Analyzing Different Mobile Applications in Time and Space: a City-Wide Scenario

Armin Okic*, Alessandro E. C. Redondi*, Iacopo Galimberti†, Francesco Foglia†, Luisa Venturini†
*Dip. Elettronica, Informazione e Bioingegneria, Politecnico di Milano
†Vodafone Group, Network Engineering and Delivery
Email: name.surname@polimi.it or name.surname@vodafone.com

***Abstract*—We analyze a city-wide dataset of 4G mobile network traffic obtained directly from user-side logs, allowing fine-grained analyses of different application services over time and space. We group applications in classes and analyze their traffic patterns: the analysis reveals great heterogeneity in the usage of different applications and in their space/time correlations, with important implications for future networking services such as network slicing and resource allocations.**

I. Introduction

The massive increase of mobile cellular data traffic (7-fold from 2016 to 2021, according to Cisco) has pushed the entire telco community to rethink completely the traditional mobile network architecture and to introduce novel hardware and software technologies to support and optimize the delivery of different application services. To this end, the upcoming 5G networks will strongly rely on a series of virtualized and cloudified tools to provide novel and flexible functionalities both in the core (Network Function Virtualization (NFV) and network slicing) and at the edge/access (Multi-access Edge Computing (MEC), Cloud-RAN). Orchestrating such a complex set of heterogeneous technologies is very challenging and many research efforts are ongoing to provide working solutions to the problem. As an example, we mention here the SPOTLIGHT project[1], which aims at improving the performance of nowadays architectures by exploiting parallelization of network functions in the cloud.

In this complex scenario, it is envisioned that big data analysis and machine learning/data-driven methodologies will play a major role in all phases of the process. Indeed, in the last few years, network operators have started collecting massive data sets from their networks and analyzing it in order to obtain a deep understanding of the communication patterns of users and its implications on social dynamics (user interactions, demographics, epidemics, etc.), user mobility (mobility models, traffic prediction, etc.) and network planning (resource management, energy efficiency, etc.) [1]. The majority of such works exploit Call Detail Records (CDR) [2] and focus on aggregated data, voice and messages traffic volumes exchanged in the network, sometimes distinguishing between incoming/outgoing calls or uplink/downlink data traffic, but rarely separating traffic produced by different services (e.g., video streaming, web browsing). However, the 5G vision requires to gain an even deeper understanding on the usage patterns of the different application services, rather than looking at aggregate traffic patterns. Only recently, with the advent of powerful Deep Packet Inspection (DPI) commodities, some works have analyzed traffic datasets with the goal of describing the properties of different services, either focusing on macro-service classes [3], [4] or on specific applications [5], [6]. The datasets used in such works are generally obtained from DPIs located at Gateway GPRS Support Node (GGSN) or Packet Data Network Gateway (PGW) for 3G and 4G networks, respectively. While on the one hand such a method allows to obtain very large datasets, possibly covering entire countries, on the other hand it has two weaknesses: (i) spatial accuracy is limited, since the geolocation information available at the GGSN/PGW are not updated during intra-RAT handovers but only during inter-RAT handovers or disconnections and (ii) the entire method relies on the classification power of the used DPI tools, which is not 100% accurate.

In this work we add a step in the direction of understanding the spatio-temporal usage patterns of different application services, focusing on a one-month city-wide mobile traffic dataset. The peculiarity of this dataset is that, differently from previous works, it is collected directly from the users rather than from network aggregation points. This allows a much finer spatial resolution (data is geolocated at the eNodeB level) as well as increased accuracy in the classification of the different services (which are labeled directly from users).

We focus on different types of analysis over the available dataset, including per-service traffic distributions in space and time, spatio-temporal traffic correlations and finally eNodeB clustering. Our analysis confirms some of the insights recently provided in related works on service characterization built from network-side measurements and provides important design guidelines for future works, especially the ones related to C-RAN optimization, radio resource management and network slicing [7].

