Exploration and analysis of massive mobile phone data: a layered visual analytics approach

Citation for published version (APA):

Document status and date:
Published: 01/01/2013

Document Version:
Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:
- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 31. May, 2021
Exploration and Analysis of Massive Mobile Phone Data:
A Layered Visual Analytics approach

Stef van den Elzen, Jorik Blaas, Danny Holten, Jan-Kees Buenen, Jarke J. van Wijk,
Robert Spousta, Anna Miao, Simone Sala, Steve Chan

Fig. 1. Different coherent components of the visual analytics solution for the exploration and analysis of Massive Mobile Phone Data in the context of the Orange Data for Development D4D-challenge.

Abstract— We present a system for the exploration and analysis of massive mobile phone data that enables users to gain insight. First we identify user tasks and develop a system following a visual analytics approach by tightly integrating visualization, interaction and algorithmic support. The system is then evaluated by exploring a massive mobile phone data set containing 2.5 billion calls and SMS exchange between around 5 million users located in Ivory Coast over a period of 5 months. From the use cases a number of findings are gathered, such as localized increase and decrease of calls due to major events.

Index Terms—Mobile Phone Data, Visual Analytics

1 Introduction

Four datasets on mobile phone communication were released in the context of the Orange Data for Development (D4D) challenge. The datasets are based on 2.5 billion anonymized Call Detail Records (CDR) of phone calls and SMS exchanges between five million of Orange’s customers in Ivory Coast between December 1, 2011 and April 28, 2012 [10]. In this paper we focus on the first dataset: tower-to-tower traffic. Data is provided to us in tab-separated-value (TSV) file format. For each hour in the timespan we are given the number of calls and duration (aggregated) between any pair of towers. Additionally, we are provided with the geographic location (latitude and longitude) of each cell tower. Initial data cleaning was performed by Orange Labs in Paris, such as removing double entries, new subscribers, and, communication to other providers. We further cleaned the data by removing entries that had missing tower identifiers.

In this paper we describe how we support the analysis and exploration of massive mobile phone data sets by identification of user tasks and according requirements for a visual analytics prototype. Next, this is implemented and applied to the provided real-world mobile phone data. We present a system for the exploration and analysis of massive mobile phone data that enables users to gain insight on different levels of abstraction both in time and space. The prototype provides a smooth user experience despite the massive amount of data. We implement a visual analytics approach adhering to the visual analytics mantra: analyse first, show the important, zoom, filter and analyse further, details on demand [21].

Visual Analytics is the science of analytical reasoning facilitated by interactive visual interfaces [32]. It aims at an integrated approach combining visualization, human factors and automated data analysis using methodologies from information analytics, geospatial analytics and scientific analytics to effectively support the decision making process [22].

---

- Stef van den Elzen is with Eindhoven University of Technology and SynerScope BV. E-mail: s.j.v.d.elzen@tue.nl.
- Jorik Blaas, Danny Holten, Jan-Kees Buenen are with SynerScope BV. E-mail: {jorik.blaas, danny.holten, jan-kees.buenen}@synerscope.com
- Jarke J. van Wijk is with Eindhoven University of Technology. E-mail: vanwijk@win.tue.nl
- Robert Spousta, Anna Miao, Simone Sala, Steve Chan are with Prince of Wales Fellowship at Massachusetts Institute of Technology. E-mail: {spousta, annamiao, salas, s-chan}@mit.edu
Correlations of call change behavior and local events are successfully identified using the prototype. We believe call change occurs due to major events. We found both an increase and decrease in the number of calls over locally concentrated communication channels strongly correlated with events.

The paper is organized as follows. First, user tasks and according requirements to the visual analytics approach are discussed in Section 2. Next, we discuss related work in Section 3. A layered visual analytics approach is presented in Section 4, where the different components of the system are discussed in detail. Section 5 portrays typical use cases utilizing our approach. Finally, conclusions and directions for future work are given in Section 6.

2 DESIGN PRINCIPLES

In this section we identify different user tasks, derive requirements and discuss design decisions following from this. We believe that change in call behavior occurs due to major events. Therefore, to detect events, we focus on call change derived from the first 9D4D dataset (tower-to-tower communication).

