On the Analysis and Visualisation of Anonymised Call Detail Records

Varun J, Sunil H.S, Vasisht R
SJB Research Foundation
Uttarahalli Road, Kengeri,
Bangalore, India
Email: (varun.iyengari,hs.sunilrao,vasisht.raghavendra)@gmail.com

Srinidhi Saragur, Abhijit Lele
SJB Research Foundation
Uttarahalli Road, Kengeri,
Bangalore, India
Email: ssrinidhi@sjbrf.org, abhijit.mlele@gmail.com

motiv

***Abstract*— Global mobile traffic is estimated to grow at a compounded annual rate of 40%. In order to keep telecom operators profitable, network planning and optimisation, and personalized traffic plans are critical to success. In order to do this, telecom operators need to analyse the data generated from their telecom network and derive information and knowledge that will assist them in network planning and developing personalised traffic plans.**

With this as the background, in this paper we analyse the *Call Detail Records (CDR)* provided by *Orange Telecom* as a part of the *Data for Development (D4D)* challenge. The data analysis of the CDR records is carried out using *Hadoop* and *Hive* framework. This paper focuses on analysing the data from telecom network optimisation and socio-economic perspectives. We utilise a visualisation tool to represent the derived information in a human understandable format. Based on visual inspection of the derived knowledge from the CDRs, we provide our recommendations for network optimisation.

I. Introduction

Telecommunications operating companies generate a large number of data records from switching systems. Data items are produced for every telephone call through a telephone network. *Call Detail Record* (CDR) is the fingerprint of how many seconds and at what time a customer is using a telephone and the associated infrastructure in terms of *Base Stations* used to process the call. Therefore, CDR displays a map of telephone customers behavior. The knowledge of customer behaviour obtained by analyzing CDRs has multiple applications. In [1] the authors deduce social attributes from calling behaviour. Mobile customer clustering based on CDR records from the perspective of marketing campaigns is discussed in [2]. In [3], the authors propose a method to derive transportation patterns based on CDR records. Considerable research is being carried out not only on socially relevant data, but also patterns that can be used to enhance the overall service parameters of telecom operations.

While methods that analyse CDR records and derive knowledge from it are important, it is equally important to visualise data and the knowledge derived from it in a human understandable format. This gives a quick reference point to the stakeholders in a manner that they can relate to. In this paper we share our experiences in visualization of the knowledge derived from CDR records and to some extent provide our observations on and interpretation of the derived knowledge. The rest of the paper is organized as follows. Section II describes the CDR data set. The framework of *Hadoop* [4] and *Hive* [5] along with mapreduce [6] are discussed in Section III. Our approach to analysing the data sets to derive information from them and then visualising the data sets is discussed in Section IV. We briefly attempt to provide analysis of the knowledge derived from these data sets in Section V. We finally conclude with future work in Section VI.

II. CDR Data

In this paper we use the Orange Telecom [7] *Data for Development D4D* data set. D4D is an open data challenge, encouraging research teams around the world to use four datasets of anonymous call patterns of Orange Telecom's Ivory Coast subsidiary, to help address society development questions in novel ways. The data sets are based on anonymized Call Detail Records extracted from Oranges customer base, covering the months of December 2011 to April 2012. Four datasets are provided by *Orange* in *Tabulation Separated Values* (TSV) plain text format.

1) **Antenna-to-Antenna:** Antennas are uniquely identified by an antenna identifier (A_{id}) and a geographic location. The data set aggregate hour by hour the duration of calls between any pair of antennas. The data set is represented by the following tuple `(date_hour TIMESTAMP, originating_antenna INTEGER, terminating_antenna INTEGER, number_voice_calls INTEGER, duration_voice_calls INTEGER)`.
2) **Individual Trajectories High Spatial Resolution Data:** This dataset contains the trajectories of 50000 randomly sampled individuals for the entire observation period but with reduced spatial resolution. The data set is represented by the following tuple `(antenna_identifier INTEGER, longitude FLOAT, latitude FLOAT)`, where `antenna_identifier` represents the antenna anonymized by an identifier and `(user_identifier INTEGER, connection_datetime TIMESTAMP, antenna_identifer INTEGER)`.

