Analysis of Call Detail Records of International Voice Traffic in Mobile Networks

Zagroz Aziz
Czech Technical University in Prague
Department of Telecommunications
Prague, Czech Republic
azizzagr@fel.cvut.cz

Robert Bestak
Czech Technical University in Prague
Department of Telecommunications
Prague, Czech Republic
robert.bestak@fel.cvut.cz

***Abstract*—** Over the last decade, mobile network technologies and services show a progressive growth in term of customers, amount of transmitted and generated data, both user and operational data. The operation data represents a valuable information source for mobile operators as they can be further processed and data mining algorithms applied to extract enhanced (value-added) information. In this paper, we focus on Call Detail Records of international incoming and outgoing mobile phone calls. We analyze traffic profile for different time periods (day, week, month, and year), and we discuss major factors affecting the traffic. The observed results show a long-term traffic stability and a periodicity that reflects human behavior and activities.

***Index Terms*—Big data, Data Mining, Mobile Network, Call Detail Record, International Call Analysis**

I. Introduction

The mobile network sector is one of the fastest growing sectors in telecommunication [1]. By implementing new technologies, introducing new services along with growing number of subscribers and by increasing interest in online gaming, social networking and video streaming, the amount of user and operational data are rapidly growing. Thus, the data storage and processing become more and more complex for mobile operators [2].

Beside the storing/processing data issues, retrieval of useful information itself and its proper interpretation represents another big challenge for operators. That is way nowadays Telco companies dedicate an important budget to hire data analysts and specialists in this field to extract from the data as much as possible [3].

In our paper, we deal with Call Detail Records (CDR). These operational data have a crucial importance as they characterize connections in mobile networks by providing metadata about these connections [4]. Every CDR file contains information such as a phone call scenario (connection time, release time, duration, date, called party identification, etc.), information about incoming and outgoing call-legs, and addresses of the switching systems at both sides of connections or identification of the connection type.

A CDR is created for a single call ([2]-[4]), no matter if it's a local or an international call. Providing details about an entire phone call connection makes the call easily traceable, especially for the reason of network statistics or system routine health check (drops, silent calls, etc.). The aim of this paper is to analyze a short/long-term traffic evolution. The paper focuses on international call scenarios by analyzing CDRs of these calls in terms of daily, weekly, monthly and yearly traffic profile. Statistics of working and weekend days per year is presented as well. Finally, we analyze what are original and destination countries of these calls and how the total amount of traffic is distributed among these countries.

This sort of analysis is usually quite sensitive to the analyzed time period as the investigated period should be carefully chosen based on network, and/or customer aspects. For example, selecting an ordinary working day versus weekends, holidays or a month free of holidays, events versus a month with many holidays, events, etc.

Additionally, when analyzing and interpreting data in telecommunication networks, there are usually terminology issues as various Telco companies, and technologies apply different terminology when presenting their outcomes. Though, they all refer to the same meaning. For instance, the terminologies used in a Global System for Mobile Communication (GSM) are different to the terminology used in a Voice over Internet Protocol (VoIP) network [5], e.g., termination/origination vs. incoming/outgoing, or using INVITE message in SIP vs. channel request in TDM to setup a call.

The rest of this paper is organized as follows. Section 2 provides an overview of related works. Section 3 details features of the analyzed CDR set. Section 4 discusses obtained results. Finally, Section 5 concludes our findings.

II. Related Work

Recently, there have been done a lot of works in the area of exploitation of mobile data, for different purposes, including network optimization or network statistics.

In [6], authors employ measurements generated by smartphones, in a Long-Term Evolution (LTE) network, to allocate radio resources and to adapt parameters of radio link. There are presented two approaches: i) usage of drive test and ii) employing UE reports. The study shows that the drive test approach does not provide good results as majority UEs are located in indoor environment.

In [7], authors present crime data analysis by using CDR. Authors propose to use a graph analysis tool in a crime records investigation.

In [8], a random matrix theory data model together with the machine learning concept are investigated. Authors describe an architectural framework using big data analytics in mobile networks to reduce amount of data and filter out the pointless data that is out of scope and interest.

