

Understanding User Behavior via Mobile Data Analysis

Eyuphan Bulut
Cisco Systems
2300 President George Bush Highway
Richardson, TX 75082
ebulut@cisco.com

Boleslaw K. Szymanski
RPI, 110 8th St. Troy, NY 12180, and
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
szymansk@cs.rpi.edu

Abstract—Recently, there has been a tremendous increase in demand for mobile data driven by the wide-spread usage of smartphone like devices. Besides the efforts of studies that propose ways for operators to handle this huge growth, there is also an emerging research area of analyzing this new type of user data. In this paper, we perform an analysis on mobile user data to understand general aggregated user behavior based on several parameters. We also investigate the reasons for the user behavior deviating from its general trend. Finally, we discuss several ways of utilizing the results of the analysis in this study in solving current problems operators face.

I. INTRODUCTION

Demand for mobile data is growing tremendously due to increasing usage of various mobile devices (i.e., smartphones, laptops, tablet PCs) among users. People are using these devices for different activities such as web browsing, video and music downloads and photo sharing. As Cisco recently announced [1] in the Cisco VNI Global Mobile Data Traffic Forecast, global mobile data traffic will grow 13 fold from 2012 to 2017, at a compound annual growth rate (CAGR) of 66 percent, reaching 11.2 exabytes per month by 2017. Most of the traffic will be generated by smartphones, followed by laptops, tablets and M2M communications (Figure 1). In terms of traffic type, mobile video will represent 66% of all traffic which will probably be impacted by the widespread usage of high speed 4G connections that will account for 45% of global mobile data traffic in 2017.

Facing this mobile data explosion, operators are searching for the ways of improving their architectures and radio access technologies to handle this growth without any problem. Switching to Long Term Evaluation (LTE), adding more radio towers or small-scale base stations such as femtocells or picocells and utilizing WiFi access points (AP) for offloading the burden on their network [2], [3], [4], [5] are some of the solutions operators are currently working on.

Apart from the efforts to manage this huge mobile data growth, there is also an emerging research area of analyzing

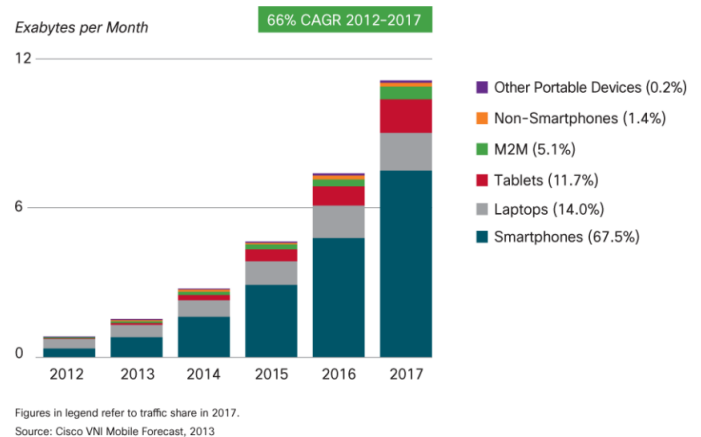


Fig. 1. Mobile data usage growth per device type [1].

and understanding this growing mobile data: its trending, anomalies etc. In this paper, our main focus is on investigating what we can learn about the user behaviors by analysing mobile data traffic and how these findings can be utilized. To this end, we analyzed aggregated user data on the gateways (i.e. PGW for LTE, and GGSN for 3G) of mobile packet core.

There are some previous work which study different aspects of cellular data traffic characterization. In [13], the individual 3G subscribers behavior in a large scale data set is analyzed and the relation between traffic load and mobility is characterized. In addition to spatial and temporal dynamics, in [14], the impact of device and application types on the traffic patterns is also studied and a Zipf-like model is proposed. Another analysis based on packet level, flow level and session level traffic is conducted in [15]. Unlike these works, we try to understand the events (like holidays or sports events that cause people to share their experience with families and friends) that cause deviations from general user behavior trends and discuss its practical implications.

