



HAL
open science

Content consumption cartography of the paris urban region using cellular probe data

Sahar Hoteit, Stefano Secci, Guy Pujolle, Zhuochao He, Cezary Ziemlicki, Zbigniew Smoreda, Carlo Ratti

► To cite this version:

Sahar Hoteit, Stefano Secci, Guy Pujolle, Zhuochao He, Cezary Ziemlicki, et al.. Content consumption cartography of the paris urban region using cellular probe data. URBANE 2012, Dec 2012, Nice, France. 10.1145/2413236.2413246 . hal-01131516

HAL Id: hal-01131516

<https://hal.science/hal-01131516v1>

Submitted on 20 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Content Consumption Cartography of the Paris Urban Region using Cellular Probe Data

Sahar Hoteit,^{*}
Stefano Secci,
Guy Pujolle
LIP6, UPMC
4 place Jussieu, 75005 Paris,
France
firstname.lastname@lip6.fr

Zhuochao He,
Cezary Ziemlicki,
Zbigniew Smoreda
SENSe, Orange labs
38 rue du G. Leclerc, 92794
Issy-les-Moulineaux, France
firstname.lastname@orange.com

Carlo Ratti
Senseable City lab, MIT
292 Main st., 02139
Cambridge, MA, USA
ratti@mit.edu

ABSTRACT

A present issue in the evolution of mobile cellular networks is determining whether, how and where to deploy adaptive content and cloud distribution solutions at base station and back-hauling network level. In order to answer these questions, in this paper we document the content consumption in Orange cellular network for Paris metropolitan area. From spatial and application-level extensive analysis of real data, we numerically and statistically quantify the geographical distribution of content consumption with per-service classifications. We provide experimental statistical distributions usable for further research in the area.

Keywords

Mobile Internet, Urban Cloud, User Mobility

1. INTRODUCTION

Nowadays, the rapid adoption of communication technologies gives us a precious opportunity to study human behaviors and mobility patterns in an efficient, timely and cost-effective manner [1]. The pervasiveness of information and communication technologies offers an ideal vehicle to study both individuals and organizations: people habitually carry their mobile phones and use them as a medium for most of their communications.

In recent years, mobile data-based research reaches important conclusions about various aspects of human characteristics, such as human mobility and calling patterns [1] [2], virus spreading [3] [4], social network properties [5] [6] [7] [8], urban and transport planning dynamics [9] [10], network design [11]. At its first days, the Internet was conceived to serving fix and sedentary usages, while current socio-technology trends clearly show that future Internet users will be increasingly mobile and nomadic. For example, as of Institut Mediametrie ([12]), more than 31% of French mobile

^{*}Work done during Sahar Hoteit's visit at Senseable City lab, MIT, USA.

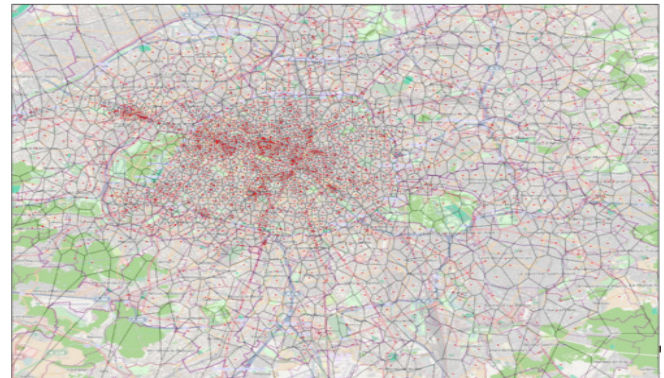


Figure 1: Voronoi Cell Distribution

users use a smartphone, and about 71% of them daily connect to the Internet. This recent trend seems to accelerate; for example, the number of mobile Internet users increased by 10% from the first to the second quarter of 2011.

On the other hand, mobile equipments are frequently used in everyday life for rather simple digital services, yet requiring increasing download volumes and resiliency levels [5]. At present, the rapid pace at which this evolution is taking place, and the often-inadequate management of broadband access networks, practically manifests with poor service availability, which represents a major bottleneck to the development of advanced services. The exponential growth of mobile internet usage is a relevant indicator of the overall bandwidth provisioning needed at the access networks, which have to be geographically and temporally distributed. Nevertheless, the dynamics of content consumption is very little known today, also because of the insufficient coordination between traffic engineering procedures and user mobility detection in nowadays telecommunication networks.