The user tasks can be categorized into higher level tasks: exploration, analysis and presentation of massive mobile phone data. The main goal of exploration is to gain insight and to form hypotheses. The main goal of analysis is to confirm or reject hypotheses. While performing analysis, visualization is not only supportive but can also raise new questions therefore, users typically switch often between exploration and analysis during data exploration. Presentation is needed to convey findings to both expert users and a broader audience. In order to support this, familiar visualizations are needed. Table 1 provides an overview of more detailed tasks and requirements. In addition to these requirements, we aim for effortless switching between the exploration, analysis and presentation process. In summary:

- data has to be shown at various levels of aggregation, both temporally and spatially;
- data has a temporal, a geospatial, and a network character; all have to be shown;
- we use multiple linked views for this;
- details have to be shown on demand;
- where possible, use automated methods to simplify analysis and reduce the amount of data;
- features like overall call behavior, call change and communities have to be clearly visible;
- where possible, use familiar mappings and metaphors for easy understanding.

In Section 4 we discuss how the system implements all requirements by detailed discussions of each of the individual components and their integration and coherence.

### Table 1. Requirements to support users in the exploration and analysis of massive mobile phone data. We acknowledge this list is not exhaustive and can be extended further, however, we believe the system should at least support these user tasks.

<table>
<thead>
<tr>
<th>Task</th>
<th>User wants</th>
<th>Requirement the system should</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exploration</td>
<td>Identification of: higher level communication channels; changes in call behavior, because we believe that a change in call behavior occurs due to major events; and communities (cell towers with similar behavior over time).</td>
<td>provide an overview of communication channels both in space and time; provide an overview of call change behavior, again, both temporal and spatial aspects should allow for exploration; provide (algorithmic and visualization) support for community identification.</td>
</tr>
<tr>
<td>Analysis</td>
<td>Perform comparison of: multiple levels of abstraction for time, space, and visualization; multiple points in time; and multiple visualizations of similar or different data dimensions.</td>
<td>enable effortless switching between different abstraction levels; enable for simultaneous comparison of multiple time points; enable for simultaneous comparison of multiple views and data dimensions.</td>
</tr>
<tr>
<td>General</td>
<td>Interactively browse through different portions of the data; being guided by complex patterns and correlations; and do all of this in real time.</td>
<td>provide appropriate visualizations for each abstraction level that; emphasize clues for further navigation to; provide a real-time smooth exploration user experience.</td>
</tr>
</tbody>
</table>

3 RELATED WORK

We briefly discuss related work as means of placing our work in context, and to motivate the development of a new visual analytics prototype: no tool exists that fulfills our requirements.

Many approaches are explored using only automated methods offering no interaction and visualization e.g., [8, 15, 23, 33]. A visual analytics system, developed by Andrienko et al. [6, 7], for extracting place histories from mobile data, combining geovisualizations, geocomputations and statistical methods allows for the exploration of temporal, spatial and thematic components of the data. However, the main focus here is on social aspects and place extraction and does not allow for the exploration of call change inferred from events. Similar approaches, not exploring call change or event detection are discussed by Kwan and Lee [24] and Sagl et al. [25].

A system developed by Correa et al. [13, 31] also mainly focuses on social behavior patterns. Also, the geographical component is not taken into account and no algorithmic support is offered. Temporal communication patterns in mobile call graphs focusing on structural network change is discussed by Ye et al. [37]. Höferlin et al. [16] focus on individual trajectory movement exploration extracted from mobile data. A more general visual analytics system for the exploration of spatio-temporal data not tailored towards mobile data is presented by Von Landesberger et al. [34].

Egocentric temporal exploration of CDR data is performed by Qi et al. [36]. However, the geographical content is not taken into account and cannot be visualized. Furthermore, the method focuses on egocentric exploration and does not allow for an overview of the (higher level) call patterns. Egocentric exploration of temporal call behavior focusing on regions rather than individual towers is explored by Blon-del et al. [9] in their web-based Geofast tool.

4 VISUAL ANALYTICS APPROACH

In this section the developed prototype enabling the effective analysis and exploration of massive mobile phone data is presented. The prototype application follows a visual analytics approach using multiple coordinated views that tightly integrates visualization, interaction and automatic computation methods. A combination of visualization and automatic methods is used, because purely visual methods fall short due to scalability issues; the data provided is large and screen space is limited. This can partially be overcome by interaction methods such as zoom, pan and filter techniques. However, this leaves less apparent patterns in the data hidden. Also, purely automatic methods fall short due to aggregation of results and loss of context. Furthermore, automatic methods are often highly focused and designed for one specific task, not allowing for the exploration and discovery of unexpected patterns. A system effectively integrating visualization, interaction and algorithmic support leverages the benefits of the individual parts.

