Use Case Requirements

Deliverable no.: **9.2**

28 April 2015

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

Deliverable Title	Use Case Requirements
Filename	asap_d9-2_final_revised_y1-review.docx
Author(s)	R. Bertoldi (WIND)
Date	28 April 2015

Start of the project:	01 March 2014
Duration:	36 months
Project coordinator organization:	FORTH
Deliverable title:	Use case requirements
Deliverable no.:	D9.2
Due date of deliverable:	28 April 2015
Actual submission date:	---

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
Final	28 February 2015	R. Bertoldi
Revised-Y1-Review	28 April 2015	R. Bertoldi

Abstract

The present deliverable D9.2 ”Use case requirements” describes the requirements for the Telecommunications Data Analytics application that will be developed and demonstrated by WIND. These are based on the preliminary version described in the Deliverable D1.1 ”Early user requirements”. Three applications will be described in 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 realize TDA applications for city administrations and telecommunication company users, in particular they will study the city dynamics, tourism behaviours and the car sharing opportunities in a specific area. Additional information gathered from Social Networks will be used to enrich the analysis.

Keywords

Telecommunications 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.

Contents

List of Figures

1 WP9 Introduction

The main objective of this Work Package is the design and development of an analytics application on WIND telecommunications customer data (Telecommunications Data Analytics, TDA), targeted towards tourism and mobility scenarios applications. The use cases will be integrated into the ASAP framework.

1.1 T9.2 Task Description

In this specific Task T9.2, which aims at producing Deliverable D9.2, starting from the Deliverable D1.1 "Preliminary use cases requirements" a more organized and clear view of the applications will be provided with the focus on the data interactions required.

2 Telecommunications Data Analytics

Generally speaking Big Data refers to the huge and increasing volume of the data available and the ways it can be processed. It is also an innovation that businesses can use to drive competitive advantage, but the value of big data lies in the insights that the companies should draw from it, rather than in the information itself. Big Data comes in many forms, structured or unstructured, and it may be generated by organizations themselves or obtained from third parties.

Analytics is the means for extracting value from this data by generating useful insights. Without analytics, businesses have no way of using their big data to establish competitive advantage. The availability of the new ASAP platform, with its high-level API, will open new possibilities in the way to develop and deploy a new generation of mobile applications, leading to a new Big Data approach for such new applications.

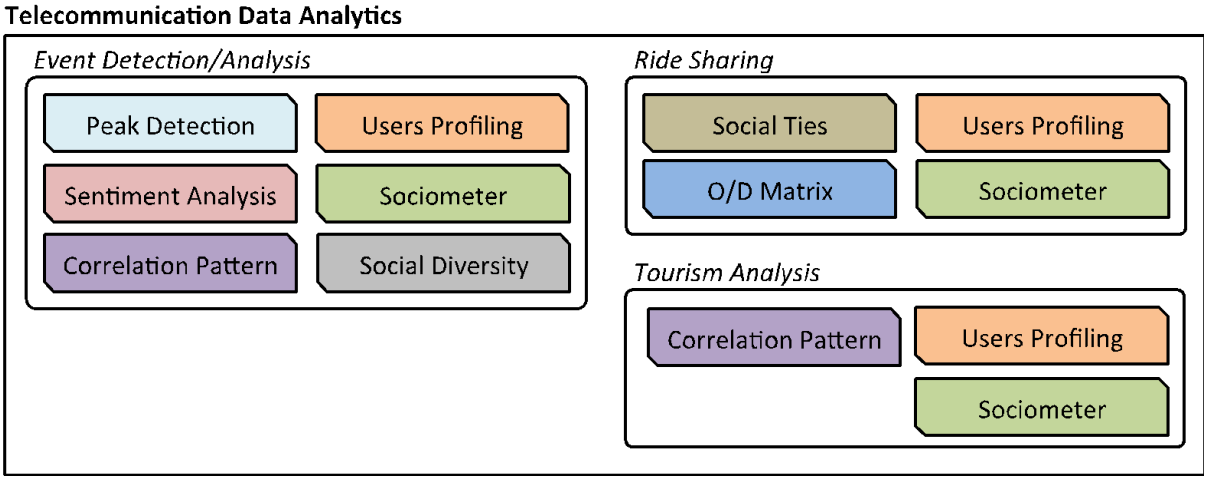

Figure 1: The Telecommunications Data Analytics applications

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 TDA application will show how a number of analytical services describing the mobility of people can be created on the basis of the structured data collected by the mobile network during routine operation. In particular three applications will be presented:

i. Event Detection,
ii. Ride Sharing, and
iii. Tourism Analysis.

Each one can be decomposed in basic analyses as shown in *Figure **1***. The requirements for each use case will be analyzed in order to understand the interaction with the data available.

2.1 Data Description

The data required by the use cases is described in the following considering three classes: structured, semi-structured and unstructured.

2.1.1 Structured Data

Customer Relationship Management (CRM) data is the clients personal information such as age, sex and billing address. This data will be fully anonymized prior to being used in order to preserve the privacy of the customers as required by the current national and European privacy protection laws [Euro 2014].

2.1.2 Semi-Structured Data

Call Data Records (CDR) related to voice, SMS and mobile traffic data used for billing purposes. This dataset will be a monthly snapshot of customers traffic in the interested area of research. Each record of this dataset contains information about the called and the callee enabling the study of the social network between the telecommunication company clients.

2.1.3 Unstructured Data

Social networks posts extracted from a specific event which can be used to evaluate the opinion of the users.

2.2 Use Cases

In this section all the basic modules needed to realize the applications will be introduced. In particular we will put a focus on the data queries and data processes required by each use case in order to highlight the basic requirement for the ASAP infrastructure.

2.2.1 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 grid, and each observation of an input dataset can be assigned to one of its cells. The number of observations that fall in a cell defines its density. The input data is partitioned into two sets: a training dataset and a test dataset. For both datasets the spatio-temporal grid of densities is 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.

Based on the densities obtained for each region and each timeslot over the training dataset, an expected density value is computed for each region, by averaging the densities measured at the same timeslot of all the periods in the time window covered by the dataset. For instance, we might obtain an expected density for each pair *(region, hour of the day)*, i.e., 24 values for each region, assuming 24 one-hour timeslots.

Then, for each region and each timeslot, the corresponding density is compared against its expected value: if the difference is significant, an event of form *(region, weight, timeslot)* is produced, representing its spatio-temporal slot and a discretized measure (*weight*) of how strong was the deviation. In particular, events are detected on the base of three parameters:

- a granularity of deviations, expressed as a percentage relative to the expected density
- a minimum relative deviation, also expressed as a percentage, used to select significant deviations
- an absolute minimum deviation, expressed as an integer number, used to discard extreme cases with very low densities.

The weights used in defining events will be multiples of the granularity, and an event for a region and a timeslot will be built only if the deviation of its density with regard to the corresponding expected density is larger than the absolute minimum deviation and in percentage is larger than the minimum relative deviation.

Requirements

In this case the primitives needed are the aggregation of the calls over the spatio-temporal partitions, an operator equivalent to the `DISTINCT` of the SQL and the possibility of performing `JOIN`.

2.2.2 Correlation Pattern

The extraction of correlation patterns (*C-Patterns*) focuses on those sequences of events that appear frequently and are highly correlated. In particular, *C-patterns* are computed as sequential patterns over the dataset of events using an algorithm called C-SPAM which extends the standard SPAM by introducing several spatial and temporal constraints. The most important constraints introduced and imposed in the *C-pattern* extraction is the minimum correlation value, in fact C-SPAM computes for each extracted *C-pattern* a correlation index, defined as follows: for a *C-pattern* $\langle D_1 , \dots , D_2 \rangle$,:

$$c - index(D) = \frac{supp(D)}{\prod_1^n \prod_{d \in D_i} supp(d)}$$

where $supp(D)$ (respectively, $supp(d)$) represents the support measure, i.e. the fraction of input sequences that contain the pattern D (resp., the single event d).