It is quite well known that the networks of many cell-phone operators recently collapsed after the release of mobile equipment software updates [13]. Moreover, cell-phone networks generally suffer during special events aggregating large masses of individuals sharing similar interests (e.g., sport events, conferences, city-wide cultural events, mass manifestations), hence accessing similar content in the same time and in the same places.

In order to better characterize these phenomena, in this paper, we analyze user content consumption habits, as a function of time and place, to determine digital consumption hotspots in the network. The purpose is to provide useful real-measurement data for the re-

search community. We focus on the Paris urban and peri-urban regions, and on two days, a normal day and a special day period when particular content consumption is expected, the one corresponding to the Eurocup Italy-Spain final.

The paper is organized as follows. Section 2 synthetically presents the dataset. Section 3 presents content consumption maps and analyzes them from a spatial distribution perspective. Section 4 focuses instead on application consumption patterns. Section 5 summarizes the findings, and Section 6 draws some perspectives and discusses possible future work in the area.

2. DATASET

The dataset used in our study consists of network probe's data, generated each time a mobile device uses wireless mobile network for Internet data exchange (not for voice calls and SMS, i.e., what is commonly referred to as "mobile Internet" service). The probe is able to distinguish the protocol used for the communication, so it is possible to categorize the traffic by application (Web, VoIP, P2P, streaming etc.). All user identifiers were irreversibly anonymized by Orange Labs before analysis. The probe collects data with six minutes interval sessions, assigning the session to the cell identifier of the last used antenna.

The data are recorded on per user basis and cover more than 1,5 million of French mobile phone users in the Parisian region "Ile-de-France", giving about 100 millions of records per day. We limit the study in the paper to the Paris Inner Ring "La Petite Couronne" metropolitan area, formed by three Ile-de-France peri-urban departments (92, 93, 94) bordering with the French capital urban department (75) and forming a geographical crown around it.

The region is decomposed as shown in Fig.1 at base station level, where red dots represent the base stations and the surrounding polygons represent the Voronoi cells. The size of a Voronoi cell depends on two basic factors: the geographical position and the coverage area (determined according to power level) of the corresponding base station.

As already mentioned in the introduction, we analyze these data for two days, one normal day with no particular expectable content consumption activity, June 25th, 2012, and one special day where a particular content consumption is expected, July 1st, 2012, the day of the Italy-Spain Eurocup football final. In the following analysis, we differentiate between the normal day and the particular period during and immediately after the football match in the special day.

3. SPATIAL DISTRIBUTIONS

User habits may be inferred through content-based analysis. We present in this section, content consumption habits, as a function of time and place, for users in the inner ring of Paris region (departments 75, 92, 93, 94), on two days, a normal day and the special event day.

Looking for difference between urban and peri-urban region is a first naturally arising research question. Does urban region attract more users during the day and especially during a special event? Does the traffic generated from urban base stations account more than from peri-urban base stations? To answer these questions, we plot in Fig. 2 and Fig. 3 the density of people expressed by the total number of people presented in a given cell to the time of observation on the two days, and the traffic density expressed by the exchanged byte volume of the base station to the time of observation.

The main observations from these figures are:

- The user density and the traffic density volume in peri-urban regions seem to be higher than those of the urban region.