The rest of this paper is structered as it follows: Section II details the dataset under consideration and the main data preprocessing operations. Section III, IV and V focus on the temporal, spatial and joint spatio-temporal analysis of different application services, respectively, while Section VI describes and discusses clustering of eNodeBs. Finally, Section VII concludes the paper.

[1] http://gain.di.uoa.gr/spotlight/

978-1-5386-7646-2/19/$31.00 ©2019 IEEE

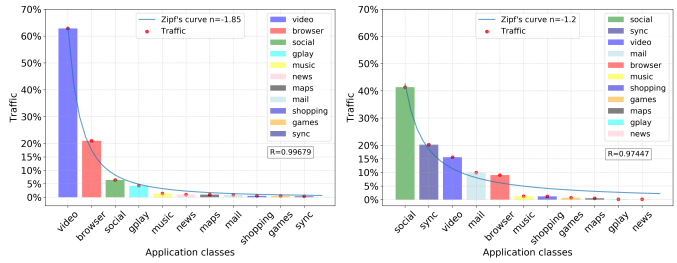


Fig. 1: Distribution of classified data traffic into application classes for Downlink(left) and Uplink(right).

TABLE I: Dataset details

Dataset size [GB]	17
Uplink to Downlink volume ratio	0.015
Number of unique users	125609
Number of eNodeBs	406
Total number of applications	7215
Classified applications	172

II. Dataset

Differently from previous works, the dataset under analysis in this paper is directly obtained from about 125k customers of Vodafone, one of the major European mobile network operators. An ad-hoc Android application is installed on the user equipments (UE) after explicit consent of their owners: the application runs in background and logs statistics relative to the different applications run by a user, including the uplink/downlink 4G traffic volume as well as the serving eNodeB indicator. All data is reported to a central server using anonymous identifiers for users and then aggregated hourly at the eNodeB level to further ensure not to raise any privacy, ethical or legal issues. In this paper we restrict the analysis to the eNodeBs of a middle-sized European city. The dataset under analysis is relative to the entire month of April 2018 and covers a fraction of the total operator customers in the city. Table I reports further details on the dataset.

A. Preprocessing

To manage the high number of unique applications contained in the dataset (more than 7k), a grouping operation is performed. Grouping is implemented considering only those applications generating at least 1 GB of traffic in the observation period, disregarding applications generating an insignificant amount of traffic or with identifiers that could not be linked to any application service. In total we observed that 95% and 93% of downlink and uplink traffic, respectively, are covered by just 172 applications, which are grouped in 11 application classes, detailed hereafter:

- Video streaming: YouTube, NetFlix, Facebook video, etc.
- Social and instant messaging: Whatsapp, Facebook messenger, Viber, Snapchat, Musically, etc.
- Browsers: Chrome, Firefox, Android built-in browsers, etc.
- Google play services
- Maps and navigation: Google Maps, Moovit, Waze, etc.
- Music: Spotify, Tidal, Deezer, etc.

TABLE II: Busy hours per app class for Downlink and Uplink

	Downlink busy hour	**Uplink busy hour**
Total traffic	18	16
Social	14	18
Video	18	19
Browsers	18	18
Maps	18	18
Music	18	7
Google Play	8	8
E-mail	12	14
Sync	19	16
News	8	8
Shopping	8	14
Games	13	14

- E-mail: Gmail, Outlook, Yahoo, etc.
- Gaming: Clash Royale, Candy Crush, etc.
- Shopping: Wish, Amazon, Ebay, etc.
- News: TGCOM24, Google News, Flipboard, etc.
- Syncing: Google Docs, Dropbox, WeTransfer, etc.

