Specification and early Prototype
Deliverable no.: 9.3
27/04/2016

ASAP is funded by the European Commission DG-INFSO
Seventh Framework Programmed, Contract no.: 619706

Table of Contents

Deliverable Title	
Filename	d9-3_y2-review_final.docx
Author(s)	R. Bertoldi, R. Trasarti, P. Cintia
Date	27/04/2016

Start of the project: 01/03/2014
Duration: 3 years
Project coordinator organization: FORTH

Deliverable title: Specification and early Prototype
Deliverable no.: 9.3

Due date of deliverable: 27/04/2016
Actual submission date: 27/04/2016

Dissemination Level

X	PU	Public
	PP	Restricted to other programme participants (including the Commission Services)
	RE	Restricted to a group specified by the consortium (including the Commission Services)
	CO	Confidential, only for members of the consortium (including the Commission Services)

Deliverable status version control

Version	Date	Author
_y2-review_final	27/04/2016	R. Bertoldi, R. Trasarti, P. Cintia

Abstract

The present deliverable *D9.3 "Specification and early Prototype"* describes the specification for the Telecommunication Data Analytics application that will be developed and demonstrated by Wind. These are based on the preliminary version described in the deliverables *D1.1 "Early user requirements"* and *D9.2 "Use Case Requirements"*. Two specific use cases used during the "Graph join" and "Peak detection" phases of the dataset processing will be described in more detail covering aspects like the workflows outline, queries used, dataset type and anonymization, visualization tools, algorithms used and users oriented functionalities and performance expected from the services implemented. The ASAP framework will be used to produce a TDA application.

Keywords

Telecommunication Data Analytics (TDA), Big Data, tourism application, mobility flows, Customer Relationship Management (CRM), Charging/Call Data Records (CDR), structured data, semi-structured data, unstructured data, social networks, Business Intelligence (BI), Multidimensional Analytical (MDA), visualization, dashboard, peak detection, user profiling, sociometer.

List of Figures

List of Abbreviations

API	Application Programming Interface
BI	Business Intelligence
CDR	Call/Charge Data Records
CRM	Customer Relationship Management
DWH	Data Warehouse
ETL	Extract Transform Load
GeoJSON	Geographic JSON
GSM	Groupe Spécial Mobile → Global System for Mobile communications
GUI	Graphical User Interface
HDFS	Hadoop Filesystem
HTML	Hypertext Markup Language
I/O	Input/Output
IReS	Intelligent multi-engine Resource Scheduler
JSON	Javascript Object Notation
JWT	JSON Web Tokens
MDA	Multidimensional Analytical
ML	Machine Learning
MR	MapReduce
POI	Point Of Interest
QB	RDF Data Cube
RDBMS	Relational Database Management System
RDD	Resilient Distributed Datasets
RDF	Resource Description Framework
REST	Representational State Transfer
SMS	Short Message Service
SNA	Social Network Analysis
TDA	Telecommunication Data Analytics
VM	Virtual Machine
WLT	webLyzard technology
WMT	Workflow Management Tool
WP	Work Package

1 WP9 Introduction

The main objective of this Work Package (WP) is the design and development of an analytics application on Wind Telecommunications customer data, targeted towards tourism and mobility scenarios. The envisaged use cases will be integrated into the ASAP framework and will be evaluated using several measurement methods. At the end of the project second year (M24) the tasks involved are three: the end of task T9.2, the task T9.3 and the beginning of task T9.4.

1.1 Task T9.2 – Use Case Requirements

This task objective is to define the requirements for the use case. Based on these requirements, the task proposes workflows of analytic queries for mining the data, deducing information and employing this information for the creation of a novel knowledge. A first definition is presented in deliverable *D9.2 "Use Case Requirements"*.

1.2 Task T9.3 – Analytics Application Development on ASAP

This task objective is to develop the analytics application that realizes the use case of task T9.2. The application includes the development of the proposed analytics tools. As described in the previous section, starting from the envisaged use case some of the subtasks are developed and integrated in the ASAP platform.

In order to fulfill task T9.3 goals, the telecommunication data analytics (TDA) application (see section 2) includes the following features: an engine for the statistical analysis of the data developed in Spark (see section 3), a visualization API used to query the dataset and a display, on a graphical dashboard, of the desired information required by the data-scientist/mobility-manager (e.g. a spatial representation of the traffic patterns mined, and an extraction instrument that extracts mined data into different formats for further analysis) (see section 3.3).

1.3 Task T9.4 – Evaluation

This task has the objective to design and conduct evaluation studies with the aim of measuring the performance and scalability of the developed method, as well as the quality of the mined patterns. Since this task is at the beginning, the major effort went to the study of the technical integration and functional evaluation that will be described in the results section (see section 4). In particular, experts in the fields are involved in the process to understand the usefulness of the methods, the quality and the validity of the results.

2 TDA Use Case Development

In the context of the ASAP project, we intend to design and implement a new application to take better advantage of the new big data approach for mobile applications. The ASAP telecommunications application (TDA) will show how a number of analytical services describing the mobility of people can be created on the basis of the data collected by the mobile network during routine operation.

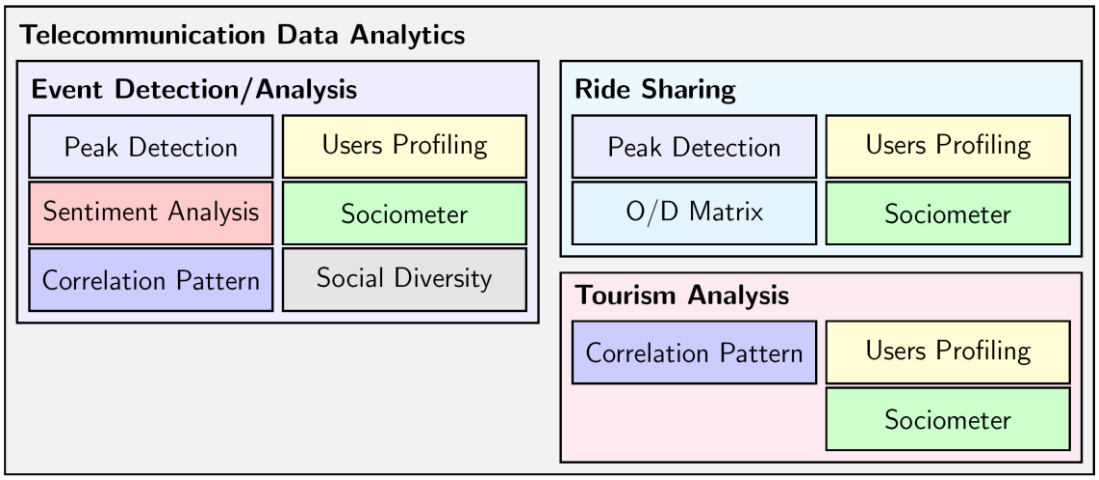

Figure 1: The Telecommunication Data Analytics (TDA) application and its decomposition in common basic analyses.