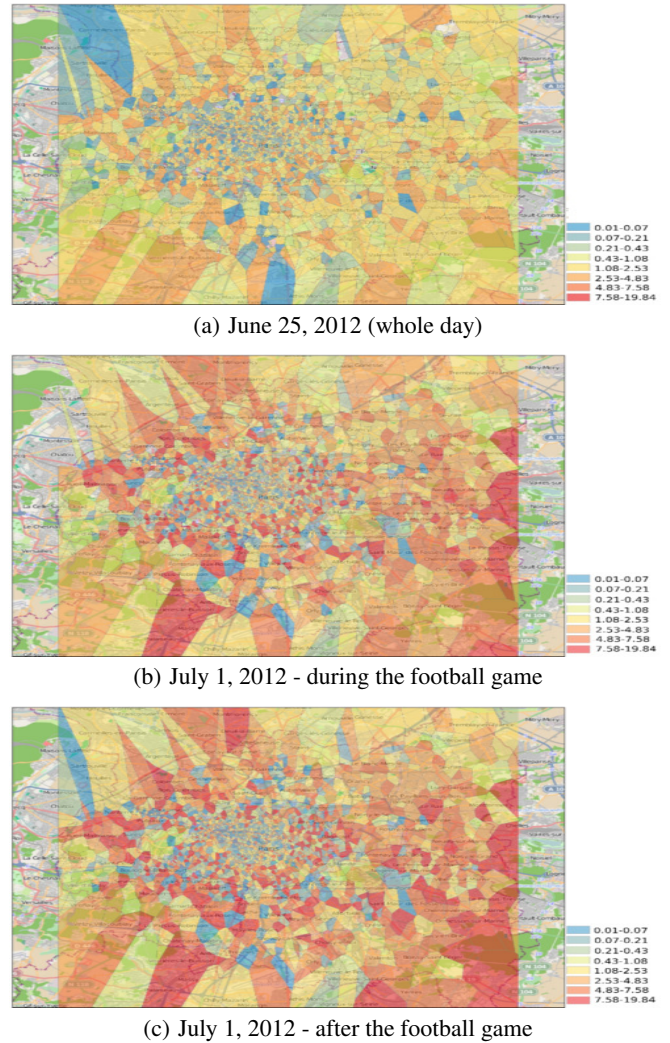


Figure 2: User density (nb. of users per minute)

This is rather counter-intuitive in that cities should attract larger people volumes, and can be partially explained by the fact that urban region is covered by a large number of base stations as we can see in Fig. 1. In fact, we have a large number of small-area cells inside the urban region (75) and a low number of wide-area cells in the peri-urban regions.

- Some cells are clearly more overloaded than others, presenting a large number of users and a large traffic volume; the reason is that they cover clearly identifiable content consumption hotspots.
- During and just after the football game a high congestion and a big number of users appear in some cells inside the capital as well as outside it. These crowded places are very likely some public spaces where people use to gather together.

In order to further explore this latter aspect and more precisely study the density distributions, Fig. 4 and Fig. 5 plot the experimental cumulative distribution functions (CDFs) of traffic and user density in the two considered days¹. We can easily notice that:

¹The provided experimental CDFs do not fit with any common

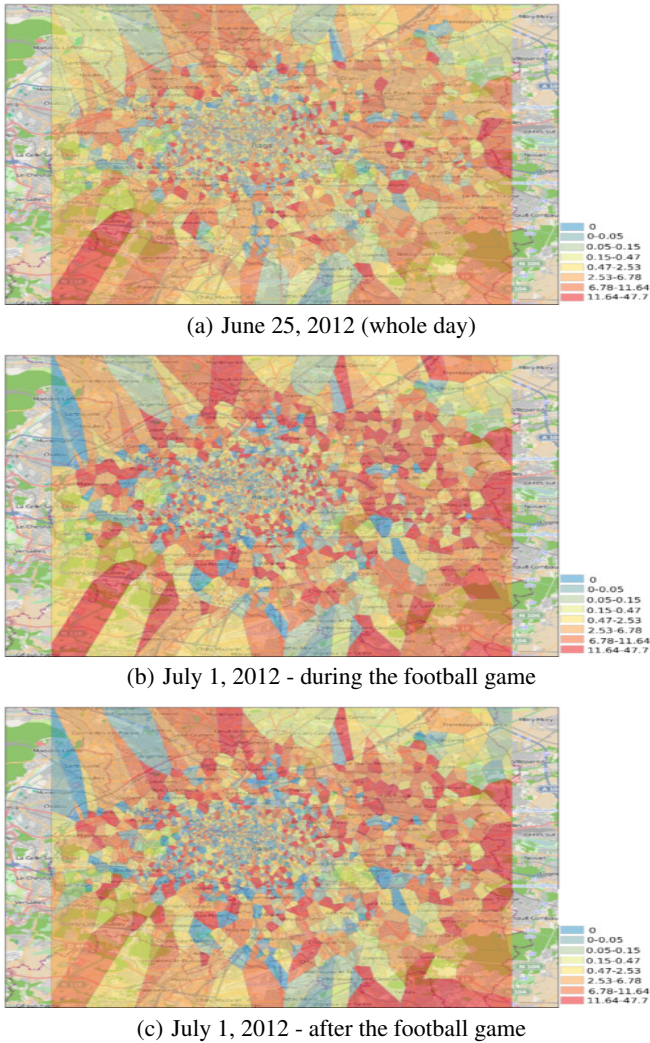


Figure 3: Traffic density (MB/minute)

