Left on read: Human behavior characterization based on messaging service applications

Geymerson S. Ramos, Gean Santos, Douglas L. L. Moura, Danilo Fernandes, Fabiane Queiroz, Rosso Osvaldo A., Razvan Stanica, Andre Aquino

Left on read: Human behavior characterization based on messaging service applications

Geymerson S. Ramos[1], Gean Santos[2], Douglas Moura[2], Danilo Fernandes[2], Fabiane Queiroz[2]
Osvaldo A. Rosso[2], Razvan Stanica[1], and Andre L. L. Aquino[2]
[1]Univ Lyon, Inria, INSA Lyon, CITI, Villeurbanne, France [2]LaCCAN Laboratory, Federal University of Alagoas, Brazil.
geymerson.ramos@inria.fr; {gean.santos, douglas.moura, dfc, fabiane.queiroz}@laccan.ufal.br;
oarosso@laccan.ufal.br; razvan.stanica@inria.fr; alla@laccan.ufal.br

I. General Problem and Motivation

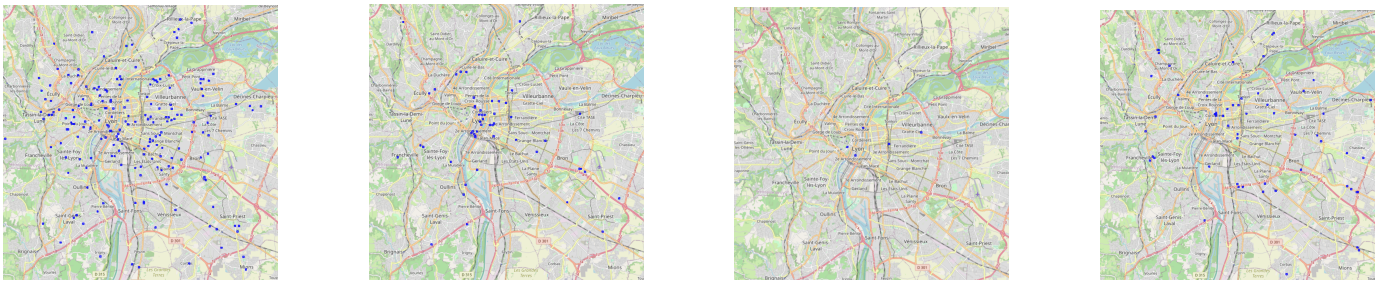
The objective of this study is to analyze mobile data from messaging applications. We aim to understand messaging app usage and gather information that can enhance infrastructure and quality of life within a smart city context. We will conduct our analysis using the NetMob23 dataset [1], which covers the period from March 16, 2019, to May 31, 2019, and provides uplink and downlink data for messaging applications such as WhatsApp, Telegram, Facebook Messenger, and Apple iMessage. Our observations pertain to the Lyon Metropolis in France. The traffic dataset provided is mapped through GeoJSON files using the WGS84 coordinate system. Each feature represents one square cell (tile) covering an area of (100×100) m^2. The Lyon metropolitan area has a total of 54013 tiles, and we used the data provided by the Grand Lyon Portal [2] to label some of these tiles, grouping them in the following classes: C1) *Education Centers* (208 tiles); C2) *Events* (74 tiles); C3) *Commerce* (67 tiles); C4) *Hotels* (58 tiles); C5) *Sports* (57 tiles); C6) *Restaurants* (51 tiles); C7) *Religious Centers* (41 tiles); C8) *Hospitals* (31 tiles); C9) *Train Station* (5 tiles). Figs. 1((a) – (d)) show the tile distribution of 4 different classes across the Lyon metropolitan area. We considered these classes to explore the following hypothesis: *"It is possible to characterize, based on information theory, the tiles where users are more likely to engage in online conversations"*. To verify this hypothesis, we conducted a detailed analysis of the tiles for each class, and looked for similar usage behavior categorized with the Complexity-Entropy Causality Plane (CECP) [3]. Only WhatsApp network traffic was considered in the study because it is the application which generates more traffic.

II. Results

Given that the NetMob dataset [1] provides 77 time series S per tile (one observation for each day), we used WhatsApp uplink data to compute the average traffic time series $\overline{S}_c = \{\overline{x}_{00:00}, \overline{x}_{00:15}, ..., \overline{x}_{23:30}, \overline{x}_{23:45}\}$ as the typical traffic signature for each class. The average network traffic at a specific time is represented by $\overline{x}_{hh:mm}$. We also computed the average traffic per day for each time series, represented by μ_S, and discarded the time series with average daily traffic below the median or above the 75th percentile (third quartile) values among all the averages. Therefore, each class has a representative and unique average time series generated from a set $D = \{S \mid median \leq \mu_S \leq Q_3\}$. This helps to mitigate the impact of anomaly events, outliers, and low traffic days that might not be representative for our analysis. We also make a distinction between weekdays (Monday to Friday) and weekends. This results in an average class behavior, which can be seen in Fig. 2 for data of the *Education Centers* class.

In Fig. 2, we can observe that the traffic typically begins to increase earlier during week days, at around 5:00, as compared to weekends (6:00). It also starts to decrease around 22:30 for weekdays and around 23:30 for weekends. One reason for this difference may be attributed to the weekly responsibilities related to studies and teaching. The *Education Centers* class suggests that people tend to wake up earlier during weekdays to attend schools, universities, and similar institutions, which might also imply exchanging WhatsApp messages. On weekends, we see a higher traffic volume. People typically have a break from these obligations, and they may wake up and go to sleep later as well. For both cases, peak usage occurs between 12:00 and 21:30. The early activity on weekdays seems to occur for most of the analyzed classes, but this difference is reduced for the *Train Stations* class. This class has the highest average traffic volume, as shown in Figure 3 (C9). The peak time traffic for this class appears around 18:00. For train stations, this refers to the afternoon period when there is the highest demand for train services. This is typically when many people are commuting from work and school, a good moment to exchange messages with friends and family. Interestingly, there is no peak in WhatsApp traffic during the morning commute hours, when the train demand is even more significant.