Figure 1 shows the application classes ranked by percentage of generated traffic volume. The rankings are nicely fitted with a Zipf distribution with parameters 1.84 and 1.2 for the downlink and uplink case, respectively. We observe that such parameters are in line with the study done in [6], where Zipf distributions were fit on a one-week country-wide dataset obtained from DPI measurements. We note that Video, Browsing, Social applications and Google Play services dominate the downlink traffic, accounting for almost 90% of the total traffic. Video class alone accounts for 60% of the downlink traffic, in line with recent Cisco estimates on mobile video traffic[2]. Things are very different for uplink traffic, where social media applications, video and file uploads as well as e-mail transmissions dominate the ranking. Similar behaviours were observed in [6] and indicate that the dataset under consideration in this work, although obtained only from a subset of the operator's customers, is representative of the whole population of mobile users.

III. Temporal analysis

A. Traffic Signatures

In this section we analyze traffic focusing on its temporal characteristics, through the use of traffic signatures. Creating signatures of the data traffic is an essential process because it reduces the dimensionality of the observation space by focusing on the most peculiar features of the data traffic behavior in time. In this work, we consider the Median Week Signature (MWS) [2], obtained for each eNodeB by taking the median value of its downlink and uplink traffic over the same hours and days of a week.

B. Application classes busy hour

As a first step, we consider the whole traffic generated in the network by summing together all MWS for all eNodeBs,

[2]https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

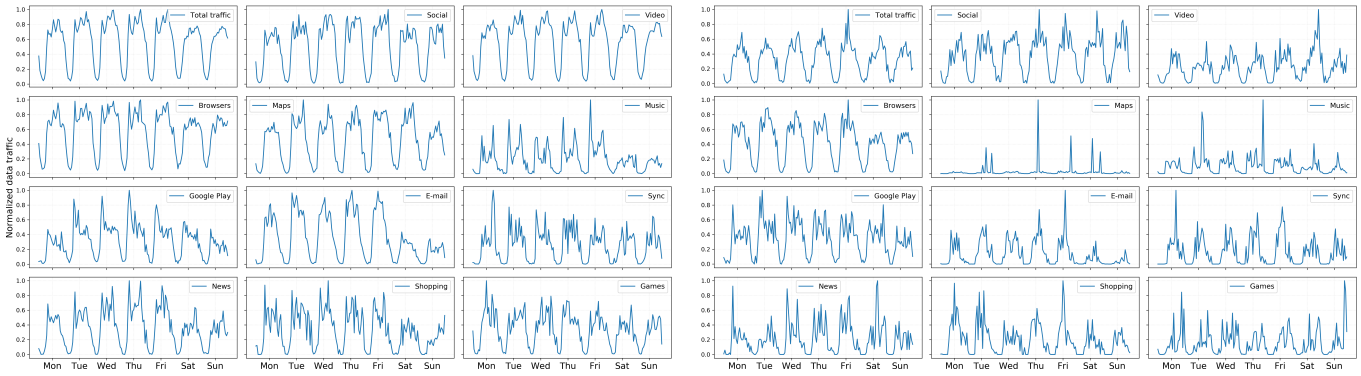


Fig. 2: Median Week Signatures per application class for Downlink(left) and Uplink(right).

as well as the traffic produced by the different application classes. The results are shown in Fig. 2, both for the downlink and the uplink, where traffic is normalized to unity not to disclose the absolute volumes as per the operator request. We notice that while total traffic exhibits the typical and well-known periodical behavior over a week, different applications have very distinctive patterns in the traffic signatures. As an example, the behavior of downlink video traffic is almost the same every day, while for the uplink has a strong peak on saturdays nights (mainly due to Instagram and Facebook live video upload). Conversely, music applications downlink traffic show two sharp peaks during work days, corresponding to commuting times in the morning and in the evening, while the traffic during weekends is much lower. Table II summarizes the downlink/uplink traffic busy hours per application, obtained by averaging together all median hourly values for the different days and by picking the hour corresponding to the maximum traffic. As one can see, the usage of different applications define three different downlink activity peaks times during a day: morning for applications such as Google Play, Music and Shopping, lunch break for E-Mail and Games and evening for the remaining set of applications, which also define the total traffic busy hour. Similar activity temporal patterns can be found for the uplink, with classes such as Music and News having their busy hours in the morning, E-mail, Shopping and Games during lunch break/early afternoon and the remaining applications in the evening. This analysis confirms the large heterogeneity present in the temporal usage of different applications and must certainly be taken into account during the design phase of future advanced network slicing services envisioned by 5G.

