Characterizing and Modeling Session-Level Mobile Traffic Demands from Large-Scale Measurements

André Felipe Zanella
IMDEA Networks Institute
Universidad Carlos III de Madrid
andre.zanella@imdea.org

Antonio Bazco-Nogueras
IMDEA Networks Institute
antonio.bazco@imdea.org

Cezary Ziemlicki
Orange Innovation
cezary.ziemlicki@orange.com

Marco Fiore
IMDEA Networks Institute
marco.fiore@imdea.org

ABSTRACT

We analyze 4G and 5G transport-layer sessions generated by a wide range of mobile services at over 282,000 base stations (BSs) of an operational mobile network, and carry out a statistical characterization of their demand rates, associated traffic volume and temporal duration. Our study unveils previously unobserved session-level behaviors that are specific to individual mobile applications and persistent across space, time and radio access technology. Based on the gained insights, we model the arrival process of sessions at heterogeneously loaded BSs, the distribution of the session-level load and its relationship with the session duration, using simple yet effective mathematical approaches. Our models are fine-tuned to a variety of services, and complement existing tools that mimic packet-level statistics or aggregated spatiotemporal traffic demands at mobile network BSs. They thus offer an original angle to mobile traffic data generation, and support a more credible performance evaluation of solutions for network planning and management. We assess the utility of the models in practical application use cases, demonstrating how they enable a more trustworthy evaluation of solutions for the orchestration of sliced and virtualized networks.

CCS CONCEPTS

• **Computing methodologies** → **Modeling methodologies**; • **Networks** → **Network measurement**.

KEYWORDS

Network measurement, Traffic modeling, Session traffic, App traffic

ACM Reference Format:
André Felipe Zanella, Antonio Bazco-Nogueras, Cezary Ziemlicki, and Marco Fiore. 2023. Characterizing and Modeling Session-Level Mobile Traffic Demands from Large-Scale Measurements. In *Proceedings of the 2023 ACM Internet Measurement Conference (IMC '23), October 24–26, 2023, Montreal, QC, Canada.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3618257.3624825

IMC '23, October 24–26, 2023, Montreal, QC, Canada

1 INTRODUCTION

Data-driven solutions will play an increasingly important role in the 5G mobile network ecosystem during its evolution towards 6G. This trend is stimulated by the unprecedented access to traffic indicators and statistics enabled by a plethora of new network monitoring functions: prominent examples include the Network Data Analytics Function (NWDAF) [4] and Management Data Analytics Function (MDAF) [5] that appeared in 3GPP Release 16, or the database-like Radio Network Information Base (RNIB) [33] and the consumer/producer Data Management and Exposure Services [32] for the near-real-time and non-real-time (respectively) RAN Intelligent Controller (RIC) in O-RAN. This abundance of data can feed innovative machine learning models that have been proven to yield promising performance in many complex network management tasks, including forecasting of mobile demand [43] and throughput [22], beam management in mmWave Radio Access Networks (RAN) [34], orchestration of network slices [9], classification of flow-level traffic [49] or control of virtualized RAN resources [7], just to cite a few representative examples. Overall, the combination of live data provisioning and learning-based inference is expected to pave the road for paradigms such as Zero-touch Network and Service Management (ZSM) [12] and Intent-Based Networking (IBN) [1].

In the emerging context above, the availability of vast and dependable mobile network data becomes even more critical to the development and evaluation of new network functions across all domains. Unfortunately, access to the large-scale real-world datasets that are needed to train and test original data-driven algorithms is today very limited. Broad measurements from actual production systems at city or national scales that capture the full diversity of mobile traffic demands are hard to come by, and are typically protected by restrictive Non-Disclosure Agreements (NDA) that prevent their circulation. Public traffic data is scarce and outdated [8] or gathered via small-scale client-driven experiments whose representativeness is inherently circumstantial [16].

In this scenario, trustworthy models of mobile traffic become an indispensable asset to networking research: they allow generating realistic synthetic traces to remove the data access barrier, and implicitly enable verifiability and reproducibility of results. As illustrated in Figure 1 and detailed in Section 2, current models of mobile traffic target: (*i*) fine-grained packet-level statistics, *e.g.*, about packet sizes or inter-arrival times [31]; or, (*ii*) aggregate dynamics at individual cellular base station (BS), *e.g.*, describing the total mobile data traffic demand at a given BS over time [47].

Figure 1: Graphic taxonomy of mobile network traffic models, with representative features and typical modeling timescales for models that operate at packet-level, (transport) session-level, and BS-level, respectively.

In this paper, we take a different, intermediate perspective between those considered in the literature, and explore mobile traffic statistics at the level of *individual transport-layer sessions served by one BS*. Transport-layer sessions, often also referred to as flows, are sequences of packets belonging to a same application-layer interaction[1] between a User Equipment (UE) and a server, and are aimed at provisioning one specific (portion of) service to the UE. They are uniquely identified by a 5-tuple consisting of the transport-layer protocol, source/destination IP addresses, and source/destination ports. For instance, one session may be generated by a user launching the Netflix application on their smartphone to stream an episode of a show, or by a UE retrieving in background a software update for one of its installed applications.

As also portrayed in Figure 1, (transport) session-level models target previously overlooked features of mobile traffic: the arrival process of transport-layer data flows of a specific application at a given BS, the duration of such flows, their associated load, or the distribution of average throughput that the combinations of such duration and load statistics entail. Since transport-layer sessions are associated to the one application they serve, session-level models are inherently service-specific. The transport session-level models fill in fact a gap in the space of mobile network traffic modelling, and allow generating for the first time realistic demands aligned with those observed at the BSs of a modern 4G/5G RAN infrastructure. Specifically, they can complement studies on packet-level modeling so as to reproduce fine-grained mobile traffic loads at an individual BS that dependably mimic how the users attached to the target BS request specific services and what amount of traffic each such request entails.

As such, session-level models support the design of data-driven solutions and more credible performance evaluations for many networking tasks, including planning [21], dimensioning with respect to specific services [25], scheduling [20], or energy-efficient operation [45]; they can also inform new traffic generators for modern network simulators [10].

[1]We remark that a single application may establish multiple transport-layer sessions. This can happen over time (*e.g.*, a messaging service initiating new sessions at every time the user switches to a new chat with a different contact than the current one), or in parallel (*e.g.*, a large file transfer application opening multiple FTP sessions). Multiple transport-layer sessions associated to a same application-layer session may have similar or different characteristics. However, in this paper we focus on individual transport-layer sessions only, and leave a thorough investigation of the relationships and interactions of such sessions at the higher layers as future work. Throughout the paper, we will refer to transport-layer sessions simply as *sessions* for simplicity, hence all future references to session-level models implicitly refer to transport sessions.

Overall, our study yields the following main contributions.

- We characterize transport-layer sessions recorded at over 282, 000 BSs of a nationwide production mobile network covering continental France, investigating (*i*) the arrival process at individual cellular antennas of sessions associated to a wide range of applications, (*ii*) the distribution of the traffic volume generated by each such session, and (*iii*) the relationship between such load and the duration of the session. Our analysis unveils statistical properties of session-level traffic that have not been observed before, and that are heterogeneous across different mobile applications but persistent across space, time and radio access technology.
- We develop simple but accurate models of the statistical properties above for a variety of mobile services, which we release publicly[2] so as to contribute to removing the access barrier to dependable data needed to design and evaluate networking solutions.
- We show the utility of the proposed models in two practical performance evaluation use cases, where we use them to assess solutions to (*i*) allocate computing resources in virtualized Radio Access Network (vRAN) environments and (*ii*) configure capacity requirements for network slicing. Our tests prove how the proposed models substantially enhance the accuracy of the results compared to traffic models currently available for mobile network performance analysis that are not informed by session-level statistics.

2 RELATED WORK

Models of mobile data traffic are instrumental to the performance evaluation of mobile communication technologies, and have existed since the early days of wireless networking. In particular, there is a vast body of models that aim at representing statistical properties of mobile traffic within each session, *i.e.*, at packet level. As illustrated in Figure 1, such models typically operate at timescales of milliseconds or less, and provide analytical formulas for, *e.g.*, inter-arrival times between consecutive packets or requests from a device [2], sizes of individual files or number of packets per frame [6], intervals for deterministic reporting [3], or duration of activity and inactivity periods [17]. Different packet-level models are specified for broad classes of services like web browsing, video streaming, voice over IP, gaming, downloads via FTP, or machine-type communications, among others.

The amount of proposals for packet-level models is such that condensing it in this section is not possible, and we refer the interested reader to a recent survey for a comprehensive review [31]. The key observation is that these models are primarily designed for the evaluation of low-layers technology in stationary environments, and do not capture, *e.g.*, inter-session timings, how long a session generated by a given application persists in a BS, or how much traffic it generates there. The analysis and models we propose in this paper precisely answer such questions and thus complement the extensive literature on packet-level representations.