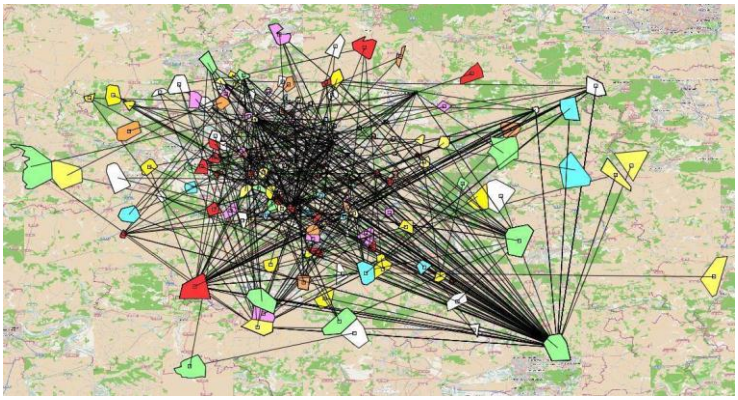

Figure 2: A set of C-patterns extracted in the Parisian area

This measure basically mimics the standard `lift` index defined for item sets, and expresses the ratio between the actual frequency of the pattern (D) and its expected frequency, computed under the assumption of complete independence between events. *C-patterns* with very high values of *c-index* can be considered surprising patterns,

therefore potentially interesting, while *C-patterns* with a very small *c-index* are probably trivial.

Requirements
The requirement here is the execution of a sequential pattern mining algorithm over the set of events sequences.

2.2.3 User's Call Profiling

The spatio-temporal profile is an aggregated representation of 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 GSM antennas 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 he is present in that area at that time.

*Figure **3*** 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 to the matrix new rows. In this case we have 3 timeslots (*t1, t2, t3*) so the matrix has 3 rows. 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 individual was present in the area of interest for 5 distinct weekdays during *Week1* in timeslot *t2* only.

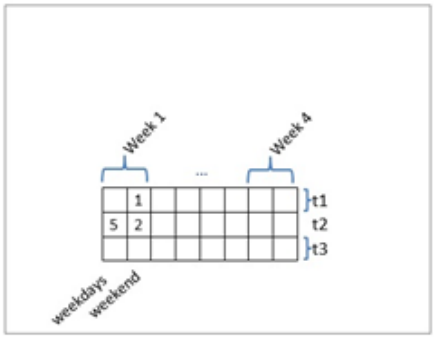

Figure 3: Structure of the aggregated spatio-temporal user's profile

Requirements
In this case the primitives needed are the aggregation of the calls over the spatio-temporal partitions and an operator equivalent to the `DISTINCT` of the SQL.

2.2.4 Sociometer

Exploiting the methodology called Sociometer [Furletti2013] 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 our mobility behaviour categories.

In particular we are interested to identify 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*.

Then, *k-means* is used to identify the representative profiles, which are then labeled with the mobility behavior categories just described. Notice that several profiles might be associated to the same category, basically representing different facets of the same class of users.

The classification phase assigns each spatio-temporal user's profile to the closest representative profile based on a proper distance measure. The output is the semantic enrichment of the set of users with tags representing the classification for each user.

This classification can be also compared with pre-existing information of the user in order to evaluate the method accuracy: e.g. having the billing address of the user it is possible to see if it corresponds to the information inferred by his/her behaviour or not.

Requirements
The requirement here is the *k-means* data mining algorithm which must summarize the profiles into a class of similarity.

2.2.5 O/D Matrix for Systematic Flows

The CDR data contains the information of the zone from which each phone call starts. Thanks to the large amount of data provided by the telephone operator it is possible to use the spatio-temporal footprint left by the users for the purpose of monitoring their movements in the territory. Several studies [Balázs 2013] assert that most people spend most of their time at a few locations, and the most important ones may be labeled as home and work.