3) **Individual Trajectories Long Term Data:** In this data set, the trajectories of 50000 randomly selected individuals are provided for the entire observation period but with reduced spatial resolution. The data set is represented by the following tuple `(subpref_identifier INTEGER, longitude FLOAT, latitude FLOAT)`, where `subpref_identifier` represents the subprefecture anonymized by an identifier and `(user_identifer INTEGER, connection_datetime TIMESTAMP, subpref_identifier INTEGER)`.
4) **Communication Sub Graphs:** The dataset contains the communication sub graphs for 5000 randomly selected individuals identified by `user_identifier`. The data set is represented by the following tuple `(Source_user_identifer INTEGER, destination_user_identifier INTEGER)`

III. Approach to CDR Data Analysis

As mentioned in section II, the CDR data is for a duration of 6 months with about twenty million records; hence, traditional methods of data analysis may not work. With this in mind we used the *Apache Hadoop* framework as a base line.

A. Framework for CDR Data Analysis

Apache Hadoop [4] is a framework for running applications on a large cluster built of commodity hardware. The Hadoop framework transparently provides applications for both reliability and data motion. In addition, it provides a distributed file system (HDFS) that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. In conjunction with *Hadoop*, we used *Hive* [5]. Hive is a data warehouse system for Hadoop that facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems. The overall framework of *Hadoop* in conjunction with *Hive* is shown in Figure 1. The major components of the framework are

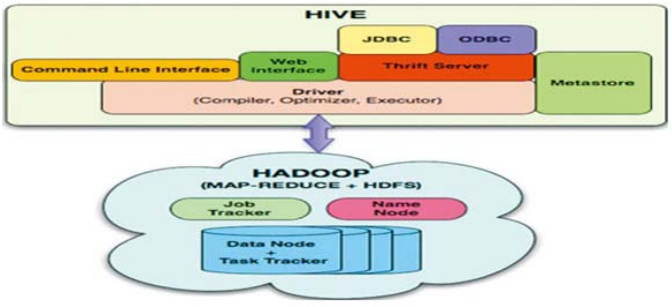

Fig. 1. Framework for CDR Data Analysis

the *Metastore* that stores all the structure information of the various tables and partitions in the warehouse and the *Execution Engine* that executes the execution plan created by the compiler. Considering the large data set *Mapreduce* technique is used to manage the complexity. For the sake of brevity, details of *Mapreduce* technique are out of scope of this paper, but for the sake of completeness, the *Mapreduce* technique takes a set of data and converts it into another set of data, where individual elements are broken down into key-value pairs. Complex operations can then be performed on these distribted key-value pairs.

B. Methodology for Analysis and Visualization

The methodology used to analyse the data using *Hadoop* is shown in Figure 2. As a first step, the data is loaded to Hadoop.

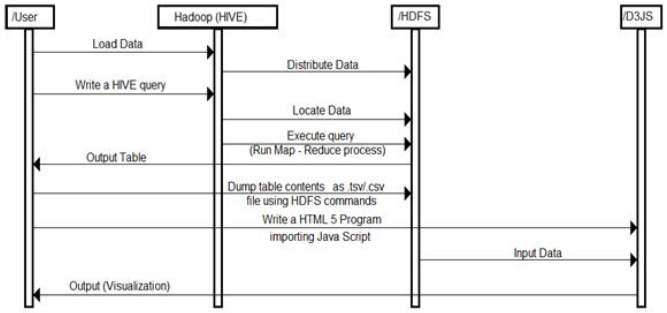

Fig. 2. Methodology used for Analysing Data

Hadoop distributes the data across HDFS in all the slave nodes. Hive query is used to fetch the required data. Hive compiles the query, locates data and executes the query. It returns the output data in a table. The table is stored in the local system in the Hadoop cluster. To visualise this stored data, HTML 5 along with java scripts based on D3JS are used. These java scripts operate on the output of the query and provide the final visualisation.

