D6.1

USEMP privacy scoring framework – v1

v1.6 2015-01-15

Georgios Petkos (CERTH), Symeon Papadopoulos (CERTH), Thomas Theodoridis (CERTH), Timotheos Kastrinogiannis (VELTI), Apostolos Kousaridas (VELTI), Theodoros Michalareas (VELTI), Yiannis Kompatsiaris (CERTH)

The current deliverable is a technical report accompanying the first version of the USEMP privacy scoring framework, a tool that aims at raising the awareness of users about the disclosure and value of their personal information. In particular, the framework defines a set of high-level privacy dimensions and a set of attributes per dimension for users of Online Social Networking (OSN) services. A number of scores are defined with respect to specific values for each attribute, and an aggregation mechanism is described for translating attribute-level scores to dimension-level scores. In addition, the attribute scores are linked to specific multimedia items and trails of user behaviour through the application of inference mechanisms.

Three such mechanisms have been implemented and experimental studies on publicly available OSN datasets are presented. The deliverable documents the usage of the accompanying prototype implementations to reproduce the reported experiments. Finally, an initial set of metrics for representing and reasoning about personal data value is described.

Project acronym	USEMP
Full title	User Empowerment for Enhanced Online Presence Management
Grant agreement number	611596
Funding scheme	Specific Targeted Research Project (STREP)
Work program topic	Objective ICT-2013.1.7 Future Internet Research Experimentation
Project start date	2013-10-01
Project Duration	36 months

Workpackage 6	User Assistance for Shared Personal Data Management
Deliverable lead org.	CERTH
Deliverable type	Prototype
Authors	Georgios Petkos (CERTH) Symeon Papadopoulos (CERTH) Thomas Theodoridis (CERTH) Timotheos Kastrinogiannis (VELTI) Apostolos Kousaridas (VELTI) Theodoros Michalareas (VELTI) Yiannis Kompatsiaris (CERTH)
Reviewers	Laurence Claeys (iMinds) Niels van Dijk (ICIS) Adrian Popescu (CEA)
Version	1.6
Status	Final
Dissemination level	RE: Restricted Group
Due date	2014-12-31
Delivery date	2015-01-15

Version	Changes
0.1	First outline ToC by Georgios Petkos
0.2	First version of related work and framework by Georgios Petkos and Symeon Papadopoulos
0.3	Privacy scoring using digital trails by Symeon Papadopoulos and

	Thomas Theodoridis
0.4	Privacy scoring based on user links by Symeon Papadopoulos
0.5	Completion of related work and framework by Georgios Petkos
0.6	Refinements and revisions by Symeon Papadopoulos
0.7	Personal data value scoring by Apostolos Kousaridas and VELTI
0.8	Updates and revisions by Symeon Papadopoulos and CERTH
0.9	Update of Chapter 5 and various refinements by Georgios Petkos and CERTH
1.0	Updates and revisions by Symeon Papadopoulos and CERTH
1.1	Inputs by Theodoros Michalareas and VELTI, refinements by Symeon Papadopoulos and CERTH and release for internal review
1.2	Refinements by CERTH after internal review by Niels van Dijk and inputs from VELTI
1.3	Additions from CERTH (Schneier's taxonomy, control scores, Annexes about prototype usage instructions) and refinements after internal review by Laurence Claeys
1.4	Refinements after second review by Niels van Dijk
1.5	Finalization after review by Adrian Popescu
1.6	Fixed formatting and presentation issues

Table of Contents

1. Introduction

The purpose of this deliverable is to document the first version of the USEMP privacy scoring framework prototype implementation. This introductory Chapter first provides an overview of the developed privacy scoring framework; it proceeds with a description of the adopted research methodology and contributions, and concludes with a multi-disciplinary view on issues related to the design, development and deployment of the presented framework. The two main objectives of the deliverable are a) to provide a detailed description of the developed framework and its role and usage in the USEMP system, and b) to expose the research insights and contributions stemming from the conducted experimental work.

1.1. Overview of framework and deliverable

In short, the goal of this framework is to provide easily interpretable risk indications and alerts by means of privacy scores. Importantly, the scores produced will be used by both use cases of the project: privacy scores will be directly utilized by the first use case that focuses on OSN presence control; and, in addition, scores expressing the value of personal data will be of central importance for the second use case.

To build a scoring model, we start by first identifying a set of aspects that characterize the personal information of users of Online Social Networking (OSN) services. This definition is based on a study of previous work about the types of personal data that may be considered private, sensitive or valuable, and of methods used to automatically extract inferences about personal data. This study is presented in Chapter 2.

The privacy scoring framework is then presented in Chapter 3. We first define a hierarchical structure that is suitable for compact presentation and management of private information. In particular, we define a set of eight privacy dimensions at the top level that summarize different aspects of private information, e.g. Demographics, Religious Beliefs, Location, etc., and for each dimension, we define a number of specific attributes that correspond to characteristics and traits of OSN users, e.g. Gender and Age under Demographics. On top of this hierarchical structure, we build a privacy scoring model which has two aspects pertaining to the two use cases: privacy scores and personal data value scores. The scoring model has a very close relationship to inference mechanisms that attempt to reveal different aspects of a user's profile. Several of these methods, such as location estimation from text posts and concept detection from images, are developed and described in WP5 deliverables. Some additional methods are also developed within Task 6.1 – based on different types of personal data compared to those of WP5. The following three chapters describe three such inference mechanisms based on complementary cues.

Chapter 4 presents an approach for inferring personal traits based on the digital trails that people leave behind when they make use of OSN services. For instance, in the case of Facebook, utilizing the set of likes that a user has performed may be used to reveal a variety of personal traits and attributes about the user. Chapter 5 describes an approach that builds topic models (also called latent topics) based on large sets of liked content (e.g. Facebook pages) and then uses the individual users' posts to infer specific personal profile attributes of interest based on the constructed latent topics. Chapter 6 present an approach that exploits the connectivity information about a user (i.e. the links between the user and other OSN users around him/her), and demonstrates that even when a particular user does not reveal information about him/her, it is still possible to infer such information based on information about his/her connections. For the three aforementioned approaches, we also present a set of experiments based on publicly available data of similar nature to those that will arise from the operation of the USEMP system.

Chapter 7 presents a framework for representing and reasoning about the value of a user's personal profile and information. The proposed framework provides estimates with respect to the relative value of users' personal data based on the influence and reactions that the data generate in the context of the OSN.

Chapter 8 concludes this document and presents the next steps towards the pilots and plans for future research.

Technical details, including detailed specification of the hierarchical privacy dimensions structure, a JSON representation of the privacy scoring framework, and instructions on how to run the accompanying prototype implementations are included in the three Annexes at the end of this report.

1.2. Research methodology and contributions

The primary focus of this work has been the generation of a privacy scoring framework that simplifies the capturing and communication of privacy risks and the value of personal data that are disclosed to OSN services. As a first step, we conducted an analysis of existing privacy scoring approaches and identified those that matched the intended usage and goals within USEMP. We also identified a key limitation of previous approaches, i.e. the capability of modelling inferred personal information based on observed user data and behaviour. Since performing inferences about user attributes and traits is one of the key awareness creation mechanisms in USEMP, the developed privacy scoring framework was largely guided by this requirement. In addition, important features of the proposed framework include the possibility to quickly summarize the "privacy state" of an OSN user, to enable the linking between inferred information and posted online content and observed behaviour, and to be easily adapted in case new requirements arise.

In addition to the framework design, a number of data mining approaches were developed and evaluated on relevant datasets. The experimental work in these cases had a dual goal: a) assess the feasibility/applicability of methods in the USEMP prototype system and explore potential problematic scenarios (e.g. availability of little training data), b) gather new insights regarding the use and implications of data mining on personal data, and c) develop new approaches for inferring user attributes based on different types of observed data.

One key methodological issue in the conducted research pertains to the use of appropriately selected datasets that are in line with the project scenarios. To this end, we made use of two publicly available anonymized datasets: a) the myPersonality dataset, which enabled us to study the inference of Facebook user traits and attributes based on their like history, and b) the SNOW dataset, comprising user interactions on Twitter, that enabled us to explore the potential of user attribute inferences with the help of social network structure. It is noteworthy that the developed inference approaches are OSN- and dataset-independent; hence the drawn conclusions should apply to other datasets and social networks.

Although much of the work performed in the first research and development iteration relied on existing approaches, we consider that it resulted in a number of important contributions and insights. In particular:

- A new privacy scoring framework was developed that captures multiple dimensions of a user's profile, and enables the modelling of inferred information and the linking to individual digital items that are responsible for increased privacy exposure. In addition, the framework supports aggregated views and is simple to visualize and communicate.
- A comprehensive experimental study was conducted on the problem of inferring personal profile attributes using observed behaviour (in the form of Facebook likes). A number of factors, including the selected features, classification algorithm and size of training set, were investigated, and a new approach was developed for quantifying the uncertainty involved in the conducted inferences.
- An approach based on latent topic models was proposed as an alternative inference mechanism that can link a user profile to topics of interest with the goal of providing easy-to-interpret inferences.
- A new graph-based approach was developed that can be used to infer user traits and attributes for a user on the basis of his/her links and interactions with other OSN users, some of which with known attributes. The proposed approach was shown to outperform a number of existing approaches in terms of inference accuracy.
- A new data value scoring method was proposed that produces estimates of the relative value of a user's personal data based on the user's audience and influence within the environment of the OSN.

1.3. Multi-disciplinary aspects of research

Although the core privacy scoring framework presented here was based on approaches from the fields of data mining, machine learning and social network analysis, the presented work was largely guided and shaped by findings and insights coming from different disciplines and work areas of USEMP. Here, we provide a concise summary of the multi-disciplinary aspects of the presented work.