In particular three services will be presented: (i) Event Detection, (ii) Ride Sharing, and (iii) Tourism Analysis. Each one can be decomposed in basic analyses or modules as shown in Figure 1.

From this initial design of the application, described in detail in deliverable *D9.2 "Use Case Requirements"*, we selected a subset of modules to be developed: the user profiling, the sociometer and the peak detection. Moreover some additional modules are defined and developed in order to build complete workflows going from the data to the publication of the results, as shown in Figure 2.

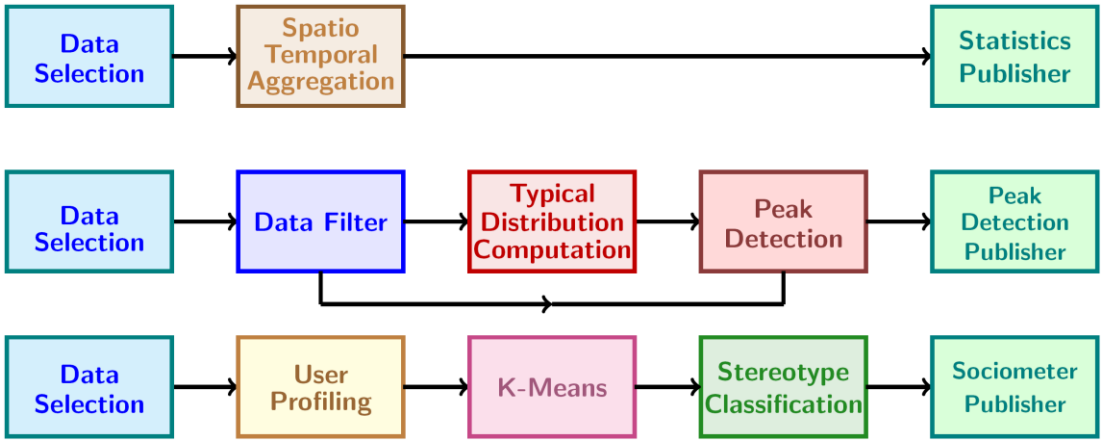

Figure 2: The workflows of the TDA application early prototype.

It is important to notice that all the modules start assuming that the data is available. This means that a phase of extract, transform, load (ETL) and data processing is needed. This step is out of the scope of the workflows because it is independent from the ASAP platform and is automatically triggered when the data is produced and stored in the Wind servers.

2.1 Data Management

In this section we will describe the type of data used by the TDA, the processes used to extract and transfer the data stored in the Wind servers to the platform as well as the techniques used to guarantee the privacy of the users.

Call Data Records (CDR)

The cellular phones are probably the most popular devices we carry everywhere nowadays. Since mobile phones functionality is based on the communication to an antenna covering a local area, the active connection to a certain antenna (e.g. a call or SMS) represents a spatio-temporal position information of the user. This information, as collected by the telecom provider, thus provides a spatio-temporal fingerprint of the users moving in an area covered by mobile telecommunication services. An example of call data record (CDR) is reported in Figure 3. The call started in *Cell_1* and ended in *Cell_2* will result in a single row in the log as follows:

< id32876, VOICE, 10/10/2015 :00:00, Cell_1, Cell_2, 5:00 >

containing the *id* of the user, the type of the event, when the event started, the initial cell, the end cell and the duration.

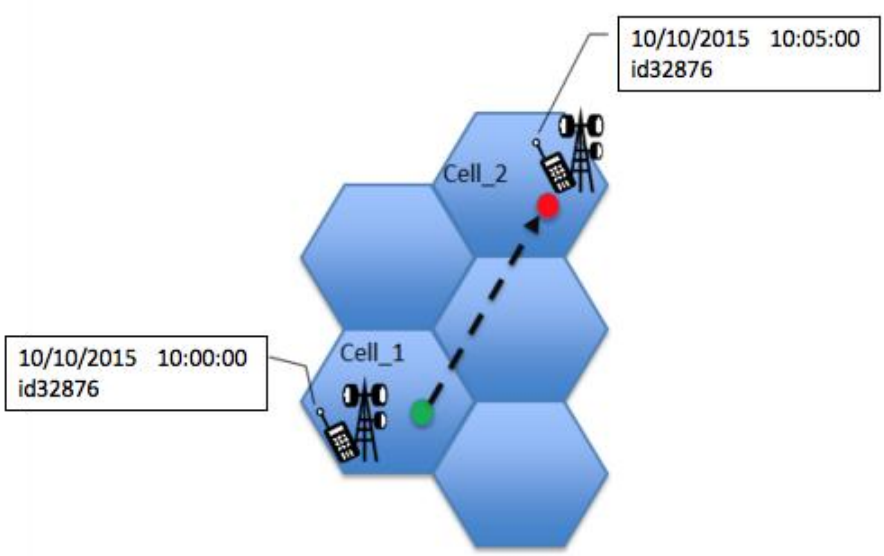

Figure 3: Example of CDR log produced by user activity.

This type of data raises three considerations: the first one is that mobile telecommunications traffic data embeds a great deal of information about the user habits and behavior; the second one is that mobile telecommunications traffic data, when collected from the provider side, comes in large volume thus posing new challenges, not only from the collection and storage point of view, but also from the analysis and mining side; the third one is that the analysis of this large volume of personal data poses several privacy issues.

Notice that the *id* of the user, as given in the call data record example above, has been appropriately anonymized in order to be compliant with current privacy rules and

regulations and thus avoiding the access to personal information about the real user. Further details on the anonymization procedure are given in section 2.2 below.

ETL Procedure

A continuous flow of data from the users is stored in the Wind data warehouse (DWH) which comprises multiple databases containing different types of user's data (CDR, CRM). The first step to realize a realistic service in the ASAP platform is to define and implement an ETL process able to update the data periodically (i.e. monthly).

The datasets being used in the ASAP project are a result of the combination of various datasets (CDR, CRM) which are extracted following the ETL process depicted in Figure 4. This ETL process takes also into account privacy aspects of the data being extracted and the resultant datasets comply with current privacy rules and regulations. The dataset contains all the call data records registered in the region of the City of Rome for several months. Each month of traffic activity (50GB of data in the average) corresponds to about 5.6 million lines of records per day. The amount of data being extracted monthly implies that by the end of the ASAP project the size of the accumulated datasets will be about 1TB.

