Implementing privacy-preserving filters in the MOA stream mining framework

Degree Final Project, major in Computer Science

Monitoring report

Author:
David Martínez Rodríguez

Supervisor:
Jordi Nin Guerrero
Computer Architecture Department

Facultat d'Informàtica de Barcelona

Universitat Politècnica de Catalunya

January 19, 2015

Contents

1 Context

This project is framed within the broad field of *data mining*, emphasizing its relation with *data privacy*. We will provide now definitions and concepts of the main topics of the project's environment.

1.1 Data mining

Today's information society produces vast amounts of data all over the world. This data comes from innumerable sources and in diverse formats, and has been stored for years in data warehouses, waiting to be processed. With the continuous increase in computing power, due to the recent advances in software and hardware technologies, the machine learning field, more commonly known as *data mining*, has arisen, allowing us to exploit this stored data and distill knowledge from it.

Data mining is, indeed, a holistic process, where many different disciplines are involved, from data acquisition and storage, through its selection, filtering and analysis up to information extraction, visualization and knowledge discovery.

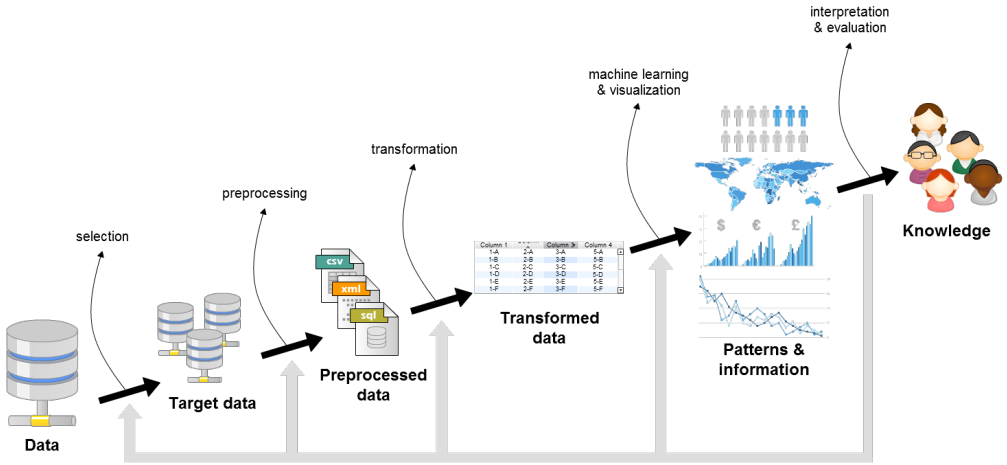


Figure 1: Data mining as a process, from data acquisition to knowledge discovery. Source: adapted from *From Data Mining to Knowledge Discovery in Databases* [6]

Data mining enables a better understanding of human or natural processes and provides us with means to identify trends, predict future events or discover useful patterns. Its uses range from scientific and medical applications to social sciences or business administration [6].

1.1.1 Facing the limits

Despite lots of effort is put into enhancing different data mining processes, there still are many cases where these techniques fail to perform correctly; mainly, it is a matter of scale.

On one hand, traditional data mining workflows cannot cope with the really massive data sets that are available nowadays, if performed on a common infrastructure. To solve this issue, clusters of hundreds or thousands of computers are used to run such analysis. It is costly and complex but, doing so, we can mine data that we couldn't some time ago.

On the other hand, we face another type of scaling problem. In some situations, data acquisition throughput is so high that it can't be stored anyway, so another approach is needed to avoid the loss of information that it could deliver us. Moreover, we might not want to store it, even when we could, but yet we want to analyze it to extract knowledge from it, as soon as we received it. Both scenarios are addressed with a series of techniques known as *stream mining*.

1.1.2 Stream mining

Stream mining or data stream mining is a process that allows us to still discover knowledge and patterns in data, even when it comes in the form of a continuous stream, or many of them [15]. Instead of processing all statically stored data, like traditional data mining does, a relatively small portion of it is kept during the analysis, and it is updated when needed - either because more resources are available to the system or because new data is acquired. A more deeper review of this research area is given in section 1.4.