The use case analysis (D2.1) and the associated requirements definition (D2.2) were instrumental in shaping the scope and objectives of the research presented here. A tech card (Personal attribute behavioural predictor) was developed to participate in the user stories, and the following requirements were taken into account in the design of the privacy scoring framework:

- [SR04] "The system may be able to make best effort associations between data placed onto OSN(s) and the profile attributes which can be inferred from such data.": The definition of several elements in the developed privacy scoring framework (e.g. confidence scores, support structures), as well as the developed inference methods largely focused on addressing this requirement.
- [SR05]"The system may be able to provide suggestions for alterations regarding the visibility of parts of the posted content in order to allow the user to make informed changes on how the profile will be outwardly perceived.": The inclusion of a "visibility" score in the privacy scoring framework was guided by this requirement.
- [SR07] "The system shall enable the visualisation of end-users digital trails and thus, the estimations of profiles (and/or profile segments/categories) she/he is placed into by different actors in the network. Thus, the system should provide information from profiling in order to show the user which entities have the greatest interest in their

data": The user attribute prediction approaches presented in this deliverable will be used by the system to address this requirement.

- [SR08] "The system shall enable to provide end-users useful insights on the value of their digital data and social footprint that are either directly shared in social networks (e.g., likes on Facebook) or are indirectly collected by various network actors that track their activities on web browsers.": This requirement guided the development of the personal data value approach presented in this deliverable.

The legal studies in WP3 have also had an important impact on the research presented here. In particular, they were essential in the definition of personal data typology, delineating the scope of personal and sensitive data, both from a data protection and an anti-discriminatory law perspective. In addition, interaction with WP3 was instrumental in making sure that the research described in this deliverable, in particular the processing and handling of personal data, was fully compliant with legal and ethical requirements. Finally, legal aspects were considered in relation to the IP rights of the developed approaches to ensure that the resulting technical components have clear IP provenance and rights of use.

Social requirements analysis work in WP4 has also had significant impact on the work of D6.1. In particular, the analysis of existing privacy enhancing tools and the list of social requirements presented in D4.1 informed the design of the privacy scoring framework to be in line with end user expectations and to address their concerns. In addition, the survey of personal information disclosure practices and the perceptions of OSN users with respect to the sensitivity of personal information, which were discussed in D4.2, contributed to the definition of the personal data typology and the privacy dimensions schema that is presented in this deliverable. In addition, the user questionnaire presented in D4.2 was shaped to a large extent by the data collection requirements of the user attribute inference methods that are presented here.

Interaction with the research on multimedia information extraction in WP5 was intense in order to define how the outputs of the text and visual mining components of WP5 will feed into the privacy scoring framework of WP6.

Last but not least, work in WP6 was also aligned with the developments and outcomes of the technical activities in WP7. More specifically, the approaches presented in this deliverable were also assessed with respect to technical considerations, e.g. need for fast and large-scale processing, ease of integration, data access limitations stemming from the use of external APIs, etc.

2. Related Work

In this chapter, we present an overview of prior research in the field that provides a basis for our research. We start with a typology of personal information, focusing on types of information that may be considered by individuals and by law to be sensitive and/or valuable. We then proceed with the presentation of existing privacy scoring approaches, and conclude with works that deal with the problem of personal trait/attribute inference based on online user behaviour.

2.1. Typology of personal information

The term personal data describes "any information relating to an identified or identifiable individual (data subject)", based on the OECD Privacy Guidelines. Characteristic categories of personal data include user generated content, activity or behavioral data, social data, locational data, demographic data, or data of an official nature, e.g., financial information and account numbers, health information (OECD, 2013).

Looking at the problem from a general viewpoint, one might argue that almost any type of personal information may be considered by different individuals to be sensitive or valuable under specific circumstances. For instance, any type of information about the profile of an individual may be valuable to a marketing company for promotion reasons. Additionally, there have been many cases in which people became victims of discrimination on the basis of their traits or past actions (Acquisti & Fong, 2012), even in cases where such discrimination was not justified. In general, our study of related work that is presented in this Chapter suggests that personal information can be considered of sensitive nature in cases that:

- The information can be used for unjustifiable discrimination in a variety of social, cultural, professional and other environments. For instance, information about the gender, age, ethnicity, political or religious beliefs, sexual preferences, and financial status of a person have been used in the past for unjustified discrimination, for instance, in the context of personnel selection (Acquisti & Fong, 2012) and for loan approval and pricing based on social media profiles (Raman et al., 2012).
- The information may be used for manipulation of the opinion/beliefs of the person him/herself or the opinions of others about him/her. In the more common case, this includes information that is often extensively used by third parties for profiling and targeting (in case of ad campaigns).
- The publication of that information may have detrimental effects on the mental, physical and economic state of the individual, e.g. threats to their residence privacy, stalking (Haron, 2010), identity theft, etc.

Given the above scenarios, one could hypothesize that almost any type of personal information could potentially be used to produce detrimental effects like those listed above.

Importantly though, in addition to the general notion of what sensitive information is, which has been presented so far, there is a legal framework for protecting personal and sensitive information. This specifies specific rules and requirements regarding the process of personal data processing. The precise legislation differs from country to country but one of its key elements that relates to the USEMP objectives defines that a user should be explicitly notified about the following:

- Type of data being collected about him/her.
- The reason for the collection of the data.
- The planned use of the data.

Additionally, the framework specifies that the user should provide his explicit consent for the collection, storage and processing of sensitive data. Types of information that are considered sensitive according to the different legislations include information about a person's:

- Racial / ethnic origin
- Political opinions
- Religious or philosophical beliefs
- Trade-union membership
- Health status
- Sex life

In addition to privacy protection law, there is also relevant anti-discrimination legislation. Types of information that are protected by the anti-discrimination law include the following:

- Sex
- Gender identity & sexual orientation
- Age
- Race and Ethnicity
- Nationality
- Disability
- Religion, Worldviews and Political opinions.

Having listed a number of types of personal information that are considered sensitive or valuable, from a general viewpoint, but also from a legal point of view in the case of sensitive data, it should be noted that different types of personal information are not perceived by individuals as equally private or valuable. As will become evident in the next Chapter, this is a very important point for the development of the framework. Different users will evaluate differently (or even disregard) information that is relevant to privacy threats as the ones listed above. At the same time there is general consensus about the criticality of some types of information as compared to others. D4.2 reviewed previous research on how users perceive the disclosure of different types of personal data. In particular, it was found that different groups of users tend to behave differently with respect to what types of data they make publicly available and therefore how private they think they are. Additionally, it was observed that disclosure behaviour can change significantly over time. Importantly though, new user studies will be necessary to obtain updated information about what types of personal data people tend to reveal in the social media. These studies will be carried out as part of the pilots and its design is also documented in D4.2. To cater for the volatility and subjectivity of what type of information is considered private, the privacy scoring framework described in the next Chapter enables fine-grained update and adaptation of these privacy scores.

Finally, in our study, we consider the taxonomy of personal data in relation to OSN services (Schneier, 2010). This taxonomy considers the source of data about a user, rather than the type of personal information, and will be useful for the development of the scoring framework. Briefly, Schneier identifies the following six categories of OSN data:

- *Service data.* This is the set of data that a user explicitly provides to the OSN service. In many cases, this includes the user's legal name, age, gender, etc.

- *Disclosed data*. This includes the content (messages, status updates, photos, etc.) posted by the user to his own page.
- *Entrusted data*. This is the content posted by the user to the page of another user. It is similar to disclosed data, with the difference that, in many cases, the user does not have full control of the content, but some other user does.
- *Incidental data*. This is the content posted about the user by some other user (e.g. when a friend of the user posts a picture depicting the user). Again, this is similar to disclosed data, but again, the user does not have control of the content.
- *Behavioural data*. This type of data includes the actions of the user in the OSN. For instance, this may include information about which profiles the user visits, what games s/he plays, what pages s/he likes, etc.
- *Derived data*. This is data about a user that may be derived from all other types of data, typically by means of algorithmic processes. We will also refer to such kind of data as inferred or inferences.

Clearly, this is not the only way by which data in OSNs can be organized. As mentioned, it primarily focuses on the source of the data; another possible organization could focus on the semantics of the data about a user, this is the type of top-level organization we adopt in the privacy dimensions framework in the next Chapter. However, Schneier's taxonomy identifies that the level of control a user has over the data that concern him/her may vary significantly depending on the above categories, for instance the user typically has full control over service and disclosed data, while limited control over entrusted and incidental data and almost no control over derived data. This fact has been considered in the development of the scoring framework.

2.2. Privacy scoring functions

Here, we review some existing work in the field of privacy scoring functions. The number of relevant papers is rather limited, indicating that the field is still relatively unexplored. Additionally, most of the work is quite recent, indicating the emerging interest in the problem.

One of the first works on the field of privacy scoring comes from Liu and Terzi (2009, 2010). The authors introduced the concept of "Privacy Scores", a function of the "sensitivity" and the "visibility" of different pieces of profile information. Sensitivity (in the sense used in the paper) denotes how important it is to protect this particular piece of information (e.g. Age, Gender, etc.) and is computed by analysing the results of user studies with respect to the disclosure of such information (the more people are willing to disclose a piece of information, the less sensitive it is considered). Moreover, visibility quantifies the extent to which it is accessible to other users. Sensitivity essentially reflects the concept that was briefly discussed at the end of the previous section. Formally, assuming a user j has n profile items, the Privacy Score PR is computed as follows:

$$PR(j) = \sum_{i=1}^{n} PR(i,j) = \sum_{i=1}^{n} \beta_i \cdot V(i,j)$$

where β_i is the sensitivity of profile item *i* and *V(i,j)* is the visibility of profile item *i* for user *j*. In order to compute the scores, Liu and Terzi use as input a *n* x *N* response matrix *R* (with *n* being the number of profile items and *N* the number of considered users), which expresses how willing a user is to disclose some profile item. Given this formulation, a simple statistical model grounded on the Item Response Theory and Maximum Likelihood Estimation is used to compute the Privacy Scores. Note that, while the sensitivity of profile item *i* (*βi*) is assumed to be the same across all users in (Liu & Terzi 2009; 2010), in the USEMP framework, as will be shown in the next chapter, we decided to let each user specify a different sensitivity score for each item.