The rest of the paper is organized as follows. The background on mobile core is given in Section II. In Section III, we introduce the data analyzed. In Section IV, we present the results of data analysis. In Section V, we discuss the paper contributions and outline the future work. We close with conclusion in Section VI.

¹This work was supported in part by the Army Research Laboratory under Cooperative Agreement Numbers W911NF-06-3-0001 and W911NF-09-2-0053 and by the EUs 7FP Grant Agreement No. 316097 and by the Polish National Science Centre, the decision no. DEC-2013/09/B/ST6/02317. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies either expressed or implied of the Army Research Laboratory or the U.S. Government.

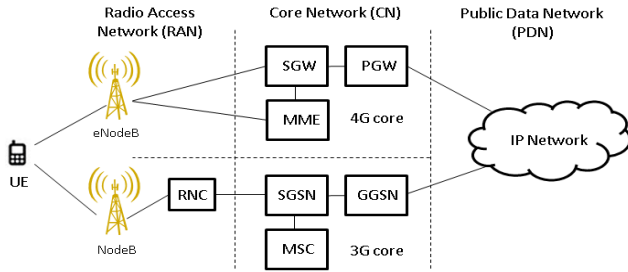


Fig. 2. 3G/4G cellular network architecture.

II. BACKGROUND ON MOBILE PACKET CORE

In Figure 2, the overall architecture of 3G Universal Mobile Telecommunications System (UMTS) and 4G Long Term Evolution (LTE) cellular networks is illustrated. A User Equipment (UE) communicates with a base station through a Radio Access Network (RAN) wirelessly. A RAN consists of a set of base stations which are called NodeBs in UMTS, and eNodeBs in LTE (4G). After the base station receives the data from a UE, it forwards the data towards the core network (CN). In LTE, main core network elements are Serving Gateway (SGW), PGW (Packet Data Network Gateway) and MME (Mobility Management Entity). SGW in LTE acts as a mobility anchor for UE handovers between eNodeBs, and forwards packets to PGW, which serves as a gateway router to IP networks which are called public data networks (PDNs). MME is in charge of many control-plane tasks such as paging and tracking. In 3G UMTS core, Serving GPRS Support Node (SGSN) takes the role of SGW and MME, while Gateway GPRS Support Node (GGSN) acts as PGW in LTE. MSC is responsible for routing circuit switch voice calls and SMS.

III. MOBILE DATA TRACES

We analyzed aggregated user data traffic from both 3G and 4G gateways in the mobile packet core. The data is collected not only from gateways in different regions/markets of a single mobile phone operator but also from gateways of different operators in different parts of the world. The data do not reveal any operator specific information and it only shows aggregated behavior of a ‘user group’ whose data pass over the same gateway, which often means that it is usually from the same geographic region¹ Our goal of analyzing such a wide range of data traffic is to see similarities and differences in different geographic regions, in different time frames and for different operators. We also look at the deviations from regular behavior and try to understand the reasons for them.

IV. ANALYSIS OF MOBILE DATA

In this section, we compare the characteristics of aggregated data usage of different user groups during different time periods (hours, days etc.). To this end, we do comparisons in terms of average data rate (kbps) per session (user), total number of sessions (on a specific gateway), and total throughput (sum of data rates of all sessions on the gateway) in a specific

¹Thus, a single user changes the user group when moving from the range of one gateway to the other. Since human mobility is highly predictable [17], then such defined user groups are actually very stable.

TABLE I. STATISTICS FOR AGGREGATED MOBILE TRAFFIC

Statistic	Value
Average daily data rate	[2.47 - 7.48] kbps
Max busy hour data rate	[3.67 - 9.79] kbps
Minimum data rate	[0.63 - 3.01] kbps
Max/Average ratio	[1.30-1.55]
Max/Min ratio	[3.24 - 6.41]
Standard deviation	[1.00 - 2.38] kbps
Correlation coefficient	[0.908 - 0.983]
Weekend/Weekday ratio	[0.99 - 1.07]
Busy hour	4:00 pm-5:00 pm, 8:00 pm-9:00 pm
Minimum usage time frame	4:00 pm - 5:00 pm

time frame. For example, if the time frame of interest starts at time t_1 and ends at t_2 , we compute the average data rate per session during this time frame as:

$$\Delta\tau = \frac{\tau(t_2) - \tau(t_1)}{(t_2 - t_1)(N_2 - N_1)}$$

where, $\tau(t_i)$ shows the total user throughput passed over the gateway until time t_i , and N_i is the number of sessions passed through the gateway until t_i .