C. Correlation between application classes

We also compute the Pearson's correlation coefficient for all application classes considering the downlink and uplink traffic and additionally splitting the analysis in working days and weekends. Results are shown in Fig. 3 and 4. This allows to have a quick overlook on which applications exhibit similar temporal behaviors, which may be useful for allocating resources dynamically in future network architectures [7]. As one can see, applications are generally more correlated in downlink than in uplink, and in working days than in weekends. The stronger correlations are among video, browsing and social applications, which are as well the classes which generate most of the traffic. Other application classes such as Music or Google Play services for the downlink, and Maps for the uplink are in general less correlated with other classes, indicating that they exhibit unique temporal behaviors.

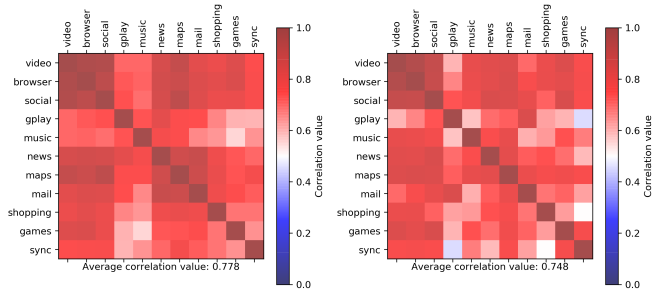


Fig. 3: Correlation matrix between classes for working days (left) and weekend (right) in Downlink.

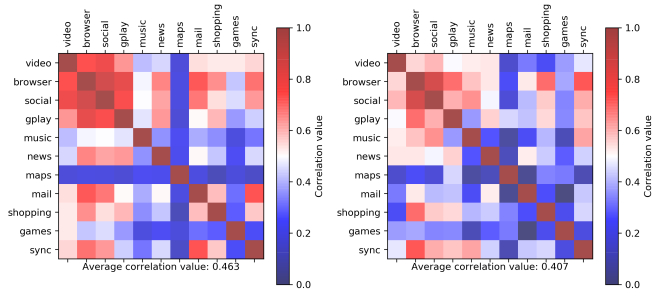


Fig. 4: Correlation matrix between classes for working days (left) and weekend (right) in Uplink.

IV. Spatial Analysis

The dataset under consideration is annotated at the eNodeB level, therefore it allows to perform a fine-grained analysis of how different applications are used in different locations in the city. We focus here only on downlink traffic, although similar considerations can be done for the case of uplink.

There are 406 eNodeBs in the dataset, for which we plot in Fig. 5 (left) the Cumulative Distribution Function of the total downlink generated traffic. As one can see, 20% of eNodeBs are generating more than 50% of the total traffic, while 95% of the total traffic is generated by less than 70% of the eNodeBs. We consider only this subset for the analysis that follows.