The work by Liu and Terzi inspired other works in the field. Importantly, it has introduced the concepts of sensitivity and visibility. Our own formulation in the next Chapter will also use the concepts of sensitivity and visibility and will also compute aggregate privacy scores as products of these factors. However, the model by Liu and Terzi focuses only on information that may be explicitly available on the profile of the user, whereas we also need to produce scores also for the case that information is implicitly made available or inferred. More details about the differences of our approach to that of Terzi and Liu will be exposed in the next Chapter, after we have presented our approach.

A second approach that introduces an extension to the Privacy Scores of Terzi and Liu is presented by Srivastavaand Geethakumari (2013). Apart from information explicitly provided by the user as part of his/her profile, Srivastava also examines textual messages that may contain sensitive information such as address, e-mail, location, etc. Srivastava uses the same sensitivity and visibility computation that Terzi and Liu used to produce a score; however, they call it Privacy Quotient. They however introduce an additional metric, called *privacy leakage*, which is applied to a single message/piece of content and measures how much of the privacy exposure for some user is due to that particular message. It is computed by dividing the sensitivity for the message by the sum of sensitivities over all messages.

A very interesting study comes from Domingo-Ferrer (2010). Domingo-Ferrer introduces the Privacy-Functionality Score which quantifies how much information a user reveals compared to other users. It utilizes the Privacy Score, as defined by (Liu and Terzi, 2009), and is defined as follows:

$$PRF(j) = \frac{\sum_{x=1,x\neq j}^{n} PR(x)}{1 + PR(j)}$$

This ratio allows the users of a social network to be easily ranked in terms of how much they reveal compared to other users. Interestingly, Domingo-Ferrer also studied various scenarios of evolution of a social network based on this score, utilizing concepts from game theory.

Another work that examined the problem of privacy scoring comes from (Nepali & Wang, 2013) and (Wang et al. 2014). Nepali introduces the concept of the Privacy Index. In particular, considering that there are items that are published and others that are not, it defines the Privacy Index as:

$$PIDX = \frac{\sum_{k\in K} S_k}{\sum_{i\in I} S_i}$$

Where *K* is the set of published items, *I* is the set of all items and *S* is the sensitivity of the item. The Privacy Index is somewhat similar to the leakage score of Srivastava. It is clear though that although the leakage score described how much of the information leakage about a user is due to some specific message/piece of content, the Privacy Index actually describes how much of the information that is sensitive and the user has disclosed to the OSN operator, has been also made public.

Finally, a very recent piece of work can be found in Sramka (2015), in which the Privacy Scores of Liu and Terzi are extended to consider: a) information from multiple OSNs, and b) information posted anywhere on the Web and retrievable via search engines.

Table 1 summarizes the different privacy scoring approaches examined in our study.

Score	**Description**	**Elements**
Privacy Score (Liu & Terzi, 2009)	A score grounded on the sensitivity and visibility of the items posted by an OSN user.	profile items, sensitivity per item, visibility per item
Privacy Quotient and Leakage (Srivastava, 2013)	Extension of Privacy Score with a focus on text messages and on attributing part of score to specific messages using the Leakage score.	profile items (text), sensitivity per item, visibility per item, leakage per item
Privacy Functionality Score (Domingo-Ferrer, 2010)	Extends the Privacy Score by normalizing it over the total privacy exposure of all users, i.e. it offers a comparative privacy view.	privacy score per OSN user, privacy score of other OSN users
Privacy Index (Nepali, 2013; 2014)	Captures the portion of sensitive data posted to a social network that has also been made public.	sensitivity score per item, visibility level (public/private) per item
Privacy Scores (Sramka, 2015)	Extension of Privacy Score to include data from multiple OSNs and information retrievable via search engines.	privacy score per item per OSN, information found via search engines

Table 1. Overview of studied privacy scoring approaches

2.3. Methods for inferring personal information

In addition to a privacy scoring function, the USEMP system will also feature methods that infer personal information based on users' digital trails. Here, we review some previous work on inferring personal information on the basis of the online behaviour and connectivity of OSN users. Those are complementary to those of WP5 that focus on information extracted from multimedia content.

In the research done by Kosinski et al. (2013), 58,466 Facebook users provided their complete like history (170 likes per person on average), their profile information as well as the results of several psychometric tests. The researchers built a prediction model based on user likes, wherein they used dimensionality reduction, Singular Value Decomposition (SVD), on the user-likes matrix, keeping the 100 components with the most predictive power, and built linear and logistic regression models to respectively predict numeric and dichotomous variables. The Area Under Curve (AUC) scores for predicting the dichotomous variables were: 95% for ethnicity, 93% for gender, 88% for gays, 75% for lesbians, 85% for political affiliation, 82% for religion, 73% for cigarette smoking, 70% for alcohol consumption, 65% for drugs use, 67% for relationship status and 60% for parents being together when the user was 21. The Pearson correlation coefficient for age was 0.75.

Schwartz et al. (2013) studied a dataset of 15.4 million status updates from a total of 74,941 Facebook users, who also submitted their gender, age and Big-5 personality scores. They tested traditional techniques of linking language with personality, gender and age such as the Linguistic Inquiry and Word Count (LIWC), which uses a lexicon with pre-selected categories, but also developed a new approach, the Differential Language Analysis (DLA), which generates the lexicon categories based on the text being analyzed. The researchers first used Principal Component Analysis (PCA) to reduce the feature dimension, and then a linear

Support Vector Machine (SVM) for classifying gender and ridge regression for predicting age and personality traits. They were able to predict gender with 92% accuracy, while the root of the coefficient of determination was 0.84 for age, 0.38 for extraversion, 0.31 for agreeableness, 0.35 for conscientiousness, 0.31 for neuroticism and 0.42 for openness.

Backstrom and Leskovec (2011) developed an algorithm based on Supervised Random Walks for link prediction. Given the friendship network structure and features related to the age of links in the network, friend requests, communication and profile observation data over a one-week period, and the number of common friends between two users, the algorithm predicts which Facebook users are going to be friends with a specified user in the future. Tested on the Iceland Facebook network, containing 174,000 users (55% of the country's population) and 29 million links among them, the algorithm produced an AUC score of 83% and a 38% precision at the top-20 suggested links. This means that out of the 20 suggested friendships, 38% of them really happened in the near future.

Backstrom and Kleinberg (2014) developed a new network measure, called dispersion, which they combined with Boosted Decision Trees (BDT) to predict whether a Facebook user is in a relationship and, in case he/she is, predict which person in his/her friendship network is his/her spouse. For the first task they used a dataset of 129,000 users that contained demographic features such as age, gender, country and the number of days since they joined Facebook, as well as features from their friendship network. They were able to predict whether a user is single or not with 68% accuracy and whether she is single or married with 79% accuracy. For predicting a user's partner they analyzed 73,000 users and used structural information from the friendship network, the number of photos in which the user and each of his/her friends are jointly tagged, the number of times the user has seen a friend's profile, the number of messages the user has sent to each of his/her friends and the number of times the user has liked each friend's content and vice versa. For married users they were able to predict their spouse with 72% accuracy, for engaged with 71% accuracy and for users in relationship with 68% accuracy. Overall, they correctly predicted the user's spouse 71% of the time.

In their study, Jernigan and Mistree (2009) used the friendship network structure of 4,080 Facebook users, as well as their profile information regarding their sexual orientation. Their prediction of whether a user was gay or not was based on the percentage of the user's friends that identified themselves as gay and it was carried out using a logistic regression classifier. They were able to correctly predict whether someone was gay 78% of the time.

Researchers Zheleva and Getoor (2009) studied several online social networks for the purpose of inferring private information. For the photo-sharing website Flickr they used a dataset of 9,179 users and attempted to predict the users' country. For Facebook, they used a dataset of 1,598 users for predicting gender and a subset of that with 965 users for predicting political views. Finally, 2,632 profiles were used for predicting dog breed in the dog-oriented website dogster. They used various methods for their predictions based on group membership, links among users and machine learning (SVM). As a prerequisite for their research, some profiles, all links and group memberships in the network had to be visible (public). Assuming that 50% of the profiles were public, their prediction accuracy for Flickr was 65%, 73% for gender on Facebook and 66% for dogster. The 58% accuracy for the prediction of political views was not significantly better than assigning the most common label, which was used as a baseline.

Popescu and Grefenstette (2010) make use of the text metadata associated with photos from Flickr and manage to detect the Home location and Gender of the users posting the photos with the use of location gazetteers and gender vocabularies.

Rao et al. (2010) investigated the possibility of predicting gender, age, regional origin and political affiliation from Twitter messages. Using socio-linguistic features such as emoticons, exclamation marks, capitalized text, etc. and unigrams and bigrams of the text, they trained binary classifiers using SVMs. With a set of 1,000 users for gender, they achieved a prediction accuracy of 72%. Some interesting findings regarding gender were that OMGs, emoticons and repeated exclamation marks were respectively used 4, 3.5 and 2 times more by women than by men. Using a set of 2,000 users for age, they were able to predict whether someone is younger than 30 or not with 74% accuracy. For the regional origin task, they tried to distinguish between north and south Indian English writing users. Having a set of 1,000 users they achieved 77% accuracy. Finally, for the classification between democrats and republicans, they used a set of 400 users and achieved 83% accuracy.

Conover et al. (2011) studied 355 million tweets gathered during the last six weeks prior to the 2010 US midterm elections. They investigated the effectiveness of tweet text and hashtags, as well as the mention and retweet network at predicting the political alignment of users (right vs left). With a random sample of 1,000 users and a SVM as their classification method they achieved an accuracy of 80% when using the tweet text (hashtags, mentions and URLs were removed from the text). The accuracy when using only the hashtags present in the tweets rose to 91%. When using a label propagation algorithm on the retweet network as their prediction method they achieved an accuracy of 95%, while applying the same approach to the mentions network did not work as well.