At the other end of the modelling spectrum are BS-level demands, which are also visually illustrated in Figure 1. BS-level statistics mainly describe aggregates of the traffic volume across all devices

[2]https://github.com/nds-group/MobileTrafficDists

associated to the target antenna, and are best characterized over timescales of minutes or hours. As such, they are different and coarser than the session-level dynamics we are interested in, which instead occur at order-of-second granularity. This dissimilarity sets our study apart from, *e.g.*, works employing α-stable distributions to model heavy-tailed samples of BS-level traffic observed in real-world networks [19, 23, 24], or recent generative neural networks that mimic BS-level dynamics over space [46], time [26] or both dimensions [47], possibly per service [41].

Closer to our goal of characterizing sessions at the transport-layer level, Mucelli *et al.* [29] develop six models of individual mobile traffic consumption, by classifying the demands generated by 6.8 million subscribers based on their temporal patterns and amount of data usage. In a similar spirit, Wu *et al.* [44] identify six major temporal profiles in the weekly demand generated by mobile devices, and propose predictors to anticipate the future load of each class of user. Compared to the novel models we present in this paper, the existing ones above are much coarser, along multiple dimensions. First, they only consider overall user-level traffic, whereas we disaggregate those into more precise session-level statistics. Second, we provide models for a large variety of services, while the studies above only consider the total traffic of each user. Third, the previous models are purely temporal and aggregate information over all BSs visited by each user, whereas our focus is on behaviors recorded within a single BS. We argue that the finer granularity, added richness and per-BS viewpoint of our models make them much more informative for the validation of mobile communication technologies and systems.

3 MEASUREMENT DATASET

Our study builds upon massive measurement data collected in an operational nationwide mobile network. The target network employs 4G and 5G non-standalone (NSA) radio access network (RAN) technologies. As depicted in Figure 2, in this configuration 5G gNodeBs coexist with 4G eNodeBs in the RAN, and provide higher-capacity wireless communication to 5G-capable UEs. Yet, the lack of a dedicated 5G network core in the NSA deployment forces gNodeBs to depend on interactions with eNodeBs, via the X2 interface, for control operations towards the 4G Mobility Management Entity (MME). Also, gNodeBs connect to 4G gateways via a slightly modified S1-U interface for all data plane transmissions.

Consistently with the aim set forth in Section 1, we record in the target network data about individual transport-layer sessions observed during 45 consecutive days at the 282,000 BSs that form the whole 4G/5G RAN of the operator. The session-level statistics are produced within secure compute premises of the network operator, and for the purpose of our work we only access distributions and averages that do not contain personal or sensitive information. A more complete discussion of how our study meets the principles of ethical research are provided in the Appendix. Next, we detail the collection process and basic features of the data.

3.1 Data collection platform

Two complementary passive measurement systems are used to gather and compose the dataset, as shown in Figure 2.

Figure 2: Simplified 4G and 5G NSA mobile network architecture illustrating the combined RAN and gateway data collection setup used by the operator.

RAN probes deployed at the S1-MME interfaces of the MME capture *signaling data*. Due to the way the 5G NSA deployment operates, these probes can monitor the control planes of both eNodeBs and gNodeBs. We employ the signalling data to geo-reference and time-stamp the session information. Specifically, the probes observe all signaling events generated by each UE, *e.g.*, when it requests a service, exchanges data, performs handovers, or moves across Tracking Areas (TA), and record the BS of attachment. By leveraging this information, we can associate each UE (and the sessions it generates) to its serving BS at all times.

Gateway probes tapping at the SGi interface of the Packet Gateway (PGW) monitor all IP traffic and extract information on each transport-layer session. These probes record the total data traffic generated by the session, its start and end times, and the associated mobile service. The probes run proprietary traffic classifiers developed by the operator and based on Deep Packet Inspection (DPI) that examines headers at both network and transport layers to derive the per-flow mobile service information. While the algorithms used for traffic classification are confidential, the operator reported high accuracy in independent tests, and regularly uses the results for network management purposes.

The measurement approach above allows overcoming inherent limitations in the precision of the localization information available in the core network. Indeed, the UE location identifiers available at the PGW are updated infrequently, leading to stale positions and localization errors in the order of kilometers [48, 28]. Relying on the locations recorded by the gateway probes would thus jeopardize our capability to geo-reference session-level data at the granularity of the individual BS in a reliable manner. To overcome this problem, the UE and time information gathered by the gateway probes are crossed with the signaling data of the RAN probes so as to retrieve the BS(s) where each session occurs and assign the correct (fraction of) sessions to all BSs.

3.2 Aggregation into session-level statistics

The gateway probes collect information about individual TCP and UDP sessions, which, as mentioned in Section 1, are uniquely identified by a 5-tuple consisting of the transport-layer protocol, source and destination IP addresses, and source and destination ports.

A TCP session is typically initiated by the three-way handshake and considered to be terminated shortly after a packet with the FIN or RST bits set is observed. Expiration timeouts that are service-specific are also employed to mitigate the effect of unorthodox TCP session terminations. In case UDP sessions, they start when a new

5-tuple is recorded, and ended once a timeout period without any transmitted packets elapses. Again, this timeout depends on the application that the traffic classification routines associate to the flow. Also, it is worth remarking that, since our study is concerned with sessions served by a single BS, handovers from and to other BSs are recorded in the measurement dataset as newly established or concluded transport-layer sessions, respectively.

Data about all sessions occurring at each BS for a given service are initially aggregated at one-minute granularity by the operator, before further processing on our part; the additional transformations are performed to ensure the privacy of the data subjects as well as to strike a balance between a sufficient precision on the traffic representation and a dataset size viable for downstream analysis. Specifically, we aggregate the data about all sessions occurring at each BS for a target service on a daily basis, in the form of (*i*) the number of sessions arriving at the BS at every minute, (*ii*) a Probability Density Function (PDF) of the total traffic volume generated by one session at the BS, (*iii*) value pairs composed of the duration of one session served by the BS and the traffic volume it generates. As we will see, this is a compact, privacy-preserving representation that allows characterizing all major session-level properties, *i.e.*, the arrival rate, duration, total load, and average throughput.

Formally, data about sessions occurring at each BS $c \in C$ for service $s \in \mathcal{S}$ are aggregated over daily intervals $t \in \mathcal{T}$. For each tuple (s, c, t), we store the following statistics.

- *Counts of sessions served by the BS*, denoted by $w_s^{c,m}$, capturing the total number of sessions received at BS c for service s each minute m of day $t \in \mathcal{T}$, which is further aggregated per day into a variable $w_s^{c,t}$.
- *Probability Density Function (PDF) of the traffic volume*, denoted by $F_s^{c,t}(x)$, describing the odds that a session of service s induces a total load x at BS c during day t.
- *Value pairs of discretized duration and traffic volume*, denoted by $v_s^{c,t}(d)$, capturing the mean load associated to sessions of duration d for service s at BS c in day t.

3.3 Statistics averaging

The dataset reports statistics per BS and day. For our analyses, we need to investigate behaviors averaged over multiple BSs and days. For duration-volume pairs, we compute a weighted average of each datapoint; for instance, average pairs over all BSs and days for a service s are obtained as

$$v_s(d) = \frac{1}{\sum_{c\in C}\sum_{t\in\mathcal{T}} w_s^{c,t}} \sum_{c\in C}\sum_{t\in\mathcal{T}} w_s^{c,t} v_s^{c,t}(d), \ \forall d. \tag{1}$$

In the case of traffic volume PDFs, averaging is achieved via a finite-dimensional general mixture model. For an all-BS and all-day average PDF, this is formally expressed as

$$F_s(x) = \frac{1}{\sum_{c\in C}\sum_{t\in\mathcal{T}} w_s^{c,t}} \sum_{c\in C}\sum_{t\in\mathcal{T}} w_s^{c,t} F_s^{c,t}(x). \tag{2}$$

The expressions in (1) and (2) are straightforwardly extended to any subsets of C and $\mathcal{T}$, so as to merge statistics from any set of BSs and days. Illustrative samples of nationwide traffic PDFs $F_s(x)$ and duration-volume pairs $v_s(d)$ averaged over all BSs and days are later reported in Figure 5.

Figure 3: Real: measurement PDFs of the per-minute session arrival rate for antennas serving different loads. Nonpeak and peak: fitted distributions modelling the bi-modal sessions arrivals (see in Section 5.1 for full details).

4 CHARACTERIZING SESSION-LEVEL DEMANDS AT CELLULAR BSs

We now explore the dataset and provide both qualitative and quantitative characterizations of session-level mobile traffic demands. The insights we derive will inform the design of our proposed models, introduced in Section 5.