A. Spatial correlation between eNodeBs

In Fig. 5 (right) we compute the correlation matrix between the MWS of the total traffic for each pair of eNodeBs, where eNodeBs are sorted in decreasing order of the corresponding total traffic. As one can see, spatial correlation is generally lower than temporal correlation, with just some specific pairs of eNodeBs showing high correlation values. We observe some eNodeBs (140 and 143 in Fig. 5) which are completely uncorrelated with the rest of the network, but still producing a high amount of traffic. This means that there are eNodeBs characterized by unique temporal behaviors and whose location could be analysed for spotting possible anomalies (e.g., excessive downlink/uplink volume during nights). The analysis is continued by breaking down the total traffic spatial correlation matrix into different application classes in Fig. 6. The horizontal line inside each box represents the median spatial correlation value and the triangle mark stands for the average value, while the lower and upper edges of a box indicate the 25th and 75th percentiles. The lines outside the box are representing minimum and maximum values. The analysis reveals application classes of two kinds: Social, Video, Browser, Gplay and Maps classes show a moderate average spatial correlation, in line with the matrix shown for total traffic. The distribution of correlations seem somehow balanced, with median values in between minima and maxima. Conversely, other classes like Games, Shopping, News, Syncing, Mail and Music, have very low median value of correlation, but exhibit very high maxima values. This means that usage of these applications is not correlated in most cases in space, but there are strong locality effects in which such applications behavior is very similar. Considering such results together with the temporal correlations between classes shown in Fig. 3, we conclude that even though some classes have high correlation from the temporal point of view, the way they are used in space may be very different. This adds another design guidelines for future resource allocation and network slicing tools, which need to consider such high spatio/temporal heterogeneity in order to work efficiently.

V. Joint spatio-temporal analysis

From the results in Sections III and IV, we conclude that the analysis should combine temporal and spatial characteristics of the traffic behavior. To understand this better, we recompute the average spatial correlation between eNodeBs restricted to time intervals of 4 hours. In details, the MWS of each eNodeB is split in 6 signatures, considering only specific time interval during each day. This decomposition of time allows to observe how spatial correlations of each application class evolve during time. Results is shown in Fig. 7. We observe

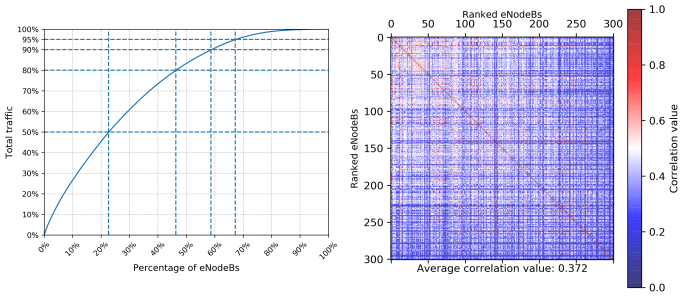


Fig. 5: Cumulative traffic over ranked eNodeBs (left), Correlation matrix between eNodeBs for total traffic in downlink(right).

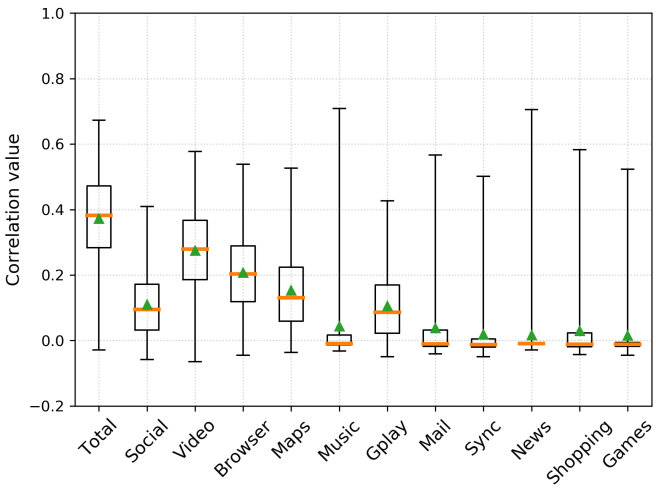


Fig. 6: Distributions of correlations between eNodeBs for each app class.

very different behaviors in time, with high values of average spatial correlation for the two first intervals mainly due to very similar activities of users (low usage in the late night and exponential increase of usage in the morning). We show here only average correlations for space reasons, but the variances shown in Fig. 6 remain persistent in each time interval. This means that operations such as spatial clustering of eNodeBs may benefit of dynamic algorithms which exploit this time-varying similarities.