- In terms of traffic density, the game period shows a higher number of cells with high density traffic than the normal day (e.g., the cells with a traffic density above 10 are 27% during the game period and 17% in the normal day);
- In terms of user density, during the normal day, dense zones show a user density three times lower than during the game (e.g., the cells with a density higher than 5 are 10% in the normal day, and 30% during the game).

These two aspects confirm the results obtained before and clearly show that users are much more concentrated and gathered during the special event. This fact leads to the creation of different spatial hotspots in the region.

Moreover, during the event opposite behaviors are exalted, i.e., users usually not surfing much, surf even less, and does usually surfing more, surf even more. This latter aspect seems being confirmed by the differential density distribution (subtracting the 25th June' density to the 1st July's one around the game) in Fig. 6 and Fig. 7. We can observe that user density difference distribution

distribution (power-law, truncated power-law, weibull), but can be oversampled for simulation purposes.

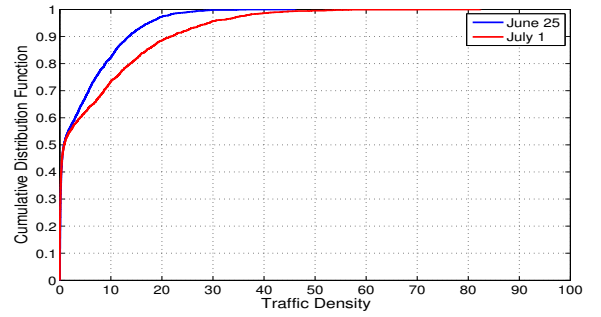


Figure 4: Traffic Density CDF

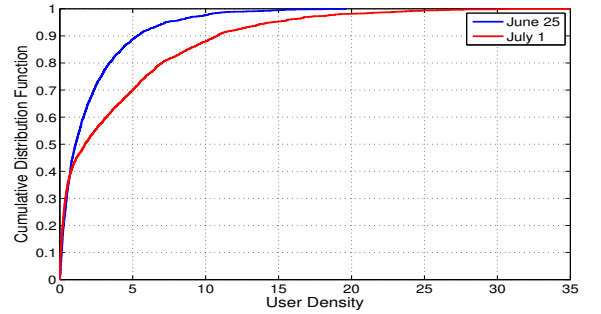


Figure 5: User Density CDF

roughly follows a normal distribution, and that traffic density difference distribution is dominated by positive differences (higher densities in normal days).

The special event is therefore impacting the cellular user distribution introducing relevant impairments that shall be considered by back-hauling network management and traffic engineering procedures, opportunely placing content and Cloud server surrogates, splitting traffic, dimensioning links toward content hotspots taking into consideration traffic and user distributions.

4. APPLICATION USAGE

In this section we characterize the applications in content hotspots during the sportive event, so as to provide a basic traffic model for these situations and draw some observations. For content consumption hotspots, Fig. 8 reports the proportion of usage of each application (i.e., the number of users connected to each application to the total number of users), while Fig. 9 represents the proportion of traffic generated from each application (i.e., the number of bytes generated from each application to the total number of bytes)². The application classification is based on network-level and transport-level protocol header information. We can notice that:

- Web applications are the most used ones in terms of number of clients during the match, attracting around 50% of the usages. This happens as people increasingly use social networks (e.g., Twitter, Facebook), using HTTP, to express reactions and impressions on the event. However the traffic generated from these applications is not so big (e.g., 26% of the total traffic) because they do not require much bandwidth.

²It is worth to note that if a user is connected to two different applications at the same time, he will be counted in both applications.