Note that the aggregated data usage (i.e., total throughput) might increase i) through the number of subscribers/sessions on the gateway, and ii) by the increase of the average data rate per subscriber. Thus, we also analyze the change in session count to show that even if the average data rate stays similar for a user group, the total session count might increase in some specific cases.

A. Daily Usage

In Figure 3, we show the average data rate per session for eight different user groups (UGs) in the last 2-3 months of 2013. The hours plotted on x-axis are the local times of user groups and the y-axis shows the average data rate per session in the given hour in kbps. We plot three different lines; (i) average data rate per session for all days of the week, (ii) average data rate per session in the weekdays, and (iii) average data rate per session at weekends.

First of all, even though the data usage patterns differ slightly, we see very similar per session data usage rates within a day for different user groups. Looking at the overall general pattern, during the work hours (i.e., 9:00 a.m. to 5:00 pm) we see a quite similar usage, with a slowly increasing trend. After work hours, the usage starts increasing and it usually reaches its maximum around 8:00 or 9:00 pm². However, in some user groups busy hour arises a bit earlier (i.e. 5:00 pm). After the busy hour of the day, the data usage rate starts to decrease until 4:00-5:00 a.m. in the morning, where it reaches its minimum. The usage then starts to increase with a high slope until it reaches its daytime period.

Table I shows the statistics from the traffic patterns of these eight user groups in Figure 3. Average, minimum and maximum data rates for different user groups show large variation. However, when we compare the ratio of max data rate in the busy hour of the day to the average data rate in a day, we see somewhat smaller variation. The maximum data rate is [1.3-1.55] times higher than the average data rate in a

²The value at 8:00 pm. shows the average result between 8:00 pm and 9:00 pm.

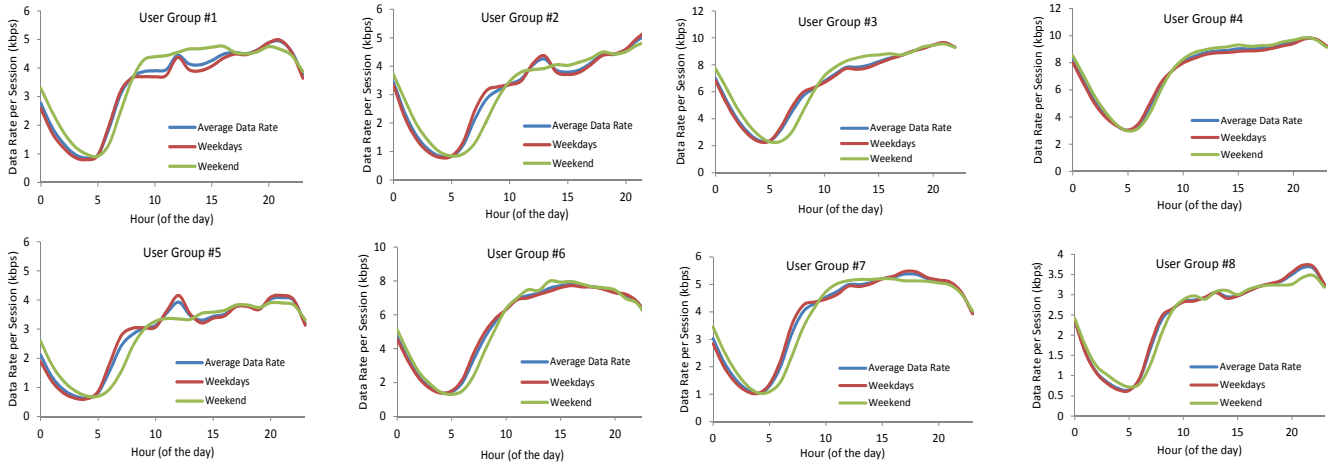


Fig. 3. Daily mobile data usage in different regions during the last two months of 2013.