Stream mining in this project:

MOA, initials for Massive Online Analysis, is an open source framework for data stream mining [20], originally developed at the University of Waikato, New Zealand. It includes several machine learning algorithms[1] to perform the analysis and tools to evaluate the quality of the results. It also deals with a problem known as *concept drift*[2]. It is related to the Weka[3] package, but it is built to perform at a greater scale for more demanding problems.

[1] Algorithms used to perform the actual data mining analysis (the "machine learning & visualization" step on figure 1) belong to the field of machine learning. In MOA, clustering, classification, regression, outlier detection and recommender systems are available.

[2] It is said of statistical properties of a target variable being analyed, when they change over time in unforeseen ways.

[3] Weka is a popular software package including classical data mining algorithms, this is, not stream mining. It is also developed at the University of Waikato. [21]

1.2 Privacy

Privacy is a concept that can be defined as the ability of an individual or group to seclude[4] themselves, or information about themselves, and thereby express themselves selectively. It is understood differently depending on the social and cultural background of each individual, but it is in fact recognised as one of the most fundamental rights of our human nature. The Universal Declaration of Human Rights' 12th article [1] states that:

> No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.

This right has been continuously violated ever since information exchange and advanced communication technologies have been developed. Despite this did not begin with the spread of the Internet, its adoption has greatly magnified both the ability to breach people's privacy and the impact that these breaches have.

1.2.1 Privacy leaks consequences

We are fully aware of the dramatic consequences that information leaks have caused these past years; for example, the *PlayStation Network* outage which left millions of video gamers without access to this service [16] or the more recent celebrity intimate photographs leaks [3]. Even though they are referred to as "IT security breaches" in the media, they are indeed privacy right violations, because sensitive personal information was compromised in both examples.

Economical and social consequences are at stake too: identity thefts are performed on a daily basis thanks to the vast amount of data about individuals available in the web. Back in 2005, the cost of these thefts was estimated to be in the order of billions of dollars [17].

1.2.2 Legal framework

Efforts are being carried out to develop legal frameworks to help protect people's privacy, at many levels. One such example is the spanish LOPD[5] [2], a law that, among other things, defines different data privacy levels and, for each of them, mandatory proceedings, protocols and data protection methods.

There are some pitfalls to these legislative efforts, though. Firstly, it is really hard to assess their accomplishment in the IT sector and, thus, it is sometimes a matter of

[4] "Seclusion is the act of placing or keeping someone away from other people." Source: Merriam-Webster online dictionary [11]

[5] LOPD stands for *Ley Orgánica de Protección de Datos*, a law that was approved by the spanish courts in 1999. It has been modified several times, being the law enforcement regulation approved in 2007.

confidence in the developers good practices and ethic behaviours. Another important drawback is that online services, such as social networks, can be accessed globally, but, on the other hand, their legislative framework is that of the country to which the backing company offering the service belongs to - jurisdiction definition in the Internet is still a matter of intense debate, nowadays.

1.2.3 Privacy and data mining

Data mining technologies have also become a relevant debate topic nowadays, concerning what information is collected from individuals, who owns it and what are the purposes behind its gathering. Information technologies deliver us many benefits at many levels - safer streets, cheaper communications, better health systems, more convenient shopping - but at the high cost of losing our privacy.

Knowledge discovery processes need data to work and, in most cases, it is sensitive and personal. Moreover, it is massively collected and stored and analyzed without us knowing much about it. Besides the lack of consent in this data acquisition stage of the process, data mining poses a bigger thread on individuals: information disclosure. Sensitive data must be treated accordingly, which involves not only good IT security practices to avoid leaks like the ones described before, but a responsible treatment when research results are published.

Statistical Disclosure Control (SDC) is the name that the statistical community has given to what the data mining community calls Privacy-Preserving Data Mining (PPDM). This field, whatever its preferred name is, deals with controlling that information about specific individuals is not extracted from statistical summary results. Also, if full datasets are to be released, PPDM methods should be applied to data in order to preserve user's privacy, whilst maintaining the statistical significance of it, i. e., the amount of information - knowledge - that this data can provide.

Further details of SDC methods and approaches can be found in the State of the Art section (section 1.4).

1.3 Involved stakeholders

To further understand the motivation driving this project we can examine which third parties are involved in it, giving a general overview of their possible interests in this work.