IV. Statistical Analysis and Visualisation

Based on the methodology discussed in Section III the CDR data was analysed and an appropiate visualisation provided. Since the focus of this paper is on visualization, we target four themes for data analysis.

A. Geographical Understanding

The CDR records are analysed to find the geographical location of antennas in terms of latitude and longiture. This information is correlated and overlayed with an actual *Georaphical Information System* using *Openmaps* [8]. The resulting visualisation is given in Figure 3. As a second step, the sub-prefecture data was then overlayed on the knowledge acquired from antenna locations, resulting in an approximate zonal coverage of a set of antennas as shown in Figure 4.

B. Call Patterns

The most important parameter when monitoring call patterns is the *Call Density*. *Call Density* is defined as the number of calls originating from a *Base Station* within a cellular network. From a telecom operator perspective, the knowledge about the call density and specifically the correlation of the call density to the geography is an important call pattern. This call pattern helps the telecom operator to tune the cellular network

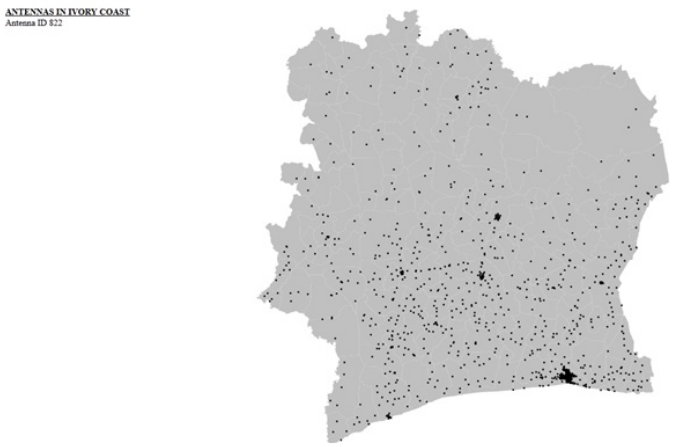

Fig. 3. Geographical Spread of Antennas

Fig. 4. Sub-prefecture based zone coverage

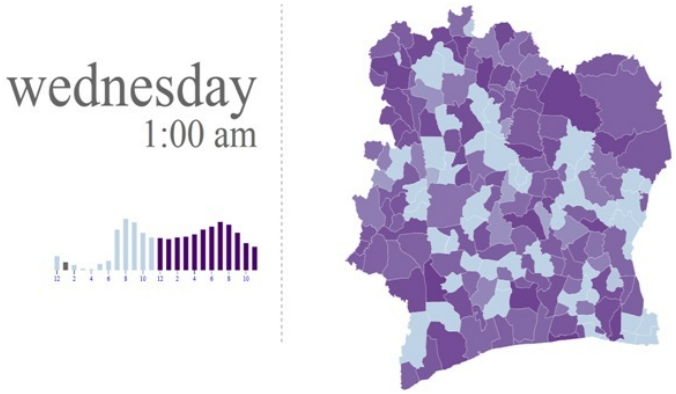

Fig. 5. Sub-prefecture based call density on Wed at 1:00 am

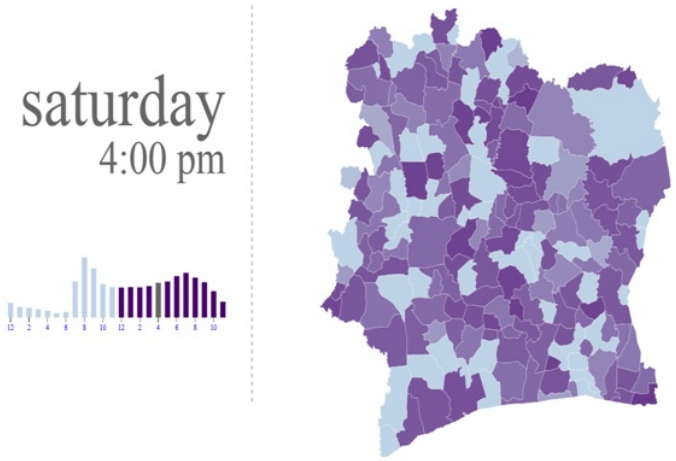

Fig. 6. Sub-prefecture based call density on Saturday at 4:00 pm

to minimise call drop rates and thus maximise revenue. Figure 5[1] and Figure 6 give a snap shot of the call density per sub-prefecture. The darker the shade higher the call density. Such a visualisation gives the telecom operator a quick mental picture of the call patterns. Such a visualisation when combined with actual data can assist the telecom operator to optimise the cellular network for optimal performance.