Pennacchiotti and Popescu (2011) used a variety of features regarding Twitter users in order to predict their political affiliation, their ethnicity and whether they are Starbucks fans. The features contained information about the users' profile, tweeting behaviour, the linguistic content of their tweets and their friends and replies network. The prediction process involved two components: a machine learning component based on Gradient Boosted Decision Trees and a label updating component, where the machine learning model applied to the user is also applied to his/her friends; the final label for the user depends on his/her original label and the labels assigned to his/her friends. The study was carried out on the totality of tweets exchanged between September and October 2010. The prediction accuracy for political affiliation was 89%, while the results of the two other tasks were reported in terms of the F-score. The score for classifying ethnicity was 70% and 76% for Starbucks fans.

Wagner et al. (2013) studied 3.7 million tweets from 7,121 users attempting to predict their professions and personality related attributes. Because these were the top users of various WeFollow lists, they also analyzed a second dataset of random users. They created features based on user activity, tweets and bio contents, user lists and linguistic style. Using Random Forests as their classification algorithm, they achieved an AUC score above 90% for the personality attributes advertising, ecological, funny, informational, inspirational and religious; and above 80% for creative and innovational. The scores for the random dataset were lower; above 80% for the same six attributes as before and above 70% for creative and innovational. For classifying professions, the scores were above 90% for all but business and finance, which had scores above 80%. In the random dataset the scores were again lower, but still above 80% for most professions. Writers and politicians had scores above 70%.

Table 2 summarizes the private information inference approaches discussed above. Note that the approaches presented in Chapters 4-6 are similar to or extend several of the approaches presented in the Table. For instance, the approach of Chapter 4 builds upon the one in the first row (Kosinski et al., 2013), the approach described in Chapter 5 is similar to the one in the second row (Schwartz et al., 2013), as well as to other approaches based on the text content of posted messages (Rao et al., 2010). Finally, the approach of Chapter 6 is comparable to approaches the leverage the friendship or interaction network of OSN users, e.g. the approaches of (Backstrom & Leskovec, 2011), (Backstrom & Kleinberg, 2014), and (Zheleva & Getoor, 2009).

Method	**Description**	**Input**	**Inferences**	**Dataset**
Kosinski et al., 2013	SVD + Liner/Logistic regression	Likes	Demographics, Psychometrics, Habits, Preferences	Facebook
Schwartz et al., 2013	Differential Language Analysis (text) + PCA + SVM	Text of posts	Demographics, Psychometrics	Facebook
Backstrom & Leskovec, 2011	Supervised Random Walks	Friendship network, communication profile	Future friendship	Facebook
Backstrom & Kleinberg, 2014	Dispersion + Boosted Decision Trees	Friendship network, Demographics	Single/Married, Spouse	Facebook
Jernigan & Mistree, 2009	Friend attributes + Logistic regression	Friendship network, Sexual Preferences of friends	Sexual Preferences	Facebook
Zheleva & Getoor, 2009	Social network features + SVM	Friendship network, social features (membership to groups), partial labels	Country (Flickr), Gender, Political views (Facebook), dog breed (dogster)	Flickr, Facebook, dogster
Popescu & Grefenstette, 2010	Location Gazzetteer and Gender Vocabulary	Photo title and tags	Gender, Home location	Flickr
Rao et al., 2010	Text features + SVM	Text of tweets	Demographics, Political views	Twitter
Conover et al., 2011	Text-interaction features + SVM + label propagation	Text of tweets, retweet and mention network	Political views	Twitter
Pennacchiotti & Popescu, 2011	Text-social features + Gradient Boosted Decision Trees + label updating	Text of tweets, profile information, friends and replies network	Ethnicity, Political views, Starbucks fans	Twitter
Wagner et al., 2013	Text-social features + Random Forests	Text of tweets, profile bio, twitter lists	Profession, Personality features	Twitter

Table 2. Overview of personal information inference approaches

3. USEMP Privacy Scoring Model

This Chapter presents the USEMP scoring model. We start by compiling an explicit list of personal data attributes that may be considered sensitive or valuable. We organize these attributes in a number of high-level categories that we refer to as privacy dimensions. This organization allows for a more clear and intuitive presentation and handling of the different aspects of a user’s personal information. For instance, one of the privacy dimensions that will be considered is demographics, which includes user attributes such as age, sex, etc., and another is about health factors, which includes attributes such as smoking and drinking, etc. Such a grouping has multiple benefits for the end user. First, it enables him/her to form a succinct, easy to grasp mental model of his/her private information and to prioritize its different parts. Second, it enables the use of different compact visualization methods that will further augment the user’s awareness with respect to his/her private information (as will become obvious in the upcoming Deliverable 6.3).

On top of this privacy dimensions framework, we develop the USEMP scoring model, by enriching it with privacy and data value scores. Privacy scores are about quantifying the potential negative impact entailed by the disclosure of different parts of the information of a user. This is directly related to the first use case. On the other hand, the value of a user’s data (e.g. posts) is inferred by measuring their impact on the user’s social graph, i.e. audience (in terms of reactions, e.g. likes, shares, comments). This is related to the second use case and is discussed in more detail in Chapter 7.

At a glance, the proposed privacy scoring framework consists of multiple scores that reflect quantities such as the sensitivity, visibility, etc. of different privacy dimensions and attributes. In addition to maintaining a number of distinct scores, each of which reflects a distinct aspect of privacy, we also compute aggregate privacy scores in order to end up with concise and simple-to-grasp privacy indicators

3.1. Privacy dimensions

Based on the study of prior work presented in Chapter 2 and discussions with all consortium partners, we have defined eight key categories of personal attributes, which we name *privacy dimensions*. These eight privacy dimensions are: A) Demographics, B) Psychological Traits, C) Sexual Profile, D) Political Attitudes, E) Religious Beliefs, F) Health Factors & Condition, G) Location and H) Consumer Profile.

These dimensions cover a wide variety of personal information, which OSN users would in many cases consider of private nature, and also encompass information that is considered sensitive from a legal perspective. In addition, based on current business practices (mainly stemming from the marketing industry), the identified dimensions are associated with certain value levels, i.e. they carry a certain level of utility for (marketing) companies interested in targeting consumers. Table 3 summarizes the eight identified privacy dimensions, along with the value levels associated with them.

This set of eight privacy dimensions constitutes the current top-level *schema* of the USEMP privacy model, and although we do not foresee considerable changes at this level, the overall privacy scoring framework is generic enough and can accommodate such changes if needed (e.g. addition of a new dimension, splitting of an existing dimension into more).

#	Name	Description	Threats-Sensitivity	**Value** (for advertisers)
A	Demographics	Personal data, such as Gender, Age, Nationality, Racial background, etc.	Discrimination in a variety of settings. The most frequently used type of information.	**High:** advertisers wish to target users of certain demographic criteria
B	Psychological Traits	Defined by psychologists (extraversion, openness, etc.)	Discrimination, e.g. in personnel selection	**Low**
C	Sexual Profile	Relationship status, preferences, habits	Discrimination, e.g. in workplace, education, housing	**High:** advertisers wish to target consumers based on their relationship status/lifestyle related to their sexual profile
D	Political Attitudes	Supported politicians, parties and stance	Discrimination, e.g. in workplace or personnel selection	**High:** advertisers wish to target consumers based on the political affiliations since these are related to their general profile
E	Religious Beliefs	Religion (if any) and beliefs	Discrimination, e.g. in the sale or rental of housing, job selection, workplace.	**Moderate:** advertisers wish to target consumers based on their religious and cultural beliefs
F	Health Factors & Condition	Habits (e.g. smoking, drinking), medical conditions, disabilities, health factors (exercise)	Discrimination, e.g. health insurance denial or discriminatory pricing.	**High:** advertisers wish to target consumers based on their habits
G	Location	Characteristic locations of the individual and history of previous locations	Discrimination, e.g. house insurance, stalking	**High:** advertisers wish to target consumers based on their current location or their home location
H	Consumer Profile	Preferred products and brands	Ad targeting and discrimination in online price-setting	**High:** advertisers wish to target consumers based on their consumer profile attributes like the devices the use to access digital content

Table 3. Overview of USEMP privacy dimensions

Under each privacy dimension, we define a set of attributes that can be used as important cues to characterize a user along this dimension. For instance, the Health Factors and Condition dimension includes variables such as smoking, alcohol consumption, and use of drugs. Within WP5 and WP6, we explore the hypothesis that it is possible to infer several values of such attributes by monitoring the online traces of OSN users (including those made visible through their online connections).

To give a more concrete view on such attributes, Table 4 presents a list of the attributes identified for the Demographics dimension. The specification of attributes for all privacy dimensions of our model are detailed in Table 20-Table 27 of Annex I. While effort was made to come up with a relatively extensive list of representative attributes for each dimension, it should be clear that these lists are not aiming to be exhaustive, but rather to support the

USEMP user stories and scenarios as described in Deliverable 2.1. In case new user stories and needs arise, it is straightforward to update the respective lists.