At the moment the datasets being used contain structured and semi-structured data. It is foreseen to include also datasets containing unstructured data coming from a data crawl of Facebook public pages in order to perform social network analysis (SNA) on this unstructured data. However, it will be possible to make use of these datasets only after further privacy and compliance analysis, which is still being done.

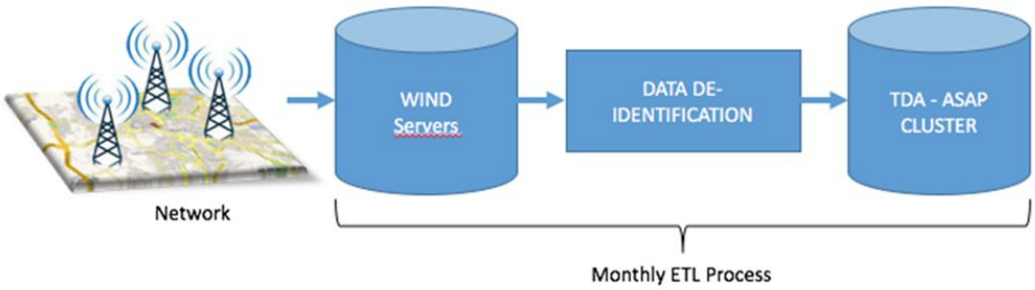

Figure 4: The implemented ETL process to extract the data from the Wind servers.

2.2 Data Anonymization Procedure

Due to the sensitive nature of the data, we have taken into account the privacy issues during the entire process of analysis, customizing and applying the privacy risk analysis method presented in [5] and already tested in the work presented at CPDP in 2013, [7]. This methodology implements and satisfies the constraints issued by the European Union for data protection in [6] and follows the principle therein given.

In summary the risk analysis follows the idea that, given a dataset and a specific application, it is possible to define the set of possible attacks with regard to different levels of knowledge.

The risk of "linkability", inference and single out can then be evaluated. After a risk is detected, a technique for anonymizing the data is chosen, i.e. masking or perturbation, realizing a good trade-off between privacy guarantee and quality of service. In Figure 5 the general framework presented in [4] is shown.

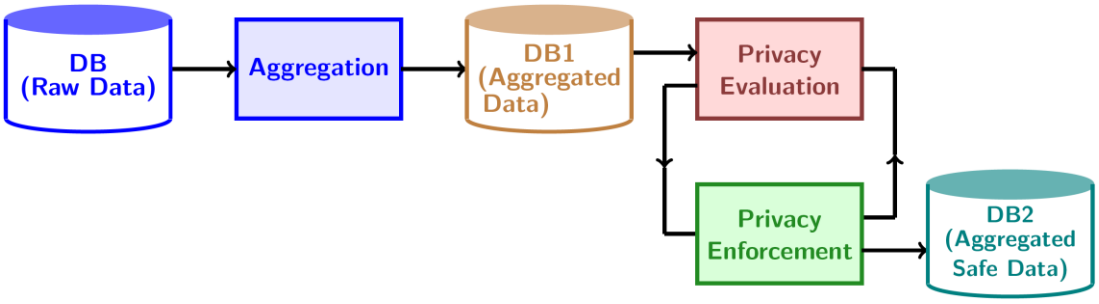

Figure 5: The Privacy Risk framework.

Here the data is firstly aggregated to fit the requirements of a specific application, then the privacy is evaluated in order to detect the portion of data having a high risk and finally that data is deleted or transformed to generate the safe data.

In ASAP the idea is to use this approach in order to cut part of the process from the workflows and put it directly at the sources of the data, in this way the data provided can be accessible to the analysts in a safe way. At this moment the algorithms are developed but not integrated yet in the process. For this reason the other partners are working with a simulated dataset and only Wind have the possibility of running the workflow over the real data.

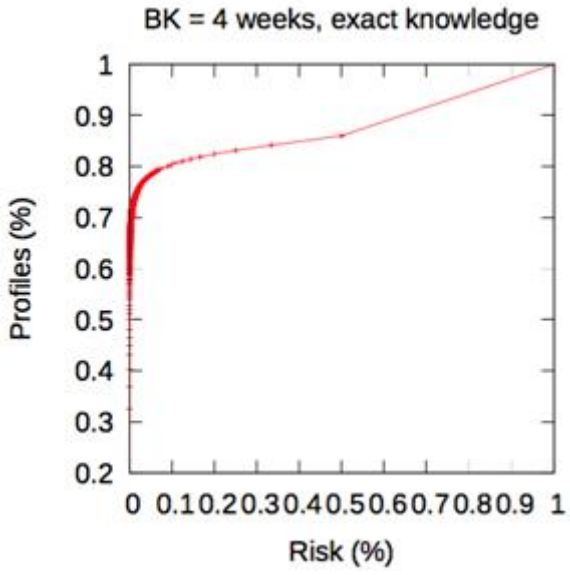

Figure 6: Cumulative curve of the privacy risk in disclosing users profiles computed from Wind Call Data Records (City of Rome, November 2015).

In the workflows presented there are two aggregation functions: (i) a sum over a spatial and temporal partitioning and (ii) the generation of user profiles. For the first case the privacy is guaranteed by eliminating all the regions with a number of call below a certain threshold, in the second case the information disclosed is more complex (user profiling) and therefore more discriminative for the re-identification of the user.

The implementation of the methodology presented in [4] allows us to check the privacy risk over the Wind dataset and the results are reported in Figure 6. The results show that, considering an acceptable risk of 10% (a user is not distinguishable in a group of 10 users) the 80% of the profiles are safe. The other 20% of user profiles lead to a real risk on the privacy of the users, and therefore a privacy enforcement method must be applied on them. The most radical solution is to delete them. The real effects of this deletion on the results of the *Sociometer* are evaluated empirically with good results, but a formal definition of them is still under study as well of other enforcing methods in substitution of the simple deletion.

2.3 TDA Workflows

In this section a detailed description of the subtasks performed by the developed modules is presented.

Spatio-Temporal Aggregation

The call data records are aggregated in space and time in order to obtain time series of the number of calls or SMS. This basic statistic is very useful to have an overview of the data feeding the platform without disclosing the raw data.

Peak Detection

With this analysis we want to detect relevant peaks representing an event. Comparing the density of population within a region in a given moment against the expected density for that area at that hour of the day can do this. This process is part of the methodology presented in [3]. In detail it is realized by means of two modules: typical distribution computation and peak detection.