1.3.1 Users: organizations and developers

The most direct users of the results of this project will be developers, data scientists[6], researchers and the companies and organizations that employ them. Using privacy preserving filters they are allowed to release their results or data sets without compromising their users privacy. National statistical agencies[7] and private enterprises might find it useful as well.

1.3.2 End users: data providers

Data providers (anyone that allows collection of their data) get the most important benefit from privacy preserving data mining: their data is secure against information disclosure attacks. Because all kinds of sensitive data is kept and analyzed, this happens to be the main motivation for the project - even though data providers might not be aware of the specific technologies that keep their data safe, they still have the right to demand this protection.

1.4 State of the art

1.4.1 Stream mining

Data stream mining is a relatively new field. Even though its theoretical foundation is based in well-established statistical and computational approaches, it has not been until recent years that this research area has experimented such growing interest.

The main problem when dealing with streaming data is the high throughput of data being analyzed, under computational resources constraints. Variable data rates is another problem that has to be addressed too. Once these problems are resolved, efforts are done so the same kind of data mining analysis as in the case of batch data processing are available: classification, regression or clustering tasks, as well as outlier detection and recommendation systems. We will not cover these techniques here, because they are not related to this project, by themselves. Instead, we will have a look at some different stream mining solutions, because their working principles do affect the way the project's algorithms will be implemented.

Solutions provided in this field can be categorized into *data-based* and *task-based* ones [10], depending on their approach.

- **Data-based stream mining solutions:** The idea behind these solutions is to use

[6] The term *data scientist* is used to designate the evolution of data analysts or business analysts job titles in the world of *Big Data* or data mining. They are trained in computer science and its applications, modeling, statistics, analytics and mathematics.

[7] The ONS (Office for National Statistics, in the UK) team is using statistical disclosure control methods as part of their process methodology [7].

a subset of the original dataset to perform the required analyses. Diverse techniques that have been used in this sense can further be split into two more categories:

  - *Sampling methods:* either by randomly picking samples of the data stream or by randomly selecting chunks (subsets) of the stream, sampling methods discard part of the incoming data, while performing the knowledge discovery processes with the sampled data. The main problem with this approach is that is hard to know when to pick a sample or which records should be stored, because there is no previous knowledge of the dataset size or its information structure.
  - *Summarizing methods:* they use aggregated data or calculated statistical measures (that are continuously recalculated) to provide the information needed for the data mining algorithms. In this case, it is the loss of information and accuracy and the inability to control data distribution fluctuations what renders these methods not so usable as it was desired.

- **Task-based stream mining solutions:** The solutions that fall into this category are based not on performing data transformations, but on changing the data mining methods to enable their use on data streams.
  - *Approximation algorithms:* these are a kind of algorithms that are designed to solve computationally hard problems, by giving an approximate result. Instead of computing exact solutions, they just guarantee a certain error bound. The problem with these methods is, again, the high received data throughput, which they cannot cope as well. Additional tooling is therefore needed if one wishes to use them.
  - *Sliding window method:* this method, a common pattern in many online[8] applications, maintains a *sliding window* in which the most recent data is kept. As data is received from the incoming streams, this window "advances" so new observations are kept inside. The data mining analyses are then performed using the data available inside the window and summarized versions of the older records, in the form of statistical measures or aggregated data. This particular method is the one that the MOA package uses - thus its name: Massive **Online** Analysis. This solution scheme enables dealing with concept drift, which would not be possible if just aggregated data was used.
  - *Algorithm output granularity:* this method is a resource-aware data analysis approach that can perform the local analysis on resource constrained devices, by adapting to resource availability and data stream rates - when resources are completely running out, the results are merged and stored.

Data stream mining software: because it is an incipient field, stream mining

[8] In computer science, an *online algorithm* is one that can process its input piece-by-piece in a serial fashion, i.e., in the order that the input is fed to the algorithm, without having the entire input available from the start.

software packages are quite uncommon. Even though specific applications have been developed [12], MOA remains as one of the few generic[9], free and open sourced systems. In relation to MOA, a new project called SAMOA [22] is being developed too, based on top of MOA itself, and a couple of streaming processing engines: Apache S4 [8] and Apache Storm [9], developed by the Apache Software Foundation.