Mobile network traffic prediction by employing spectral estimation techniques as an alternate method for Time Serious Database (TSDB) is analyzed in [9]. The paper shows the benefits of method in isolating the structures of signal that delivers useful information.

In [10], a novel approach is proposed to estimate location of mobile users by using an enhanced Kalman filter together with mobility models. The estimated user's locations are then compared with locations obtained via GPS.

In paper [11], authors discuss and compares two approaches for analytic: i) Parallel Database Management System (DBMS), i.e. a SW for database management, and ii) MapReduce, i.e. a simple programming model used to process large-scale data in parallel and distributed architecture. The paper demonstrates that the Parallel DBMS approach provides better results in querying and predefined processes scenarios, while the MapReduce approach offers more flexibility in case of time consumption process.

Authors of [12] introduce characteristics of mobile network behavior that is based on big data generated in telecommunication networks to obtain important insights using big data analysis. The authors classify the data set based on user-oriented and network-oriented. At the end, they present two case studies; i) using Erlang measurement and ii) Call Detail Record to study the Base Station behavior.

In [13], authors present a framework to analyze a wide-ranging set of CDRs to define mobile call profile categories and to classify the network usages accordingly. To evaluate the proposed framework, 5-month period, consisting of 300 million call records in an urban area is considered in the analysis.

In comparison of previous works, we use CDRs of mobile international calls, and we analyze traffic profiles how the calls are distributed among countries.

III. Data processing and features

In general, a telephony network architecture, including a mobile network, consists of several network elements, such as transmission, switching, supervision, or billing system. A network is interconnected to national and international networks, located abroad (Fig. 1). Interconnection of these networks are typically done by using Internet Protocol (IP) to deliver the voice traffic, no matter if Session Initiation Protocol (SIP) or Time Division Multiplexing (TDM) approach is used.

Operational data concerning local/international phone calls are collected via CDRs and store in the CDR database. The CDRs are firstly imported using a CDR mediation server in their raw formats as they are recorded by network switching system. Then, the mediation server parses, stores and formats the CDRs based on requirements of network system (e.g., Billing system, Least Cost Routing system, Authentication system, etc.).

In our study, we use one-year CDR data set (July 2016 to June 2017). The total amount of investigated CDRs corresponds to 39 million phone calls (incoming and outgoing). The traffic, presented in our graphs, is calculated on hourly basis, i.e. the number of calls is counted in 60-minute window period. At first, we extract the information by transforming the raw data into a dataset. Next, we filter out data attributes that are appropriate for our study. Finally, the dataset is rearranged by adding missing entries and dropping the unnecessarily data.

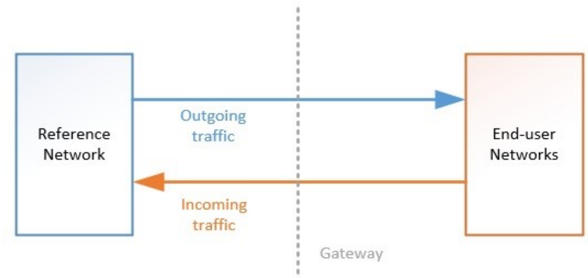

Fig. 1. Telephony network interconnection scheme.

The CDR set represents international traffic of incoming and outgoing calls of geographical region, which represents several countries, where the incoming/outgoing calls from the perspective of reference network is considered as shown in Fig. 1.

IV. Results and discussions

This section describes obtained results. Fig. 2 shows a daily evolution of outgoing call traffic. The daily profile analysis is done for 4 consecutive Wednesdays in October (5, 12, 19, and 26. 10. 2016). We have picked up for the daily traffic analysis Wednesday as it is a typical working day in the middle of a working week; in the studied geographical region, and in majority regions all over the world. Similarly, October represents an ordinary month, without any specific holidays or events.

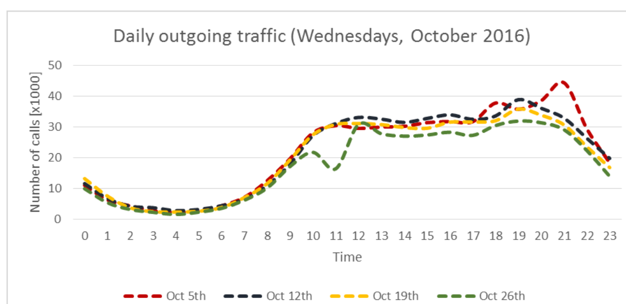

Fig. 2. Daily outgoing traffic (Wednesdays, October 2016).