4.1 Session arrivals

We start by analyzing the arrival process of sessions at a BS. Figure 3 reports the distribution of the number of new sessions established at every minute at different categories of BSs, *i.e.*, the PDF of $w_s^{c,m}$ at all BSs $c \in C_i$ of category i, and aggregated over all services $s \in \mathcal{S}$. The x-axis values are then normalized by the cardinality of set C_i, so as to obtain the typical number of sessions arriving in one minute at a single BS of category i. Namely, categories $i \in \mathcal{I}$ tell apart BSs experiencing different loads: we compute the distribution of total traffic served by each BS during the whole measurement

Figure 4: Services ranked by the fraction of sessions they generate, along with their normalized total traffic.

time, and separate BSs based on the decile they pertain to. Thus, each set C_i includes 10% of the BSs, with growing mobile traffic demands from the first decile to the last one. The rationale for this categorization is that it allows observing how the session arrival process is affected by the target BS load.

In fact, the plots in Figure 3 show that the behavior of the arrivals is semantically similar across all classes of BSs, or, equivalently, *the traffic volume served by the BS has no significant impact on the high-level statistics of the arrival process*. Indeed, the shape of the overall distribution is the same for all plots, apart from the obvious difference of scale in the abscissa induced by the growing demand across deciles. More precisely, *all PDFs of values $w_s^{c,m}$ show an evident bi-modal distribution*, which a close inspection reveals to be due to the well-known circadian rhythm of mobile network traffic, with low traffic (hence small number of sessions per minute) overnight and much increased demands (hence more frequent session arrivals) during daylight hours. Transitions between these two phases are very rapid, which leads to a negligible probability of having intermediate arrival rates.

Since sessions are naturally service-specific, a relevant follow-up question is how such arrivals are distributed across different mobile services. Figure 4 offers a first result in that sense, as a ranking of the top 100 services based on the fraction of total sessions they generate. The curve predominantly follows a negative exponential law (with a very high coefficient of determination R^2 of 0.97), implying that *the number of sessions generated by each service is very heterogeneous*: the top 20 services are responsible for over 78% of the sessions recorded overall. The imbalance is less dramatic than that in traffic, which is known to follow an even more skewed power law [39, 27]; nonetheless, it suggests that the probability that a newly established session belongs to a given application is far from uniform.

4.2 Qualitative analysis of service sessions

Figure 4 also shows the total normalized traffic produced by each service. While some correlation with the number of sessions exists, the load dots are fairly scattered (on a logarithmic scale) for similarly ranked services: hence, *different applications entail a very varied traffic volume per session*. This motivates an in-depth investigation of such session-level traffic dynamics on a per-service basis, which is indeed the target of our next analysis.

Samples of session-level traffic volume PDF and duration-traffic pairs are portrayed in Figure 5 for six representative mobile services. All statistics are averaged over the whole set of BSs and days, using the methodology described in Section 3.3, hence they capture archetypal behaviors of the demands of each application.

A first qualitative observation is that *total traffic volumes and duration values are highly heterogeneous, among sessions of a same service and even more so across different services*. Indeed, intra-service statistics show how sessions belonging to a same application can generate very diverse traffic volumes, spanning several orders of magnitude, over intervals that can range from seconds to hours; and, the shapes of PDFs and duration-traffic pairs are completely different at inter-service level. By looking at each subfigure, we note that the traffic volume PDFs present multi-modal shapes with an overall smooth Gaussian-like trend (over the logarithmic abscissa) interrupted by abrupt and marked spikes of probability. Both the main statistics (such as the mean, standard deviation or skewness) and the probability peaks are not comparable across the selected services. Interestingly, *heterogeneous probability peaks also tell apart applications that ostensibly belong to the same class*, *e.g.*, messaging services like Snapchat and Whatsapp, or video streaming services like Netflix and YouTube.

A closer look to the PDFs of each service reveals unique facets linked to the nature and usage of the mobile application. For instance, Netflix, the leading platform for movie streaming, has a clear mode around 40 MB, and a drop of probability just after the 200 MB mark. When a user is connected to a mobile network, Netflix adopts an automatic balancing of data usage and video quality, allowing 4 hours of playback per GB of data in typical cases. Considering this setting, the first peak occurs at around 10 minutes of streaming, and the drop after around 50 minutes: both values are consistent with intuition, as they match the duration of one short episode of a series, and a full episode of a longer show.

The session-level traffic dynamics change substantially when looking at a different video streaming service, *i.e.*, Twitch, which, unlike Netflix, focuses on live content. The main mode, around 20 MB, and the main knee, at 800 MB, are shifted to the right with respect to the Netflix case; also, the amount of traffic per minute is much higher. The data indicates that Twitch users engage in long sessions with a high bitrate, suggesting that live streams tend to be consumed in more stationary conditions than on-demand movies.

Another example is Deezer, a popular audio streaming service, which shows two main traffic modes that map to the highest probability values: one is located around 3.5 MB and the other at 7.6 MB. At the standard bit rate of 128 kbit/s [11], the two modes translate to 3:40 minutes and 8:00 minutes of listening time, respectively. These roughly match the duration of one and two songs, including advertisements: according to the data, Deezer users most often listen to a couple of tunes while connected to a same BS, and longer listening times, while possible, are less likely.

Applications that mainly rely on relatively short message exchanges, such as Amazon (an archetypal web browsing service), Pokemon Go (a popular location-based game) or Waze (a navigation service generating floating car data), show a completely different behavior than streaming services. Loads per session are much lower, with traffic PDFs flattening to a zero value early on. Yet, the distributions and duration-traffic pairs are completely different also among these applications, highlighting once more the unique behavior exhibited by diverse services at the session level.

It is worth recalling that, in all PDFs, the duration of a session and the volume of traffic it generates are not only the result of the application or user's behavior, but also of the UE mobility. Indeed,

(a) Netflix (b) Twitch (c) Deezer

(d) Amazon (e) Pokemon Go (f) Waze

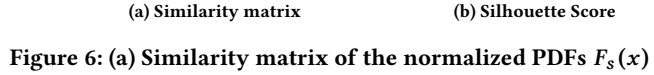
Figure 5: Probability density functions of the traffic volume $F_s(x)$ (top plots in each subfigure), and value pairs of discretized duration d and traffic volume $v_s(d)$ (bottom plots in each subfigure) for a selection of mobile services. PDFs and duration-traffic pairs are aggregated over working days (Monday through Friday) and weekends (Saturday and Sunday) separately.

many sessions of mobile users occur only in part within a same BS, and generate a smaller-than-expected volume of traffic for a complete sessions of the same application. This explains the presence of many very short sessions generating reduced traffic loads in the left part of the distributions of all services. Also, it allows interpreting the main mode of a streaming service like Netflix, which matches 3 MB and less than one minute of content: this is a reasonable mean dwell time in the BS for in-transit UEs running the application. Although frequent and thus important for a credible evaluation of mobile network performance, transient sessions have been ignored by traffic models proposed in the literature so far.

4.3 Quantitative analysis of service sessions

The qualitative analyses above unveil interesting aspects of session-level mobile traffic dynamics, which are however based on a close inspection of a few representative cases. To substantiate our observations, we perform a quantitative study of the traffic volume distributions $F_s^{c,t}(x)$; we consider for now data aggregated over all BSs $c \in C$ and days $t \in \mathcal{T}$, and compare different services s, *i.e.*, the PDFs $F_s(x)$ from (2), as per the following steps.

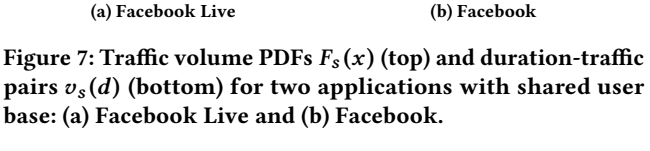
(i) We normalize $F_s(x)$, for each service s, so that all PDFs have zero mean. This removes the impact of the sheer volume of traffic generated by each application, enabling a comparison of less obvious dynamics, such as the standard deviation or the modes of the PDFs.
(ii) We compute pairwise earth mover distances (EMD) [35] among normalized $F_s(x)$, compiling a similarity matrix.
(iii) We run a centroid hierarchical clustering algorithm [30] on the similarity matrix, so as to identify classes of services characterized by similar PDFs. This algorithm iteratively groups the two PDFs at minimum distance, computes their average via (2), adds it to the set of PDFs in place of the original pair, and recomputes distances from the aggregate to all other PDFs in the set. By doing so, it builds a hierarchy of PDFs based on their similarity.



(a) Similarity matrix (b) Silhouette Score

Figure 6: (a) Similarity matrix of the normalized PDFs $F_s(x)$ of all services, with the three major clusters highlighted. Distance values closer to zero (dark red) indicate more similar PDFs. (b) Associated Silhouette score.

The result of this process is summarized in Figure 6a. Three main clusters emerge; by looking at the services in each, we relate these groups to (A) streaming services, (B) low-duty-cycle services relying on short messages, and (C) outliers.



(a) Facebook Live (b) Facebook

Figure 7: Traffic volume PDFs $F_s(x)$ (top) and duration-traffic pairs $v_s(d)$ (bottom) for two applications with shared user base: (a) Facebook Live and (b) Facebook.