1.4.2 Privacy preserving stream mining

Many different methods have been developed to help prevent information disclosure when data mining datasets or results are released. These algorithms pursue the generation of results or data that have particular properties concerning privacy preservation.

Privacy preserved data properties: some of the desirable properties of privacy-protected data are:

- First described in 2002, by Latanya Sweeney, a release of data is said to have the *k-anonymity* property if the information for each person contained in the release cannot be distinguished from at least $k-1$ individuals whose information also appears in the release [18].
- The evolution of the concept of k-anonymity is *l-diversity* and adds further privacy preservation by adding intra-group diversity, so to avoid the flaws of the k-anonymity privacy model [13].
- Further on, the *t-closeness* property definition adds attribute-based privacy enforcement to the l-diversity model: to better preserve privacy, all values (all observations) from a particular attribute must not be too much different - instead, they should be close up to a certain threshold [14]. This is needed to preserve the privacy of those records that are more easily identifiable because their attribute values are more distinguishable.

Privacy preserving algorithms: the algorithms being used nowadays to achieve effective privacy preserving properties to datasets can be categorized into the following groups [5]:

- *Non-perturbative data masking:* these kind of methods do not perform data values transformations. Instead, they are based in partial suppressions of records or reductions of detail of the datasets. Some examples[10] are:
  - Sampling
  - Global recoding
  - Top and bottom coding

[9] MOA is not focused towards any particular application scenario: it is a base tool with which we can build such specific systems.

[10] We can't cover every algorithm in detail, because some of them are not relevant and because those which are to be implemented in this project will be described in detail in the final project report.

  - Local suppression

- *Perturbative data masking:* these methods do release the whole dataset, if required, but it is perturbed, this is, values are changed by adding them noise. This way, records are diffused and reidentifying individuals is harder. Some examples are:
  - Noise masking
  - Micro-aggregation
  - Rank swapping
  - Data shuffling
  - Rounding
  - Re-sampling
  - PRAM
  - MASSC

Privacy preserving *Stream* Mining: many of the previously listed methods are already implemented in many classical data mining frameworks and software systems; for example, the `sdcMicro` package for the **R** statistical package [19]. However, privacy preserving methods are still not widespread in the stream mining ecosystem - that is another motivation for this project.

1.5 Environmental impact

No relevant direct environmental impact is related to this project, neither tied to its development nor its further deployment. No use of massive resources is done and the results of the work will not, presumably, result in a significant environmental change of any kind.

It is still true, however, that data mining, as a discipline and its broad use, does consume a lot of resources, in terms of technological infrastructure and energy. We cannot forget that collecting, storing and processing data at the industry scale needs entire data centers fully dedicated to the data mining process. Power consumption is a big concern with nowadays information technology, as it is the huge amount of rare materials that electronic devices contain. These are derived or indirect effects of the data mining process. This issue deserves to be examined more closely, in the project's final report.

1.6 Social impact

Concerning social impact, this project's strength is related to users privacy preservation through the implementation of algorithms that help anonymize their data within the data mining process. This topic will be explained more extensively in the project's final report, but it is of paramount importance nowadays. Privacy is being relegated to the

background with the advent of new information technologies, devices or platforms such as social networks or banking applications. Data is gathered from everyone and there is an increasing need of methods to protect it. Together with other fields more focused in securing the access to data, privacy-preserving data mining is designed to keep data safe at the analysis stage of the data mining process.

Not only ethical concerns are addressed by protecting the users' privacy, but economical issues too. Industrial-scale information theft has a huge impact on enterprise economies, because of distrust and because disclosed sensitive data can be used to make profit of it.

2 Project management

We present herein a discussion of all deviations from the initial project constraints concerning its management.

2.1 Scope

2.1.1 Requirements analysis

There have been no major changes in the scope of the project. Both the functional and non-functional requirements sets remain the same as the ones defined in the final report of the Project Management course.

We provide now a summary of the current state of the project, in terms of what requirements have already been fulfilled throughout its development phase (see the Schedule section 2.2):

- **Functional requirements:**
  1. **[In Progress]** Implement privacy preserving stream mining *filters* for the MOA stream mining framework. The suggested algorithms to be implemented are:
     a) **[Completed]** Noise addition [5, p. 54]
     b) **[Not started]** Multiplicative noise [5, p. 57]
     c) **[Completed]** Microaggregation [5, p. 60]
     d) **[Completed]** Rank swapping [5, p. 72]
     e) **[Not started]** Rank shuffling [5, p. 73]
     f) **[Not started]** Differential privacy [4]

Concerning the non-functional requirements of the project, there are some of them that should be emphasized from now on, in order to correctly achieve them at the end of the development phase.