The first step of the process consists in defining the geographical area to analyze and to partition it into a set of regions. The same must be done for the time, where a timeframe is chosen (for instance, a month), partitioned into periods (for instance, days) and then into smaller timeslots (for instance, hours). Timeslots are described by a parameter T, while the regions that cover the area of analysis are described by a parameter S, both parameters being provided by the user. These two parameters, then, allow defining a spatio-temporal partition, and each observation of an input dataset can be assigned to one of its region. The number of observations that fall in a region defines its density. In the workflow we can notice that the input data is partitioned into two sets: a training dataset and a test dataset. For both datasets the spatio-temporal densities are computed. The first is used to compute the densities of a typical period for each region. The second dataset is then compared against such typical period in order to detect significant deviations.

User Profiling

The spatio-temporal profile is an aggregated representation of the presence of a user in a certain area of interest during different pre-defined timeslots. This profile is

constructed starting from the CDR and with reference to a particular spatial representation. The CDR spatial coverage describes the distribution of the antennas used by the mobile telecommunications operator on the territory, which can be used to estimate the corresponding coverage. A spatio-temporal profile codes the presence of a user in the area of interest in a particular time (or timeslot) identified by the information in the CDR. The idea is that if a person makes a call in the area *A* at time *t*, it means that the user is present in that area at that time. The complete procedure is presented firstly in [1]. Figure 7 shows the typical matrix structure of a spatio-temporal profile. In this case the temporal aggregation is by week, where each day of a given week is grouped in weekdays and weekend. Given for example a temporal window of 28 days (4 weeks), the resulting matrix has 8 columns (2 columns for each week, one for the weekdays and one for the weekend). A further temporal partitioning is applied to the daily hours. A day is divided into several timeslots, representing interesting times of the day. This partitioning adds new rows to the matrix. In this case we have 3 timeslots (*t1, t2, t3*) so the matrix has 3 rows.

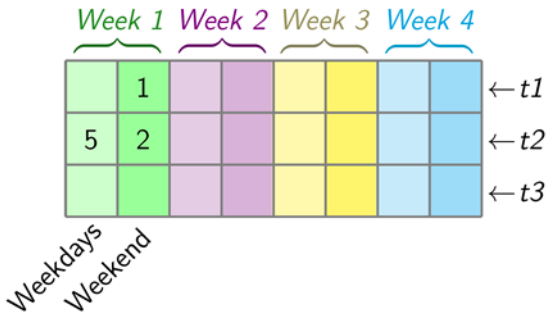

Figure 7: Structure of the aggregated spatio-temporal user profile.

Numbers in the matrix represent the number of events (in this case the presence of the user) performed by the user in a particular period within a particular timeslot. Take for example the number 5: it means that the user was present in the area of interest for 5 distinct weekdays during *Week1* in timeslot *t2* only.

Sociometer

Exploiting the methodology called *Sociometer* it is possible to classify the users using the presence of cellphone users. Once the profiles have been created, the *Sociometer* classifies them implementing a set of domain rules that describes the mobility behavior categories. In particular we are interested in identifying residents, commuters and visitors:

- A person is *Resident* in an area *A* when his/her home is inside *A*. Therefore the mobility tends to be from and towards his/her home.
- A person is a *Commuter* between an area *B* and an area *A* if his/her home is in *B* while the work/school place is in *A*. Therefore the daily mobility of this person is mainly between *B* and *A*.
- A person is a *Dynamic Resident* between an area *A* and an area *B* if his/her home is in *A* while the work/school place is in *B*. A *Dynamic Resident* represents a sort of "opposite" of the *Commuter*.

- A person is a *Visitor* in an area *A* if his/her home and work/school places are outside *A*, and the presence inside the area is limited to a certain period of time that can allow him/her to perform some activities in *A*.

In the workflow this module is implemented in terms of two subtasks: the *K-Means* algorithm which groups the profiles into similar behavior groups and the stereotypes classification which finds the most appropriate class for each cluster as presented in [2].

Publisher Modules

Each workflow ends with a publisher module which is able to produce a report of the result obtained. In particular they translate the results in a JSON [1] or GeoJSON [2] format readable by the webLyzard visualization tool. More details will be given in the next section.

Example of a GeoJSON collection:

```
{"type": "FeatureCollection",
      "features": [
            {"type": "Feature",
            "geometry": {"type": "Point", "coordinates": [102.0, 0.5]},
            "properties": {"prop0": "value0"}
            },
            {"type": "Feature",
                  "geometry": {
                  "type": "LineString",
                  "coordinates": [
                  [102.0, 0.0], [103.0, 1.0], [104.0, 0.0],
                  [105.0, 1.0]
                  ]
            },
                  "properties": {
                  "prop0": "value0",
                  "prop1": 0.0
                  }
            },
                  {"type": "Feature",
                        "geometry": {
                        "type": "Polygon",
                        "coordinates": [
                              [ [100.0, 0.0], [101.0, 0.0],
                              [101.0, 1.0], [100.0, 1.0],
                              [100.0, 0.0] ]
                        ]
                  },
                  "properties": {
                  "prop0": "value0",
                  "prop1": {"this": "that"}
                  }
            }
      ]
}
```

[1] http://json.org/
[2] http://geojson.org/

3 System Deployment and Integration

The modules are implemented in *Spark* [3] and deployed on a Wind cluster composed of 4 machines with 12 hyper-threading processors. Spark is an open source cluster computing framework, in contrast to Hadoop two-stage disk-based MapReduce (MR) paradigm, and Spark multi-stage in-memory primitives provide performance up to 100 times faster for certain applications. This technology is suitable in all the applications where big quantities of data should be processed while considering a small portion of data. In our case the *spatio-temporal partitions* as well as the concept of user activities are very effective in dividing the jobs.

Moreover, the Workflow Management Tool (WMT) and the IReS platform are installed on the master node of the cluster. In addition to that, thanks to the publisher modules, all the workflows are able to send the results of the computations to the webLyzard Visualization tool which is capable of rendering the results.

3.1 Workflow Management Tools

The Workflow Management Tool (WMT) is a component of the ASAP system. It is used for workflow creation, modification, analysis and optimization. All the modules described in the previous sections are integrated into the system specifying the parameters needed for the execution as well as the technologies required for the execution. In this way the user is able to build the workflows presented. Once they are defined the system translates them into an execution plan which is readable by the IReS platform.

3.2 IReS Platform

The IReS platform is a core component of the ASAP system architecture. Its role is to manage complex analytics workflows in an intelligent way. Its main task is to "mix-and-match" diverse execution engines and data stores in order to optimize a workflow with respect to multiple, user-defined criteria.