Initial data transformation steps were taken to be able to implement the requirements as defined in Section 2. These steps are discussed next, followed by a detailed description and discussion of design decisions of the individual coordinated views and their integration into the system (see Figure 2 for a screen-shot of the graphical user interface).
4.1 Analyze first, show the important

The data provided consists of 3.9GB (zipped) and 33.5GB (unzipped) TSV files. Clearly, this does not fit into memory. In order to provide a real-time exploration experience to users different techniques are employed. Here we choose for a combination of pre-computation, divide-and-conquer, and, load-on-demand strategy. The tower-to-tower communications, mainly focused on in this paper, were provided in ten separate TSV files each spanning a period of two weeks. One of the requirements is to provide an overview, therefore all files need to be loaded into memory, however due to size constraints this is not possible. In order to acquire manageable data, we first processed this data using scripts taking an advantage of a line-by-line streaming approach that divided the large files into smaller files each containing the data for one day. Furthermore, the data lines in the smaller files are sorted descending on number of calls between any pair of towers. Finally, separate files were created for calls and duration. This process is repeated to create files on different abstraction levels such as weeks and months. Instead of loading everything into memory at once, smaller chunks can now be loaded on demand if detailed information is requested. Due to the relative small file sizes this allows for real-time exploration. In addition, data can also partially be loaded, requesting only the most important data because the files are internally sorted.

In addition to splitting and organizing each file, different measures are identified for which an overview needs to be provided. These measures were then pre-computed to be shown as a line graph in the measure overview (more on this in Section 4.2). The pre-computed measures are number of calls, duration, call-change, duration-change, and, combined-change. Each of the measures are pre-computed for the different abstraction levels (days, weeks and months). The number of calls measure and duration measure aggregates per abstraction level the total number of calls and total duration respectively. The call-change, duration change and combined change are computed based on the extended Jaccard index, rather than taking the absolute difference of the number of calls for two points in time. This has the advantage that a large change is reported when the number of calls goes from 100 to 1000 but also if the number of calls goes from 1 to 10 for a certain communication channel. Let $E$ be the set of all cell tower pairs having communication on one or more points in time. For each pair of cell towers involved ($e \in E$), at two points in time $t_i$ and $t_j$ we compute the individual call change $ICC_e$:

$$ ICC_e(t_i, t_j) = \frac{M_e(t_i) \times M_e(t_j)}{M_e(t_i^2 + M_e(t_j)^2 - M_e(t_i) \times M_e(t_j))} $$

where $M_e(t_i)$ gives the value of the according measure (here number of calls) at time point $t_i$ for cell tower pair (communication channel) $e$. If $M(t) = 0$, 1 is used to prevent a final value of 0. Next, all individual call changes $ICC$ are summed and divided by the number of involved tower pairs $E$ to provide a final call change value $CC$ for two points in time:

$$ CC(t_i, t_j) = \frac{1}{|E|} \times \sum_{e \in E} ICC_e(t_i, t_j) $$

Duration change and combined change (calls + duration) are computed in a similar fashion. In addition to the change values $CC$ we also store the individual change values $ICC$ and again sort these descending. These files are later loaded-on-demand if additional information is required on the aggregated measures e.g., provide information on the
contribution of each communication channel to the aggregated measure. By pre-computation of these measures we can provide multiple overviews that allow for a smooth real-time exploration of the data.

4.2 Measure overview

In the measure overview line graphs are shown. Users can select what to show on the x-axis and what to show on the y-axis. On the x-axis different aggregation levels of time can be set; users are able to choose one from days, weeks, and months. On the y-axis (a combination of) different measures can be conveyed such as number of calls, duration, call change, duration change, and, combined change. Showing multiple measures in the overview enables users to explore correlation. For example, in Figure 3, we see that the number of calls in the b40-data highly correlates with duration.

The measure overview provides a high-level overview of interesting points in time that require further investigation. Depending on the chosen measure one can focus on (any combination of) curves, peaks and dips. For example, if call change is selected users can identify points in time where change is high by focusing on peaks. The identified points can then be further explored in detail using the different linked views.

On mouse hovering the according measure value is highlighted and the actual value is shown. Furthermore, the aggregated measure for the selected time point is broken down into individual values that are shown in the measure contribution view.