- **Non-functional requirements to be emphasized:**
  1. **Test coverage:** measures and tests will be performed to assess the quality of the developed software, as well as its scalability and performance, which is paramount in this project's context.
  2. **Documentation:** MOA is an *open source* data mining framework, which means that its community can assess how is it built and how to improve it. One of the benefits of the open source development model is that software can be safer, more robust and efficient, by receiving contributions from different

developers. If people are to continue improving the work done, it has to be well documented.

2.1.2 Methodology

Even though no significant changes have been made to the methodological framework being used, an agreement has been reached to apply it *more rigorously*, in order to enhance the project monitoring and to avoid further schedule deviations.

In this respect, the usage of the Trello[11] task management system will be emphasized from now on, as it provides all necessary features to overcome the detected methodological errors.

2.2 Schedule

There have been significant deviations concerning the initial project schedule. Not only the global duration has been lengthened, but more phases have been layed out, as was needed. As a positive contrast, early detection of such alterations has been sometimes possible.

2.2.1 Overall duration

The original total duration has been extended from 5 months to 8 months, approximately. Thus, the final report and its defence is now scheduled to be in April, which is the next available lecture shift in the Faculty. We believe that this extended duration will allow us to fulfill all requirements defined in the scope of the project.

2.2.2 Deviation analysis

There are several possible reasons behind this schedule deviation:

- The Project Management module lasted longer than expected, forcing the development phase of the project to begin later.
- During the definition of the project initial schedule, we expected to begin developing it while the Project Management module endured, which was, definitely, a planning error. Such tasks concurrency was not possible at that time.
- At the beginning of the development phase, we explored different technological alternatives, before deciding which approach was mostly suited to our needs, but this exploration delayed the actual development process for a couple of weeks.

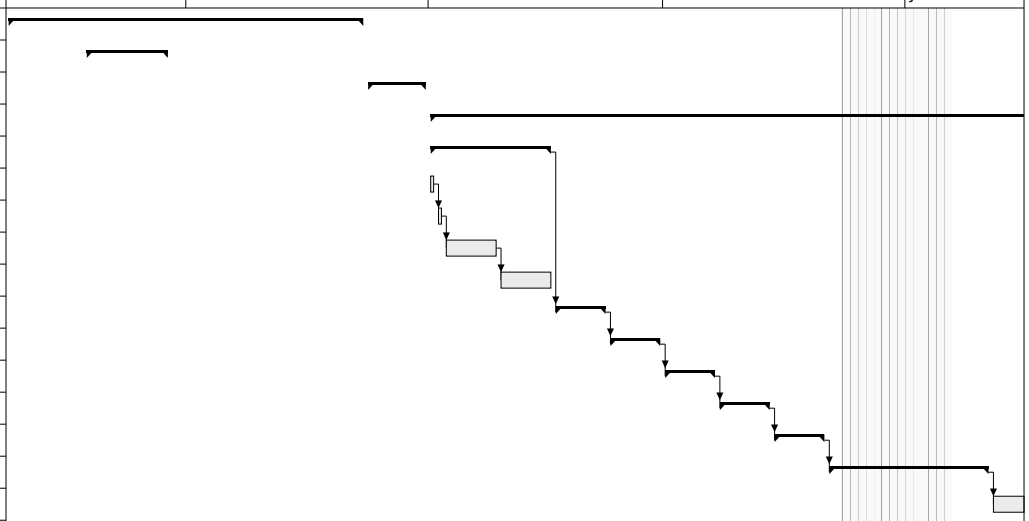
[11] Available at: `https://trello.com/`