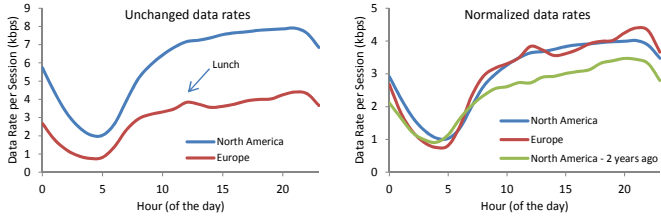


Fig. 4. Comparison of North America and Europe sites (user groups) based on unchanged (left) and normalized (right) average data rates.

day. When we exclude user groups with max and min ratios (UG#2 and UG#6), this range gets even narrower, [1.38-1.44]. Thus, we believe this quite stable ratio could be used as one of the important invariant parameters for traffic pattern analysis and predictions. Moreover, even though there are big differences between the exact data rates between different user groups, as the correlation coefficient matrix in Table II shows, there is a high correlation (in the range [0.908, 0.983]) between the data rate trends of each pair of user groups.

When we look at the average data rates in weekdays and weekends, we also notice some expected behaviors. At weekends, we see that the start of increasing trend between 4:00 a.m.-10:00 a.m. is one to two hours delayed. This is probably because on Fridays and Saturdays people go to bed late compared to other days and subsequently wake up late on Saturday and Sundays. Therefore, the increase in the data usage starts with a delay compared to weekdays. Moreover, even though it is not shown in graphs, this delay is much larger on Sundays than the delay on Saturdays, which is probably due to the fact that some people work on Saturdays so they get up on that day as early as on weekdays.

It is also important to note that since some people go out Friday and Saturday nights, we observe higher data usage at weekends after midnight period (see data usage between midnight and 4:00 a.m. at weekends).

In weekdays, another expected data rate change compared to weekends is at lunch times. As it is clearly seen in some of the graphs in Figure 3, there is a noticeable increase in data

TABLE II. CORRELATION COEFFICIENT MATRIX

Statistic	UG1	UG2	UG3	UG4	UG5	UG6	UG7	UG8
UG1	1							
UG2	0.943	1						
UG3	0.993	0.944	1					
UG4	0.957	0.936	0.953	1				
UG5	0.941	0.988	0.936	0.954	1			
UG6	0.914	0.978	0.909	0.953	0.985	1		
UG7	0.985	0.916	0.978	0.981	0.924	0.908	1	
UG8	0.971	0.989	0.968	0.957	0.982	0.972	0.951	1

rate around noon time. This could be because people might find more time to access Internet with their phones during lunch breaks which are their free times. On the other hand, such a behavior is not observed at weekends in any user group.

Comparing the total data usage at weekends and weekdays, we see that at weekends people may use up to 7% more mobile data than their usage on weekdays. This might be also expected since people have more free time at weekends. However, some of the free times are spent for extra sleep, additional activities etc. (which are not usually done on weekdays), therefore, the increase is not very pronounced.

We also grouped these user groups as North American (3,4,6,7) and European (1,2,5,8) carrier based user groups. As Figure 4 shows, even though the exact data rates in North America are almost double of the data rates of European user groups (this is affected by the current radio technology used), when we normalize³ the average data rates in both groups, we see similar data usage trends in a day. The clear exceptions are the lunch times and busy hour time periods for European user groups.

Another interesting observation we made is the comparison of data rates of North American users today and two years ago. As it is illustrated in Figure 4, general trend for both normalized North American data seems similar (correlation coefficient is 0.971). Moreover, even though the average data rates have increased compared to two years ago, the max/average data rate ratio is very similar. Thus, in addition to the high stability of this parameter for different users from different regions, we

³The data rate of North American user groups is multiplied by the ratio of average user data rates in Europe and North America.