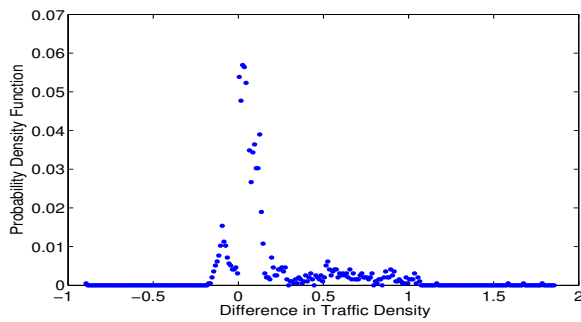


Figure 6: Probability density function of the differential traffic density (normal day - special day).

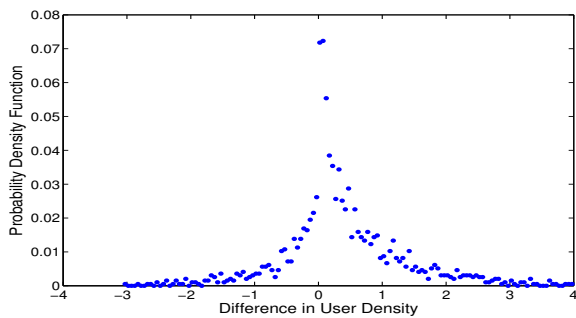


Figure 7: Probability density function of the differential user density (normal day - special day)

- The trend toward social communications is also confirmed by the large usage of chat and messaging applications covering together around 17% of the usages. However, as seen in Fig.9, this sort of applications has a negligible contribution in the total traffic since their required bandwidth is very limited.
- A third class of application is the one of bulk transfers, file and mail download applications (e.g., relying on the POP, SMTP, FTP protocols), which cover around 26% of the usages and contribute in 24% of the total traffic.
- Real-time applications, video streaming, gaming and VoIP, occupy a small proportion, around 7%, likely because these interactive applications are hardly compatible with the attention that is devoted to the event but in the other hand, they are the most bandwidth consuming applications (i.e., contributing in 50% of the total traffic) as seen in Fig.9.

Another way to interpret these data is to classify them with respect to the transport-level protocol only, i.e., TCP (Transmission Control Protocol) and UDP (User Datagram Protocol), used for applications needing or not, respectively, flow control and packet re-transmissions upon loss, so roughly corresponding to non-interactive and interactive real-time services. It is worth noting that video streaming here mostly falls into the first category as web streaming is nowadays mostly based on HTTP. Fig. 10 and Fig. 11 show the distribution difference between UDP and TCP, in Paris city (department 75, restriction due to computation time issues). Fig. 12 and Fig. 13 show the corresponding CDFs. From their analysis, we can evidence that:

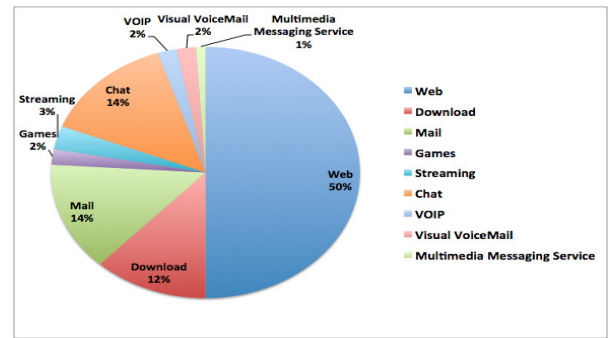


Figure 8: Application usage ratios

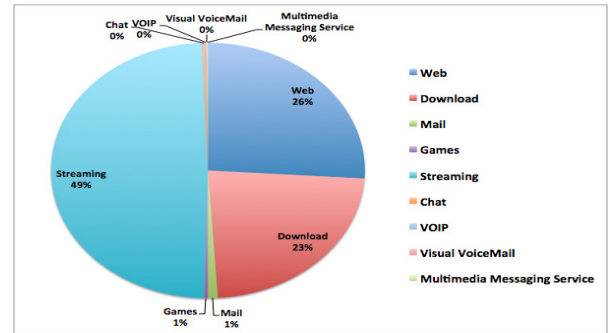


Figure 9: Application traffic ratios

- in terms of traffic volume, TCP applications are much more used and bandwidth harvesting than UDP ones;
- in terms of relative location of UDP vs. TCP users, we have an almost perfect spatial match between TCP and UDP hotspots.
- the statistical experimental distributions confirms these statements, with coinciding user density CDFs and distant traffic density CDFs (e.g., for traffic density over 1, 2% of cells are based on UDP protocol whereas more than 22% cells are based on TCP).