4.3 Measure contribution view

The measure contribution views shows the individual contributions to the aggregated measure of the selected time point in the measure overview. The individual contributions of communication channels (antenna-to-antenna) are shown as horizontal bars (see Figure 4). The horizontal bars are sorted from highest contribution (most important) to lowest contribution to the aggregated measure value. Each bar shows the region, department, sub-prefecture and identifier of both the sending and receiving cell towers. Furthermore, the number of calls (or a different selected measure) over this communication channel is shown along with the measure of the previous day for comparison purposes. This difference is also encoded as bar color; red or green indicates that on this point in time there were less or more calls compared to the previous point in time. By default only the fifty most important contributors are shown. Users are enabled to adjust this value to their likings. In addition we provide users with filtering options. Filtering is possible on the minimum number of calls required or the minimum difference of the number of calls between the previous and current point in time. This enables users to focus on communication channels with low, average or high activity.

All communication channels shown in the measure contribution view are also shown in the geospatial view for spatial identification. By hovering over an individual communication channel in the contribution view, the according channel is also highlighted in the geospatial view.

4.4 Geospatial view

The geospatial view displays the map of Ivory Coast. On top of this map all communication channels and involved cell towers are rendered that are currently shown in the contribution view. The communication channels are rendered as arcs. The direction of the communication is encoded clockwise (see Figure 5). Also here, color depicts whether the number of calls (or a different measure) is lower (red) or higher (green) compared to the previous point in time. The opacity of the arcs depend on the contribution value, similar to the length of the bars in the contribution view; the more important a link is, the higher its opacity. This emphasizes the more important communication channels for easy identification.

On mouse hovering, the according region, department and sub-prefecture are shown. Zoom and panning mechanisms can be used to navigate the map and focus on specific regions.

4.5 Matrix view

The matrix view provides a holistic overview of the behavior over time for each individual cell tower. On the vertical axis all cell towers are shown. The horizontal axis denotes time. Each row represents the behavior of one cell tower over time. On the intersection of a tower and point in time a rectangle is drawn. This rectangle represents a measure value, e.g., number of calls for one antenna at one time interval. The

Technical details: The arcs are rendered as quadraticBezier curves. First the vector from source to destination point is determined. Next we compute the vector orthogonal to this vector with half the length and position it halfway between the source and destination point. Now we take the endpoint of this orthogonal vector as the control point for the quadratic Bezier curve. Due to the computation of the orthogonal vector, taking source and destination point into account, the clockwise direction is automatically inferred. The towers are rendered as white dots with a radial gradient from white opaque (innermost) to full transparent blue (outermost). Finally, additive blending techniques are used to render the towers and arcs on the map to create a subtle aesthetically pleasing glow effect in dense areas. In addition, this allows users to differentiate between increasing and decreasing traffic; precisely overlapping arcs where one is increasing (green) and the other decreasing (red) are rendered as yellow.
rectangles are rendered using a heat-map technique; each rectangle is colored based on the according value, here we use a dark-red to yellow to white colormap; dark-red represents the lowest value, white the highest. Zoom and panning techniques are provided to navigate and explore the matrix.

The matrix enables users to identify towers of similar behavior over time. However, because of limited screen resolution this poses serious difficulty on the task. Therefore, users are enabled to interactively apply clustering methods to the rows shown in the matrix. Once clustered, the rows of the matrix are re-ordered (see Figure 6). Towers that belong to the same cluster (all having similar behavior) are grouped together. In addition, the clusters themselves are also sorted based on cluster-size. We offer cluster parameters to users, which can interactively be adjusted. The result of a parameter change is directly reflected in the matrix view by reordering. The parameters available to users are cluster method (e.g., hierarchical, k-means, k-medians), distance metric (e.g., Euclidean, Manhattan, Pearson, Spearman, Kendall), number of desired clusters, and, time period to take into account while clustering.

If a clustering is applied on the matrix view, users are enabled to only show clusters of interest. Furthermore, the geographical location of the towers belonging to a cluster can be shown or hidden in the geospatial view.

4.6 Linking and integration

From each of the initial views additional views can be opened for the inspection of details. For example, if one communication channel is identified in the contribution view, an additional overview can be created showing the number of calls (or different measure) for the entire timespan to verify outlier behavior. Similar, the number of calls for a specific cell tower can be shown from the matrix view. The creation of new views enables comparison both in time and space to verify hypotheses. Finally, data for a combination of point (or period) in time, range of cell towers, and, range of regions, can be exported to a file for further investigation in external tools such as SynerScope [1] (see Figure 7). Also, facilities for easily searching the internet for events on a specific date and region are built-in. On double clicking in the geospatial view the platform-specific default browser is opened with according constructed search strings.