The number of calls (y-axis) is counted on minute basis, i.e. the number of calls that begins in a given minute. If a call takes n minutes, the call is counted only once, at the starting minute.

As can be observed, the traffic follows the daily human behavior and activities. During the night and early morning, from 0:00 to 05:00, the users' activities are very low, as majority of people are sleeping, and therefore the traffic is nearly zero. From 05:00, the traffic progressively starts to grow, as people wake up and start their daily activities, such as go to works, schools, etc. The traffic keeps growing until it reaches the peak, which occurs on average around 11:30.

In mid of the day and forward, the traffic slightly goes down due to lunches/afternoon break times. Afternoon, the traffic is more or less constant, with slightly changes, until reaching evening.

The traffic again slightly starts to grow around 18:00 and reaches the peak on average averagely at about 19:30. This traffic increase reflects calling of people to their friends and relatives, who live abroad. From that point, the traffic starts to decrease till almost zero after the midnight.

There are two obvious abnormalities seen in the graph. The first one, October 5th (drawn in red color, in Fig. 2), reflects higher calling activities of people in the evening than usual. The second abnormality, October 26th (which is green color in Fig. 2), is due to technical issues faced at the reference network around 11:00. This applies also on incoming traffic abnormalities for those two mentioned specific hours (Fig. 3).

Otherwise, as can be observed from the figure, there are typically two peaks in the daily traffic; midday and evening peak. The midday peak occurs around noon (due to people working obligations), and the evening peak is in the evening (which is more related to people personal activities).

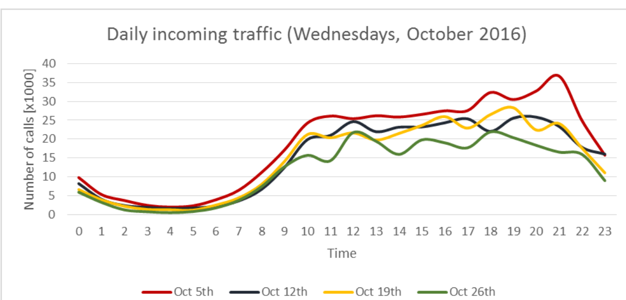

Fig. 3. Daily incoming traffic (Wednesdays, October 2016).

The incoming traffic profile is very similar to the outgoing traffic as shown in Fig. 3. During the night the traffic is nearly none. From 05:00, the traffic starts to grow and keeps growing till reaches the peak on average at about 12:00. Since that moment, the traffic is more or less stable until the evening, at about 19:00, then it again starts to decrease.

By comparing Fig. 2 and Fig. 3, we can notice a higher spread in the incoming traffic curves, comparing to outgoing traffic curves. We analyzed the data set and no specific network technical issues, events, or policies were observed during the analyzed period. Thus, we assume this spread is due to the fact that the incoming traffic comes from all around the world, i.e. the pool of potential users is much bigger, comparing to users of reference network and their generated (outgoing) traffic.

As can be observed from Fig. 2 and Fig. 3, the daily profiles follow very similar pattern, no matter which Wednesday, or other day it is.

In both outgoing and incoming traffic, the season plays an important role in the midday and evening's peaks. These are shifting according to the season as well as daylight saving time and country of origin. The midday peak is slightly affected, but the evening peak usually occur around 19:00-20:30, while in winter the peak is around 17:30-18:30. Fig. 4 illustrates the weekly traffic, from Monday to Sunday. The analyzed week is the first week of May, 2017, that can be considered a regular week of the month without any abnormal behavior of people. The number of calls (y-axis) indicates the unique number of calls in a given day. In case a call takes place over a midnight, the call is counted into the day when the call begins.

The traffic is more or less constant from Tuesday to Thursday, which are typical working days in all regions of the world. The Friday traffic drop is due to the official day off in the studied region (in the studied geographical region, the weekends are Fridays and Saturdays).