In this section we will explain the methodology used for the extraction of such important locations, which we will call *L1* (most important one) and *L2* (second most important one) using the frequency of calls made by users. The location *L1* relates to the antenna from which the user made the greatest number of phone calls. For both technical and infrastructure reasons due to the load balancing of the antenna, it may happen that the serving antenna for different calls made at the same place, may be different, even though such antennas are usually close to each other. To mitigate this effect, we have redefined *L1* as a bigger area that also includes the adjacent cells. Once the two locations are discovered, the temporal information can be derived from the user's activities in the long period. The focus of this analysis is the detection of systematic movements considering two separated timeframes: a morning timeframe, and an afternoon timeframe, in which the users usually move, respectively, from home to work and from work to home. The first step is to identify the movements performed by individuals from *L1* to *L2* (*L1* → *L2*) and from *L2* to *L1* (*L2* → *L1*). It is important to notice that we are looking for movements between these two areas even if they are not contiguous, i.e. other areas were traversed between them, as shown in *Figure* ***4*** (*A* is distinct from *L1* and *L2*).

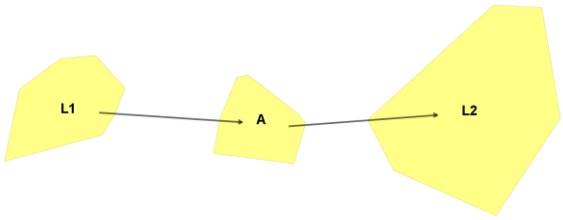

Figure 4: Example of flow from L1 to L2

The second step consists in selecting only the systematic movements, which is done by applying two different constraints: (i) request a minimum number of movements between the pair; and (ii) request a minimum value for the `lift` measure of the pattern *L1* → *L2*, which we define as:

$$LIFT(L1 \rightarrow L2) = \frac{P(L1 \cap L2)}{\left(P(L1) * P(L2)\right)}$$

`Lift` measures the correlation between *L1* and *L2*, resulting high if they appear together often with regard to the frequency of *L1* and *L2* taken separately. The main purpose is to normalize the frequency of *L1* → *L2* with regard to the frequency of calls of the user, since otherwise the candidate movements of frequent callers would be excessively favoured in the selection. The constraint on the number of movement is usually adopted in literature to exclude extreme cases where the `lift` (or other correlation or relative frequency measures) is not significant. In our case, after a preliminary exploration we chose to select only pairs that appeared at least 3 times. The threshold for the `lift` measure was chosen based on the study of its distribution, selecting the value where the slope of the cumulative distribution begins a sudden drop. As previously mentioned, the final goal of our analysis is the synthesis of O/D (Origin/Destination) matrices that summarize the expected traffic flows between spatial regions. Our O/D matrices will focus on systematic mobility, which represents the core (though not the only) part of the traffic.

Requirements

In this case the requirement for the ASAP platform is the possibility of searching a specific sequence (i.e. of *L1* → *L2*) over the sequence of places visited in each day.

2.2.6 Social Diversity

The diversity of individuals' relationships has been revealed as an important feature to determine social dynamics of a territory ([Eagle 2010]).

Given a social graph obtained from the CDR call graph, social diversity for an individual *i* is defined as the Shannon entropy associated with individual *i*'s communication behaviour, normalized by the total number of *i*'s relationships. Details on social diversity computation are given in [Eagle 2012].

Requirements

Essential requirement for social diversity computation for each individual of a given call graph is a function to evaluate the set of calls recipients for any individual, together with the calls volume for any recipient. Such function is an abstraction of the SQL constructs `GROUP BY` and `Count()`.

2.2.7 Social Media Sentiment Analysis

Sentiment analysis or opinion mining is the computational study of opinions, sentiments and emotions expressed in text. By taking advantage of specific social networks APIs, an

event manager has the possibility to collect posts and activity from the event's social network page. This leads to a classification task aiming to divide users into “promoter” and “detractor”. The former are the ones who expressed positive opinions while the latter are the remaining ones, who expressed neutral or bad opinions about the events.