5 USE CASES

In the following sections, typical use cases are presented that demonstrate the power of the the visual analytics approach to the exploration and analysis of massive mobile data in the context of the D4D challenge. First some background knowledge on Ivory Coast is discussed to provide a context. This background knowledge is assembled based on United Nations reports [26, 27, 28, 29, 30]. Next we provide general findings and interesting correlations are extracted from complex patterns found while browsing through the data using the prototype.

5.1 Background knowledge

On November 28, 2010 elections were held to choose a new president for Côte d’Ivoire. There were two candidates to be chosen from, the current president Mr. Gbagbo (leader of the FPI party) and the opponent Mr. Ouattara (leader of the RHDP party). On December 2, 2010 the Independent Electoral Commission announced that Alassane Ouattara garnered 54.1 per cent of the votes while Laurent Gbagbo received only 45.9 per cent of the votes. That same day, the constitutional council declared the electoral results to be invalid, due to missing the deadline for announcing the provisional results. The next day, December 3, 2010 the constitutional council proclaimed the final results of the presidential elections. This time, however, Laurent Gbagbo received 51.45 per cent of the votes while Alassane Ouattara received 48.55 per cent of the votes. These results of the first announcement were later certified as the rightful outcome of the elections by the UNOCI [28]. However, Laurent Gbagbo did not step down. This started the post-electoral crisis, resulting in violent attacks, killing of civilians, rape, torture, displacement and inhumane and degrading treatment. While human rights abuses have been committed by both sides, most of the killings have been carried out by elements of the forces loyal to Mr. Gbagbo [28]. The situation continued to deteriorate until former President Gbagbo was apprehended on April 11, 2011 [26]. However, pro-Gbagbo militias, mercenaries and FDS (former army) elements continued fighting. Some 50 of those elements surrendered to FRCI (new army) on April 29, 2011, the rest fled towards the Liberian border area, where they continued to kill civilians and loot property in south-western Côte d’Ivoire. Clashes between the FRCI and pro-Gbagbo militias and mercenaries continued to be reported there, as well as violence against civilians in the west and south-west [26]. On December 11, 2011 legislative elections were held in a generally calm and peaceful manner, however, the country is still struggling to recover from the devastating crisis [27]. Here our data starts. We are provided with CDI-data covering the period December 5th, 2011 until April 22nd, 2012. Because media is government controlled and many reports are made that journalists are oppressed and newspapers are banned [26, 27, 28, 30], we solely rely on U.N. reports and reports of the International Crisis Group [18, 19, 20] as our source of major events that occurred during the data-period. During this period, the situation remains particularly fragile in western Côte d’Ivoire, where large numbers of weapons, armed elements, former combatants, militias and dozos, as well as competition over the control of resources are significant sources of insecurity [29]. Most of the incidents occurred in the west, although insecurity has increased in other parts of the country. Law enforcement, while present throughout the country, remains ineffective, and some areas are still under the protection of the dozos, which increases insecurity [29].

5.2 General findings

In the measure overview events that generate a peak in the number of calls and also in call change behavior are directly visible, such as the celebration of new year (see Figure 8).
From the inspection of the highest contributors to the number of calls on any day it becomes clear that the highest number of calls is very local. This appears in the geospatial view as many predominantly self-loops (see Figure 9(right)). By plotting all contributors using a high transparency value for the individual edges, higher level communication is revealed (see Figure 9(left)). We see for example, that there is a strong link of communication between Bouake and Abidjan, but significantly less strong between Bouake and Yamoussoukro.

From the clusters in the matrix view, some clusters can directly be explained. For example, some clusters show a high week-weekend pattern, with less traffic in the weekends. These towers are based in Abidjan, more specific in the Plateau and Adjame region where most companies are located, thus indeed less traffic in the weekends (see Figure 10).

5.3 Local event increased call correlation patterns

During the following events there is a local increase of cell phone traffic. There is a clear correlation of call change and events that are directly visible when exploring the data. Below, these correlations are discussed in chronological order.