On Saturday, traffic begins to grow and keeps increasing till reach the maximum on Tuesday. The reason of the lower Monday traffic comparing to Tuesday-Thursday traffic is a fact that Monday is the first working day of week in many countries around the world, and people are just back in their work after the weekend.

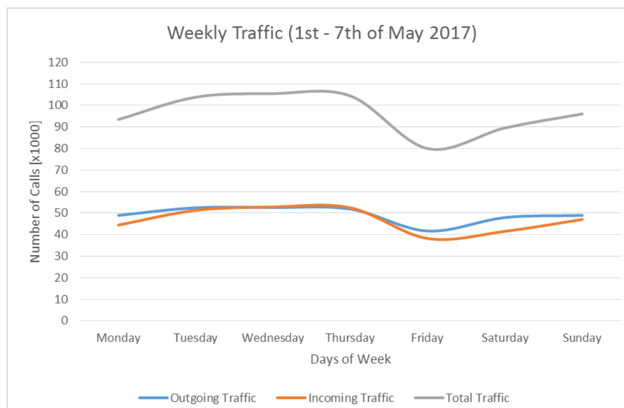

Fig. 4. Weekly traffic (1st – 7th of May).

As can be seen from Fig. 4, the incoming and outgoing weekly traffic profiles are almost the same. Both incoming and outgoing depend on working days and days-off, origin of calling and called party. People they typically respect each other's timing zone when calling each other.

Fig. 5 shows the monthly traffic, where we again have selected May 2017 for the analysis.

Obviously, the monthly traffic consists of the weekly traffic repetition, and the traffic is higher during the weekdays and lower during the weekends. Similarly, to the weekly profile, both incoming and outgoing monthly traffic is nearly same.

The first two weeks show higher amount of calls comparing to the third and fourth week. This is due to the fact that May is the last academic month (in the studied geographical region) followed by summer holiday. Usually during summer holidays, less traffic is expected because of no school period, many people having vacations and their off-days.

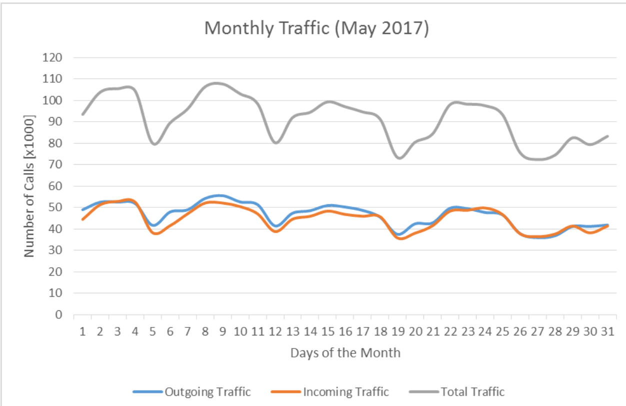

Fig. 5. Monthly traffic (May 2017).

Finally, Fig. 6 illustrates yearly traffic, from July 2016 to June 2017.

During the first four months, from July 1st to November 30th, i.e. days 1-123 in Fig. 6, the incoming traffic is lower than the outgoing. This is due to company's policy and restrictions on the incoming traffic, from the End-user networks.

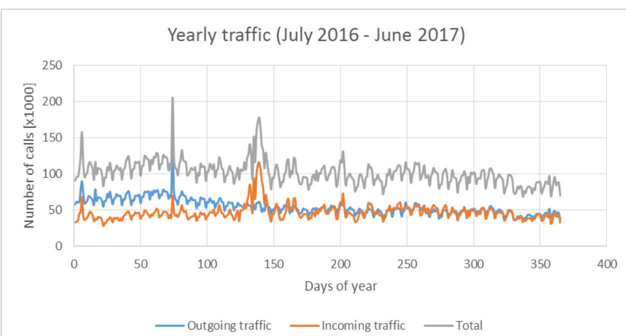

Fig. 6. Yearly traffic (July 2016-June 2017).

In Fig. 6, we can observe 3 peaks that match to specific events in the analyzed region. The 1st peak (day 6), reflects religious holiday in Islamic countries, celebrating the end of Ramadan, which usually results in high traffic during that time. The 2nd peak (day 74) represents a feast day in Islamic countries, where the traffic doubled comparing to a normal day. Finally, the 3rd peak (days 138-148) corresponds to a regional holiday where the incoming traffic is more affected than the outgoing traffic.