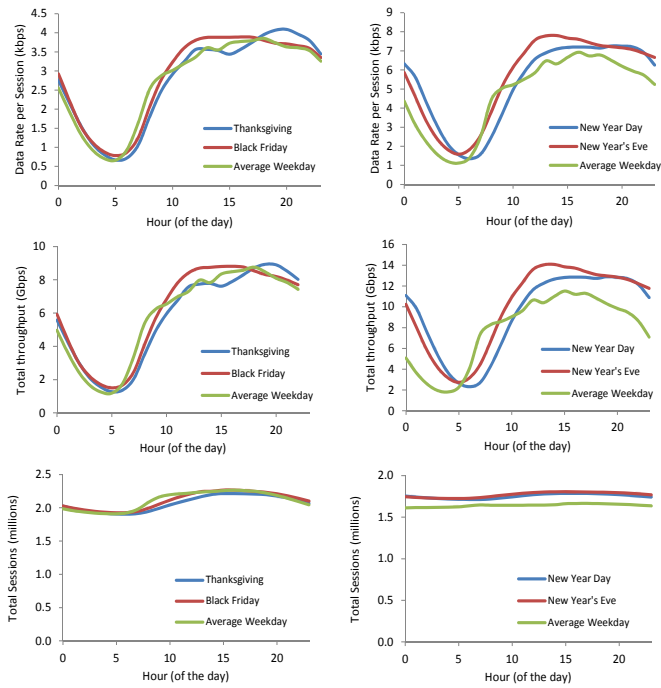


Fig. 5. Changes in data rate, session count and total throughput on Thanksgiving and New Year days.

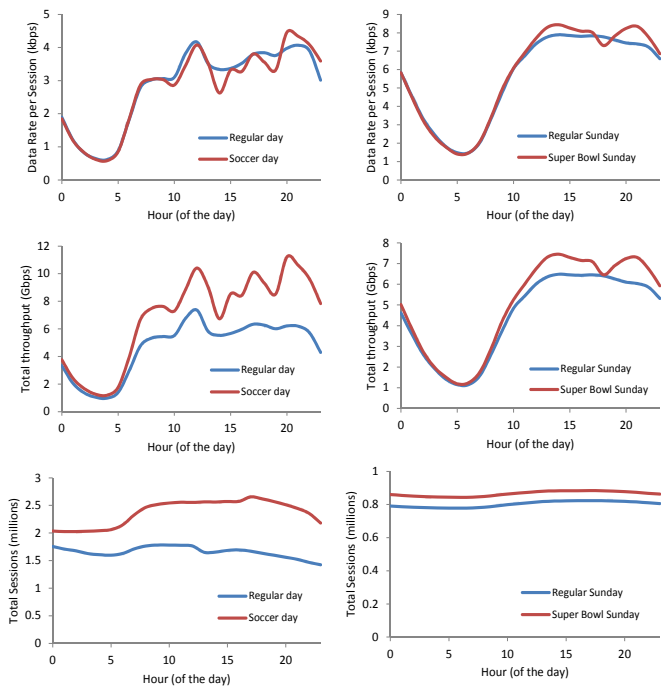


Fig. 6. Impact of an important soccer game and Superbowl game on data rate, session count and total throughput.

observe that it also varies little in different time frames. As a result, it could be used as a good invariant parameter for future network traffic design.

B. Holidays

To understand the deviation in data usage trends, we first checked the usage on some holidays. The left graphs on

Figure 5 shows the data rate, session count and total throughput information on the Thanksgiving Day and following day (Black Friday) compared to regular average weekday usage. The first observation is, the number of sessions does not change too much in Thanksgiving period. However, the data rate per session, thus the total throughput, shows different behavior. Since on Thanksgiving day in US, there is a football game in the afternoon, the data usage right at game start time is lower compared to surrounding hours. However, as the game continues, at later hours, the data usage increases compared to regular weekday. Moreover, at the night of Thanksgiving day, since some of the shops open their gates for Black Friday deals around midnight, the increase in data usage at late hours of the day might be explained by high data usage with more awake people. Additionally, family gatherings for Thanksgiving dinners might have triggered this behavior too. On the Black Friday day, the data usage during day time is also higher compared to regular weekday and Thanksgiving day usage. This is probably again due to people being more active motivated by the shopping for Black Friday deals, thus, causing people to use their phones more actively.