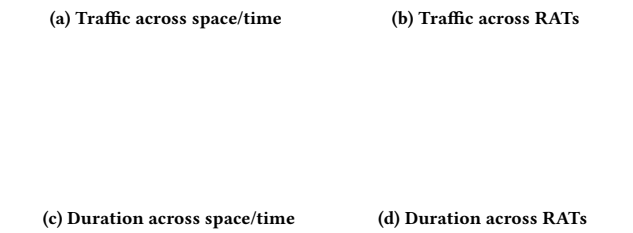
The emergence of two major behaviors is aligned with early observations in Section 4.2 about the clear difference between the dynamics of streaming applications like Netflix, Twitch and Deezer and those of less demanding services like Amazon, Pokemon Go or Waze. We can also confirm that this polarity does not depend on the user base but it is inherent to the nature of the service. Indeed, as shown in Figure 7, it affects services like Facebook Live (video streaming, cluster A) and Facebook (social media, cluster B), which have a largely common user population: the former has $F_s(x)$ and $v_s(d)$ aligned with that of the streaming applications in Figures 5a–5c, whereas the latter has flattened-out PDF and low-bitrate pairs as in Figures 5d–5f. We conclude that *session-level traffic is marked by a main dichotomy between video and audio streaming services and applications that rely on short or lightweight message exchanges.*

However, clustering services beyond the two major groups above is not possible. Figure 6b shows the evolution of the Silhouette score [37] over progressive splits of the services into a growing number of clusters. This index is widely used to identify meaningful clustering levels, where values closer to 1 indicates no overlaps and zero indicates overlapping clusters; the ideal cluster level is identified by a major drop of the score in the following level, as this indicates that breaking down the set into more classes generate significant overlap. Apart from the substantial change of value after the first 3 clusters, the Silhouette score stays nearly flat for all subsequent splits: finer-grained grouping of services is haphazard and does not reveal any informative pattern. Therefore, apart from a very macroscopic separation of streaming versus non-streaming traffic, *session-level statistics of mobile traffic demands cannot be characterized for whole classes of applications, but must be studied for specific services independently.*

4.4 Impact of space, time and technology

We now break down the analysis over the temporal and spatial dimensions, by looking at PDFs $F_s^{c,t}(x)$ and pairs $v_s^{c,t}(d)$ that are not aggregated over all BSs $c \in \mathcal{C}$ and days $t \in \mathcal{T}$.

In the time dimension, mobile traffic workloads are known to differ primarily between working days and weekends [14], hence we explore if the same distinction exists in session-level dynamics. We generate new aggregations of $F_s^{c,t}(x)$ and $v_s^{c,t}(d)$, over all BSs c but telling apart two sets of days: working days and weekends. We then compute the earth mover distance (EMD) [35] between the two traffic volume PDFs of a same service s for the two types



(a) Traffic across space/time (b) Traffic across RATs

(c) Duration across space/time (d) Duration across RATs

Figure 8: (a,c) Boxplots of differences in session-level traffic for (*i*) different services, and for each service across (*ii*) working days and weekends, (*iii*) urban, semi-urban and rural regions, and (*iv*) different cities. (b,d) Boxplots of differences in session-level traffic (*i*) for the same service across 4G and 5G RATs, and for difference services relying on (*ii*) 4G or (*iii*) 5G. Differences PDFs $F_s^{c,t}(x)$ in (a,b) are computed via EMD, while those between pairs $v_s^{c,t}(d)$ are computed using SED. Whiskers indicate the 5-th and 95-th percentiles, while the boxes outline the first, second and third quartiles.

of days. EMD compares a pair of PDFs by calculating the minimum cost of displacing samples of one distribution to match the other, returning a value zero for identical PDFs. For duration-traffic pairs we use a simple squared Euclidean distance (SED) of value vectors.

The distribution of these EMD and SED values is condensed in Figures 8a and 8c, under the 'Days' tag. As a reference, we also report the distances between different services, *i.e.*, the values in the matrix of Figure 6a, under the 'Apps' tag. By comparing the two boxes, it is evident that *the dynamics observed for a same service yield negligible differences across working days and weekends*, whereas inter-service heterogeneity is much more pronounced. Visual examples of the lack of impact of the day type on session-level traffic are also in Figure 5 and Figure 7, where measurements collected in workdays and weekends does not show clear differences.

From a spatial perspective, we experiment by aggregating $F_s^{c,t}(x)$ and $v_s^{c,t}(d)$ for each service s, over all days t but separately over BSs that belong to different regions and cities. At a region level, we compute PDFs and pairs for BSs that are located into (*i*) dense urban, (*ii*) semi-urban and (*iii*) rural regions; we employ urbanization level information provided by the local national institute for statistics to tell apart the three types of regions. Concerning cities, we derive statistics for each of the 5 largest metropolitan areas in the country.

We then repeat the test used before for different days of the week, by calculating for a same service the EMD of the traffic volume PDFs and the SED of the duration-traffic pairs among different regions as well as among diverse cities. The results are reported in Figures 8a and 8c, under the 'Regions' and 'Cities' tags. Again, distances are very small when confronted to those that affect diverse services under the 'Apps' tag. We conclude that *the geographical location of the BS has very limited impact on the session-level traffic statistics, hence a single model would generalize well across urbanization levels.*

Finally, we investigate the impact that different radio access technologies (RATs) have on the session-level statistics. For each service, we compute separate traffic volume PDFs and duration-traffic pairs for all sessions served by 4G eNodeBs and 5G gNodeBs. This allows studying if the statistics of a same application change when a user is connected via 4G or 5G. The result is reported in terms of EMD and SED in Figures 8b and 8d, under the 'RATs' tag, and shows that the diversity entailed by different RATs is negligible if compared to that determined by the service itself. The latter is reported in Figures 8a and 8c, under the 'Apps' tag, but is also broken down by technology in Figures 8b and 8d, under the 'Apps (4G)' and 'Apps (5G)' tags: there, we observe that difference across applications remain stable no matter if those are served by 4G and 5G BSs. Our conclusion is that *RATs do not impact in a significant manner the way users consume a same mobile service within a single transport-layer session.*

4.5 Key insights

Our characterization of session-level traffic yields a number of takeaways relevant to modeling, as summarized below.

A) The arrival rates of newly established sessions at a given BS follow a bi-modal distribution, independently of the load served by the BS, hence a same modelling strategy can be applied to arrival processes of all BSs.
B) The fraction of total sessions generated by each service is not uniform, but follows a negative exponential law, calling for a suitable breakdown of arrivals on a per-service basis.
C) Services are characterized by unique multi-modal distributions of per-session traffic volume, which present varied probability peaks at specific load values. Apart from a broad distinction between streaming and best effort services, applications cannot be grouped on the basis of class or using statistical clustering methods: each service requires dedicated session-level modeling of the load and duration of the session they induce.
D) The statistics of session-level traffic and duration of a given service do not vary significantly across days, urbanization level, metropolitan areas or RATs; hence, a single model suffice to represent the dynamics of a service at a BS.
E) Transient, partial sessions generated by users crossing the BS coverage area for a short time period occur with significant frequency and should be properly modelled.

5 MODELING SESSION-LEVEL TRAFFIC

We build upon the insights above to develop original models of mobile network traffic at the session level. Insights A and B offer pointers on how to model arrivals of sessions $w_s^{c,m}$ at one BS. The remaining ones provide indications on the modeling of the traffic volume PDFs $F_s^{c,t}(x)$ and duration-traffic pairs $v_s^{c,t}(d)$. Specifically, insight C implies that dependable models need to target each service $s \in \mathcal{S}$ separately. However, following insight D, we do not need to further specialize these per-service models for individual BSs $c \in C$ located in different regions and cities, for different days of the week $t \in \mathcal{T}$, or even across 4G and 5G NSA RATs. Ultimately, models of the aggregate $v_s(d)$ from (1) and $F_s(x)$ from (2) are enough to capture typical session-level mobile traffic reliably.

Based on these considerations, we adopt the following modeling approaches for $w_s^{c,m}$, $F_s(x)$, and $v_s(d)$, respectively.

- For session arrivals $w_s^{c,m}$, we use simple fittings of theoretical distributions on the bi-modal PDFs observed in Section 4.1, using a constant measurement-driven breakdown to associate each arrival to a specific service s.
- For the traffic volume PDFs $F_s(x)$, we present a novel algorithm to decompose and approximate the distributions as log-normal mixture models. Our model achieves good estimation of the original $F_s(x)$ for a wide set of services s with a small set of components (hence parameters). The approach operates over the full PDF domain, thus including short-lived transient sessions and abides by insight E.
- For duration-traffic pairs $v_s(d)$, we show that a regression using a power law model fits well all services s. Interestingly, these models let us comment on how throughput varies non-linearly with the duration of a session, in ways that are unique to each service.