#	Attribute	Description	Example values and range
A.1	Age	Rather than using the absolute number of years, it is typical to use age groups.	The following is a preliminary set of age groups: 6-12, 12-18, 15-25, 25-35, 35-45, 45-55, 55-65, 65-75, older than 75 years
A.2	Gender	The gender of the user	Male, Female
A.3	Nationality	Nationality of the user	French, Belgian, Greek
A.4	Racial origin	The racial background of the user.	Asian, African, Caucasian, Latino/Hispanic, Other
A.5	Ethnicity	The ethnic origin of the user that could relate to a combination of racial background, language and religion.	List of target ethnicities, e.g. Arabic, Easter-European, etc..
A.6	Literacy level	Literacy level will be represented by the highest degree or level of school attended.	None, Nursery school, High school, Bachelor’s degree, Master’s degree, Ph.D., Other
A.7	Employment status	The employment status of the user	Employed, Unemployed, Retired, Other
A.8	Income level	This could be represented as the perceived deviation from the average national income.	Qualitative ranges of monthly income. e.g., a 5-scale range from low to high.
A.9	Family status	Marital status of the user.	Single, in a relationship, married, separated, other

Table 4. Demographic attributes

The above effectively creates a hierarchy in which, the top level represents the OSN personal data profile, at the next level there is a number of privacy dimensions, each privacy dimension has a number of attributes and each attribute can take one or more out of a set of possible values. This formulation will be the basis of our privacy scoring model. In short, we will perform inferences using a variety of mechanisms, e.g. multimedia information extraction techniques from WP5 and inference techniques such as the ones of Chapters 4-6, and OSN presence data (typically in the form of observed user activities, e.g. likes, posts, user interactions, or volunteered profile information) in order to fill (or predict) the values that these attributes can take. Each value will be characterized by a number of scores expressing e.g. the sensitivity of this piece of information, the confidence of our prediction etc. Subsequently, we aggregate those low-level scores in order to compute scores for attributes, dimensions and eventually for the overall user profile.

3.2. Privacy scoring

Structure: The USEMP privacy scoring mechanism enriches the privacy dimensions hierarchy with a number of scores, each reflecting a different aspect of personal information disclosure. Additionally, overall privacy scores are computed at each level of the hierarchy. Clearly, the two important characteristics of this framework are the following: a) it is tailored to the hierarchical structure of the privacy dimensions, b) there are multiple scores associated with the elements of each level of the hierarchy. Hence, the framework enables the following two kinds of user awareness: a) navigation through the levels of the hierarchy and understanding of how the scores for some particular value affect or are affected by the levels above and below it, and b) focus on specific aspects of the factors that are related to privacy; e.g., it will be possible to focus on visibility, the overall privacy score, etc.

Here, we consider an additional level at the root of the framework, which contains any type of data that is generated as a result of a user's behaviour and interaction with the services of an OSN operator. This includes posted content (text, images), explicitly declared profile information, user network data, sets of likes, etc. We call this the *OSN presence data layer* and consider it as the primary source for populating the privacy scores for the given user. Naturally, between the privacy values level and the online presence data, there is a layer of modules that perform various mining and inference procedures. The overall framework is visualized in Figure 1.

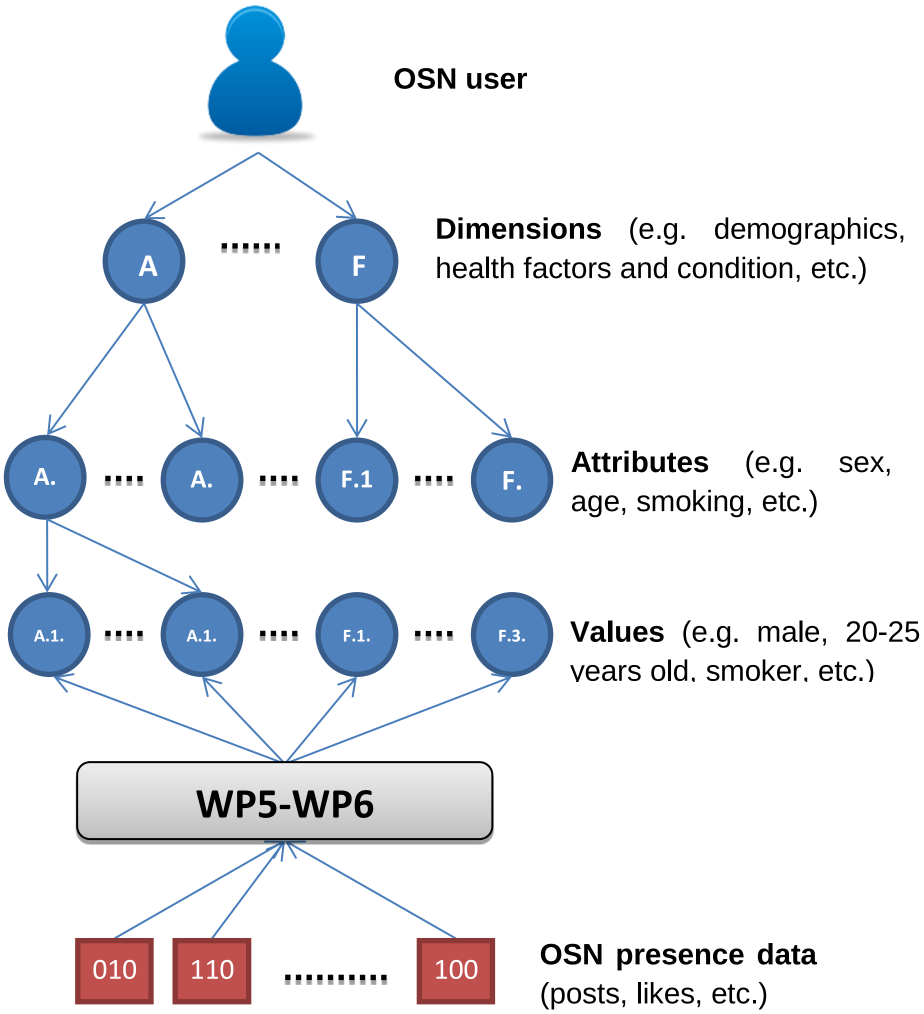

OSN user
A
F
Dimensions (e.g. demographics, health factors and condition, etc.)
A.
A.
F.1
F.
Attributes (e.g. sex, age, smoking, etc.)
A.1.
A.1.
F.1.
F.3.
Values (e.g. male, 20-25 years old, smoker, etc.)
WP5-WP6
010
110
100
OSN presence data (posts, likes, etc.)

Figure 1. Overview of privacy scoring mechanism

We will now present the various scores that will be assigned to the nodes of the four layers of the hierarchy (user, dimensions, attributes, values). It should be noted that some of those cannot be considered as scores per se, they may rather be considered as additional fields that enrich the representation of the privacy scoring model. These appear at the values layer;

for instance, there is a field called "Declared/Inferred" that simply states whether knowledge about the particular value has been explicitly provided by the user or if it has been inferred.

Starting from the level of values, the scores that characterize each value are the following:

a) *Confidence*. This is a continuous value in the range from 0 to 1. It represents how confident we are that the corresponding value is true and is typically computed by the inference algorithm along with the produced inference (in Section 4.2 we propose a new method of deriving such a score). It needs to be noted that the confidence values under the same attribute should sum to 1 (except for the case that an attribute can take multiple values simultaneously).

b) *Sensitivity*. A continuous value in the range from 0 to 1, with higher scores corresponding to higher sensitivity levels. It reflects how important it is to protect this piece of information. An important decision we made was to define sensitivity scores at the values level, rather than at the attributes level. The reason for this is that for many attributes, the sensitivity of some value will be different than of some other value; e.g. the sensitivity of the value "homosexual" is (typically) higher than the sensitivity of the value "heterosexual".

c) *Visibility*. This reflects how accessible this specific piece of information is to other people. We will express this score with three individual sub-scores. The first is the *overall visibility score*, a continuous value in the range from 0 to 1. A score of 1 denotes that this piece of information is accessible to everyone, whereas a score of 0 denotes a piece of information that is accessible only to the user (private). The overall visibility score will depend on the OSN privacy settings on the specific content that has been used to identify the relevant value. The second visibility sub-score is a qualitative label that is related to the overall privacy score and expresses the widest possible audience to which this information is accessible. For instance, a value whose overall visibility score is 0 will have a visibility label of "Private", a value whose overall visibility score is 1 will have a visibility label of "Public" and an intermediate value will denote the widest group of people that have access to the value, e.g. "Friends" or "Friends of friends", etc. This sub-score will be called "visibility label". The third visibility sub-score will express an estimate of the *actual audience* that sees this value and we will refer to it as "actual visibility". It will be an integer number representing the actual number of users that are aware of that value and will depend on the estimates of the actual audience of the content that has been used to infer that value. It should be noted that the estimation of this score entails high uncertainty. Importantly, as will be explained later, out of these three sub-scores, only the first will be used for computing the overall privacy score and the other two will be used only for visualization.

d) *Declared/Inferred*. This is a binary value that defines whether our knowledge about the particular value comes from explicitly provided information that the user has provided or has been inferred (derived) by the WP5-WP6 algorithms. It is not an actual score but reflects information that is important for maintaining a complete view of privacy with respect to some particular value. Declared values will have a confidence of 1; however, there may be values with confidence 1 that are inferred. Additionally, in some cases a value may be both declared and inferred. In such cases, the value will be considered as declared (i.e. declared will override inferred).

e) *Support*. Just as Declared/Inferred, this is not a score per se, it is rather a structure that points to the OSN presence data based on which the other score dimensions have been filled. If the value is declared, then this will point to a single item; however, if the value is inferred this may point to a list of items. More particularly, if the value is inferred, this may point to different types of data, depending on the employed inference mechanisms. For instance, in some cases this may point to textual data, in some other it may point to images or to the network around the user. This field will allow the user to obtain a justification about why the system believes that certain variables apply to him/her.

f) *Level of control.* This score represents the ability of a user to control the disclosure of data about him/her. It ranges from 0 to 1; low values will denote a limited ability to control the disclosure of this particular data about the user, due to the involvement of entrusted and incidental data (cf. Schneier's taxonomy in Chapter 2).

g) *Privacy score.* This is a score that reflects the overall privacy exposure of a user with respect to a particular privacy value. It summarizes the other scores (confidence, sensitivity, visibility). It will range from 0 to 1 and higher values will denote a higher privacy exposure. Note that although the privacy score essentially summarizes the other scores, the model maintains a separate list of the individual scores (confidence, sensitivity, visibility) in order to support richer visualization and analysis capabilities (i.e. separate visualization of visibility and sensitivity).