In the following the integration of the User Profiling and Sociometer workflow is presented. All individual operator implementations that are used in the workflows have been described in individual description files, that contain all the aforementioned fields (since they constitute materialized operators). Additionally, for all input datasets (which are also materialized ones, since they are existent), the respective description file has been produced. Starting from the workflow of the Sociometer in Figure 2, the abstract modules are transformed into concrete modules as shown in Figure 8: the green concrete module is the optimal, i.e. the one with minimum execution time, since execution time is the metric for which an optimal plan is requested.

[3] http://spark.apache.org

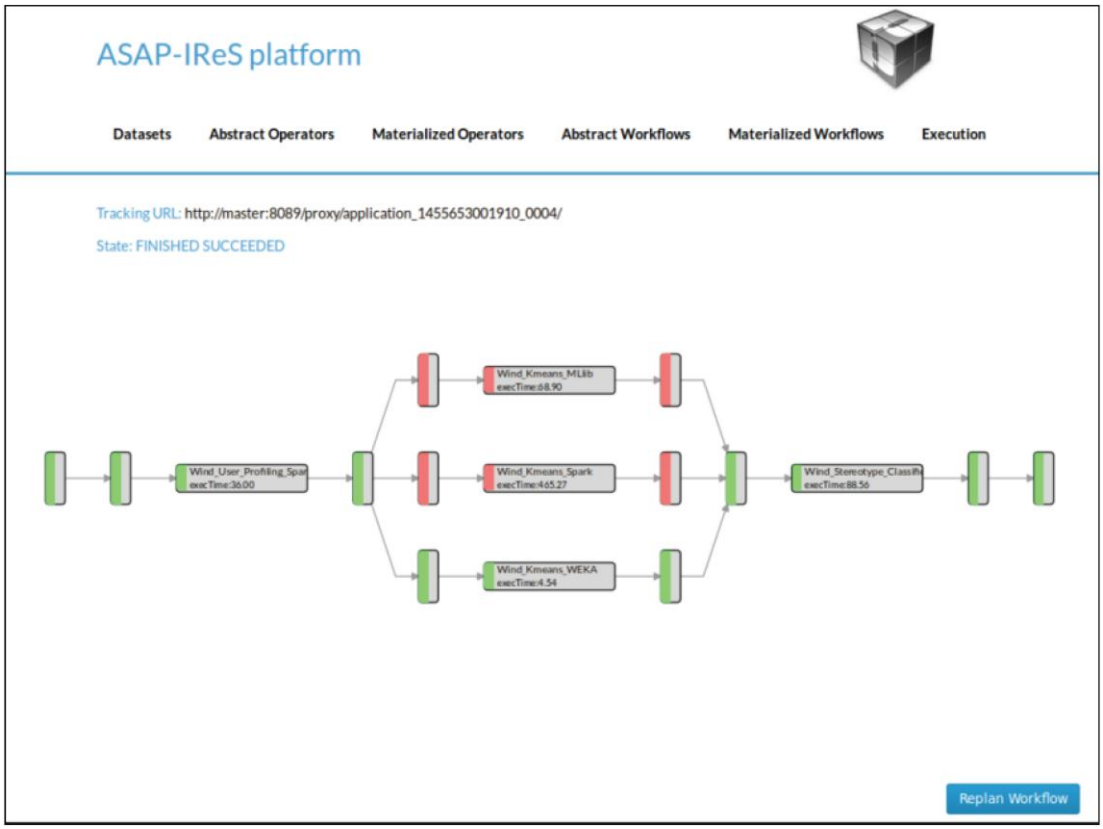

Figure 8: Modules workflow in IReS platform with three different concrete implementations of the K-Means (in green the best solution).

IReS in general considers all alternative implementations of an abstract operator, adding, when necessary, move (from one engine to another) or transform (from one format to another) operators, as long as they exist in the IReS operator library. The cost and execution time of these auxiliary operators are naturally taken into account during planning. Experiments on the performance of the IReS platform have shown that for medium-sized workflows the best execution plan is discovered in less than 2 seconds.

Specifically in the case of the TDA workflows, only the Sociometer contains an operator with more than one implementation: k-means has 3 implementations, all of which share the same input/output format, reading from and writing to HDFS. Thus, no move/transform operator is required. IReS was able to select the execution plan within a few tenths of milliseconds (40 ms on average).

3.3 Dashboard and Visualization Tools

The webLyzard dashboard shown in Figure 9 pursues a multiple coordinated view approach in order to offer a feature-rich visual analytics solution that integrates telecommunication data with semantic search and Web intelligence [8]. The dashboard provides content streams about Italy and the region of Rome from (i) international news media outlets, (ii) Google social media channels – YouTube and Google+, and (iii) Twitter postings that were published with a geospatial annotation. These unstructured data streams are combined with the *Call Data Records* (CDR) to show spatio-temporal aggregation and peak detection results, as well as the *Sociometer* classification. The data

points are visualized along temporal and geospatial dimensions – in the form of stacked bar chart and geographic maps, respectively.