C. Market Analytics

Operating a cellular network is a CAPEX and OPEX intensive business. Telecom operators are continuously striving to optimise both the CAPEX / OPEX cost. One of the factors that influences the opex cost and determines the return on investment (profits) the telecom company can make is the *Active Air Time*. *Active Air Time (AAT)* is the aggregate time in which the air-waves are used. In other words AAT is the total number of calls made in a day, and monitored across all days / months of the years. In order to give the telecom operator a pictorial view of this data, CDR records were analysed and daily aggregated call information plotted as shown in Figure 7 and Figure 8 respectively. While the value of visualizing the same data in two different views might be debatable, from a telecom operators' perspective it provides two different sets of information. Figure 7 gives a time series perspective of the data and assists the telecom operator to get a visual perception of the number of calls made across different days in a month. On the other hand Figure 8 gives the telecom operator an easy visual reference to compare the number of calls made on the same day across different months. These inputs are important to fine tune the market stratagy of the telecom operator.

D. Sentiment Analysis

One important business objective that every telecom operator tries to achieve is *personalized services*. The personalisation can be in terms of customized tariff plans, personalised greetings and such. Sentiment analysis plays an important role in personalisation. While a lot more can be derived from CDR records from a sentiment analysis perspective, as a first

[1]We thankfully acknowledge Paradigma Labs for the Visualisation concept and code.