For such scenarios, we face rapid increment in traffic. It is sometimes expected especially for some predicted events. In contrast, there are some high traffic moments are out of our scope. For that, the network cannot handle such huge traffic, which leads to network congestion. In this case, these unpredicted incidents should be studied in order to avoid future network congestion.

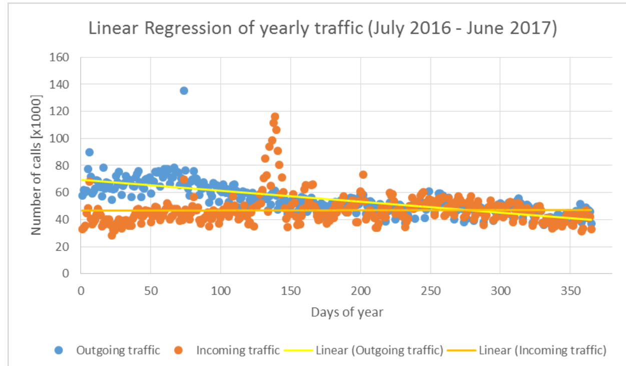

Fig. 7. Linear Regression of yearly traffic (July 2016-June 2017).

Fig. 7 depicts linear regression for the yearly traffic. The incoming traffic is quite stable with no changes, except the predicted peaks explained previously. In the other hand, the outgoing traffic show noticeable changes in the beginning and then gradually goes back to the level of incoming traffic. As already mentioned, the decrease is due the operator's policy.

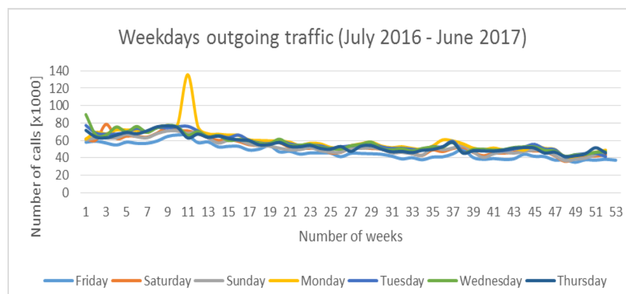

Fig. 8. Weekdays outgoing traffic during the year.

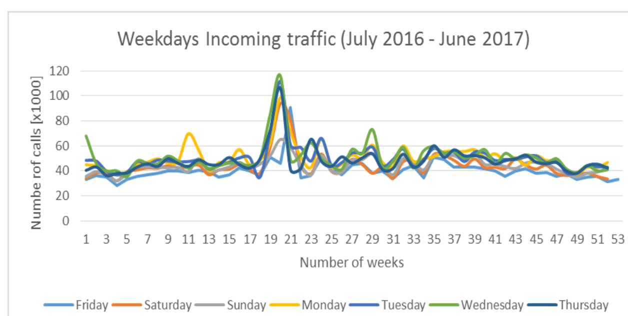

Fig. 9. Weekdays incoming traffic during the year.

The traffic changes during the whole year (July 2016-June 2017), for each day per week, is illustrated in Fig. 8 and Fig. 9. The figure describes how the traffic changes for 52x Saturdays to Thursdays, and 53x Fridays. On the figures, we can observe the abnormalities explained above (religious holidays, and the decrease of outgoing traffic through in the first part of the year).

The following table, Table 1, summarizes the total number of calls per year for each day (from Monday to Sunday). The mean (μ) and standard deviation (σ) for each days is provided as well.

TABLE I. STATISTICS OF CALLS PER YEAR.

Days	*Total calls per year*	*μ [calls]*	*σ [calls]*
Monday	5725454	108027	18898
Tuesday	5590481	105481	14609
Wednesday	5624193	106117	16934
Thursday	5401519	101915	13508
Friday	4625490	87273	11869
Saturday	5066378	95592	14301
Sunday	5065466	95575	10340

The lowest, resp. highest, μ is for Fridays (weekend day), resp. Monday (week day). In the table, we see Monday has the biggest number of calls among other days. This is because two of the major holidays at the studied geographical region occurred on Mondays, which have a quite important impact on number of calls. If we discard these exceptional calls due to the two events, Tuesdays and Wednesdays are days with the highest number of calls, which corresponds to our above discussion when describing the weekly traffic profile.

