both. As users arrive on campus and their phones announce their first location, they switch on their laptops. This issue bears further research through a future census study.

To measure the stay of devices throughout a day, we look at 1-hour intervals, and measure the number of hours a device accessed an AP [9]. Fig. 11b depicts $S_h - L_h$, where S_h and L_h are total number of flutes and cellos respectively, with at least one record per hour, as a function of the number of hours online h. Flutes are predominant for short visits and very long stays, but the difference drops significantly at 9 hours, then increases. The rise after 9 hours is likely due to students living on campus, with always-on connected phones.

B. Visitation preferences



(a) (b)

Fig. 12: Time spent at preferred building.

Fig. 12b shows a scatter of the median time spent at a user's preferred building. Each dot represents this value for a given location. This plot shows that *academic*, *police* and *museum* buildings tend to have laptops staying longer, which makes sense intuitively, with students using laptops during lectures and staff working at the other two categories. On the contrary, for *social* and *housing* buildings, there is a higher probability of having flutes staying longer, hinting at a tendency to use mobile devices more in such places. Finally, *administrative*, *sports* and *library* buildings tend to have both types of devices staying for similar amounts of time. Analysis of inherited differences in browsing of online services given by this heterogeneity among buildings is left for future work.

Fig. 12a depicts the time devices spend at their *preferred building* in a day.

C. Return probability

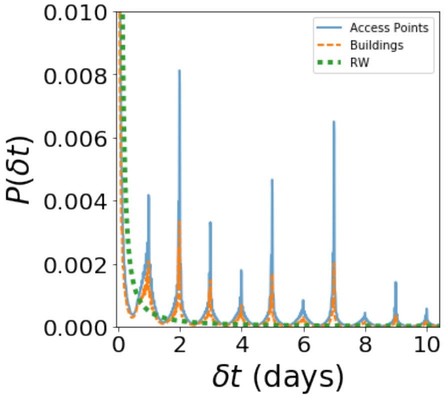


Fig. 13: Probability to return to a previously visited location.

We compare empirical values for devices to return to previously visited APs or buildings in Fig. 13. We observe returning spikes at every 24 hours, with the highest peaks at 48 and 168 hours (2 and 7 days). This can be explained by the schedule of classes at the university.

IV. Traffic Analysis

In this Section, we further discuss references from Section VI of the conference paper.

A. Flow sizes

This metric is the sum of bytes for all packets within a single flow. First, outlier data points are removed using a robust measure of scale, based on inter-quartile range (IQR). Looking at individual flows of each device type shows that size of flows that originated from smartphones are significantly different that laptop flows (p-value< .05). [10]

On weekdays, the average size for smartphone flows is 2070 bytes and 822 bytes for laptop flows; with no significant changes on weekends (CDF in Fig. 14). The difference in medians is more pronounced, on weekdays, for smartphones it is 678 bytes while it is 142 bytes for laptops (similar values on weekends).

[9] P. Widhalm, et al., "Discovering urban activity patterns in cell phone data".

[10] Flow metrics do not fit Gaussian distribution (based on Shapiro-Wilk test for normality, goodness-of-fit test and Q-Q plot results, not included for brevity). Thus, we use Mann-Whitney statistical test[11] to compare two unpaired groups (laptops vs smartphones), and Wilcoxon signed-rank test to compare two paired groups (each device type on weekdays vs weekends) [12].

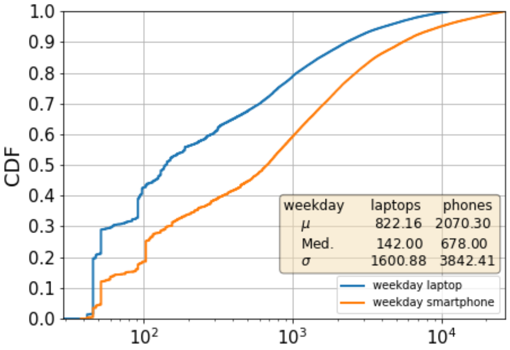


Fig. 14: CDF of individual flow sizes (bytes, log-scale x axis), similar pattern on weekends

B. Lognormal plots

For flow sizes in our dataset, a Lognormal distribution is the best fit, regardless of device type (Fig. 15). Many models assume flow sizes are static, or follow an exponential distribution but real world data provides no supporting evidence. Such simplifying assumptions fail to accurately account for very large flows obtained from a Lognormal distribution.