As evidenced, with a traffic engineering perspective, it would not make much sense to implement different traffic management solutions separating TCP from UDP traffic, while it would rather make sense to aggregate using categories with similar requirements and content consumption volumes, as for instant the web traffic, messaging applications, bulk transfers and real-time applications.

5. CONCLUSION

Mobile network operators are currently assisting to a very rapid evolution of services and their usages. "Cellphones now used more for data than for calls" titled the New York Times already on May 13th, 2010: indeed, people use the mobile phone more and more to connect to the Internet, but smartphones also frequently exchange data with distant applications automatically (e-mail checks, synchronization, etc.). Actually, bandwidth-greedy applications and the increased mobility of data-plan users are stressing the management of cell-phone access networks. At present, mobile phones provide the best means of gathering information about user movements and content consumption behavior on a large scale, indeed. However little is publicly known about how content is consumed.



(a) July 1, 2012 - TCP - bytes

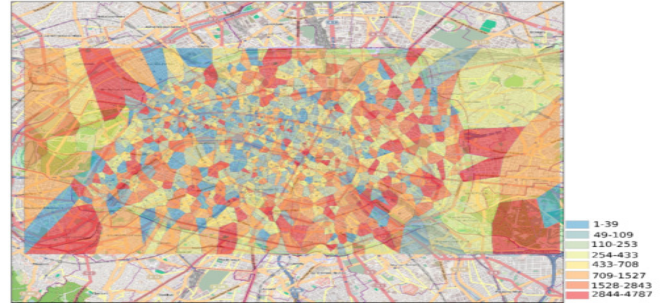


(b) July 1, 2012 - UDP - bytes

Figure 10: Spatial distribution of TCP/UDP usages - MB



(a) July 1, 2012 - TCP - clients



(b) July 1, 2012 - UDP - clients

Figure 11: Spatial distribution of TCP/UDP usages - Number of clients

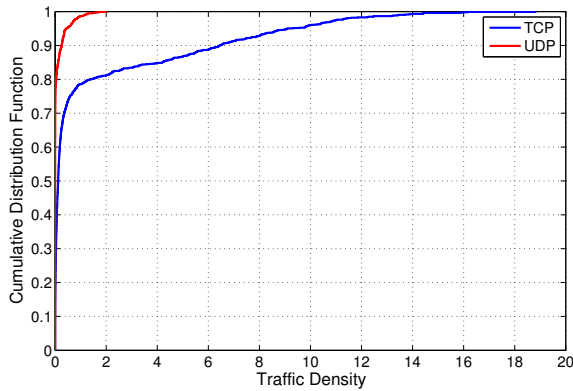


Figure 12: Traffic Density CDF: TCP vs UDP

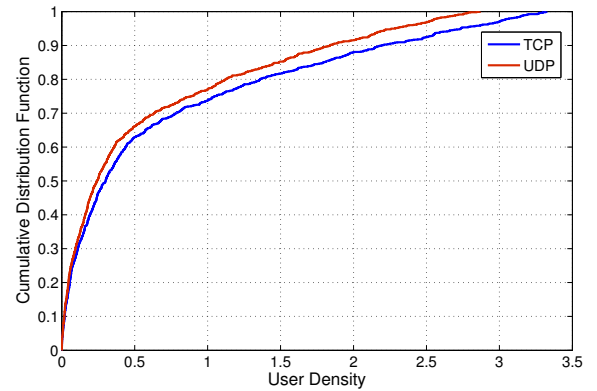


Figure 13: User Density CDF: TCP vs UDP

In this paper, we studied content consumption habits, as a function of time and place, for users in the inner ring of the Paris metropolitan area on two periods, a normal day and a special one corresponding to the Eurocup final Italy-Spain. We determine digital consumption hotspots, characterized their distribution, and further investigated the content consumption in those spots.