An example of noon/evening peaks movement for the incoming and outgoing traffic within a 3-month period (October-December 2016) is illustrated in Fig. 10 and Fig. 11. The peaks are calculated with 60-minute window, which is shifted minute by minute to find out the maximum values at noon and in the evening for each day. The points in Fig. 10 and Fig. 11 show the beginning of the max. 60-minute window.

Fig. 10. Peaks of outgoing traffic in case of October-December 2016

In case of the outgoing scenario (Fig. 10), the obtained result shows oscillations around 12:00h, resp. 19:15h, with a slow movement towards 13:00h, resp. 19:00h, by the end of December. Similar values can be observed for the incoming traffic (Fig. 11), initial oscillations occur around 12:15h, resp. 19:00h, with a slow movement towards 13:15h, resp. 19:15h, by the end of December.

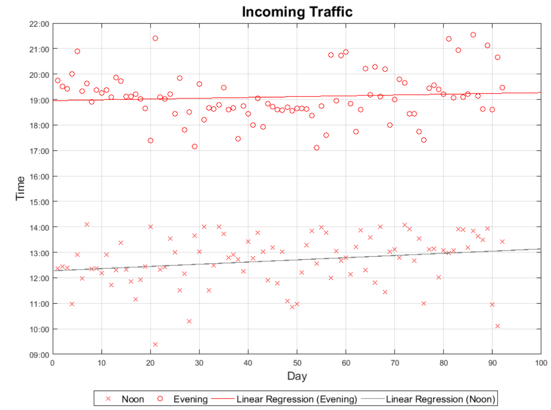

Fig. 11. Peaks of incoming traffic in case of October-December 2016

Apart of traffic profiles for different time periods, an operator needs to also know how the traffic is distributed among the end-user networks and what countries are with the highest traffic for incoming/outgoing. Having this knowledge, an operator can better plan for its network capacity.

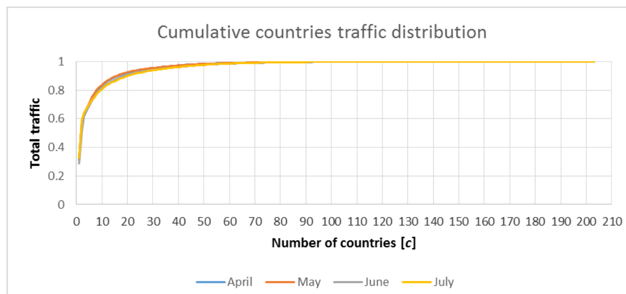

Fig. 12. Cumulative countries traffic distribution of April-July 2017.

Fig. 12 represents the traffic distribution among all interconnected networks, more precisely countries, for 4 months (April-July, 2017); similar shape of curves can be observed for all months per year. In our study, the reference network is interconnected in total to 204 countries. We arrange the countries based on their traffic, from the highest to lowest one, and we count the Cumulative Traffic Distribution function using (1):

$$CTD = \sum_{i=1}^{c} \frac{Traffic\ of\ country_i}{Total\ traffic\ of\ countries} \quad (1)$$

In Fig. 13, we show top 10 countries with the highest traffic. All 10 countries, and their order, are same for the analyzed 4 months. By analyzing the whole year, we observed that the given top 10 countries, and their order, is more or less same during all 12 months.

As can be seen, the top 10 countries in total take over 80% from the total amount of traffic, where about 60% represents the outgoing and 40% the incoming traffic. It means, there is 0.8 probability that an international call comes from one of the top 10 countries. On the other hand, majority of countries only show 1-2 calls per month (see Fig. 12).

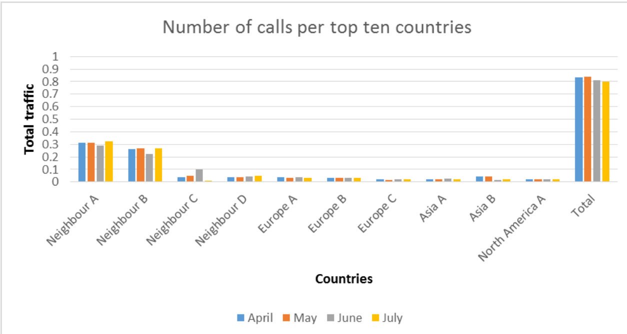

Fig. 13. Number of calls per top 10 countries, for April-July 2017.