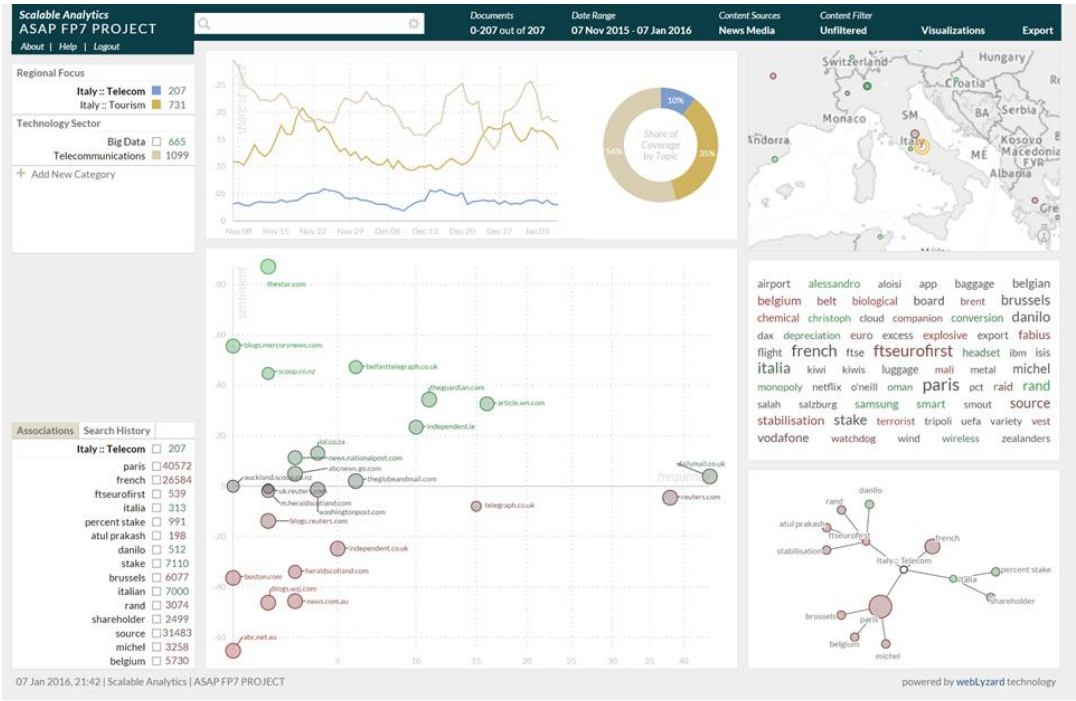

Figure 9: Screenshot of the ASAP dashboard as of January 2016, synchronized in real time following a multiple coordinated view approach.

The three datasets (Call Data, SMS, Sociometer) are ingested using webLyzard Statistical Data API , which uses a JSON representation of the QB [4] Vocabulary (RDF Data Cube Vocabulary) for including data into the webLyzard visualization engine. The main object of this API is an observation and corresponds to a measurement (or a set of measurements) taken in a known moment of time. The required fields from this format describe datasets and observations with the QB vocabulary (dataset, observation URI, observation value, date, etc.), while optional fields can accommodate dataset-specific information such as geographic location or measurement unit.

The API uses the *Flask* and *Vert.x* frameworks and supports CRUD operations (Create/Read/Update/Delete) for observations. Authentication is required for third-party users, and the generated security tokens are valid for eight hours.

To add observations to a repository, one can issue a request with a structure similar to the following:

Adding an observation to a repository:

```
$ curl -XPOST 'https:// ... /0.1/observations/<repository_name>/<cdr' -d
'{
"_id": "111",
"uri": "cdr/111",
```

[4] http://www.w3.org/TR/vocab-data-cube

```
"added_date": "2015-10-10T15:00:02.294083",
"date": "1982-01-01T00:00:00",
"indicator_id": "cdr",
"target_location": [
            {
            "name": "Rome",
            "point": {
                  "lat": 41.54,
                  "lon": 12.30
                  }
            }
      ],
...
}'
```

Using the floating menu of the dashboard trend chart module, users can switch from the default line chart to a stacked bar chart that supports the display of aggregated Call Data Records (CDR) data and Web intelligence metrics in a joint and customizable representation. In the example of Figure 10, blue areas represent the number of *phone calls* and *text messages* sent, green areas above the horizontal axis the association with *desired topics* and the number of *positive references*; the grey area the number of *neutral references*; red areas below the axis the *number of negative* references and the association with *undesired topics*.

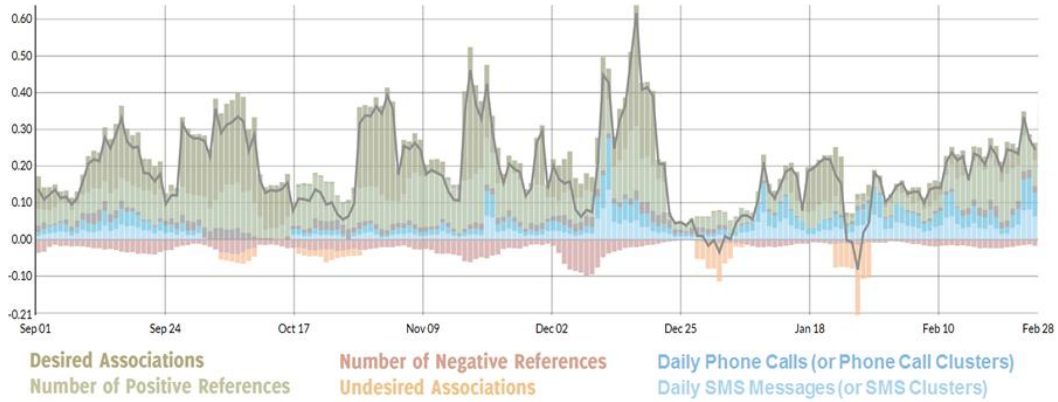

Figure 10: Stacked bar chart for the hybrid display of aggregated Call Data Records (CDR) data and Web intelligence metrics.

In addition to the fully synchronized dashboard [9], the geospatial map and the trend chart have also been made available as stand-alone versions, embeddable in external applications based on the new webLyzard REST API, which also supports the upload of structured datasets. A simple software container-based setup offers all the necessary data transformation and visualization steps. The package contains an API endpoint, a data store including an analytics engine for data aggregations and a client for the visualization. Based on *docker-compose*, the whole container setup can be deployed and initialized with a single command. The geographic distribution in the left part of Figure 11 represents Twitter postings ingested using the API and annotated with *Point of Interest* (POI) coordinates within the City of Rome as of February 11, 2016.

4 Analytical Results

In this section we describe some results on the real data in order to assess the usefulness of the analyses. In order to create statistics and models over the City of Rome it was important to define some meaningful partition of the space. In order to reach this goal, the Mobility Agency has been contacted, and they remained strongly impressed by the tools proposed, providing as feedback a spatial partitioning of the City of Rome as shown in Figure 11.

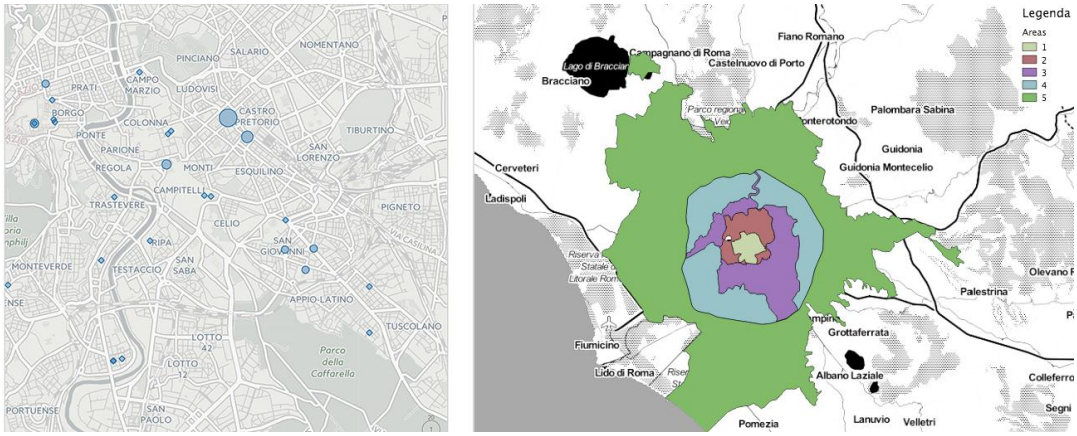

Figure 11: Geographic distribution of Twitter postings (left); partitioning into five administrative areas defined by the City of Rome Mobility Agency (right).