The right graphs on Figure 5 shows the same parameters on a world-wide holiday, the New Year Day, and the day before it compared to regular average weekday usage⁴. It is clear that overall data usage is remarkably higher on both days. More people are awake during night for celebrations, thus, there is higher data usage at nights. The data usage on the New Year's Eve is higher than the usage on the New Year Day, probably due to the fact that most people rest on the New Year Day. Note that the number of sessions from the same group of users does not increase too much but the data rate per session increases remarkably. This clearly is the result of more active usage of mobile devices on these days.

C. Special Days

To see possible behavior change in mobile data usage, we also plotted the usage on some special days. Specifically, we looked at the usage during an important soccer game in Europe and the data usage during Superbowl game in US. In Figure 6, the left graphs shows the usage changes on soccer day, while the right graphs show the changes on Superbowl day. Comparing the impacts of these two events on data usage trends, we observe different behaviors. The bottom graphs show that soccer game had caused remarkable session count increase (this can be due to more people in the service area of the gateway, or due to the increase in multiple session creation from the same group of users). In contrast, the Superbowl did not cause too much change in session count. However, when we look at the average data rates per session, we observe that data rates change visibly both for soccer game and Superbowl days, with different patterns however. There is only one slight decrease in the shape of 'V' in Superbowl corresponding to the start of the game. Similar 'V' for the soccer game is followed by several picks probably corresponding to post-goal or game break discussions among fans.

V. DISCUSSIONS AND FUTURE WORK

In this section, we discuss several ways of using the mobile data analysis results presented in the previous section.

⁴The statistics for left and right graphs are from different user groups.

A. New Pricing Mechanisms for Data Usage

At the very early stage of smartphone usages by subscribers, operators have offered unlimited data access opportunity for flat rate monthly fees. The subscribers were actively using their devices to access Internet without worrying about their bills. However, this brought some problems to operator networks. As the mobile data usage has increased, the burden on the networking architecture of operators has also risen. Therefore, besides the physical solutions (such as femtocell deployments) operators are working on, one critical challenge operators have started to propose to their subscribers is tiered pricing on the data usages. This type of pricing not only provides operators with control over the usage on their networks but also their revenue can increase. Operators might have the risk of losing some of their customers to their rivals, but the fact is the data usage limits offered by operators do match with what most (i.e., 98%) of their customers need [8]. Moreover, some operators noticed that only a small portion (0.1%) of their customers generates around 30% of their total traffic. With the tiered and limited pricing, they may not satisfy such customers, but they can relieve the burden on their network.

Even though this type of pricing mechanism offers a relief on an operator's network, we can still see the same patterns on the mobile data usage during a day. An elegant extension of this could be adding the time parameter to the tiered pricing mechanism. For example, since the overall data usage (so does the burden on the operator's network) is low during night, the operator might encourage the data usage during that period by setting a lower price per unit (Megabit) than it is in busiest time of the day. Even though this may not help increase the number of users active during the night due to its inconvenience, some periodic data usage such as the automatic updates of applications and their information could be scheduled to be done at these times, thus, users still might get benefit from lower prices during that period.

In some recent studies, different pricing mechanisms have already been studied. In [11], a time dependent pricing scheme is proposed where the users are encouraged for data usage in non-peak durations. It has been shown that both the operators and customers can benefit from the proposed approach since the temporal fluctuation of data usage demand could flatten and users can save money by selecting the time and usage volume. In [10], the workload history of the basestations is also analyzed to differentiate the prices geographically. A survey of different pricing mechanisms can be found in [12]. Even though several different pricing algorithms are proposed, none of the previous work proposes a mechanism for time frames where the data usage deviates from regular behavior. In our future work, we plan to study the pricing mechanisms for such cases, since we believe there is a great potential for the benefit to both the users and the operators.