As expected, the biggest impact on total number of calls are from neighboring countries. This is due to typically strong relationship among the neighbor countries in term of business, culture, trading, tourism, etc.

V. Conclusion

In the paper, we analyze one year CDR data set (July 2016-June 2017), representing testing international traffic of incoming and outgoing calls, for which we describe daily/weekly/monthly/yearly traffic profiles.

The studied area covers a geographical region, comprise of several countries, Statistic of working and weekend days per year is provided as well. Furthermore, we depict how the incoming/outgoing calls are distributed among end-user networks and what are countries with the highest traffic.

The obtained results show a long-term traffic stability and a daily/weekly traffic periodicity that reflects human activities. Additionally, as expected, we observe that major part of the incoming/outgoing traffic comes mainly from neighboring countries.

In our future work, by using the CDR data set, we would like to determine the distribution function of the duration of calls, time distribution between calls, apply queuing system approach to the analysis of waiting time, etc.

Acknowledgment

This research work was supported by the Grand Agency of the Czech Technical University in Prague, grand no. SGS18/181/OHK3/3T/13.

References

[1] Shui Yu, Song Guo (2016). Big Data Concepts, Theories, and Applications.

[2] Md Salik Parwez, Danda B. Rawat, Moses Garuba (2017). Big Data Analytics for User-Activity Analysis and User-Anomaly Detection in Mobile Wireless Network. IEEE Transactions on Industrial Informatics (Volume: 13, Issue: 4).

[3] Yojna Arora, Dinesh Goyal (2017). Big data: A review of analytics methods & techniques. IEEE, 2nd International Conference on Contemporary Computing and Informatics (IC3I).

[4] Gajendra Kumar Vaikar, Prateema Gautam (2016). Data Mining Method Use in Crime Investigation Network CDR Analysis. International Journal of Advances in Computer Science and Cloud Computing, Volume-4, Issue- 1, ISSN: 2321-4058.

[5] Nicola Chemello (2016). Correlating CDR with other data sources. IEEE International Conference on Cybercrime and Computer Forensic (ICCCF).

[6] Peter Chiu, Jussi Reunanen, Riku Luostari & Harri Holma (June 2017). Big Data Analytics for 4.9G and 5G Mobile Network Optimization. Vehicular Technology Conference (VTC Spring), 2017 IEEE 85th.

[7] Dr. Manish Kumar, Dr. M. Hanumanthappa (2017). Crime Investigation and Criminal Network Analysis Using Archive Call Detail Records. Advanced Computing (ICoAC), IEEE.

[8] Ying He, Fei Richard Yu, Nan Zhao, Hongxi Yin, Haipeng Yao, Robert C. Qiu (2016). Big Data Analytics in Mobile Cellular Networks. IEEE, pp. 1985-1996.

[9] Anish Kurien, Barend Jacobus Van Wyk, Yskander Hamam (2008). Mining Time Series Data in Mobile Cellular Networks. IEEE, Third International Conference on Broadband Communications, Information Technology & Biomedical Application.

[10] Artjom Lind, Amnir Hadachi, Oleg Batrashev (2017). A new approach for mobile positioning using the CDR data of cellular networks. IEEE, 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS).

[11] Sara B. Elagib, Aisha-Hassan A. Hashim, R. F. Olanrewaju (2015). CDR analysis using Big Data technology. IEEE, International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering.

[12] Sihai Zhang, Dandan Yin, Yanqin Zhang, Wuyang Zhou (2015). Computing on Base Station Behavior Using Erlang Measurement and Call Detail Record. IEEE Transactions on Emerging Topics in Computing Year: 2015, Volume: 3, Issue: 3, Pages: 444 – 453.

[13] Diala Naboulsi, Razvan Stanica, Marco Fiore (2014). Classifying Call Profiles in Large-scale Mobile Traffic datasets. IEEE Conference on Computer Communications (IEEE INFOCOM).