To understand better how different applications' space usage vary over different time intervals, we plot in Fig. 8 heat-maps of traffic generated by eNodeBs for specific classes. The behavior of applications usage is changing among time intervals, as well as between different applications. The overlap of heat-map layers created by different applications shows that some of the most active areas, mainly in the central part, are persistent among application classes, while the differences in the location of peaks indicate distinctive activity patterns of applications in space. These two observations underlines the fluctuations in application usage and general dynamics in the network on a daily basis.

VI. Clustering

Based on the outcomes from our temporal and spatial analysis, we decided to proceed with a clustering of eNodeBs

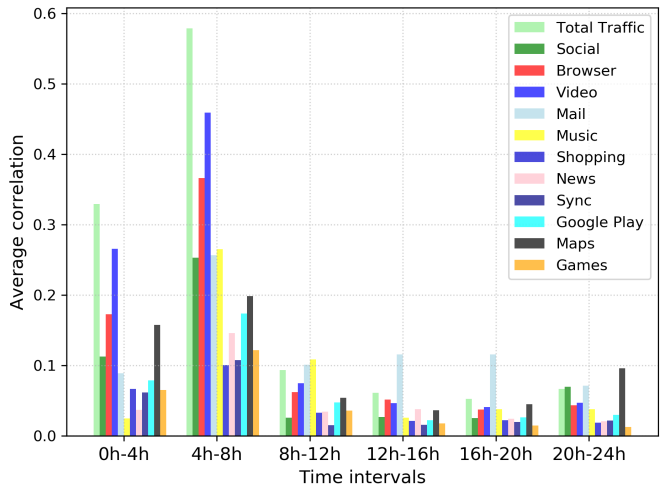


Fig. 7: Average correlation values between eNodeBs for different classes and time intervals.

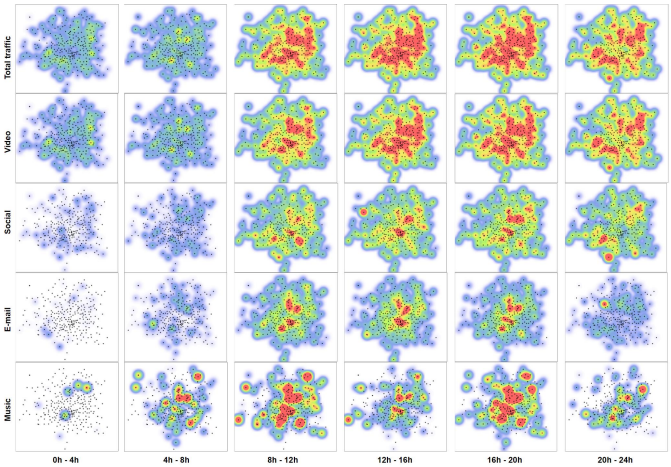
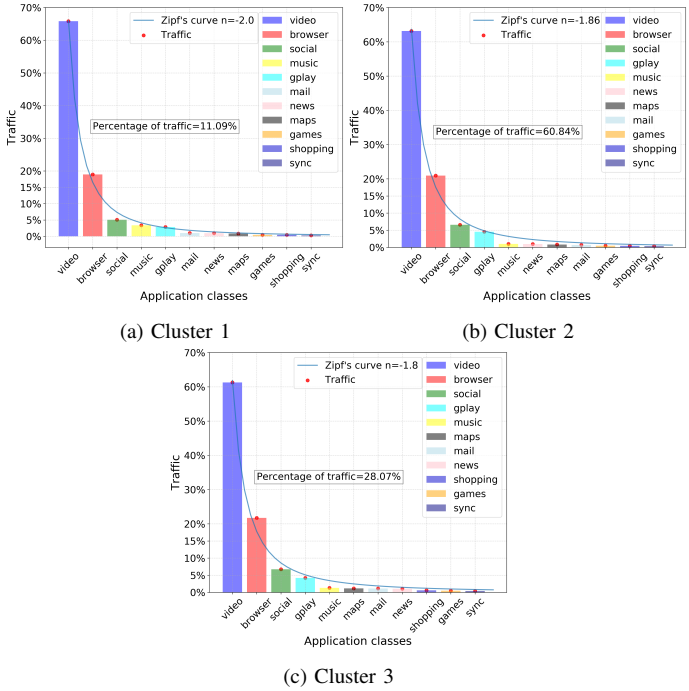