The first analysis performed was the spatio-temporal aggregation of the data considering the telecommunication towers included in the different regions, the result of which is shown in Figure 12. Here it is possible to notice how all the regions follow the same global behavior over the hour of the days, having a decrease during the end of the month (November 2015).

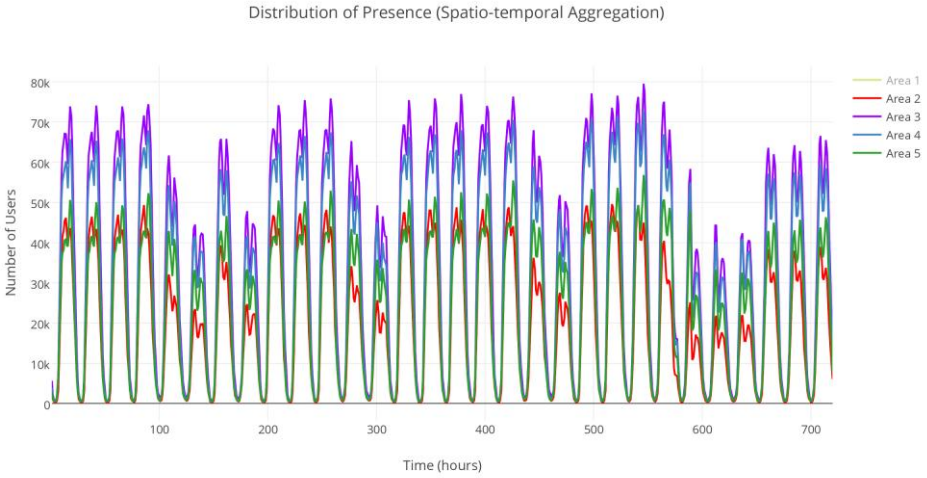

Figure 12: Result of the spatio-temporal aggregation. Each colored line represents the number of users in a specific area over time (City of Rome, November 2015).

Another interesting aspect is the fact that the Areas 1 and 2 have the same amount of calls, while the Areas 2 and 3 have a greater number of events, but the Area 5, which is the biggest area in size, is similar to the first two. Clearly this follows the fact that the real city is covered by the first four areas and the rest has a lower density of population.

A more informative result is shown in Figure 13, where the peaks over the network are reported. In practice it is possible to see when the number of people is larger or smaller than the typical scenario. During the month of October all the five areas follow a similar behavior reacting to the number of people that are present in the city.

In Figure 13 we can see that there are three distinct timeslots during the month of October where the mobile network traffic is outside of the typical range: from the 1st to the 4th hour it is lower, from the 8th to the 10th hour it is higher and from the 16th to the 18th hour it is higher again.

Figure 13: Peak Detection results in the five administrative areas of the City of Rome (October 2015).

It is important to notice that several peaks are caused by a general increase in the number of people and therefore they did not identify with a real event. For this reason we computed the relative deviations as the deviation from the median of all the peaks detected in all the areas in a single hour. A high value characterizes the peaks occurring in a specific area which are relevant with regard to the global population. The result is shown in Figure 13.

From this analysis it is possible to detect peaks generated for some *events* in a specific region. Four of those are highlighted with circles in Figure 14:

- ***A.*** 3/10, Saturday between 22:00 and 02:00 in Area 5, Positive.
- ***B.*** 10/10, Saturday between 16:00 and 02:00 in Area 1, Positive.
- ***C.*** 10/10, Saturday between 20:00 and 02:00 in Area 5, Negative.
- ***D.*** 17/10, Saturday between 11:00 and 22:00 in Area 4 and 2, Positives.

In the case of the peaks *B* and *C* a possible explanation is the festival called "Cucine di strada per le vie di Roma" (Street food on Rome streets) that took place in the city center, which is well known to attract a lot of locals, explaining also the negative peak in the Area 5 which are probably people who moved to the city center until late night.

Clearly this is the initial step of the analysis towards the Event Detection service which will use other external sources and indicators to discover and semantically enrich those peaks.

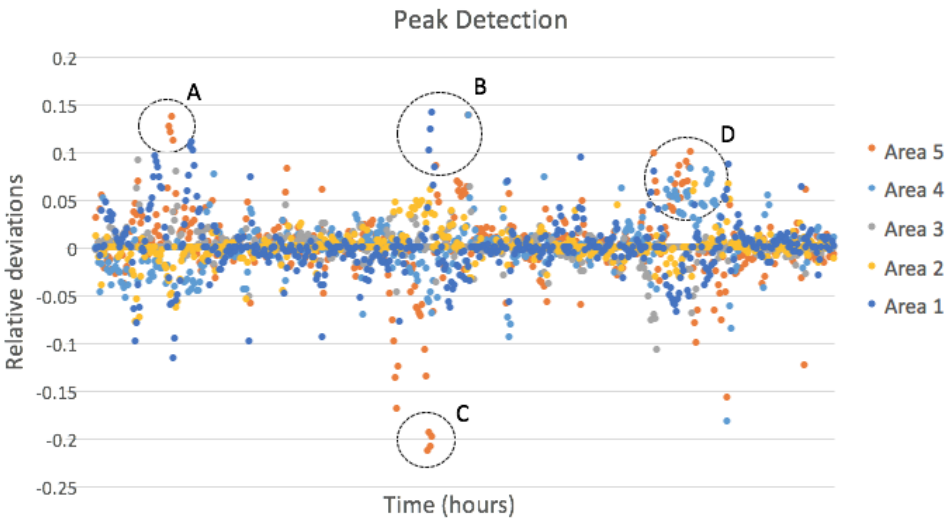

Figure 14: Peak Detection relative deviations (City of Rome, October 2015).