Requirements
Requirements for this task are the execution of a preliminary training task in order to obtain a classification model, to be used with all the single comments collected from the event’s social network page.

2.2.8 Social Ties

Social ties have been widely identified as a high valuable information to be used in many fields, from social network analysis to decision support systems. CDR data contain this kind of information in the underlying call graph: from who-calls-whom graph we can assess, for each pair of individuals (*u, v*), the following aspects:

- How connected are *u* and *v* in the social network. For this purpose, we adopt several well-established measures of network proximity, based on the common neighbours or the structure of the paths connecting *u* and *v* in the who-calls-whom network.
- How intense is the interaction between *u* and *v*. For this purpose we use the number of calls between *u* and *v* as a measure of the strength of their tie.

One of the common measures adopted to accomplish this task is Adamic-Adar, based on the number of common neighbours between *u* and *v*, together with their network degrees. Details on Adamic-Adar and more other social ties measures are highlighted in [Wang 2011].

Requirements
Requirement to compute the strength of social ties between two individuals is the availability of SQL-like operation such as `JOIN` and aggregated functions like `Count()`.

2.3 Privacy Risk Analysis

Due 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 [Furletti2014] and already tested in the work presented at CPDP in 2013 [Mascetti2013]. This methodology implements and satisfies the constraints issued by the European Union for data protection in [Euro2014] and follows the principle given in [Euro2001]. 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 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. An example of study on GSM data is presented in [Monreale2014] and it will be used as a roadmap for the methodology that will be implemented in ASAP. In the following *Figure **5*** the general privacy risk analysis framework is represented, where 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.

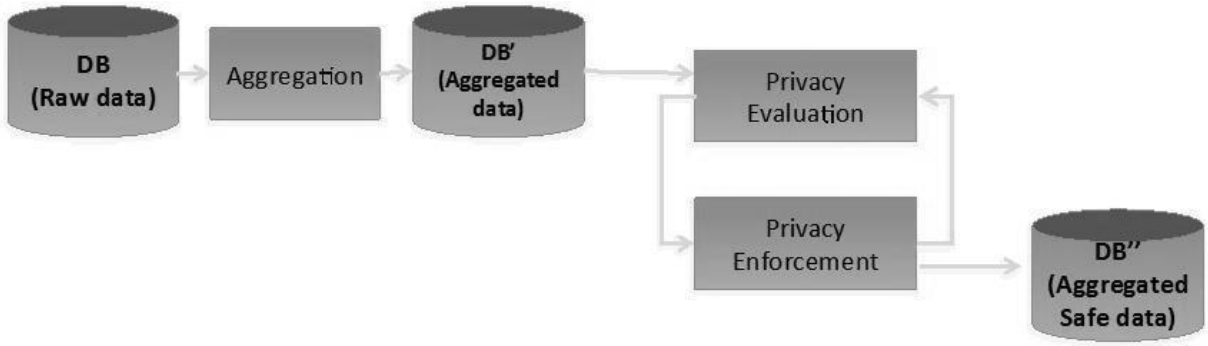

Figure 5: The privacy risk framework

It's important to notice that the aggregation function may be also an identity function, in this case the *aggregated data* is the same as the *raw data* detaching the process from the specific application but in this case the level of information in the database during the privacy evaluation is very detailed and therefore the enforcement may lead to stronger deletion or transformation.

To better explain the process in the following we illustrate the process instantiated on the *User Call Profiling* use case applied on a CDR dataset.

The question to answer is the following: Is it possible to infer private information about a user by accessing the set of SCT profiles? Is this form of data enough for protecting the individual privacy of each user in the system? If the answer is yes, we could have both individual privacy protection and perfect quality of the analytical results.

First of all, we should note that a user profile can be seen as a spatio-temporal generalization of the CDR data of that user. Clearly, this form of data is more aggregated with regard to the CDR logs because it cannot reveal the history of the user movements, the number of calls and the exact day and time of each call. Moreover, this profile is constructed by considering a specific area (such as a city), so it is impossible to infer where exactly the user went. The only information that can be inferred is that the user visited the area in a certain period. Clearly, to infer this kind of information on a specific user, two conditions need to be met: i) one needs to have access to a set of user profiles and b) one needs to have a priori knowledge on the user to find his/her user profile .