The three upper levels of the proposed privacy scoring structure, namely the user, the dimensions and the attributes, are all associated with the following set of scores: a) Visibility, b) Privacy score and c) Level of control. These have similar meaning to the corresponding scores at the value level. We summarize the above in Figure 2. In addition, the top level (user) is also associated with an overall personal data value score.

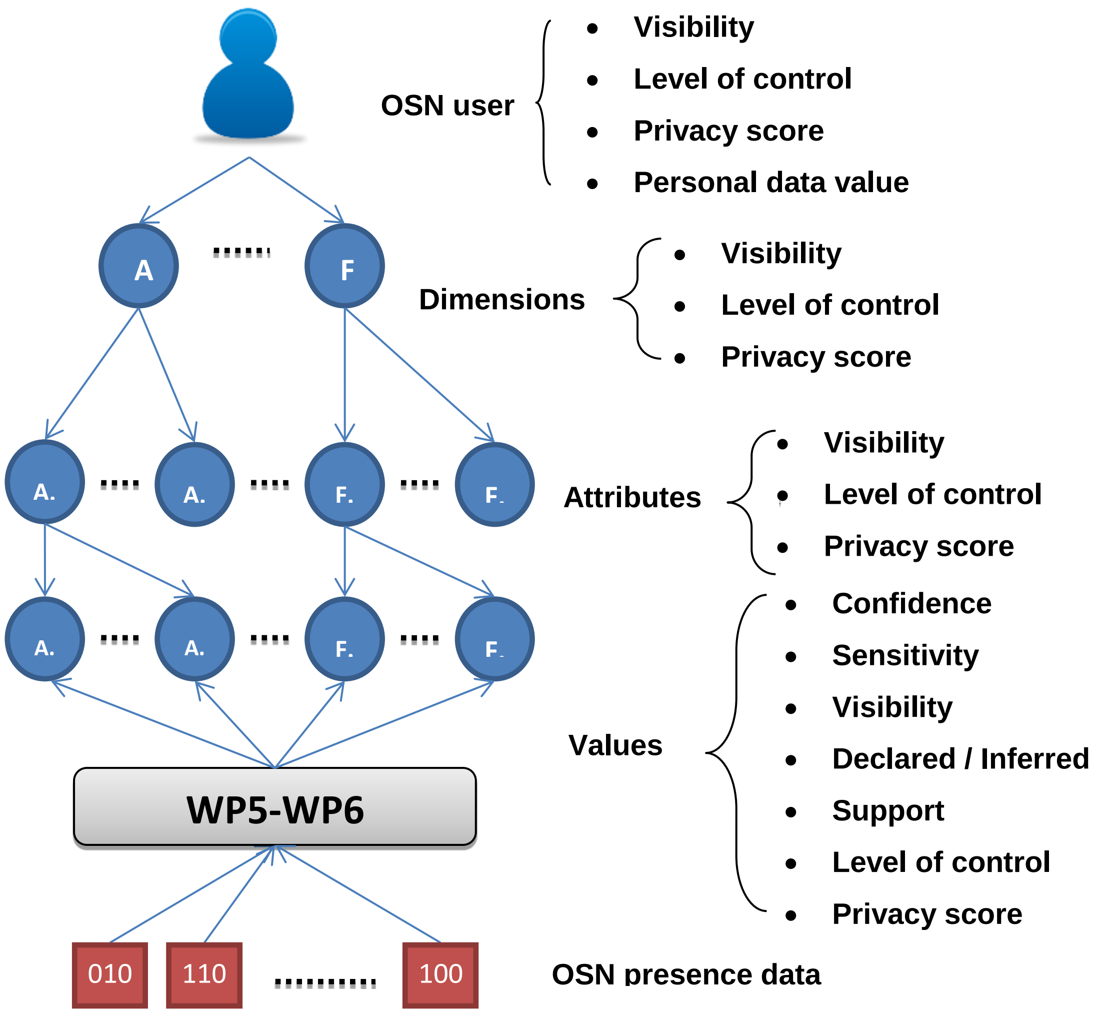


Figure 2. Privacy scoring framework levels and scores

Computation: In the following we examine how these scores are computed. Since the available data come at the bottom of the hierarchy, we will follow a bottom-up aggregation strategy. Thus, we examine how each component of the score at each level is computed from data that is available from the level below it. Let us start with the computation between the raw OSN data and the values level:

Confidence. There are two cases for filling in the confidence value. The first is when we examine a declared value, in which case the confidence is set to 1, regardless of any inferences made with respect to the same value. The second is when we examine an inferred value. Some of the inference mechanisms are described in the respective WP5 deliverables (D5.1, D5.2 and D5.3), whereas some are developed in WP6 and are described in the next chapters. Importantly, there are multiple inference mechanisms each of which will process a subset of the available data, e.g. Natural Language Processing (NLP) techniques will be utilized to process the user's posts, deep learning approaches will be utilized to process images posted by the user and label propagation techniques will be utilized to process information coming from the network around the user. Moreover, some inference algorithms will be applied on multiple pieces of data (specifically WP5 algorithms, e.g. concept detection algorithms will be performed on all images of a user). Therefore, we need a mechanism for aggregating the predictions of the multiple inference mechanisms and for different subsets of the data.

It can be observed that different privacy values may be reflected on only a subset of the data. Therefore, if for example only a small subset of the data reflects the fact that the user is homosexual, then the confidence obtained from examining this subset should not be significantly decreased by examining other pictures that do not reflect this. An additional complication arises from the fact that the confidence values for most of the attributes should sum to one. If we had a single inference mechanism working on an individual set of data, this constraint would be straightforward to handle. However, considering that we will have many different inference mechanisms and some will operate on multiple data items, we should be careful with respect to how we ensure the normalization constraint. Let us make the issue more clear with an example. Consider the attribute “sexual orientation” with possible values “heterosexual”, “homosexual” and “bisexual”. Some algorithm may provide the confidence scores 0.1, 0.8 and 0.1 for the three possible values, whereas some other algorithm that considers some other piece of data (that possibly reflects something different) may give confidence scores of 0.8, 0.1 and 0.1. If we perform aggregation using the maximum operator at value level, then we will end up with confidence scores 0.8, 0.8 and 0.1. These scores could be normalized by their sum (we would then obtain the scores 0.47, 0.47 and 0.06 respectively); however, this is likely to not be realistic in many cases. That is, given that a privacy attribute may not be reflected in all content related to a user, it makes sense to give more weight to some predictions than others. The currently proposed solution is to use only the confidence scores provided by the single inference mechanism or piece of data that maximizes an intermediate privacy score of the form:

$$\sum_i sensitivity_i \cdot confidence_i$$

where *i* runs through the different values an attribute may take. Finally, it should also be noted that this overall scheme allows us to easily plug more inference mechanisms and data. This assumes that sensitivity values are defined for all different values of an attribute.

Sensitivity. Sensitivity scores of different values will be obtained in two ways. The first is to use direct user input. The user will be given the possibility to either set explicit sensitivity scores for each privacy value or – for efficiency - to do it in a top-down manner where for example he/she will give a sensitivity score for some privacy dimension or attribute, which will then be propagated down the hierarchy. The possibility for users to directly set their sensitivity scores in different parts of the model is an important empowerment tool, directly linked to the user requirements (cf. last social requirement defined in D4.1), and turning privacy management into an ongoing and organic process of negotiating the boundaries of disclosure, identity, and time (Hong et al., 2004). The second way by which sensitivity scores can be obtained involves prior knowledge about the sensitivity scores of an average user. As already mentioned, there are several user studies that consider the relative importance of different attributes, and those could be used to set initial values to the attributes contained in the USEMP framework. However, in order to better capture all the privacy values that we examine in USEMP, we are going to ask the users that will fill in the questionnaire (as part of the upcoming pre-pilot study) to also provide sensitivity scores. Moreover, the use of an approach similar to that of Liu and Terzi (2009) will be considered. That is, we may opt for estimating the sensitivity scores automatically from a response matrix R. As mentioned, the response matrix expresses how willing a user is to disclose information about some personal item. In our case, these values may be explicitly provided by the users during the pilots.

Visibility. Appropriately setting the visibility scores is crucial as a means to enhance users' awareness with respect to institutional privacy exposure risks (cf. first two social requirements defined in D4.1 and SR05 in D2.2). Driven by these requirements, the overall visibility score will primarily be computed by taking into account the privacy settings of the OSN data that support a particular value. If the support comes from declared or inferred data that is publicly available, then the visibility will be set to 1, and if the support comes from data that is private, then it will be set to 0. Intermediate scores will be computed using a monotonically increasing function – such as a logistic function - that will depend on the size of the neighbourhood of a user to which the data is visible. The visibility label will be computed by considering the size of the audiences to which the data used to infer the value is accessible. For instance, if an inference relies on the use of two pieces of content: a piece that is accessible only to friends and another piece that is publicly available, then the visibility label will be "Friends", since only Friends have access to both pieces of content in order to perform the same inference. The actual visibility score, which quantifies the size of the audience that actually is aware of that value, will be based on estimates about the actual audience of the content that is posted. The actual audience estimation method to be used for the computation of this score has not been investigated yet. Different proxies based on the number of likes and comments that a post received could be tested.

Declared/Inferred. This field can be directly filled simply by checking whether the value of specific attributes is provided by the user in his/her profile information or not.

S*upport*. In the case of a declared value, this field will directly point to the relevant field in the profile of the user. In the case of an inferred value, this will point to the data that was used to draw the inference. It is reminded though that multiple inference mechanisms will be used and that some inference mechanisms will work on multiple data items. Therefore, it makes sense to keep pointers to all data that support some value with a significant confidence (above a threshold that will be determined empirically). Therefore, the support field will be filled by the inference mechanisms and will be a list consisting of the used inference mechanisms, pointers to the data items used, and the associated confidence levels.