We can take a deeper look at the underlying nature of C9 and the other classes by analyzing the Complexity-Entropy Causality Plane in Fig. 4. The *Commerce* class (C3) has the highest permutation entropy and statistical complexity, which means more randomness and hard to predict usage behavior if compared to the other classes. This behavior usually implies chaotic traffic, with the least existing patterns and minimal information. Uncorrelated stochastic processes have $H \approx 1$ and $C \approx 0$. If we look specifically at the *Train Station* and the *Commerce* classes, one possible explanation to why C3 is more chaotic than C9 is that train stations coordinate travels, and people arrive and leave stations at specific and



(a) Education + Leisure (b) Hotels (c) Train Stations (d) Commerce

Fig. 1. Tile location and distribution for some of the mentioned classes in Lyon metropolitan area.

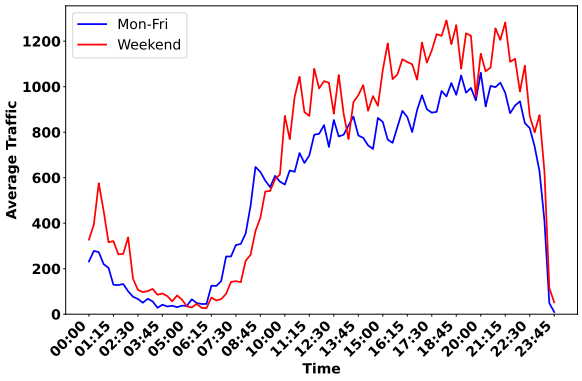


Fig. 2. The average uplink traffic during weekdays and weekends for the *Education Centers* tiles class.

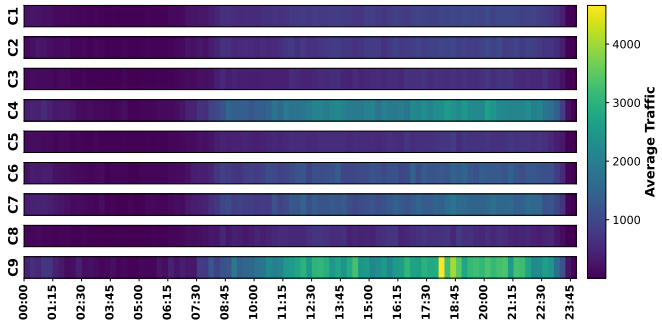


Fig. 3. The average uplink traffic during weekdays for all the classes.

somehow controlled times every day. This kind of behavior can create certain underlying traffic patterns. In the other hand, the *Commerce* class contain very distinct types of business, with distinct client profiles with different daily routines. The clients might go shopping whenever they feel like it, with no specific schedule or planning. It is also worth mentioning that spatial factors might influence user behavior, and the *Commerce* class has significantly more tiles, which are distributed in a greater area, as shown in Fig. 1(d).

III. Conclusion and Future Work

This work analyzed network traffic from messaging apps to identify user texting behavior. We defined classes based on

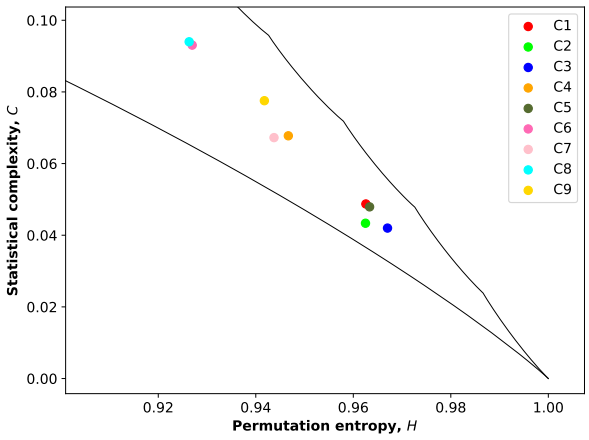


Fig. 4. Complexity-Entropy Causality Plane for the tiles classes.

specific locations in the Lyon metropolitan area and computed an average network traffic signature for each class using WhatsApp uplink traffic. The *Train Station* class exhibited the highest traffic volume, and we also identified a peak-time traffic at around 18:00. It is interesting to observe that the morning commuting rush hour does not generate a peak of messaging traffic. Transportation authorities can use this information to imagine adapting evening trains and services to this intense messaging usage. By analyzing Entropy-Complexity features, we observed that the *Commerce* class demonstrated more randomness compared to the other classes. We intend to continue our efforts by investigating temporal features to identify texting behavior during specific time periods and exploring information theory quantifiers such as permutation entropy and statistical complexity.

References

[1] O. E. Martínez-Durive, S. Mishra, C. Ziemlicki, S. Rubrichi, Z. Smoreda, and M. Fiore, "The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography," 2023.

[2] Data Grand Lyon, https://data.grandlyon.com/portail/fr/accueil, last visited on September 18, 2023.

[3] C. G. Freitas, O. A. Rosso, and A. L. Aquino, "Mapping network traffic dynamics in the complexity-entropy plane," in *2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2020, pp. 1–6.