5.1 Fitting of session arrivals $w_s^{c,m}$

Based on the analysis carried out in Section 4.1, we model the peak daylight arrival process of sessions at a BS and its off-peak nighttime counterpart separately. This gives a degree of freedom in emulating either day or night traffic.

By looking at Figure 3, the mode during peak hours can be described by a simple Gaussian distribution. The mean $\mu^{c,w}$ of the Gaussian fitting is necessarily different across classes of BSs characterized by different loads, which observe diverse arrival rates: it ranges from 1.21 sessions/minute for the first decile class up to 71 sessions/minute for the busiest BS decile. For the standard deviation $\sigma^{c,w}$, we observe a pattern emerging across all classes of BSs, such that $\sigma^{c,w} \sim \mu^{c,w}/10$ in all cases: this lets us automate the setting of $\sigma^{c,w}$ and simplify the models. The second mode, representing off-peak hours, is better modeled by a Pareto distribution, represented by $b^{c,w} \cdot (s^{c,w})^{b^{c,w}}/x^{b^{c,w}+1}$, where $[b^{c,w}, s^{c,w}]$ are the shape and scale parameters, respectively. The measurement data is well fitted by fixing the shape to $b^{c,w} = 1.765$ and modify only the scale $s^{c,w}$ across antennas. In fact, the growth of $\mu^{c,w}$ and $s^{c,w}$ across BSs in increasing load decile classes is similar, *i.e.*, exponential with akin rate. Examples of the resulting fittings are also shown in Figure 3.

According to the results of Section 4, it is important to model arrivals associated to different services, which are not uniform. We opt for a simple yet effective way to break the aggregate arrival distributions above on a per-service basis. Our approach stems from the consideration that the share of sessions induced by each service is relatively constant across different BSs and over time. Specifically, Table 1 presents the expected fraction of sessions and traffic volume generated by 28 popular mobile applications. The table also report the corresponding Coefficient of Variation (CV), *i.e.*, the ratio of standard deviation to the mean, across BSs and minutes. The CV thus represents the expected diversity of session and traffic shares yielded by each service. While the CV of the traffic share tends to fluctuate, that of the session share is fairly stable at around 1% across applications. In light of this observation, we use the session shares in Table 1 as probabilities to assign to a specific service a newly established session obtained from the fitted arrival rate PDFs.

(a) Main component and residuals (b) Residual selection (c) Final model

Figure 9: modeling steps for the log-normal mixture model of the traffic volume PDF $F_s(x)$, for a sample service, *i.e.*, Netflix. (a) Decomposition of the measurement distribution (light blue) into a main log-normal component (dashed) and residual probability peaks (red). (b) Identification and characterization of the residuals to be modelled (light grey areas), using their first derivative (orange). (c) Final residual components used by the mixture model (red), and reconstructed PDF $\tilde{F}_s(x)$ (black).

Table 1: Percent contribution to the total number of transport-layer sessions and to the total mobile traffic volume, for 28 applications and with associated CV.

Service	Sessions %	(CV)	Traffic %	(CV)
Facebook (FB)	36.52	±1.15	32.53	±1.68
Instagram	20.52	±1.27	31.48	±2.13
SnapChat	18.33	±1.17	9.52	±2.12
Youtube	4.94	±1.14	0.24	±1.39
Google Maps	2.76	±1.14	0.10	±2.82
Netflix	2.40	±1.29	11.10	±1.66
Waze	1.63	±1.39	0.62	±1.75
Twitter	1.46	±1.43	0.45	±1.49
Apple iCloud	±1.45	1.04	3.24	±4.20
FB Live	1.42	±1.17	1.80	±1.08
Spotify	1.12	±1.28	0.12	±2.54
Deezer	1.08	±1.91	1.59	±1.81
Amazon	0.96	±1.17	0.25	±1.11
Twitch	0.91	±1.22	3.67	±0.96
WhatsApp	0.85	±1.27	0.41	±2.91
Clothes	0.83	±1.23	0.85	±1.58
Gmail	0.54	±1.16	0.02	±1.17
LinkedIn	0.51	±1.23	0.54	±1.41
Telegram	0.44	±1.16	1.08	±3.27
Yahoo	0.32	±1.18	0.10	±2.40
FB Messenger	0.23	±1.25	0.01	±1.85
Google Meet	0.22	±1.11	0.14	±2.16
Clash of Clans	0.18	±1.25	0.09	±3.31
Microsoft Mail	0.11	±1.31	0.01	±4.48
Google Docs	0.09	±1.21	0.02	±3.58
Uber	0.07	±1.92	0.01	±1.55
Wikipedia	0.06	±1.30	0.01	±3.01
Pokemon GO	0.04	±1.21	0.01	±2.33

5.2 Log-normal mixture modeling of $F_s(x)$

The modeling approach for $F_s(x)$ is in three steps, which are illustrated in Figure 9 for one representative service, *i.e.*, Netflix. In the first step, exemplified in Figure 9a, we fit the experimental $F_s(x)$ using a log-normal distribution, *i.e.*,

$$LogN(x;\mu_s,\sigma_s^2) = \frac{1}{\sigma_s\sqrt{2\pi}} \cdot \exp\left(-\frac{(\log_{10} x - \mu_s)^2}{2\sigma_s^2}\right), \qquad (3)$$

which let us represent the broad trend of session-level traffic volume for each service s, denoted by $f_s(x)$. The rationale behind the choice of a log-normal fit is that it is the single function best representing the whole $F_s(x)$ for the vast majority of services: indeed, we can observe in all plots of Figure 5, Figure 7 and Figure 9a that the PDFs yield a resemblance to Gaussian-like shapes when the traffic is represented in a logarithmic scale. In this stage, we also subtract the fitted PDF $f_s(x)$ from the measurement PDF $F_s(x)$, bounding the result to positive values and obtaining a residual probability.

The second step, depicted in Figure 9b, focuses on analysing the residuals; these represent the unique peaks of session-level traffic of each service, and are thus instrumental to a realistic modeling of $F_s(x)$. We automate the process of identifying the most representative residual modes as follows.

- We compute the first derivative of the residual, using a first-order Savitzky-Golay filter [38] that smooths the resulting curve and helps the subsequent steps.
- We check the derivative against a threshold,[3] and record all continuous intervals of traffic values within which the derivative stays seamlessly above the threshold.
- We rank the aforementioned intervals based on the residual probability they contain, simply computed as the integral of the residual curve within each interval.

This method employs the change rate of the derivative to single out the residual peaks of actual interest for the modeling process; these are characterized by a high rate of change over a short traffic interval, such as the (zoomed-in) light grey regions identified by our algorithm in Figure 9b.

In the third and final step, we model the retained residual peaks. As those resemble low-variance Gaussian PDFs in log scale, we represent the n-th peak as a log-normal function

$$f_{s,n}(x) = k_{s,n} \cdot LogN(x;\mu_{s,n},\sigma_{s,n}^2), \qquad (4)$$

where $LogN(\cdot)$ is defined in (3). We set $\mu_{s,n}$ to the traffic value with maximum probability in the associated interval, so as to properly center $f_{s,n}(x)$; $\sigma_{s,n}$ is then set to $(0.997 \cdot \ell_{s,n})/3$, where $\ell_{s,n}$ is the span if the n-th interval, so that 99.7% of the modeled probability lays inside the interval. Finally, $k_{s,n}$ is the residual probability used to rank the intervals, and allows scaling the log-normal distribution. Samples of modeled residuals are in Figure 9c for the case of Netflix.

The final mixture model for a service s, denoted by $\tilde{F}_s(x)$, is obtained by composing the main and residual functions:

$$\tilde{F}_s(x) = \frac{f_s(x) + \sum_{n=1}^{N} f_{s,n}(x)}{1 + \sum_{n=1}^{N} k_{s,n}}, \qquad (5)$$

[3] Upon extensive tests, we find the algorithm to be robust to the choice of the derivative threshold, which avoids misinterpreting tiny oscillations as peaks. This allows using a same value, *i.e.*, 10^{-5}, to model any service s.

Figure 10: Power law exponents of the fitted $\tilde{v}_s^{c,t}(d)$ for a subset of services. Coefficients R^2 are in bold.

where N is the number of modelled residual peaks during the third step above, and the normalization factor at the denominator ensures that the expression in (5) is a distribution. Figure 9c provides an example of real $F_s(x)$ and its modeled counterpart $\tilde{F}_s(x)$, for the Netflix service.

To conclude, we note that other approaches to derive $\tilde{F}_s(x)$ are possible, *e.g.*, using traditional mixture models that automatically find the best decomposition of a PDF into multiple distributions of a given type. With respect to such alternative solutions, our algorithm not only produces models that are compact and accurate, but outputs components with a clear semantic (*i.e.*, the main trend, and a set of characteristic peaks), easing results explainability.

In this regard, it is worth noting that, when applied to the measurement data, our procedure identifies and models at most 3 residual peaks for the majority of services; the rare additional peaks have negligible weight $k_{s,n}$ below 10^{-4}. Therefore, we align all models and avoid irrelevant components, by limiting the maximum number of residual contributions to 3.