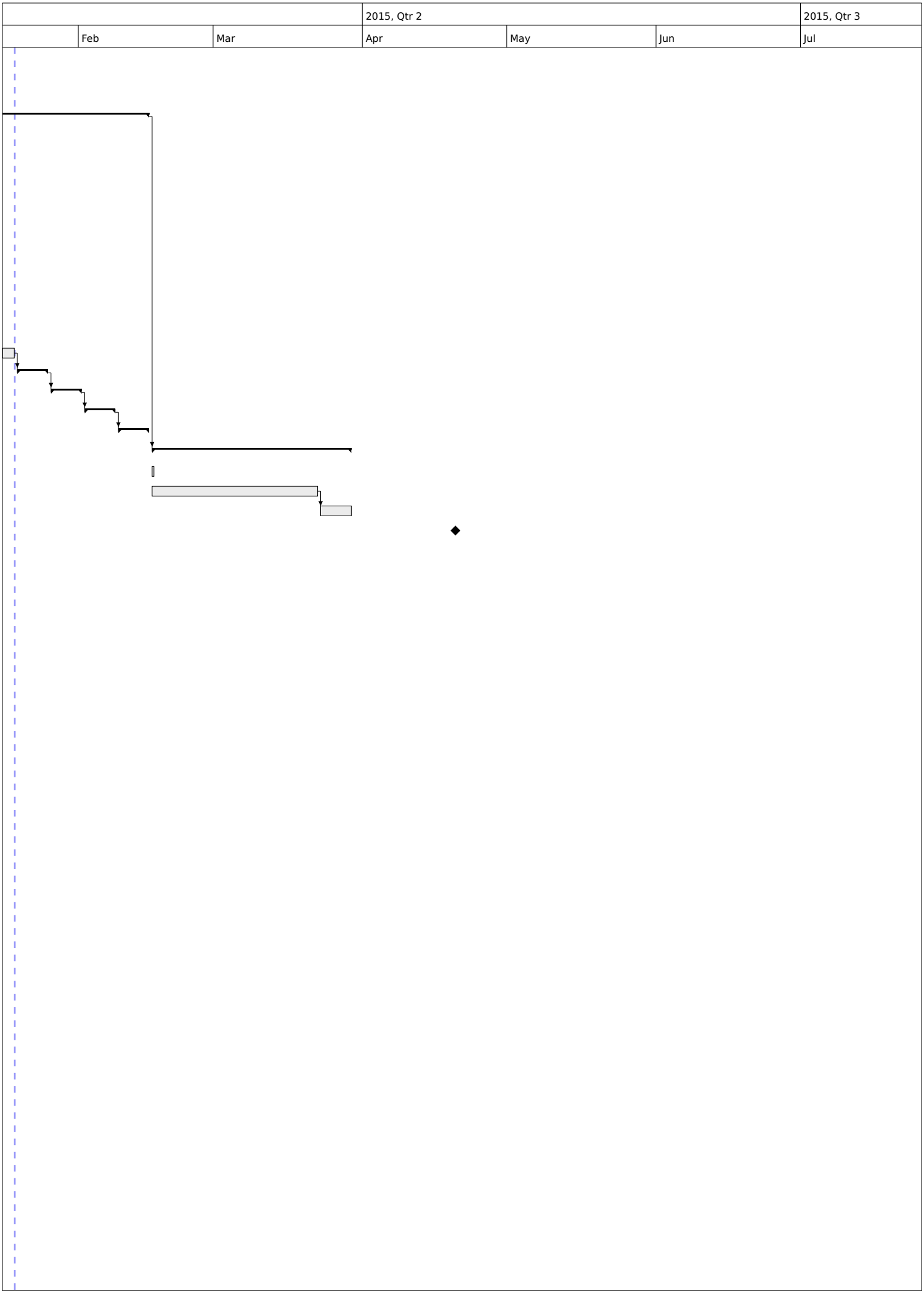
- As was already stated in the Project Management report, some of the requested features have posed to be more complicated than was expected, consuming some more time than that assigned to them.
- For personal reasons, no work could be carried out during the Christmas vacations, which lasted two more weeks, furtherly delaying the project's development.

2.2.3 Current detailed schedule

Considering the previous analysis, a new Gantt chart has been built, with the new project's schedule, which is detailed in the following pages.