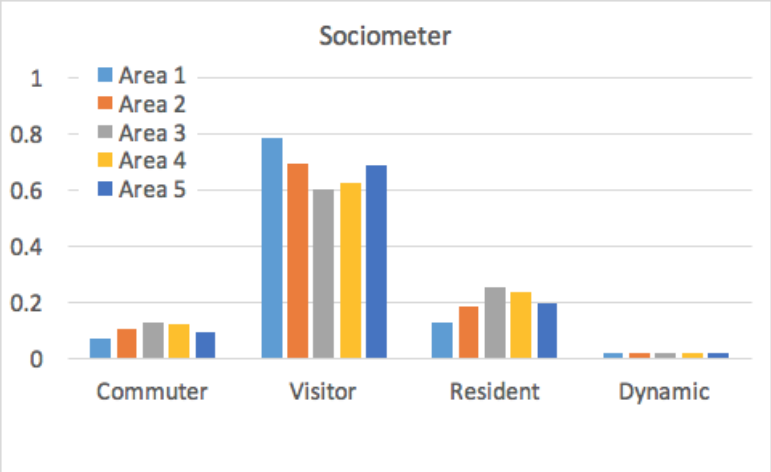

Figure 15: Results of the Sociometer in a single month (City of Rome, November 2015).

From peak detection to the classification of the users, with the *Sociometer* it is possible to estimate the kind of users present in the different areas. In Figure 15 the composition of the population during November 2015 is shown. Interesting is the fact that there is a high number of users *Passing by* in all the areas, this means that almost 50% of the users seen in the city are people producing only a single event (call or SMS). This is also confirmed by the Mobility Agency that considers them as people who come to Rome for business (city center) or pass through the city traversing the highways (outer areas). For the *residents* the different distributions confirm that the guess about the density of populations seen before is correct. In fact the Areas 2 and 4 have the higher number of *residents*, but looking at the *commuters* it is possible to notice that those areas are attractive even from the point of view of workplaces. In fact those areas have the higher number of people moving there for work. This is also highlighted by the small number of *dynamic residents*, which means that people tend to move in the same area where they live, in particular for the Area 3.

Moreover looking at the *visitors* it is possible to see that the first two areas have the higher number of *visitors*, in fact they are the areas where usually the tourists focus their attention. The number increases again for Areas 4 and 5 due to the fact that the highways are there and a lot of users "visit" the place just for passing through (approximately 30%). Those results should be investigated further and for this reason we contacted the City of Rome Tourism Office in order to better understand the configuration of the areas with their attractions and peculiarities.

Focusing the analysis on a single area it is possible to see how the proportions of different classes change over time. Analyzing Figure 16 it is possible to see how the trends decrease for *visitors* (leading to an increase for *residents* and *commuters* due the fact that they are proportions) in particular between October and November. This is interesting because it highlights the fact that during the beginning of the Jubilee (December 8, 2015) there was an increase in the number of *visitors* (less than 2%) which is not the huge increase predicted by the experts.

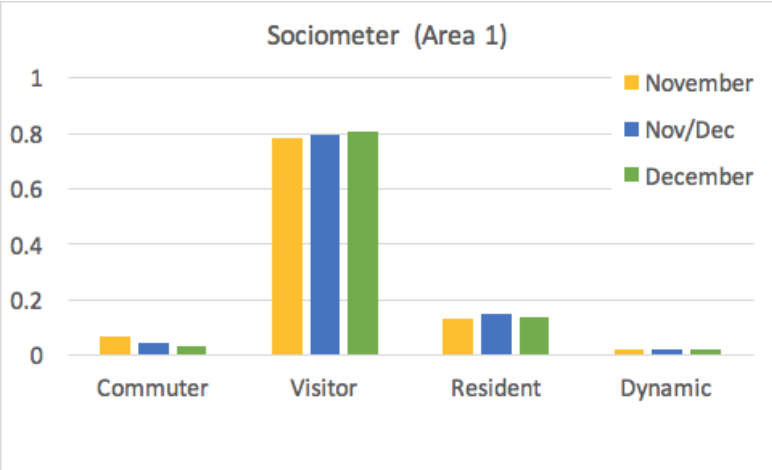

Figure 16: Results of the Sociometer for the city center considering a sliding window of two weeks.

This is confirmed by the City of Rome Mobility Agency which declared: "Due to the organization of this Jubilee, which is not concentrated only in the City of Rome, and the fact that there was a very high risk level after the Parisian terrorist attack (November 12, 2015), there have been practically no visitors, as if the Jubilee never started". Moreover we can see how the *commuters* decrease going toward the end of December, because the people stop working for the holidays, becoming *residents* for the classifier or they are no longer taken into account by the classifier if they move out of the city.

References

[1] *Analysis of GSM calls data for understanding user mobility behavior.* **Barbara Furletti, Lorenzo Gabrielli, Chiara Renso, Salvatore Rinzivillo.** Big Data, 2013.

[2] *City users' classification with mobile phone data.* **Lorenzo Gabrielli, Barbara Furletti, Roberto Trasarti, Fosca Giannotti, Dino Pedreschi.** Big Data, 2015.

[3] *Discovering urban and country dynamics from mobile phone data with spatial correlation patterns.* **Roberto Trasarti, Ana-Maria Olteanu-Raimond, Mirco Nanni, Thomas Couronné, Barbara Furletti, Fosca Giannotti, Zbigniew Smoreda, Cezary Ziemlicki.** Telecommunications Policy Journal, Volume 39, Issues 3–4, 2013.

[4] *Privacy-by-design in big data analytics and social mining.* **A. Monreale, S. Rinzivillo, F. Pratesi, F. Giannotti and D. Pedreschi.** EPJ Data Science, 3:10, 2014.

[5] *Assessing the Privacy Risk in the Process of Building Call Habit Models that Underlie the Sociometer.* **Furletti B., Gabrielli L., Monreale A., Nanni M., Pratesi F., Rinzivillo S., Giannotti F., Pedreschi D.** Technical report CNR ISTI, Italy, 2014.

[6] *Article 6.1(b) and (c) of Directive 95/46/EC and Article 4.1(b) and (c) of Regulation EC (No) 45/2001.* European Union for Protection of personal data.

[7] *Anonymity: A Comparison Between the Legal and Computer Science Perspectives.* **Sergio Mascetti, Anna Monreale, Annarita Ricci, Andrea Gerino.** European Data Protection: Coming of Age 2013: 85-115.

[8] *Visualizing Statistical Linked Knowledge Sources for Decision Support.* **Adrian M.P. Brasoveanu, Marta Sabou, Arno Scharl, Alexander Hubmann-Haidvogel, Daniel Fischl.** Semantic Web Journal 2016: Forthcoming.

[9] *Scalable Knowledge Extraction and Visualization for Web Intelligence.* **Arno Scharl, Albert Weichselbraun, Max Göbel, Walter Rafelsberger, and Ruslan Kamolov.** 49th Hawaii International Conference on System Sciences (HICSS-2016). Kauai, USA: IEEE Press. 3749-3757.