On January, 21, 2012, there is a meeting of the FPI (pro-Gbagbo) in Abidjan. This meeting is violently disrupted by supporters of the RHDP (pro-Quattara). One person was killed, several were injured and property was damaged. Also, national police officers were assaulted. On this day we see, Figure 11, that there is an heavy increase in telephone calls in the west (pro-Gbagbo), also noticeable is the increase in traffic from Abidjan to the western region (probably supporters calling their friends and family, informing them of the disruption).

On February, 11, 2012 and 13, clashes between communities are reported in Arrah. During these days (especially 12 and 13) there is indeed a local increase in the number of calls to the Arrah region, directly visible in the geospatial view (see Figure 12). If we bring up detailed information on the number of calls from the specific communication channels (antenna-to-antenna) there is indeed a remarkably high spike at these days, confirming something is going on.

On February, 21, 2012, the village of Zriglo is attacked, killing six persons and wounding many more. There is indeed an increase in the number of calls to this village for this day. This also becomes apparent if the individual antenna-to-antenna communication is inspected, showing an unusual peak (see Figure 13).

Cocoa is a key commodity in Ivory Coast. The country is the world’s largest producer of cocoa beans, accounting in 2010/2011 for a total 35% of the world’s total production [17]. Because of such a high economic value cocoa has already a driving factor for conflict in the country [12, 35].

Cocoa trees in Ivory Coast are harvested twice a year: a main harvest happens between September to December, while a second minor harvest happens from April to June [11]. April/May is a particularly important time for cocoa farmers, as two important events other than
the aforementioned harvest happen: (a) it is one of the two times when pesticides are applied on cocoa plants; (b) if precipitation has been abundant farmers can establish new cocoa plantations or expand existing ones (starting field operations in May). Moreover, yam varieties growing in forest areas are planted in April and May [11]. Two main hypotheses can be thus made:

- Events happening in April/May (i.e. harvesting, marketing and input supply) are likely to produce an increase in telephone traffic from the western cocoa-growing regions towards urban areas and other (market, logistics) hubs.

- Correlation with abundant (i.e. above normal) rainfall in March/April is likely to produce additional increase in telephone traffic from the western cocoa-growing regions towards agricultural inputs supply hubs.

Two positive rainfall anomalies that may have impacted agricultural activities and may be linked with increased phone calls have been identified in two regions: Bas-Sassandra and Dix-Huits Montagnes (see Figure 14). In the Bas-Sassandra region it was reported a positive rainfall anomaly during the first decade of April (see Figure 15) and an increase in telephone calls on April 8, 2012, was noticed in two different areas: (a) the area of Sassandra and San Pedro subprefectures; and (b) the area of Tabou, Grand-Bereby and San Pedro subprefectures.

Higher data granularity was available only for the area of Sassandra. Rainfall was absent during the first decade of April, with the exception of an highly-anomalous storm on April 3 reported in Sassandra [5]. Still this event does not explain the increase of telephone calls on 8 April. It is important to highlight that April 8 2012 matched with Easter. Some sort of correlation with religious events may hence be assumed.

In the Dix-Huits Montagnes region it was reported a positive rainfall anomaly during the second decade of April (see Figure 16) and an increase in telephone calls on April 13 and 16, 2012. Particularly the increase in phone calls was localized in the Man sub-prefecture, which is an important center for both cocoa and coffee production national and is the most important production area of coffee in the whole country.

5.4 Local event decreased call correlation patterns

In the events described below we found a strong local decrease of call activity (majority of complete shutdown of call activity). Again, these local decreases were clearly visible in the visualizations while browsing the data. Once more, these events are discussed in a chronological fashion.

On December 15th, 2011 all cell towers in the western region appear to be shut down (see Figure 17). A large number of antennas is connected to the electric grid of Ivory Coast. On this day half the country was shut down due to electric failure. Next, we identify additional towers by applying a hierarchical clustering on call behavior over time (see Figure 18). These cell towers have similar call behavior over time. If we inspect one of these towers (typically they all have this pattern) we see that these towers are not entirely shut down, but they remain to have an unusual low number of calls. Then, this low activity remains until the 4th of January, when they appear to be turned on again.

On January 15th, 2012, there were confrontations between communities in Gagnoa resulting in the deaths of 16 people, injuries to many more and the burning of several houses. On this day there is indeed a regional call change. The calls in this region drop to 0 (see Figure 19). A possible cause is the fleeing of locals and damaging of the cell towers.