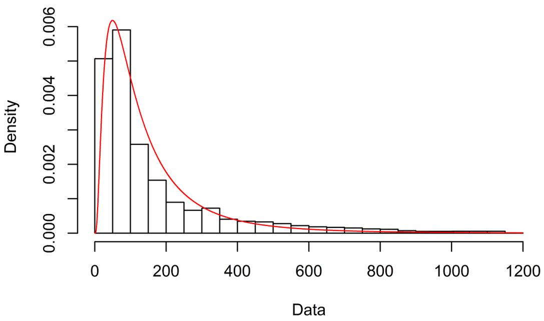


Fig. 15: Lognormal distribution.

C. beta distribution

Inter-arrival times (IAT). Our results show that the flow IAT, regardless of device type, does not follow an

exponential distribution. Flow IAT matches a beta distribution well (Fig. 16) with a very high estimated kurtosis and skewness (estimated at 58 & 6.9 respectively). The high estimated kurtosis illustrates that there are infrequent extreme values, which explains the observed highly elevated standard deviation of IAT[13]

V. First Modeling Steps

More details of the KS test and GMM model are provided here. We found that providing both mobility and traffic features to train a GMM results in lower average KS statistic. Fig 17 shows a sample CDF of *TAT*. KS statistic details can be found in Table V.

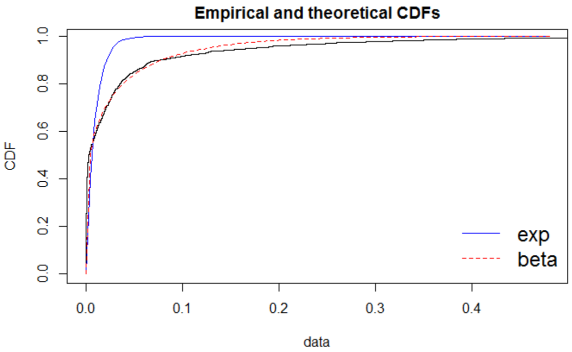


Fig. 16: Exponential and Beta distributions.

TABLE V: KS-statistic summary

KS statistic	Flutes	Cellos
Average	0.150	0.140
Min	0.052	0.027
Max	0.380	0.350
Std	0.086	0.0787

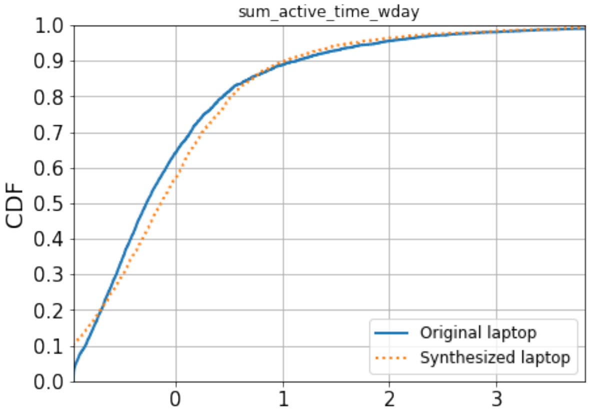


Fig. 17: Synthetic vs. Original *TAT* feature for flutes.

VI. Lessons Learned and Modeling Insights

Our above findings provide further (but surely not yet comprehensive) insights into considerations relevant to the design and parameterization of mobility and network traffic models. While we leave devising and validating a concrete candidate model for future work, we can readily identify the following important elements:

It is crucial to differentiate flutes vs. cellos for both mobility and traffic due to their very different nature. More specifically, flutes exhibit continuous presence whereas cellos are on/off with jumps between locations. Beyond differences in continuity, the traffic patterns (flow sizes, arrival times, etc.) should be specified by device class. Moreover, the traffic generation, spatial locations, and temporal behavior can be linked per device type and per user "community" (e.g. students of different disciplines at various buildings).

[13] In the research community, packet IAT and its Fourier transform are considered important features in traffic analysis. They are used extensively in simulation and modeling of networking protocols as well as internet traffic classification [17]. Realistic modeling of IAT is required for accurate simulation and measurement of congestion control mechanisms [24]. Due to limited availability or staleness of most packet-level datasets, although our NetFlow is on a higher abstraction layer (flow-level vs packet-level), analysis of flow IAT can still be used for measuring delay and jitter effects.