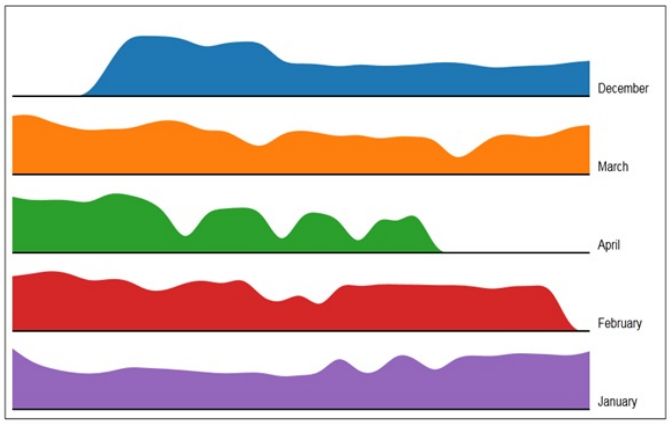

Fig. 7. Snapshot of Call Density variation across months as a Time series

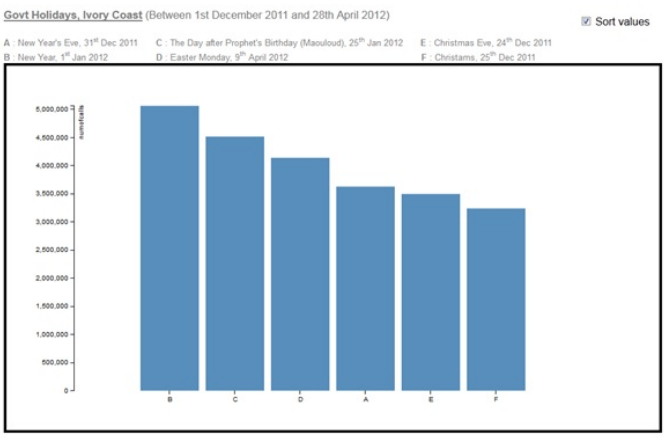

Fig. 9. Aggregated Calls around festive period

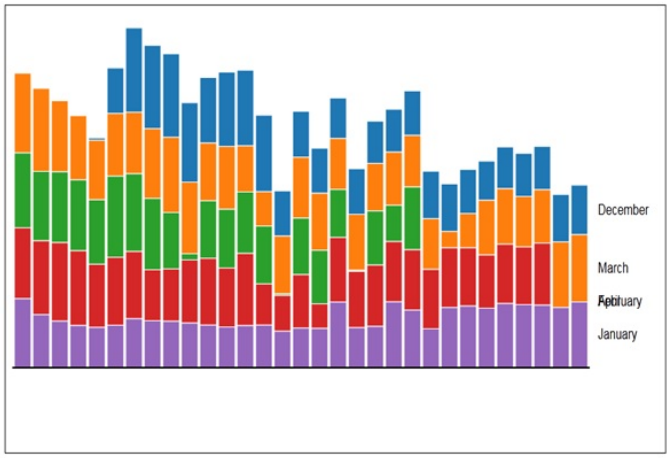

Fig. 8. Snapshot of Call Density variation across months as bar chart

step we have tried to analyse the call patterns around festive periods. Figure 9 shows a pictorial view of the number of calls around the festive period. Observe that on New Year's day the total number of calls is more than on New Year's eve. One interpretation of this data point is possibly indicative of a positive outlook of people of Ivory Coast. This interpretation will be discussed in detail in the following section when we correlate these data sets.

V. Interpretation

Section IV discussed the visual representation of some of the information derived from the CDR. In this section we attempt to provide an interpretation to the data with possible applications of the derived knowledge. While more work needs to be done to have concrete recommendations based on the derived knowledge, this is the first step in that direction. While there can be numerous interpretations of the derived knowledge, we focus on two key aspects *viz.*, *Telecom Operator perspective* and *Socio-Economic perspective*.

A. A Telecom Operator Perspective

From Figure 3 observe that the antenna density is more towards the south east of *Ivory Coast*. Correlating this data with Figure 5 and Figure6 one would expect a higher call density, but a combination of light and dark shades indicates that this is not the case. Our experiments with the CDRs show a consistent low call density in the region. It cannot be conclusively said whether this observation is true all year long since the CDRs are only for five months. However, based on the available data, it can be inferred that some amount of network planning to optimise antenna density to match call density is advisable[2].

Another interesting observation is the correlation between the antenna density on the north east of *Ivory Coast* (Figure 3) and the call density (Figure 6). While the antenna density is low, the call density is relatively high. This is likely to result in higher call drop rate, since there isn't enough network capacity available. Based on this inference, increasing the network capacity by increasing the number of antennas and in turn base stations is advisable.

The other observation is the correlation between the sub-prefectures and the antenna density from Figure 3 and Figure 4 respectively. Observe that the sub-prefectures are smaller towards the southern region of *Ivory Coast* and grow to be larger towards the northern region. However, the antenna density decreases as we travel north. While we are still in the process of deriving the call drop rate across different sub-prefectures, such a decreasing antenna density as the sub-prefecture size grows larger is non-intutive.

From Figure 8 if we compare the call density across different months, and days of the month, it is interesting to note that the call density towards the middle of the month is lower in *January* than in other months. On the same lines note that call density in the second week of *December* is higher than other months, but reduces in the following weeks. Based on

[2]These are not recommendations, but only observations based on visual inspection

this observation, we are attempting to find the mobility pattern of mobile phone users and to determine whether the mobility pattern influences this call density pattern.