In the following, we define an "attacker" as a person who uses some a priori knowledge in order to derive new, sensitive information about a user.

We can assume that it is possible to possess different levels of a priori knowledge on a specific user. Clearly, more detailed a priori knowledge can lead to higher privacy risks, as it facilitates the re-identification of a user in the profiles. As a consequence, the analysis and the interpretation of the privacy risk for every user can be done with respect to the different levels of a priori knowledge:

1. *Exact knowledge of the mobile phone activities of a user concerning a defined period*. Assuming that an attacker, for a determined period *p*, knows the exact set of calls performed by a user between certain hours and in the investigated area.
2. *Uncertain knowledge on the user's mobile phone activity.* Assuming that an attacker, for a determined period *p*, has no exact knowledge on the phone calls of a user, but the attacker knows exactly where and when the user was in the investigated area

Starting from these two definitions of a priori knowledge, it becomes possible to compute the probability to identify a profile in the set of users' profiles in the database. At the end of the procedure, the data manager can evaluate the percentage of risk associated to each profile. If the number of profiles at risk is small, the data manager can isolate these profiles and use only the harmless profiles, without compromising the utility of the analysis. In other words, this procedure allows the selection of a set of profiles based on the risk you are or are not willing to run.

Applying this process to a set of CDR data in an Italian city covering one month the results are the following (see *Figure 6* and *Figure 7*):

Re-identification probability		
% Users	**A priori knowledge: 2 weeks**	**A priori knowledge: 3 weeks**
30%	$P \leq 0.004\%$ (|CI|=25,000)	$P \leq 0.006\%$ (|CI|=15,000)
40%	$0.004\% < P \leq 0.02\%$ (|CI|=5,000)	$0.006\% < P \leq 0.04\%$ (|CI|=2,500)
20%	$0.02\% < P \leq 0.2\%$ (|CI|=500)	$0.04\% < P \leq 0.4\%$ (|CI|=250)
9.4%	$0.2\% < P \leq 0.8\%$ (|CI|=125)	$0.4\% < P \leq 1\%$ (|CI|=100)
0.6%	$0.8\% < P \leq 25\%$ (|CI|=4)	$1\% < P \leq 50\%$ (|CI|=2)

Figure 6: Privacy Risk in the case 1

% Users	**A priori knowledge: 4 weeks**
10%	$P \leq 0.003\%$ (|CI|=33,000)
60%	$0.003\% < P \leq 0.017\%$ (|CI|=5,800)
30%	$0.017\% < P \leq 0.025\%$ (|CI|=4,000)

Figure 7: Privacy Risk in the case 2

Analyzing these results, we can notice that the profiles at risk are only the 0.6% of the total amount of users in the first case. The suppression of these high-risk users permits to release users that in the worst case are indistinguishable from a set of at least 100 other users. In the second case the attacker has an a priori knowledge which does not

lead to any risk: each person is guaranteed, as |C| values are always quite high. The worst-case scenario here is that a user is indistinguishable from 4,000 other users.

This example of analysis can be adapted to all the use cases presented and will be part of the pre-processing of the data in order to guarantee that all the data stored is safe. If the portion of data which is unsafe data is larger and cannot be removed, specific anonymization techniques, e.g. *perturbation,* may be used to transform the data.

Requirements
The requirements of the privacy risk analysis are two:

i. the possibility of performing search over the aggregated data,
ii. the possibility of executing a transformation considering as input the entire dataset or specific group of elements (i.e. group of users).

The aggregation of the data is application dependent and it will be described in the following section.

2.4 Applications

In the following sub-sections the three foreseen applications will be described as a composition of the previously presented use cases as illustrated in *Figure **1***.