Fig. 8: Map of eNodeBs activities during different time intervals for different applications. Red is showing high traffic, blue is for low traffic. The traffic is scaled for each class.

per application class (e.g., grouping together those eNodeBs whose temporal behavior per application is most similar). The main goal of such process is to understand whether eNodeBs are clustered in the same groups when considering different application classes. Clustering is performed starting from the normalized MWS of each eNodeBs: several techniques can be used to cluster the signatures of similar eNodeBs together, including k-Means, k-Shape and DBSCAN [8], [9]. All clustering techniques produce similar clusters, although here we report results only for k-Means, which outperformed other approaches in terms of both Silhouette [10] and Davies-Bouldin [11] clustering evaluation indicators. The same metrics were also used to obtain the value k of clusters, which for most applications was equal to 3. Cluster centroids resulting from the clustering process for some representative cases are shown in Fig. 10, while the locations of eNodeBs on the map of the city are shown in Fig. 11. We can observe that even though clustering is done on a per-application basis,



(a) Cluster 1 (b) Cluster 2

(c) Cluster 3

Fig. 9: Distribution of downlink traffic inside extracted clusters.

the centroids extracted are always very similar. Three main behaviors may be observed:

1) *Commuting-specific pattern*: the first centroid highlights strong peaks during morning and evening and very low usage during weekends, and it is common to all applications. Observing the locations of the corresponding eNodeBs on the maps, they tend to be localized in few spots, relative to big commuting hubs like main train and subway stations.
2) *Daily periodic pattern*: in the second pattern the daily behavior remains quite constant throughout the week, with different applications having small differences. As an example Social class tends to have pronounced activity peaks during evenings. Observing the locations of eNodeBs in the city, they tend to be localized in residential areas. Note that some applications related to business activities (e.g., e-Mail) do not show this behavior.
3) *Working days / week ends pattern*: the third centroid shows the well known daily differences between working days and week ends typical of business areas [2]. The eNodeBs grouped in this cluster are generally located in the city center where all business activities take place.

Figure 9 shows the distribution of applications inside each cluster when eNodeBs are clustered using the MWS of total traffic. The three different clusters have very similar distributions, with Video, Browser and Social classes dominating the ranking. However, the rankings for different applications in each cluster are different. For example, in the second and third cluster (residential and business locations), Google Play services are generating more traffic than Music apps, while in the first cluster (commuting areas) the ranking is inverted.

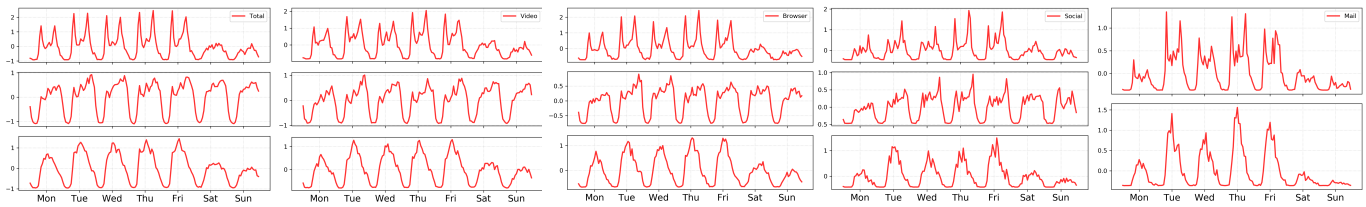


Fig. 10: Cluster centroids obtained by k-Means clusterting of eNodeBs per application class.