B. Socio-Economic Perspective

From Figure 9 observe that the call density is considerably higher on *New Year* day as compared to any other festival days. This is an indication of the importance of the New Year in the lifestyle of the people of *Ivory Coast*. What is worth noting is that aggregated call density on New Year's day is about 30% higher than any other normal day. On an average the call density on festival days is about 10-15% higher than normal. While it might be premature to infer conclusively, this seems to indicate a strong emotional bias in the people of *Ivory Coast*. We say this becuase the economic condition of the people of *Ivory Coast* is not very good, but at the same time during festive periods they do not mind spending on calls. We intend to correlate this data with region-wise socio-economic conditions of the people to infer the telecom spending index and then extrapolate this data to determine region-wise socio-economic conditions. For the sake of completeness and to provide a visual reference related to the geography of *Ivory Coast*, Figure 10 is a map of *Ivory Coast*. Based on our literature survey of the socio-economic conditions of the people of *Ivory Coast* [9], the southern part of *Ivory Coast* (Figure 10) is the the economic capital of the country, but the political capital of the country is *Yamoussoukro* which is in the central region. Comparing Figure 10 and Figure 3 observe that the antenna density is lower in the central region where the capital is located than in the business district of *Abidjan* towards the southern region. Also observe from Figure 5 and Figure 6 that the average call density is higher in the business districts than around the captial of the country. This is indicative of the fact that more of the socio-economic growth is focused in and around the business districts than around the capital. It will be interesting to determine the mobility profiles of users and check whether there is a distinct movement towards the business districts.

Fig. 10. Map of Ivory Coast

Note that all the above observations were made possible only because of the sophisticated visualisation techniques used to represent the derived information. When correlated with the actual inferred data, this can potentially be a handy tool for the telecom operators to optimise their cellular networks.

VI. Conclusion and Future Work

This paper analyses anonymised CDRs obtained from *Orange Telecom* for a duration of five months, one of the telecom operators in *Ivory Coast*[3]. Four data sets as discussed in section II provided by *Orange Telecom* are analysed from a telecom operator and socio-economic perspective. The analysis has been implemented on a Hadoop and Hive framework. This paper proposed a visualisation tool using HTML 5 to visualise the information derived from CDR records.

Information and knowledge related to antenna density, sub-prefectures and call density patterns are derived by analysing CDR records. In this paper we attempt to correlate the derived information and provide observations from a telecom operator perspective and socio-economic perspective. One key observation, from a telecom operator perspective, is that the antenna density does not match with the call density patterns, and there is conderable scope for improvement. Another key observation, from a socio-economic perspective, is that the call density is saturated more towards the business districts than towards the political capital of the country.

While these are just preliminary investigations, in the future we intend to derive the mobility profile of the users and correlate it with the call density and antenna density patterns to provide concrete network optimisation recommendations to the telecom operator. In the process, our larger goal is to propose and develop a tool for CDR record analysis which is configurable and scalable.

References

[1] *Chen Zhou et. all*, Activity Recognition from Call Detail Record: Relation Between Mobile Behavior Pattern And Social Attribute Using Hierarchical Conditional Random Fields, *2010 IEEE International Conference on Green Computing and Communications & 2010 IEEE International Conference on Cyber, Physical and Social Computing*, June. 2010.
[2] *Qining LIN et. all*, Mobile Customer Clustering Based On Call Detail Records For Marketing Campaigns, *International Conference on Management and Service Science* , 2009.
[3] *Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo, Carlo Ratti*, Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records, *13 International Conference onIntelligent Transportation Systems*, 2010
[4] *Tom While*, Hadoop A Definitive Guide,*O'riely Press*
[5] Introduction to Hive, *Cloudera Inc.*
[6] Mapreduce Tutorial, *Apache Software Foundation Tutorial*
[7] *http://www.orange.com*
[8] *http://www.openstreetmaps.org*
[9] *http://www.gouv.ci/Main.php*

[3]We wish to thank Orange Telecom for providing anonymised CDRs on their Ivory Coast subscriber base for a duration of five months as part of the D4D challenge