5.3 Power-law fitted modeling of $v_s(d)$

Value pairs of duration d and traffic volume $v_s(d)$ tend to align into very consistent patterns, as shown by the examples in Figure 5 and Figure 7: therefore, the relationships between the duration of a session and the load it generates are clearly not random, but follow statistical trends. Longer sessions are largely associated to higher traffic volumes, which is reasonable. Yet, the exact growth patterns are quite different across applications, as also observed in the figures above.

In order to properly represent the expression of $v_s(d)$ for each mobile service, we fit to the data varied functions from a range of families. Upon experimenting with polynomial, exponential, and power laws we find that the latter yield the best quality of fitting across all services, while limiting the model complexity. Specifically, we obtain a power-law model $\tilde{v}_s(d) = \alpha_s \cdot d^{\beta_s}$ for each application, by fitting $\{\alpha_s, \beta_s\}$ via the Levenberg-Marquardt non-linear least squares method.

The low complexity of the power law model facilitates its explainability, and in particular it let us quantify the diversity of behaviors in $v_s(d)$ discussed before. The fitted exponent β_s is especially revelatory in that sense. In a linear model where $\beta_s = 1$, then $\tilde{v}_s(d) = \alpha_s \cdot d$, and all sessions experience an average throughput α_s independently of their duration. A super-linear $\beta_s > 1$ denotes sessions whose mean throughput increases as they last longer, and a sub-linear $\beta_s < 1$ indicates that the instantaneous demand decreases for longer sessions.

(a) Twitch **(b) Twitter** **(c) Google Maps** **(d) Amazon**

(e) Facebook Live **(f) Facebook** **(g) Snapchat** **(h) Google Meet**

Figure 11: $\tilde{F}_s(x)$ and $\tilde{v}_s(d)$ (black solid lines) against measurement data (light blue) for a choice of services.

Figure 10 shows the value of β_S for a representative subset of mobile services. The exponent spans a wide range of values, from 0.1 to 1.8, so each application has quite different scaling of the average throughput to the session duration. Interestingly, when looking at the super- or sub-linearity of the models, video streaming services dominate super-linear behaviors. We speculate that this may be due to the fact that longer sessions within a same BS are generated by more stationary users, who also enjoy higher video bitrates thanks to a more stable and strong radio signal. Non-video applications have a sub-linear evolution of $\tilde{v}_s(d)$, as most require user interactions that tend to be less steady over longer periods.

5.4 Model quality and usage

Overall, we generate session-level traffic models as presented above for 31 mobile services, including all those listed in Table 1. We assess the accuracy of the models for $\tilde{F}_s(x)$ and $\tilde{v}_s(d)$ presented above by means of standard tests. For the traffic volume PDFs, we compute the error of the models by calculating its EMD with respect to the original measurement-based $F_s(x)$. Although the absolute value of EMD is not easily contextualized, we obtain in all cases results in the order of 10^{-5}, hence one order of magnitude lower than those recorded in the various tests on $F_s(x)$ in Figure 8a: we consider this a good indication of the fidelity of the models. In the case of duration-volume pairs, we compute the coefficient of determination R^2 as a measure of the quality of fit. The values are typically in the 0.7–0.9 range, which denotes a reliable fitting; in some cases, we obtain values as low as 0.5, which are still reasonable and, upon close inspection, are mainly due to noisy measurement data that creates outliers. Examples of the R^2 values are on top of each bar in Figure 10. Finally, visual illustrations of the real data and models are provided in Figure 11 for a subset of services, and show the good resemblance of $\tilde{F}_s(x)$ and $\tilde{v}_s(d)$ with the measurements.

Each model is fully characterized by a tuple of parameters $[\mu_s, \sigma_s, \{k_{s,n}, \mu_{s,n}, \sigma_{s,n}\}_n, \alpha_s, \beta_s]$, which we release publicly. This allows reproducing realistic session-level statistics for the traffic volume (extracted from $\tilde{F}_s(x)$), duration (obtained by applying the inverse function $\tilde{v}_s^{-1}$ to the traffic volume) and average throughput (computed as the ratio of the volume to the duration). Our open models can thus benefit the research community by empowering more dependable performance evaluations of mobile network systems and solutions, as demonstrated next in practical use cases.

6 APPLICATION USE CASES

We describe two use cases that showcase the critical impact that accurate session-level per-service traffic modeling can have in network management. We remark that our goal is not to derive innovative solutions for these use cases, but rather providing examples that illustrate, through simple network management scenarios, the utility of the presented models with respect to more traditional, simpler and not data-informed modeling of traffic. The first use case highlights the importance of per-service traffic characterization, while the second one shows the benefits of session-level modeling.

6.1 Capacity allocation for network slicing

Correct characterization of mobile traffic demand is crucial for resource allocation in network slicing: An accurate knowledge of the expected traffic demand for each Service Provider (SP) requesting a network slice would allow the operator to adjust the reserved resources in a more efficient manner, considerably increasing its profit margin by reducing operating expenses for each slice while accommodating more slices.

We consider a scenario in which the operator signs a Service Level Agreement (SLA) with each one of the 28 SPs included in Table 1. Each SP acquires a network slice to guarantee its traffic demand during peak hours (*i.e.*, except night from 10pm to 8am). The incoming sessions are sampled from the real data distribution, such that the share of traffic and number of sessions of each service follows the values indicated in Table 1, and the arrival time of the sessions is modeled such that the number of arrived sessions per minute at each RU follows the distribution of the model in Section 5.1. We consider that the terms of the SLAs are satisfied if the operator successfully delivers all the traffic demand from the SP's users at least the 95% of the time. In this setting, the operator must decide how much capacity it allocates to each of the SPs at each of the antennas.

6.1.1 Algorithms. We consider the derived models for the sessions' arrival time, the traffic per session, and the session's duration to determine the capacity allocation of each slice. Based on these models, we obtain the CDF of the traffic per service per antenna for different levels of demand. Considering this CDF and the average antenna load, we allocate to each slice the capacity that corresponds to its 95th percentile.

We compare this approach, which is only feasible with our derived session-level results, with two benchmarks. For that, we consider the mobile traffic models available in the literature [42], [31] that provide shares of mobile traffic for 3 service categories (Interactive Web (IW), Casual Streaming (CS) and Movie Streaming (MS)). To the best of our knowledge, there are no available models with higher level of service specification. Thus, as benchmarks, we consider BM A, which considers the three mentioned categories with the session shares derived from aggregating the corresponding values of Table 1 (IW: 49.30%, CS: 48.46%, MS: 2.24%), and BM B, which considers the three mentioned categories with the session shares from the literature (IW: 50%, CS: 42.11%, MS: 7.89%). For both benchmarks, the capacity allocated to each service within a category is split uniformly, since no information w.r.t. the intra-category session shares is available.

Table 2: Performance results for capacity allocation for network slicing averaged over antenna and service.

	Time with no dropped traffic (%)	Standard deviation
Model	95.15%	2.1%
BM A	89.8%	4.3%
BM B	87.25%	4.2%

Figure 12: Normalized traffic demand and allocated capacity to Facebook network slice at one BS over time.

6.1.2 Evaluation. We evaluate the performance of the system for a week in an area covered by 10 different antennas, for all the 28 services listed in Table 1. Table 2 shows the percentage of time for which the capacity allocated by the operator is sufficient to serve all the traffic demand, averaged over antennas and services. The solution based on the models proposed is the only one that achieves the SLA terms to guarantee the proposed Quality of Service. The other solutions suffer from the inaccuracy in estimating the share of demand of each service, and they also have bigger variability between services. An important aspect of this session-level per-service modeling is the robustness against outliers. Mobile traffic is very bursty, and dimensioning the slices based on traffic peaks may be very detrimental and lead to a waste of reserved resources. This can be seen in Figure 12, where the actual allocated capacity that satisfies the SLA terms is far below the traffic demand peaks.

6.2 Energy consumption in CU-DU

We consider a standard virtualized Radio Access Networks (vRAN) scenario, portrayed in Figure 13a, where Centralized Units (CU) located at a Telco Cloud Site (CS) serve traffic from a set of DUs at multiple Far Edge Sites (ESs), each associated to a group of Radio Units (RU). CUs run within physical servers (PS), whose energy consumption depends on the computing load. Therefore, the dynamic association of DUs to CUs within each PS, in accordance with the fluctuation of mobile data traffic at the RUs, determines the energy cost of the vRAN infrastructure for the operator. This is a major operating expense that needs to be minimized [15, 40].

6.2.1 Energy optimization model. Let us assume that all PSs at the CS are identical machines, whose capacity is limited by the maximum sum throughput of the mobile traffic they handle, up to 100 Mbps when working at full load [36]. We model the energy consumption at the each PS following real specifications of IBM servers [36, Table IV], such that a PS consumes a maximum power of 200 W when working on traffic at 100 Mbps; the power consumption is instead at 60 W when the PS is turned on but idle, and increases proportionally until the 200 W above at 100% load.