We provide experimental statistical distributions usable for further research in the area and for network modeling. It is particularly evinced that around half of the users use Web applications during such special events, that another relevant part uses messaging applications, and that real time applications are limited to less than 10% of the usage. On the other hand, real time applications generate most of the traffic against negligible part for chat and msg apps.

Moreover, TCP-based applications largely overcome UDP ones in absolute volumes (even if there is no apparent correlation between TCP and UDP applications and geographical location).

6. TOWARDS FOLLOW-ME CLOUDS

In the light of these results, it looks reasonable to dynamically take into account content consumption dynamic cartographies, ideally joint with user mobility patterns, to distribute content and services to the users. On one hand, crowd-driven as well as application-driven traffic shall be detected online so as to dynamically move relevant content close to access gateways, where specialized equipment can be used, also using advanced content caching architectures.

On the other hand, Cloud computing services also require distributed information management: these services enable remote processing and storage of personal data, i.e., of resources hosted out of the user terminal. For mobile users, Cloud applications are expected to significantly grow and stress communication networks since mobile equipment has notably limited computing and energy resources.

As reported in this paper, it is possible to capture user content consumption habits, as a function of time and place, to determine consumption hotspots. Further work is needed to define scalable online detection algorithms and to conceive mobility-aware and usage-aware traffic engineering procedures taking into account the user mobility, hence adaptively moving content and computing resources close to where large masses of individuals commonly pass.

For instance, the first iPhone software updates, being executed almost synchronously, have caused bandwidth outages in many cell-phone networks worldwide. Dynamically placing the download server, close to the denser access points, upon prediction and detection of massive people displacements, could have allowed a seamless service provisioning to the users, without performance degradation across the network. If also data processing could be performed closer to the denser access points (e.g., an itinerary computation over a map service), i.e., moving Cloud computing functionalities toward the access network, the service resiliency could further be enhanced (with lower connection latency and data loss). This is the main motivation behind this new arising paradigm one could coin "Follow-Me Cloud" for cellular network design, which is currently pushing toward novel mobile back-hauling traffic engineering and resource allocation procedures, we plan to address in future works.

7. REFERENCES

- [1] M. Gonzalez, CA . Hidalgo, Al. Barabasi "Understanding individual human mobility patterns", *Nature* 458, pp. 238-238, 2008.
- [2] H. Hohwald, E. Frias-Martinez, and N. Oliver "User modeling for telecommunication applications: Experiences and practical implications", in *Proc. UMAP*, pp. 327-338, 2010.
- [3] R. Huerta, L. Tsimring "Contact tracing and epidemics control in social networks", *Physical Review E* 66, 2002.
- [4] P. Wang et al., "Understanding the spreading patterns of mobile phone viruses", *Science* 324, pp. 1071-1076, 2009.
- [5] F. Calabrese et al., "The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events", in *Proc. of IEEE PerComp 2010*.
- [6] M. Turner, S. Love, M. Howell, "Understanding emotions experienced when using a mobile phone in public: The social usability of mobile (cellular) telephones", *Telemat. Inf.* 25:3, pp. 201-215, 2008.
- [7] R.C. Nickerson, H. Isaac, B. Mak "A multi-national study of attitudes about mobile phone use in social settings", *Int. J. Mob. Commun.* 6:5, 541-563, 2008.
- [8] M. Naaman, J. Boase, C.-H. Lai "Is it really about me?: Message content in social awareness streams", in *Proc. of ACM CSW 2010*.
- [9] M. R. Vieira et al., "Characterizing dense urban areas from mobile phonecall data: Discovery and social dynamics", in *Proc. IEEE SocialCom 2010*.
- [10] H. Wang et al., "Transportation mode inference from anonymized and aggregated mobile phone call detail records", in *Proc. IEEE ITSC 2010*.
- [11] H. Zang, J. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks", in *Proc. of ACM MOBICOM 2007*.
- [12] "L'audience de l'Internet mobile en France", Mediametrie, <http://www.mediametrie.fr/internet/communiques/telecharger.php?f=e8c0653fea13f91bf3c48159f7c24f78>, 2011.
- [13] J. Giles "Smartphone use makes cellular networks? collapse a real possibility", *The Washington Post*, <http://www.washingtonpost.com/wpdyn/content/article/2010/11/29/AR2010112904854.html>, 2010.