In early February, 2012 (no date mentioned) there were reports of confrontations between farmers and cattle breeders in Odienne. This led to injuries to several persons and the displacement of some 200 people. On February 5th, 2012 a shutdown of towers in the region around Odienne immediately show up in the call change graph (see
Figure 20). It should be underlined that farmer-pastoralists conflicts in the area of Odienne have been already reported by scholars [14], and the livelihoods of local farmers suffered additional significant stress in the recent two years. Particularly, in March 2012 FAO [2] classified Ivory Coast as a country in need of external assistance because of severe localized food insecurity, citing the northern regions as a food insecurity hotspot because of lacking support services and conflict-related damages to agricultural activities. We performed an assessment of disasters and weather for the time covered to test for correlation. However, no disasters were reported during the whole period according to the International Disaster Database [3]. Also, no significant weather events in terms of rainfall that may have disrupted telecommunications were recorded both in Gagnoa [5] and Odienne, based on Interpolated Estimated Dekadal Rainfall provided by NOAA/FEWSNet1 [4] for the whole Denguele region (see Figure 25).

On March 3rd, 2012, near Daloa and neighboring big cities Man and Duekoue there was a drop in the number of calls. The number of calls on these communication lines dropped from the normal level of 100-200 to 10-20 on this day (see Figure 21). On March 5th, 2012, in Agboville, we again see a decreased call activity (no calls) of at least two towers in the area, that directly pop-up in the call change behavior (see Figure 22). A thunderstorm was recorded on March 3rd, 2012 in Daloa [5]. This event may have seriously impacted telecommunications activity in the neighboring areas. The same weather conditions were reported in Abidjan on March 5, 2012 and we can fairly assume that Agboville (65 km distant from Abidjan) was affected as well by the thunderstorm.

On March 13th, 2012 cell towers in the western region are shut down (see Figure 23). This again might be the result of electric failure. Finally, if we cluster towers on the number of calls over time, several more interesting clusters are revealed, all having a different shutdown period (see Figure 24). To our opinion the shut down of towers can have a number of reasons: (a) this is missing data, Orange could or did not register calls, (b) external factors making communications more difficult, like weather and disasters, (c) sabotage on the antennas, (d) electric failure or diesel replenishment problems for off-grid cell towers, or (e) other technical problems.

1Quantitative estimate of rainfall combining METEOSAT derived Cold Cloud Duration imagery and data on observed rainfall (GTS-Global Telecommunication System by the NOAA Climate Prediction Centre)
The first explanation can be ruled out as it is stated in the D4D data information report that indeed there is missing data but this only covers a period of about 100 hours (~4 days) [10]. However, we notice shut downs (or significantly lowered communication) for periods more than 15 continuous days. As we have seen weather conditions explain some of the local decreased cell tower activity. However, due to time constraints no explanation is found for the clusters of towers that have a significant period of lowered (or none at all) activity. These cases are currently being investigated in a collaborative effort with Orange.

5.5 Socio-economic development

We believe that by focusing on call change, local events can be detected as we have shown in the use cases. Both the local increase and decrease of call behavior provides insight in complex patterns. By focusing on clusters of cell towers having similar call behavior, events can be detected. These events are of different nature, such as weather (heavy rainfall in important cocoa area), social (new years eve), political (party meetings), disorder (clashes between communities) etc. and can only be explained by domain experts by enriching the visual analysis process with external data to gain insight in complex correlations, anomalies and communities both in time and space. This in turn enables event detection which is an important first step towards prediction, improving early intervention through development, aid and other civil initiatives.

6 Conclusions

We aimed at developing a tool for the exploration and analysis of massive mobile data supporting all aspects of the process. We identified user tasks and requirements from which appropriate visualization, interaction and automated support techniques are selected. Next, we implemented these in a highly interactive prototype. We next showed the effectiveness of our visual analytics approach by applying the prototype on massive mobile phone data containing 2.5 billion calls and SMS exchange between around 5 million users located in Ivory Coast over a period of 5 months, provided by France Telecom within the context of the Orange D4D challenge. From the typical use cases obtained while browsing the data, we extracted significant and interesting events by cross-correlating these using UN reports and weather information.
6.1 Future Work

In the context of the D4D challenge we mainly focused on the first dataset, containing detailed information of tower-to-tower communication due to time constraints. We believe it would be valuable to incorporate additional visualizations and automated techniques that enable also the exploration and analysis of the remaining datasets. Also, we believe exploration and analysis of non-aggregated data provides even more insight in complex patterns.

REFERENCES