(a) System model (b) Performance error of traffic models (c) Power consumption over time

Figure 13: (a) vRAN system model considered in 6.2. (b) APE with respect to the measurements traffic in terms of active PSs and power consumption, for our model and the benchmarks. (c) Power consumption sample.

The operator employs then a dedicated algorithm for orchestrating the resources in the CU, executed at every time slot (TS) of one second. Due to the considered energy consumption model, minimizing the energy consumption of the system is equivalent to minimizing the number of active PSs. Thus, the algorithm is a bin-packing heuristic [18] that minimizes the number of PSs based on the current state of served sessions and the new session arrivals during the TS. While this model is relatively simple, it offers reasonable performance; more importantly, it provides a basis to assess the impact of traffic models on the compute resource management results, which is our goal.

6.2.2 Mobile traffic models. We assume a vRAN system with one CS serving 20 different ESs, each handling 20 RUs. The arrival time of the sessions is modeled such that the number of arrived sessions per minute at each RU follows the modeled distribution in Section 4.1. We consider that the sessions are generated according to three different strategies: (*i*) using the measurement data presented in Section 3.1, by sampling $F_s(d)$ and matching the traffic volume values to $v_s(d)$ to derive duration and average throughput; (*ii*) using our proposed models as described in Section 5.4; (*iii*) from traditional mobile traffic models available in the literature [42, Table II], [31, Table XVII] that provide throughput and session size/duration for three service categories.

For all cases, the share of per service sessions are extracted from Table 1. For (*iii*), we map our 28 classes (services) into the 3 categories that their model considers, and again generate sessions for each category according to Table 1. As we use the model in (*iii*) as a term of comparison, we generate in fact three different benchmarks from it: BM A fully adheres to the original models, BM B normalizes the generated data so that the total system throughput matches that observed in the measurement data, and BM C normalizes the throughput of each service class so that it matches that recorded in the measurement data. Clearly, BM B and BM C would not be feasible with information from the existing literature only, but they let us highlight the advantages of our models.

6.2.3 Performance evaluation results. We run experiments for several emulated days, orchestrating CS resources via the described strategy for all traffic models above. We employ the same realization of class-level session arrivals in all tests to avoid biases. Figure 13b summarizes the results, expressed as the distributions of the number of active PSs and of the power consumption. We report the absolute percentage error (APE) with respect to the same figures obtained by feeding the optimization model with the measurement data. Our model tightly approximates the real scaling of the compute resources at the CS, with median APE well below 5% and very small deviation for both metrics. The difference is apparent with respect to the benchmarks, which incur into APE of 100%–1000%, hence leading to performance results that are completely off. Clearly, the traffic generated by the benchmarks fails to capture real-world session-level statistics, which completely undermines the reliability of the performance evaluation. Figure 13c offers a close-up view of the temporal evolution of the power consumption with real data, our model and BM C: the result further highlights the quality of our contribution in mimicking real-world traffic.

7 CONCLUSIONS AND LIMITATIONS

We presented a first-of-its-kind exploration of mobile traffic at a transport-layer session level. Our study builds on substantial measurement data and reveals new facets of traffic, which we model accurately as a contribution to more reliable performance evaluations of mobile system. Our work presents a few limitations: the granularity of our data does not allow for fine grained or intra-session simulations (i.e. packet level generation); since our models are at service level, they will require updates over the years to consider changes in popularity and new services that emerge; due to the aggregation of sessions at BS level to comply with user privacy regulations, we lose the ability to study sequences of TCP/UDP flows a user may generate through the use of a mobile service, which limits an expansion of our study to application-layer dynamics. We plan to continuously collect data to provide updated models to the community, and in future works explore the patterns of specific network protocols and application-layer dynamics, as well as analyze the impact of user mobility on our models.

ACKNOWLEDGMENTS

The work of A.F.Z. was supported by BANYAN project, which received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 860239. The work of M.F. was supported by NetSense, grant no. 2019-T1/TIC-16037 funded by Comunidad de Madrid, and by the research project CoCo5G (Traffic Collection, Contextual Analysis, Data-driven Optimization for 5G), grant no. ANR-22-CE25-0016, funded by the French National Research Agency (ANR). The work of A.B.N. was supported by the Regional Government of Madrid through the grant 2020-T2/TIC-20710 for Talent Attraction.

REFERENCES

[1] 3GPP Technical Specification Group Services and System Aspects. 2020. TR:28.812 – Study on scenarios for Intent driven management services for mobile networks, Telecommunication management.

[2] 3GPP TR 36.814 V9.2.0. 2017. 3rd Generation Partnership Project; technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects (Release 9). (Mar. 2017).

[3] 3GPP TR 36.888 V12.0.0. 2013. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE (Release 12). (June 2013).

[4] 3GPP TS 23.288 v16.1.0. 2019. Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services (Release 16). (June 2019).

[5] 3GPP TS 28.533 v16.0.0. 2019. Management and Orchestration of Networks and Network Slicing; Management and Orchestration Architecture (Release 16). (June 2019).

[6] 3GPP TSG-RAN1#48 R1-070674. 2007. LTE physical layer framework for performance verification. (Feb. 2007).

[7] Jose A. Ayala-Romero, Andres Garcia-Saavedra, Marco Gramaglia, Xavier Costa-Perez, Albert Banchs, and Juan J. Alcaraz. 2019. Vrain: a deep learning approach tailoring computing and radio resources in virtualized RANs. In *ACM MobiCom '19*. ISBN: 9781450361699. https://doi.org/10.1145/3300061.3345431.

[8] G. Barlacchi et al. 2015. A multi-source dataset of urban life in the city of Milan and the province of Trentino. *Scientific Data*, 2.

[9] Dario Bega, Marco Gramaglia, Marco Fiore, Albert Banchs, and Xavier Costa-Perez. 2020. Aztec: anticipatory capacity allocation for zero-touch network slicing. In *IEEE INFOCOM '20*, 794–803.

[10] Biljana Bojovic and Sandra Lagen. 2022. Enabling NGMN mixed traffic models for Ns-3. In *Proc. Workshop on Ns-3*. ACM WNS3 '22, Virtual Event, USA, 127–134. ISBN: 9781450396516. DOI: 10.1145/3532577.3532602.

[11] Deezer Support. 2022. Deezer audio quality. https://support.deezer.com/hc/en-gb/articles/115003865685-Deezer-Audio-Quality. Accessed: 2022-05-31. (2022).

[12] ETSI. 2019. GS ZSM 001 V1.1.1 – Zero-touch network and Service Management (ZSM); Requirements based on documented scenarios.

[13] European Union. 2016. Eu general data protection regulation (gdpr): regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Retrieved October 18, 2021 from https://gdpr-info.eu/.

[14] Marco Helbich, Jamal Jokar Arsanjani, and Michael Leitner, editors. 2015. *Towards a comparative science of cities: using mobile traffic records in new york, london, and hong kong. Computational Approaches for Urban Environments*. Springer International Publishing, Cham, 363–387. ISBN: 978-3-319-11469-9. DOI: 10.1007/978-3-319-11469-9_15.

[15] Himank Gupta, Mehul Sharma, Antony Franklin A., and Bheemarjuna Reddy Tamma. 2020. Apt-ran: a flexible split-based 5g ran to minimize energy consumption and handovers. *IEEE Transactions on Network and Service Management*, 17, 1.

[16] Jin Huang and Ming Xiao. 2020. Mobile network traffic prediction based on seasonal adjacent windows sampling and conditional probability estimation. *IEEE Transactions on Big Data*, 1–1. DOI: 10.1109/TBDATA.2020.3014049.

[17] IEEE 802.16m-08/004r2. 2008. IEEE 802.16m evaluation methodology document (EMD). (July 2008).

[18] David Johnson. 1973. *Near-optimal bin packing algorithms*. PhD thesis.

[19] A. Karasaridis and D. Hatzinakos. 2001. Network heavy traffic modeling using /spl alpha/-stable self-similar processes. *IEEE Transactions on Communications*, 49, 7.

[20] Hatem Khedher, Sahar Hoteit, Patrick Brown, Véronique Véque, Ruby Krishnaswamy, William Diego, and Makhlouf Hadji. 2020. Real traffic-aware scheduling of computing resources in cloud-ran. In *ICNC '20*, 422–427. DOI: 10.1109/ICNC47757.2020.9049679.

[21] Daegyeom Kim, Myeongjin Ko, Sunghyun Kim, Sungwoo Moon, Kyung-Yul Cheon, Seungkeun Park, Yunbae Kim, Hyungoo Yoon, and Yong-Hoon Choi. 2022. Design and implementation of traffic generation model and spectrum requirement calculator for private 5g network. *IEEE Access*, 10, 15978–15993. DOI: 10.1109/ACCESS.2022.3149050.