Name	Work
GEP module	**55d**
Contextualization	**14d**
Environment setup	**8d**
Development	**93d**
Alternatives exploration	**16d**
sdcMicro exploration	1d
R-Java bindings options	1d
JRI	7d
Renjin	7d
Sprint 1 - Noise Addition	**7d**
Sprint 2 - Microaggregation	**7d**
Sprint 3 - Microaggregation framewor	**7d**
Sprint 4 - Rank Swapping	**7d**
Sprint 5 - Privacy Filter framework	**7d**
Sprint 6 - Rank Shuffling	**7d**
Monitoring report	7d
Sprint 7 - Differential Privacy	**7d**
Sprint 8 - Benchmarks & Results	**7d**
Sprint 9 - Test & Packaging	**7d**
(Optional) Sprint 10 - Results publicat	**7d**
Documentation	**43d**
Project deployment & finalization	1d
Project report	35d
Presentation & rehearsal	7d
Project defense	




2015, Qtr 2
2015, Qtr 3
Feb
Mar
Apr
May
Jun
Jul

3 Current project status

We will now assess the current status of the project, giving a more thorough overview of the job done and of that to be carried out.

3.1 Legal framework analysis

One of the first aspects to bear in mind when developing a technological project is the legal environment in which it is framed.

Concerning the actual code base of the project, we must implement all necessary intellectual property protection measures. Because it is an *open source* project, an internationally recognised software license will be included in the public code repository, hosted at GitHub[12]. The chosen license is the MIT License, which has proven to be easy to understand, relatively widespread and quite permissive.

Within the project, no personal data has yet been used to perform any benchmarking process nor to assess the quality of the developed methods - random data generators are being used instead. However, if such data was ever used, it is clear that is should be under the terms of the spanish LOPD law and that, therefore, protection and security measures should be taken accordingly. It is unclear, at the moment, that any sensitive data set might be used throughout this project; many benchmark data sets available are free from any kind of sensitive, personal data.

We have not detected any other kind of legal consequences or regulations bound to this project's development.

3.2 Technology alternatives

Before we started developing privacy preservation filters, a number of existing technologies was evaluated as possible solutions or basis for the goals of the project.

As for the proposed SDC methods to be developed, there is a popular library which implements all of them: `sdcMicro` [19]. The main problem is that this piece of software is written in R, whereas the MOA framework is written in Java. An evaluation was performed to decide whether to use this library or to develop the privacy filters ourselves, from scratch.

As can be seen in figures 2 and 3, the `sdcMicro` library would be used from MOA by developing an adapter (a wrapper) to interconnect the Java and R execution environments. There are some benefits of taking this approach, but also some drawbacks:

- **Faster development:** the only new piece of software to be developed is the wrapper or interconnect between MOA and the `sdcMicro` package.

[12]The project is available at `https://github.com/necavit/moa-ppsm`.

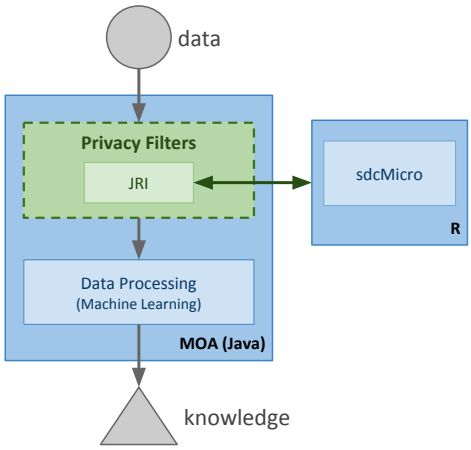


Figure 2: Data flow: R/Java hybrid solution.

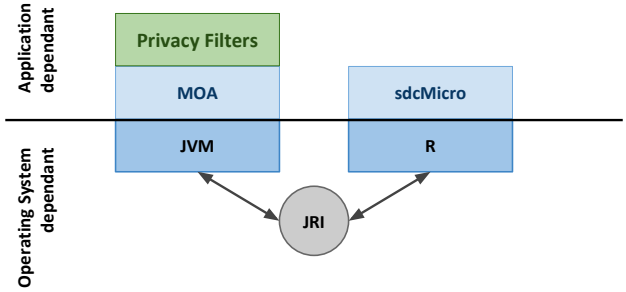


Figure 3: Hybrid solution architechture: strong dependencies.