2.4.1 Event Detection/Analysis

The event detection application is designed to let the user analyze different features of an event: spatio-temporal characteristics, social aspects and statistical properties. By controlling input parameters such as time-window, spatial area and CRM filters, the user is able to identify, thanks to a smart visualization, all the events occurred.

Storyboard: *the analysis of the last month of data provided by the telecommunication company covers a festival which lasts 4 days. The peak detection is able to spot 3 major peaks: the first during the grand opening of the event and the other 2 during the second and third days. The number of participants involved in the peaks gives the administrator an idea of how the mass of people distributes during the days of the festival which can be used for the next year. Moreover the analysis of the participants classification using the Sociometer is able to highlight that the first two peaks are mostly composed by visitors coming from outside the city. Looking at their social diversity the analyst discovers the existence of a large community coming during the second peak. These insights are used by the marketing to push specific offers to this large community of users.*

2.4.2 Ride Sharing

The ride sharing application is useful both for mobility managers and single drivers. A mobility manager can visualize all the routine trips of a specific input area, together with an optimized car sharing solution for those trips. A driver uses this application as a recommendation system to get information about the possibilities to share his/her trips.

Storyboard: *The user has the app installed on his/her phone which starts learning his/her behaviour from his/her activities. Once the learning phase is complete the app may suggest possible options to share the car with others anonymous similar users. The suggestion takes into consideration their social ties and the mobility behaviour. Once both the users give their permission to send personal information, the two are connected by the system.*

2.4.3 Tourism Analysis

The analysis of tourist flows is another valuable application for a mobility manager. From a map-based dashboard, the manager can identify the common movement patterns of tourist/visitor travelers. As for the applications introduced above, the users can provide spatio-temporal constraints as input.

Storyboard: *The mobility manager is able to visualize the interconnections between the principal touristic places by means of presence variations. Using this knowledge the system may highlight when a specific area becomes crowded and the possible consequences in the next hour. The dashboard in his/her hand will be a tool to preventively put in action measures to control the touristic sites in order to avoid problems or more critical situations.*

2.5 General Requirements

The development of the described applications implies the ASAP framework to match some general requirements in terms of data-manipulation operators. The following operations are fundamental for the use cases presented in the previous section:

- aggregation of the calls over spatio-temporal partitions; an operator equivalent to the `DISTINCT` of the SQL and the possibility of performing `Joins`
- sequential pattern mining algorithm over the set of events sequences; operator equivalent to the `DISTINCT`, `GROUP BY` and `Count()` of the SQL
- *k-means* data mining algorithm to summarize the profiles into a class of similarity
- search for a specific sequence (i.e. of $L1 \rightarrow L2$) over the sequence of places visited in each day
- text classification model

These high-level operations represent the primitives that the ASAP framework should provide to mobility analysis applications developers.

3 Visualization Tools

- **Event Detection/Analysis:** The geographic map [D6.2] will be a central component of the ASAP dashboard [D6.3]. The user can visualize *peaks* using an interactive trend chart [D6.1], and *correlation patterns* through color-coded position markers (e.g. reflecting social diversity) on the geographic map, using tooltips to show additional metadata [D6.2]. The Sociometer related to the map could be represented as a pie chart, either to be drawn directly on the map or synchronized using multiple coordinated view technology in a separate view. Using an advanced search dialog in conjunction with the date range selector of the ASAP dashboard [D6.1], users can specify the time interval and CRM-based filter criteria, e.g. *"visualize all the events for the period 1-15 January 2015 related to people in the age group 18-25 years old"*.

  The sentiment analysis conducted on social media postings that relate to the event could be represented by a line chart indicating the time-series of "promoter" and "detractor" users. If there is more than one event in a given time interval, the advanced search of the dashboard will allow focusing the results on this particular event.

  The *Figure* ***8*** shows a toy sample of the Event Analysis dashboard. Once the spatial area related to the peak is identified, the application is able to visualize related information.