[22] Jinsung Lee et al. 2020. Perceive: deep learning-based cellular uplink prediction using real-time scheduling patterns. In *ACM MobiSys '20*, 377–390. ISBN: 9781450379540.

[23] Rongpeng Li, Zhifeng Zhao, Chen Qi, Xuan Zhou, Yifan Zhou, and Honggang Zhang. 2015. Understanding the traffic nature of mobile instantaneous messaging in cellular networks: a revisiting to α-stable models. *IEEE Access*, 3.

[24] Rongpeng Li, Zhifeng Zhao, Jianchao Zheng, Chengli Mei, Yueming Cai, and Honggang Zhang. 2017. The learning and prediction of application-level traffic data in cellular networks. *IEEE Transactions on Wireless Communications*, 16, 6.

[25] Yu-Ting Lin, Thomas Bonald, and Salah Eddine Elayoubi. 2018. Flow-level traffic model for adaptive streaming services in mobile networks. *Computer Networks*, 137, 1–16. DOI: https://doi.org/10.1016/j.comnet.2018.01.027.

[26] Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. 2020. Using gans for sharing networked time series data: challenges, initial promise, and open questions. In *ACM IMC '20*. Virtual Event, USA, 464–483. ISBN: 9781450381383. DOI: 10.1145/3419394.3423643.

[27] Cristina Marquez, Marco Gramaglia, Marco Fiore, Albert Banchs, Cezary Ziemlicki, and Zbigniew Smoreda. 2017. Not all apps are created equal: analysis of spatiotemporal heterogeneity in nationwide mobile service usage. In *ACM CoNEXT '17*. Incheon, Republic of Korea, 180–186. ISBN: 9781450354226. DOI: 10.1145/3143361.3143369.

[28] Florian Metzger, Albert Rafetseder, Peter Romirer-Maierhofer, and Kurt Tutschku. 2014. Exploratory analysis of a ggsn's pdp context signaling load. *Journal of Computer Networks and Communications*, 526231. DOI: https://doi.org/10.1155/2014/526231.

[29] Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, K.P. Naveen, and Carlos Sarraute. 2017. Mobile data traffic modeling: revealing temporal facets. *Computer Networks*, 112, 176–193. DOI: https://doi.org/10.1016/j.comnet.2016.10.016.

[30] Daniel Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*.

[31] Jorge Navarro-Ortiz, Pablo Romero-Diaz, Sandra Sendra, Pablo Ameigeiras, Juan J. Ramos-Munoz, and Juan M. Lopez-Soler. 2020. A survey on 5g usage scenarios and traffic models. *IEEE Communications Surveys & Tutorials*, 22, 2, 905–929.

[32] O-RAN.WG2.Non-RT-RIC-ARCH-TS-v01.00. 2021. O-RAN Non-RT RIC Architecture 1.0. (Oct. 2021).

[33] O-RAN.WG3.RICARCH-v02.01. 2022. O-RAN Near-RT RAN Intelligent Controller Near-RT RIC Architecture 2.01. (Mar. 2022).

[34] Michele Polese, Francesco Restuccia, and Tommaso Melodia. 2021. Deepbeam: deep waveform learning for coordination-free beam management in mmwave networks. In *MobiHoc '21*. ACM MobiHoc '21, Shanghai, China, 61–70.

[35] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. 2017. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19, 2. DOI: 10.3390/e19020047.

[36] Soha Rawas. 2021. Energy, network, and application-aware virtual machine placement model in SDN-enabled large scale cloud data centers. *Multimedia Tools and App.*, 80, 10.

[37] Peter J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. DOI: https://doi.org/10.1016/0377-0427(87)90125-7.

[38] Abraham. Savitzky and M. J. E. Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36, 8, 1627–1639. eprint: https://doi.org/10.1021/ac60214a047. DOI: 10.1021/ac60214a047.

[39] M. Zubair Shafiq, Lusheng Ji, Alex X. Liu, and Jia Wang. 2011. Characterizing and modeling internet traffic dynamics of cellular devices. In *ACM SIGMETRICS '11*. San Jose, California, USA, 305–316. ISBN: 9781450308144. DOI: 10.1145/1993744.1993776.

[40] Rajkarn Singh, Cengis Hasan, Xenofon Foukas, Marco Fiore, Mahesh K. Marina, and Yue Wang. 2021. Energy-efficient orchestration of metro-scale 5G radio access networks. In *IEEE INFOCOM '21*, 1–10. DOI: 10.1109/INFOCOM42981.2021.9488786.

[41] Chuanhao Sun, Kai Xu, Marco Fiore, Mahesh K. Marina, Yue Wang, and Cezary Ziemlicki. 2022. Appshot: a conditional deep generative model for synthesizing service-level mobile traffic snapshots at city scale. *IEEE Transactions on Network and Service Management*, 19, 4, 4136–4150. DOI: 10.1109/TNSM.2022.3199458.

[42] Ilias Tsompanidis, Ahmed H. Zahran, and Cormac J. Sreenan. 2014. Mobile network traffic: a user behaviour model. In *2014 7th IFIP Wireless and Mobile Networking Conference (WMNC)*, 1–8. DOI: 10.1109/WMNC.2014.6878862.

[43] X. Wang, Z. Zhou, F. Xiao, K. Xing, Z. Yang, Y. Liu, and C. Peng. 2019. Spatio-temporal analysis and prediction of cellular traffic in metropolis. *IEEE Trans. Mobile Comput.*, 18, 09, (Sept. 2019), 2190–2202.

[44] Jing Wu, Ming Zeng, Xinlei Chen, Yong Li, and Depeng Jin. 2018. Characterizing and predicting individual traffic usage of mobile application in cellular network. In *ACM UbiComp '18*. Association for Computing Machinery, Singapore, Singapore, 852–861. ISBN: 9781450359665. DOI: 10.1145/3267305.3274173.

[45] Shangbin Wu, Yue Wang, and Lu Bai. 2020. Deep convolutional neural network assisted reinforcement learning based mobile network power saving. *IEEE Access*, 8, 93671–93681. DOI: 10.1109/ACCESS.2020.2995057.

[46] K. Xu, R. Singh, H. Bilen, M. Fiore, M. K. Marina, and Y. Wang. 2022. Cartagenie: context-driven synthesis of city-scale mobile network traffic snapshots. In *IEEE PerCom '22*. Los Alamitos, CA, USA, (Mar. 2022), 119–129. DOI: 10.1109/PerCom53586.2022.9762395.

[47] Kai Xu, Rajkarn Singh, Marco Fiore, Mahesh K. Marina, Hakan Bilen, Muhammad Usama, Howard Benn, and Cezary Ziemlicki. 2021. Spectragan: spectrum based generation of city scale spatiotemporal mobile network traffic data. In

ACM CoNEXT '21. Virtual Event, Germany, 243–258. ISBN: 9781450390989. DOI: 10.1145/3485983.3494844.

[48] Qiang Xu, Alexandre Gerber, Zhuoqing Morley Mao, and Jeffrey Pang. 2011. Acculoc: practical localization of performance measurements in 3G networks. In *ACM MobiSys '11*. Bethesda, Maryland, USA, 183–196. ISBN: 9781450306430. DOI: 10.1145/1999995.2000013.

[49] 2020. *Microscope: mobile service traffic decomposition for network slicing as a service*. ACM MobiCom '20, 14 pages. ISBN: 9781450370851.

ETHICS

Our work builds on mobile network traffic generated by users of a nationwide cellular infrastructure. Specifically, we employ session-level statistics at the level of individual BSs, which are generated from network measurements carried out in the target infrastructure as explained in Section 3.1.

The traffic measurements used to derive the session information were collected by the operator for network management and research purposes, and temporarily stored within a secure platform at their own premises. The aggregation into session-level statistics was also carried out in the same platform by personnel of the network operator, in full compliance with Article 89 of the General Data Protection Regulation (GDPR) [13] of the European Commission. The data collection and processing was approved by the Data Protection Officer (DPO) of the operator, and authorized by the French National Commission on Informatics and Liberty (CNIL), within the context of a collaborative research project.

We remark that the original network measurements contained personal identifiers (*e.g.*, the International Mobile Subscriber Identifier, or IMSI) and sensitive data (*e.g.*, locations of visited BSs, or mobile services consumed) about individual users, and were deleted upon aggregation. Instead, the aggregated session-level statistics consist of distributions and averages computed over hundreds of sessions at least, and do not contain personal identifiers or sensitive information, such as the device type, preference in terms of application consumption, or trajectories. In addition, the level of spatiotemporal aggregation ensures that no data subject can be re-identified, and that the statistics do not configure as personal data in the GDPR acceptation.

The researchers involved in the work presented in this paper only had access to such aggregated and privacy-preserving statistics for the purpose of carrying out the study. Ultimately, our dataset and research do not involve risks for the mobile subscribers, while they provide new knowledge about the dynamics of session-level traffic demands, which will benefit an improved design and more dependable validation of technical solutions for mobile network operations.