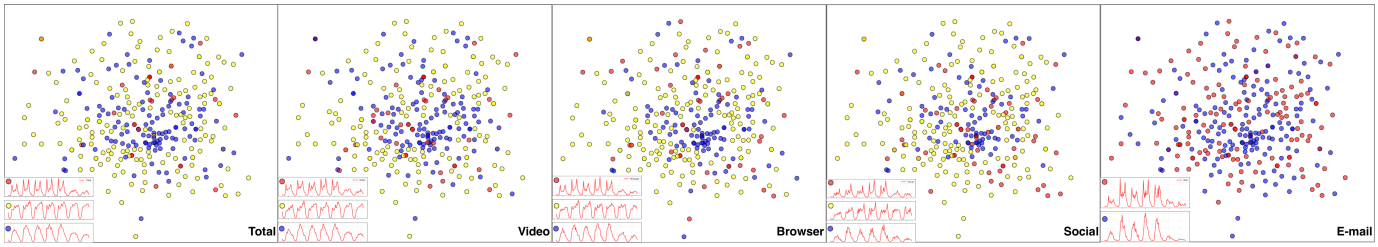


Fig. 11: Clustered eNodeBs on the map. Each dot represent real location of the eNodeB. Different color is indicating different cluster.

From Figure 11 we also observe that there are differences in how eNodeBs are clustered according to different applications. This means that even clustering resulted in the creation of very similar temporal centroids for different applications, the spatial behavior of clusters is not always the same.

The observations in this section indicate the importance of fine-grained spatio-temporal analysis for eNodeBs clustering. At the same time, the existence of standard and well-known temporal profile patterns common to all applications can greatly simplifies the design of advanced networking solutions, which can be built using template signatures to be adapted to each application / network slice in a space-dependent fashion.

VII. Conclusion

In this paper we analyzed the time and space characteristics of the network usage of different mobile applications. The dataset under consideration is derived directly from user terminals, therefore allowing very fine-grained spatial as well as application classification accuracy. Although the dataset is coming from a subset of users, we confirmed results recently presented in the state of the art, indirectly validating the available data. The analysis reveal that (i) different applications are used very differently in both space and time, (ii) the correlation between eNodeB usage of different application has great spatio-temporal variance and (iii) clustering eNodeBs based on temporal usage produce similar centroids for all applications. Future research directions will explore how to apply the results from this paper on dynamic clustering of distributed units in the C-RAN architecture as well as on the combination of the results with emerging 5G technologies like network slicing or MEC orchestration.

Acknowledgment

This research has been partially supported by the H2020-MSCA-ITN-2016 framework under grant agreement number 722788 (SPOTLIGHT).

References

[1] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 124–161, 2016.

[2] A. Furno, M. Fiore, R. Stanica, C. Ziemlicki, and Z. Smoreda, "A tale of ten cities: Characterizing signatures of mobile traffic in urban areas," *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2682–2696, 2017.

[3] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman, "Identifying diverse usage behaviors of smartphone apps," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, ACM, 2011, pp. 329–344.

[4] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3g cellular data network," in *INFOCOM, 2012 Proceedings IEEE*, IEEE, 2012, pp. 1341–1349.

[5] Y. Zhang and A. Arvidsson, "Understanding the characteristics of cellular data traffic," in *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*, ACM, 2012, pp. 13–18.

[6] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, and Z. Smoreda, "Not all apps are created equal: Analysis of spatiotemporal heterogeneity in nationwide mobile service usage," in *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*, ACM, 2017, pp. 180–186.

[7] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "How should i slice my network? a multi-service empirical evaluation of resource sharing efficiency," in *ACM MobiCom*, vol. 18, 2018.

[8] T. W. Liao, "Clustering of time series data a survey," *Pattern recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[9] J. Paparrizos and L. Gravano, "K-shape: Efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ACM, 2015, pp. 1855–1870.

[10] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[11] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.