B. Efficient Usage of Network Resources

The data rates of users change drastically throughout a day. The ratios of maximum per session data rate to minimum data rate shown in the graphs in Figure 3 are between 3.24 and 6.41. Moreover, the standard deviation of data rate during the day is in the range from 1.0 to 2.38. This huge variation of data rate

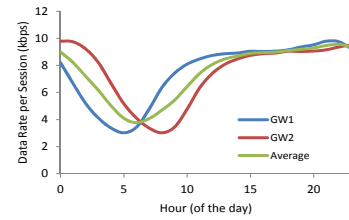


Fig. 7. Distributing the sessions of two different gateways may result in better utilization of resources.

in a day also triggers an inefficient and unbalanced usage of network resources (gateways etc.). While the user behavior can be altered to some extent with ideas like tiered pricing and low prices during non-busy durations, it cannot be changed drastically. However, some carriers operating in countries with multiple timezones (like US) can leverage this information to their benefit⁵. Consider the two lines with legends 'GW1' and 'GW2' in Figure 7. They show the daily data rate of users who are at locations with 3 hour time difference. Their data rate trends are similar to each other with respect to their local times. However, when we consider the data rates at the same global time, there is a difference (i.e. 3 hour shifted behavior). If such a carrier can direct the half of the traffic of GW1 to GW2 and half of traffic of GW2 to GW1, both should see a normalized data rate pattern similar to line with legend 'Average'. Note that directing traffic between gateways is possible with fiber optic cables. It may bring some extra work, however, the resulting benefit may make it worth performing this extra work. Moreover, the benefit of this distribution activity can be more pronounced if the timezone between gateways increases.

C. Dynamic Advertisement Pricing

With the increase of mobile data usage, we progress from voice-dominated to data-dominated mobile Internet. As a result of this change, the revenue generated per byte of data (and hence per unit wireless resource) reduces significantly [6].

Another way to monetize the outputs of the data analysis in this study could be to have the operators or the other involved companies (application developers etc.) adjusting the schedule and pricing of the advertisements that are presented to the users in their devices. The advertisement fee can be defined according to user availability and the data usage at the current time frame. For example, a day can be divided into different periods according to the data usage, and the advertisement fee can be charged accordingly, having the highest fee at the busiest time and lowest fee at the time of least usage. A similar adjustment can also be made for different days of a week, for holidays and for some special events/days.

D. Efficient Social Content Caching Scheme

The mobile users make requests, of which the most bandwidth consuming and least delay tolerant are requests for internet content, including videos. To save cost of uploading all requested data, the WiFi access points and the Internet

⁵Note that it is also possible that operators from different regions of the world with large time zone differences can collaborate and benefit from the proposed approach.

routers/gateways can use content-aware caching schemes. Research on social networks indicates that users with social ties have the correlated mobility patterns and the correlated interests in requested content. In [9], the authors demonstrate that the correlations between mobility and between interests can be used to improve efficiency of content-based caching network caching. It is achieved by taking into account social ties between users visiting a WiFi access point or a gateway router. The authors created a content-aware caching scheme and demonstrated that by using information about frequency of visit by members of different social communities combined with topic tags on the content requested by each community, the performance of caching can be increased remarkably compared to its performance with random caching.

E. Split Billing

A novel idea of splitting the bill [16] allows content providers to pay for the data usage by mobile users visiting their websites or using their services. Even though this creates an additional revenue for carriers, it also forces them to update the capacities of their networks. On the other hand, users can benefit from this opportunity since part of their bills will be paid by the service providers.

We believe that one way of flattening the data usage rates in a day could be achieved via dynamic split billing policy. In other words, when the data usage is low in the system (i.e., night period), a high percentage of user bills could be paid by service providers (to encourage night time (non-peak) duration usage) while only a small percentage of their bills are paid by service providers in the high data usage times. This is another way of applying dynamic pricing mechanism, however, the control of user reflected pricing is now determined by service providers. Carriers should work with service providers in such a way that both their networks will be relieved and both parties will benefit from this feature.