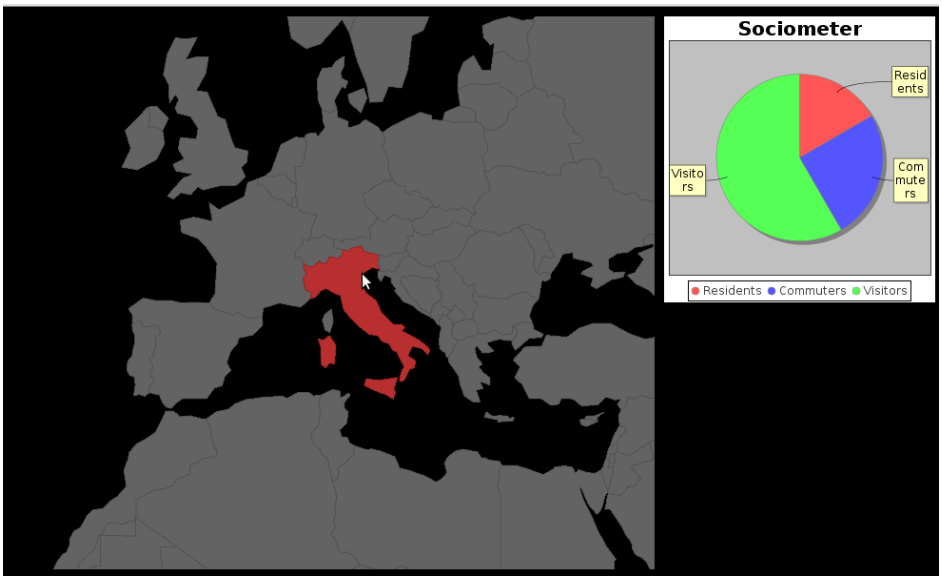

Figure 8: A toy sample of the Event Analysis dashboard

- **Ride sharing:**
  - *Mobility manager view*: a circular graph representing the O/D matrix, regarding the movements of commuters, extracted from Sociometer. For each O/D pair, a car pooling arrangement is given based on social ties between commuters: this could be visualized by simply indicating the pool of commuters when analyzing the O/D pair. Another interesting result is the percentage of shareable rides and the strength of the solution proposed (i.e. how optimal is the pool of travelers with regard to the strength of their ties).
  - *Driver view*: a map with the common user's routines and a list with suggested users to share the highlighted trip. Furthermore, a share possibilities list for an input trip – different from the default ones – could be useful.

- **Tourism Observation**: This is also a map-based dashboard, where the data to be visualized are correlation patterns of visitors, selected by Sociometer. Users will need to set a time interval and a specific spatial region as input for the visualization.

References

[Furletti2014] *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.

[Euro2014] *Opinion 05/2014 on Anonymisation Techniques*. **European Union for data protection.** April 2014.

[Mascetti2013] *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.

[Euro2001] *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.**

[Monreale2014] *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.

[Furletti2013] *Analysis of GSM calls data for understanding user mobility behaviour.* **B. Furletti, L. Gabrielli, C. Renso, S. Rinzivillo** Proceedings of Big Data 2013.

[Balázs 2013] *Exploring the mobility of mobile phone users.* **Balázs Cs. Csájia, Arnaud Browetc, V.A. Traagc, Jean-Charles Delvennec, Etienne Huensc, Paul Van Doorenc, Zbigniew Smoredae, Vincent D. Blondelc** Physica A: Statistical Mechanics and its Applications Journal 2013.

[Wang 2011] *Human Mobility, Social Ties, and Link Prediction.* **D. Wang, D. Pedreschi, C. Song, F. Giannotti, A.-L. Barabási** Proceedings of KDD 2011

[Eagle 2010] *Network Diversity and Economic Development.* **N. Eagle, M. Macy, R. Claxton** Science, 2010

[D6.1] *Deliverable D6.1 - InfoViz Services Early Design.* **A. Scharl, A. Weichselbraun, A. Hubmann-Haidvogel, and W. Rafelsberger** ASAP, February 2015.

[D6.2] *Deliverable D6.2 - InfoViz Services v1.* **ASAP**

[D6.3] *Deliverable D6.3. - InfoViz Services v2.* **ASAP**