Level of control. As mentioned this score expresses the ability of the user to control the disclosure of data about him/her. Referring to Schneier's taxonomy that was presented in Chapter 2, a user does not have full control over entrusted and incidental data. Thus, the level of control will be computed as one minus the ratio of the number of support items that are entrusted or incidental over the number of all support items. More formally:

$$1 - \frac{entrusted + incidental}{all}$$

Privacy score. The overall privacy score summarizes the other scores. It is a monotonically increasing function of aggregated sensitivity, visibility and confidence. In its simplest form it could be the product of these three scores. In addition, it could be modulated with the help of an appropriately shaped logistic function.

Once scores are computed at the values level, computation at the three upper levels is straightforward. In particular, only the overall privacy score, the level of control and visibility are considered for the three upper levels. For the privacy and visibility scores, the same strategy is employed for computations between any pair of levels. In particular, these scores will be computed as the averages of the corresponding scores at the level below. Another option is to pass the averages through an appropriately shaped nonlinear function, possibly with the goal to boost the privacy score a bit, and thus to increase privacy awareness. The visibility labels will be aggregated in a different manner. In particular, they will be aggregated using a max operator. Finally, the level of control will be handled by a min operator, since we are interested in highlighting attributes and dimensions, where the user has reduced (minimum) control.

As will be described in Annex II, there will be two main scoring procedures. The first is about the initial computation of the scores for a new user. The second is about the updating of the scores when the OSN content related to a user changes. In both cases, the computation will be performed in a bottom-up manner with only the affected branches being re-evaluated.

It is useful to compare the described framework with the popular scoring framework of Liu and Terzi (2009). There are some important differences between the two frameworks. The first is that the USEMP framework also considers inferred (derived) attributes, rather than only explicitly declared attributes. In order to deal with this, we have one additional score – confidence – as well as two additional fields – declared/inferred and support. Hence, the overall privacy score is not only a function of sensitivity and visibility, but also of confidence. Moreover, the USEMP framework features a score pertaining to the level of control that a user has over the exposure of individual items, attributes and dimensions. Finally, an important characteristic that distinguishes our framework from that of Liu and Terzi is that personal information and privacy scoring are organized in a hierarchical manner.

3.3. Personal data value

In addition to privacy scores, a set of personal data value indicators are also developed so that the end-users can be informed about the value of the data they are sharing. The initial set of proposed personal value indicators are based on the activities of the end user in the OSN environment and his/her OSN social graph. Two basic indicators are initially proposed:

a) a measure of influence for a specific person, referred to as Influence score and denoted with *I*, that is based on the history of the objects that the specific person has created in the OSN;
b) a measure of the importance of an object (picture/video/post), denoted with *M*, that is posted to the OSN. *M* is calculated taking into consideration the type of action on the specific object of the first- and second-hop friends of the object creator.

These two basic indicators are computed over the set of actions and the social graph of OSN users and they are combined to a single indicator of value for an object (referred to as *V*) equal to their product. The details of the overall framework for personal data value and its computational methods are presented in further detail in Chapter 7 of this document.

The presented indicators can also be computed over collections of user actions on OSNs that may be associated to different attributes and dimensions of the privacy scoring framework. This relation can be further investigated - along with additional indicators of personal data value - by future research work within WP6.

The privacy scoring framework presented in this chapter, comprising the hierarchical privacy structure, the scores associated with each level of the structure, and the personal data value of a user has been materialized in a JSON representation that is presented in Annex II.

4. Privacy Scores using Digital Trails

Through the daily usage of the applications and services offered by OSNs, their users leave a variety of digital trails that are indicative of their preferences, interests, opinions and even personality traits. These digital trails have many forms and are not limited to textual or visual content directly posted by the user. Here, and in the next two chapters, we will examine particular inference techniques that work on such data: e.g., the set of pages liked by a user or the network of friends around a user.

In this chapter, we present a fundamental approach, building on the framework of (Kosinski et al., 2013), that can be used for inferring user attributes based on a particular type of trails: Facebook likes. Although the discussion of the approach and the associated experimental study is specific to likes, the approach is applicable to other types of trails either within Facebook, e.g. commented pages, shared pages, or in other OSNs, e.g. tweeted and retweeted links, hashtags, and even in generic Web settings, e.g. visited webpages. In each case, in order to build appropriate inference models, it is necessary to have access to sets of training data, i.e. sets of users for whom we know the attributes of interest (so that we can apply the trained model on "unknown" users). In the experimental study described here, we made use of fully anonymized datasets from the myPersonality project . In order to build and deploy similar mechanisms in the USEMP system, we will make use of the user responses that will be collected during the Databait research pilot that will take place before the pilot studies (as described in D8.1).

4.1. Approach overview

In our work, user attribute inference is formulated as a machine learning problem, in which users correspond to data points, their likes correspond to features, and the user attributes correspond to the target labels or values. Table 5 presents an artificial example in order to illustrate the format and nature of the data of interest. In mathematical notation, given a set of *n* users and the likes for each one of them (to a total of *m* Facebook pages), a very sparse user-like matrix *L* ($n \times m$) is first created where L_{ij} is set to 1 if user *i* likes page *j* and 0 otherwise.

user-id	like-1	like-2	like-3	…	like-N	cat-1	cat-2	…	num-1	num-2
34354521	0	1	0		1	1	0		34	2000
32458134	1	0	1		0	0	1		42	1500

Table 5. Artificial toy example illustrating the inference problem at hand. Columns of type "like-X" correspond to Facebook Pages that may be liked by users (in which case the corresponding entry is set to 1). Columns of type "cat-Y" correspond to Categorical attributes (e.g. Male/Female), while columns of type "num-Z" correspond to numerical attributes (e.g. Age, Monthly Income).

It becomes obvious that for real-world datasets, the feature space (the number of *m* columns corresponding to likes) can become extremely large (e.g. millions of features) and sparse, i.e. users typically like only a very small fraction of all possible pages. In addition, the number of target categorical and numerical attributes can be quite large (typically tens to hundreds). Due to resource constraints and limitations in available data, only a subset of selected attributes are explored in the experimental study of the next Chapter, and will be supported by the prototype USEMP system. Attributes are selected in such a way that they enable sufficient coverage for achieving the project goals, i.e. raising the awareness of users' with respect to the capabilities of data-driven profiling, by presenting them with the inferences made based on their like history.

The basis for the aforementioned user profiling approach is the work of (Kosinksi et al. 2013). Starting from the matrix *L*, similar to the one of Table 5, the method first performs a dimensionality reduction step using Singular Value Decomposition (SVD). Through this step, the original sparse features (likes) are transformed to a small number of *k components* that are expected to capture higher-level semantics of users, and then each user is represented as a vector of numerical values, each of which expresses the degree to which the particular user matches with the respective component. The number of original features (likes) may range between a few hundred thousand to several millions (10^5-10^7), while the number of components after the SVD typically ranges in the order of tens to hundreds (10^2).

Given the new representation of users, the next step is to build a predictive model for the attributes (categorical or numerical) of interest. In the original work of (Kosinski et al., 2013), categorical variables (e.g. single vs. in-relationship) are predicted with the help of *logistic regression* classifiers, while numerical variables (e.g. age) are predicted with the help of *linear regression* models. In their original work, the authors use 10-fold cross-validation (i.e. split the dataset to 10 parts and use 9 parts for training and one for testing, and then repeat 10 times, each one with a different test part) to derive estimates with respect to the attribute prediction accuracy.

4.2. Studied extensions and contributions

We explored a number of research questions that arose in the context of USEMP, and were not sufficiently covered by the work of (Kosinski et al., 2013). In particular, we looked into the following:

- *Role of feature selection.* Due to the particularity of the original feature space (likes) one may expect that a well-designed feature selection $L^* = fs(L)$ (L^*:n×m^*,m^*<<m) may lead to sizeable improvements in terms of prediction accuracy and robustness. At the same time, discarding (non-informative) features may lead to having no features at all for a certain number of users. In our experiments, we investigated this trade-off with the goal of informing the decision for an effective feature selection strategy for the integrated system.

- *Role of training size.* Due to the sensitivity of several of the target attributes discussed here and to the limited number of users that are expected to take part in the research pilot study, we expect that little data will become available to train the inference model. To this end, we are interested in investigating what is the effect of having less training data available on the accuracy of the predictive model.

- *Role of predictive model.* In order to explore further ways of increasing prediction accuracy, we tested a number of different predictive algorithms for classification, expecting that more sophisticated models would enable the creation of higher quality inferences at the cost of more expensive computations.

- *Confidence of predictions.* A crucial requirement of making inferences about private user traits and attributes is to also produce scores reflecting the confidence (or reliability) of the predictions, as part of the privacy scoring model of Chapter 3. To this end, we devised a novel model building and testing methodology, which we explain below, and explore the extent to which the inferences about different user attributes are robust for different sets of users.

In particular, in order to capture the risk involved in misclassifying particular data points of the collection, we devised an alternative evaluation approach to cross-validation, which is grounded on the concept of bagging (Breiman, 1996). In particular, we independently sample from the available training set *B* bags, each covering *α*% of the training set, and we use them to train *B* classification models for the same target attribute. We then apply the ensemble of all *B* model outputs on the test set to derive the final predictions using majority vote. Comparing the individual model predictions with the ensemble one, we then quantify the extent to which the prediction for user *x* is reliable with the following reliability score.

$$S_x = \frac{|\sum_i^B (m_i(x)=+1) - \sum_i^B (m_i(x)=-1)|}{B},$$

where S_x stands for the reliability score for the prediction about user *x*, $m_i(x)$ is the prediction of model m_i for *x*, and -1, +1 denote the two possible labels for the classification at hand. In the case that all models agree on the decision, the score takes a value of 1, while in the case that the decision is made on the basis of just one vote in favour of the majority (*B* should be always an odd number), the score takes a value of *1/B*. Hence, analyzing the distribution of these scores for all test users, we can reason about the reliability of the performed inferences for different groups of users.