VI. CONCLUSION

In this paper, we analyzed the mobile data usage of different user groups in different geographical regions and different time frames. We have observed that there are interesting daily trends in the data usage. Moreover, some specific events and holidays have significant impact on the data usage. We also made some recommendations to operators about using the results of the analysis in this study. The summary of our analysis of the results is as follows:

- General aggregated user behavior is predictable. The average correlation coefficient between the daily data usage of all pairs of user groups is 0.95.

- Max/average data rate ratio is pretty stable for different regions and time frames and could be used as a good invariant parameter.
- There are some events/days that can cause deviations from general usage trends and create a good potential for monetizing these deviations.

REFERENCES

- [1] Cisco Visual Networking Index, http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf, February, 2013.
- [2] E. Bulut, and B. Szymanski, *WiFi Access Point Deployment for Efficient Mobile Data Offloading*, in Proceedings of ACM International Workshop on Practical Issues and Applications in Next Generation Wireless Networks (PINGEN 2012) at Mobicom 2012., Istanbul, Turkey, August 26, 2012, pp. 45-50.
- [3] K. Lee, I. Rhee, J. Lee, S. Chong and Y. Yi *Mobile data offloading: how much can WiFi deliver?*, in Proceedings of the 6th ACM International Conference, ser. Co-NEXT 10, pp. 26:126:12, New York, NY, 2010.
- [4] S. Dimatteo, P. Hui, B. Han and V. O. K. Li, *Cellular Traffic Offloading through WiFi Networks*, in Proceedings of 8th IEEE Intl Conf. Mobile Ad Hoc and Sensor Systems (IEEE MASS 11), 2011.
- [5] B. Han, P. Hui, and A. Srinivasan, *Mobile data offloading in metropolitan area networks*, SIGMOBILE Mob. Comput. Commun. Rev., vol. 14, pp. 2830, November 2010.
- [6] R. Kokku, R. Mahindra, S. Rangarajan, H. Zhang, *Opportunistic alignment of advertisement delivery with cellular basestation overloads*, in Proceedings of MobiSys'11, pp. 267-280, 2011.
- [7] V. Chandrasekhar, J. Andrews, and A. Gatherer, *Femtocell networks: a survey*, Communications Magazine, IEEE, 2008.
- [8] Rethinking Mobile Data Pricing, <http://www.freshfields.com/publications/pdfs/2011/jan11/29081.pdf>, 2011.
- [9] B. O. Holzbauer, B. K. Szymanski, and E. Bulut, *Impact of Socially Based Demand on the Efficiency of Caching Strategy*, in Proceedings of IEEE PerCom 2014, IQ2S workshop, Budapest, Hungary, 2014.
- [10] A. G. M. Damasceno, R. A. de F. Mini, and H. T. Marques-Neto, *Base Station Congestion-Dependent Pricing Scheme for Cellular Data Network*, CoNEXT 2013 Student Workshop, 2013.
- [11] S. Ha, S. Sen, C. Joe-Wong, Y. IM, and M. Chiang, *TUBE: Time-dependent pricing for mobile data*, in Proc. of ACM SIGCOMM, 2012.
- [12] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, *A survey of smart data pricing: Past proposals, current plans, and future trends*, ACM Computing Surveys (CSUR), vol. 46(2), November 2013.
- [13] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, *Understanding traffic dynamics in cellular data networks*, in Proceedings of IEEE Infocom, page 882-890, 2011.
- [14] M. Shaq, L. Ji, and A. Liu, *Characterizing and modeling internet traffic dynamics of cellular devices*, In Proc. of ACM SIGMETRICS, 2011.
- [15] Y. Zhang, and . Arvidsson, *Understanding the Characteristics of Cellular Data Traffic*, in Proc. of ACM SIGCOMM CellNet Workshop, 2012.
- [16] H. Raj, S. Saroiu, A. Wolmen, and J. Padhye, *Splitting the bill for mobile data with SIMlets*, in Proceedings of HotMobile Workshop on Mobile Computing Systems and Applications, 2013.
- [17] C. Song, Z. Qu, N. Blumm, A. L. Barabasi, *Limits of predictability in human mobility*, Science 327(5968):10181021, 2010.