- BLA BLA

3.3 Implemented algorithms

3.4 Proposed methods

3.5 Results generation

3.6 Proposed publications

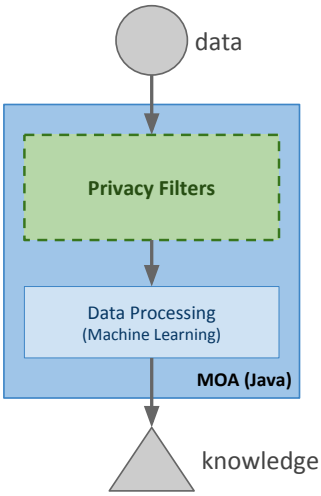


Figure 4: Data flow: pure Java solution.

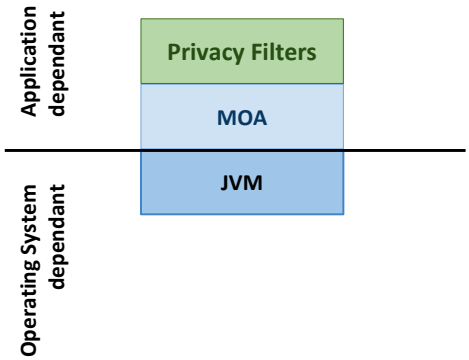


Figure 5: Pure Java solution architechture: few dependencies.

References

[1] UN General Assembly. Universal declaration of human rights, December 1948.

[2] BOE. Ley orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal, 1999. URL: `https://www.boe.es/boe/dias/1999/12/14/pdfs/A43088-43099.pdf`.

[3] Arthur Charles. Naked celebrity hack: security experts focus on icloud backup theory, September 2014. URL: `http://www.theguardian.com/technology/2014/sep/01/naked-celebrity-hack-icloud-backup-jennifer-lawrence`.

[4] Cynthia Dwork. Differential privacy. In *in ICALP*, pages 1–12. Springer, 2006.

[5] Anco Hundepool et. al. *Statistical Disclosure Control*. John Wiley & Sons, Ltd.

[6] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. Advances in knowledge discovery and data mining. chapter From Data Mining to Knowledge Discovery: An Overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996. URL: `http://dl.acm.org/citation.cfm?id=257938.257942`.

[7] Office for National Statistics. Statistical disclosure control, 2014. URL: `http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/statistical-disclosure-control/index.html`.

[8] The Apache Software Foundation. S4: Distributed stream computing platform, 2014. URL: `http://incubator.apache.org/s4`.

[9] The Apache Software Foundation. Storm, distributed and fault-tolerant realtime computation, 2014. URL: `https://storm.incubator.apache.org`.

[10] Mohamed Gaber. Mining data streams: a review. *ACM SIGMOD Record*, 34, June 2005.

[11] Merriam-Webster Inc. Seclusion, October 2014. URL: `http://www.merriam-webster.com/dictionary/seclusion`.

[12] H. Kargupta. Minefleet®: The vehicle data stream mining system for ubiquitous environments. *Ubiquitous Knowledge Discovery*, 2010.

[13] Ashwin Machanavajjhala. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 2007.

[14] Li Ninghui. t-closeness: Privacy beyond k-anonymity and l-diversity. *Proc. of IEEE 23rd Int'l Conf. on Data Engineering*, 2007.

[15] Anand Rajaraman and Jeffrey David Ullman. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA, 2011.

[16] Shane Richmond. Millions of internet users hit by massive sony playstation data theft,

April 2011. URL: http://www.telegraph.co.uk/technology/news/8475728/Millions-of-internet-users-hit-by-massive-Sony-PlayStation-data-theft.html.

[17] Sasha Romanosky. Do data breach disclosure laws reduce identity theft? In *Workshop on the Economics of Information Security*, 2008.

[18] Latanya Sweeney. K-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

[19] Matthias Templ. *sdcMicro: Statistical Disclosure Control methods for anonymization of microdata and risk estimation.* CRAN R package repository. URL: http://cran.r-project.org/web/packages/sdcMicro/index.html.

[20] New Zealand University of Waikato. Moa - overview, October 2014. URL: http://moa.cms.waikato.ac.nz/overview.

[21] New Zealand University of Waikato. Weka 3 - data mining with open source machine learning software in java, October 2014. URL: http://www.cs.waikato.ac.nz/ml/weka.

[22] Yahoo. Samoa by yahoo, 2014. URL: http://samoa-project.net/.