4.3. Experimental study

To study the research questions of the previous section, we carried out a large number of experiments on several anonymized Facebook datasets that we obtained after registering as research collaborators of the myPersonality project. This was a popular Facebook application that allowed users to take real psychometric tests, and enabled the researchers to record (with consent) the participants' psychological and Facebook profile.

Currently, the database contains more than 6,000,000 test results together with more than 4,000,000 individual Facebook profiles. The respondents come from various age groups, backgrounds and cultures. Not all attributes are available for all respondents; in fact, the number of training data points that are available for each target attribute varies considerably (for a few thousands to many tens of thousands). For our experiments, we focused on the following four dichotomous attributes: a) gay/straight (sexual preferences), b) single/married (marital status), c) liberal/conservative (political views), d) christian/muslim (religious beliefs), which constitute a representative sample of personal and sensitive attributes. For brevity, in the result tables we denote these variables with G/S, S/M, L/C and C/M respectively.

In the following experiments, we report classification accuracy in terms of the area under the receiver-operating characteristic curve (AUC), which is equivalent to the probability of correctly classifying two randomly selected users one from each class. For each experiment, we also report the number of users *n*, features *m*, selected features *m** and Class Balance (CB). In all experiments, the number of SVD components was set to *k=100*. The results are presented following the research questions of the previous section.

Role of feature selection: In this test we experimented with different feature selection techniques in order to see the effect that they would have on prediction performance. After the feature selection process we performed L2-normalization on the columns, kept *100* SVD components and used an ensemble of 15 neural networks with 2 layers and 2 neurons per layer, to be denoted as *NNet(2,2)*, as our prediction algorithm, which was empirically found to perform well over a number of different cases. After feature selection, it was not possible to apply the prediction model to all users, since there were a few users that were associated with no features. The number of users, for which it was still possible to make inferences, is denoted as *n**.

The simplest feature selection technique we used was to consider features (likes) that were associated with a single user. We denote this technique as FS-1 and report its results in Table 6. A first finding from this test is that even this extremely simple feature selection technique could lead to significant improvements to the prediction accuracy (up to 4%), while at the same time not reducing a lot the number of users that can be classified (at most 3% of users remain without features). A positive side-effect is that the original feature space is considerably compressed (only one-fifth of the original features are retained).

Variable	**n**	**m**	**CB**	**n***	**m***	**ALL**	**FS-1**
G/S	2,592	218,490	50/50	2,538 (98%)	46,371 (21%)	77.37	81.43
S/M	8,184	511,775	50/50	8,075 (99%)	114,181 (22%)	67.94	70.09
L/C	4,571	296,298	55/45	4,437 (97%)	59,959 (20%)	72.11	75.86
C/M	1,304	134,120	75/25	1,262 (97%)	28,468 (21%)	85.29	85.36

Table 6. Feature selection results using the FS-1 feature selection technique. Apart from the number of remaining users (n) and features (m*) after feature selection, we also report the percentage compared to the original number of users n and features m, respectively.*

The second feature selection technique, which is denoted as FS-95, aims at removing a larger number of sparsely used features. In particular, we determine which features are significant by creating an empty user-like matrix and filling in likes randomly according to the average number of likes in the original matrix. We use the 95th percentile of the total number of likes for each variable as our threshold value, below which features are removed. Table 7 reports the obtained results. It is noteworthy that further improvement gains (0.5-2.5% on top of the gains achieved using FS-1) and more aggressive feature compression is attained (less than one tenth of the original features are used), and as expected the number of users, for which it is still possible to perform predictions decreases further: predictions are not possible for a percentage of users between 6-10%.

Variable	**n**	**m**	**CB**	**n***	**m***	**ALL**	**FS-95**
G/S	2,592	218,490	50/50	2,412 (93%)	15,614 (7%)	77.37	82.75
S/M	8,184	511,775	50/50	7,732 (94%)	29,402 (6%)	67.94	70.92
L/C	4,571	296,298	55/45	4,106 (90%)	18,650 (6%)	72.11	78.47
C/M	1,304	134,120	75/25	1,195 (92%)	10,337 (8%)	85.29	87.20

Table 7. Feature selection results using the FS-95 feature selection technique.

Next, we also tested two approaches grounded on information theory, namely based on the Mutual Information, denoted as FS-MI, and Conditional Mutual Information criteria, denoted as FS-CMI (Peng et al., 2005). Using those criteria, we selected the top 500, 1000 and 2000 features and carried out the same experiments. Table 8 and Table 9 illustrate the obtained results when only 500 features are selected. One may conclude that although large accuracy gains are achieved for two of the four variables (G/S and L/C) compared to the previous feature selection techniques, for the other two variables the performance remains almost the same or even drops. Given the fact that the computation of these criteria (especially CMI) is very costly for large datasets, we conclude that the FS-95 approach strikes a good trade-off between accuracy and simplicity and is a good candidate for integration in the USEMP system. In addition, very aggressive feature selection (i.e. keeping less than 1% of the original features) makes it impossible to classify a sizeable percentage of the user base (about one fifth).

Variable	n	m	CB	n*	m*	ALL	FS-MI
G/S	2,592	218,490	50/50	2,050 (79%)	500 (0.23%)	77.37	85.28
S/M	8,184	511,775	50/50	6,654 (81%)	500 (0.10%)	67.94	69.01
L/C	4,571	296,298	55/45	3,514 (77%)	500 (0.17%)	72.11	82.59
C/M	1,304	134,120	75/25	1,003 (77%)	500 (0.37%)	85.29	86.32

Table 8. Feature selection results using the FS-MI feature selection technique and 500 features.

Variable	n	m	CB	n*	m*	ALL	FS-CMI
G/S	2,592	218,490	50/50	2,052 (79%)	500 (0.23%)	77.37	86.39
S/M	8,184	511,775	50/50	6,651 (81%)	500 (0.10%)	67.94	69.81
L/C	4,571	296,298	55/45	3,504 (77%)	500 (0.17%)	72.11	83.01
C/M	1,304	134,120	75/25	1,057 (81%)	500 (0.37%)	85.29	87.27

Table 9. Feature selection results using the FS-CMI feature selection technique and 500 features.

Role of training size: To study the role of training set size in the accuracy of predictions, we kept 10% of the available data points (users) for testing, and used subsets of different size from the remaining 90% of points. In particular, we measured the prediction accuracy when 5%, 10%, 50% and 90% of the data points were used for training (using a logistic regression classifier). Figure 3 illustrates this dependency. The examined dichotomous variables in this case include the following: a) male/female, b) gay/straight, c) lesbian/straight, d) single/married.

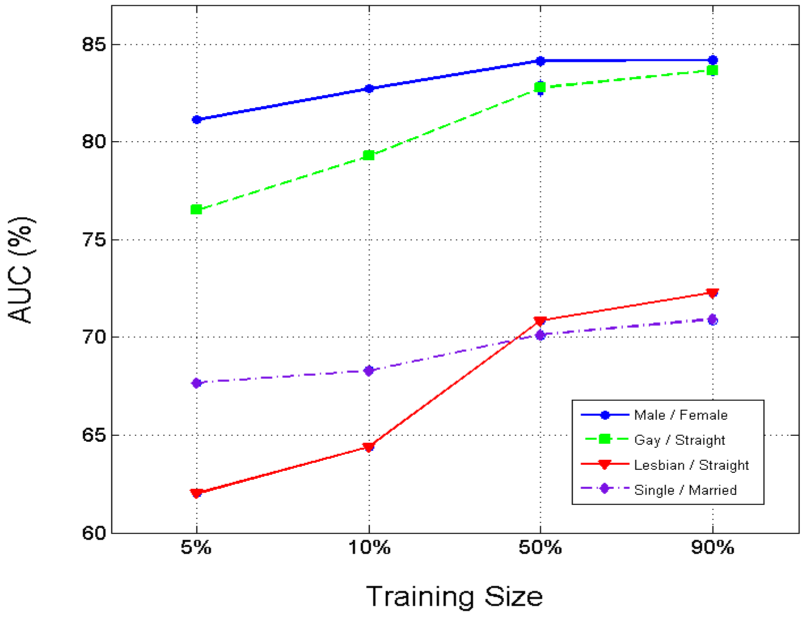


Figure 3. Role of training set size on prediction accuracy.

It is noteworthy that the effect of training size on the prediction accuracy depends greatly on the variable of interest. For instance, in the case of male/female training using only 5% versus using 90% of the dataset resulted in a drop of just 3.5% in AUC, while the respective drop for the case of lesbian/straight variable was about 10%. The main consequence of this finding for the design of the system is that for a number of variables, inferences will be highly unreliable until a certain number of training samples become available for training. For such cases, a user-centred application design would dictate refraining from presenting users with such inferences.

Role of predictive model: In the next experiment we compared the accuracy of different predictive models. In particular, we compared the following classifiers: a) logistic regression (LR), b) neural networks with 2 layers of 2 neurons each (NNet), c) K nearest neighbour classifiers (Knn), and d) linear SVM (LSVM). For each of those, we also created an ensemble model using 15 instances of the classifier trained on different subsets of the training set. Those are denoted as 15xLR, 15xNNet, 15xKnn and 15xLSVM respectively. Table 10 presents the obtained results. Note that the experiments were conducted on the full set of features for each variable; hence the values of *n*, *m* and CB of Table 6 apply in these experiments as well.

Variable	LR	NNet	Knn	LSVM	15xLR	15xNNet	15xKnn	15xLSVM
G/S	**76.35**	75.38	70.18	63.62	**77.50**	77.37	71.32	71.40
S/M	**66.21**	65.82	62.78	58.04	66.11	**67.94**	63.87	61.77
L/C	**72.77**	70.90	69.20	59.31	**72.85**	72.11	70.18	67.42
C/M	**84.19**	83.27	80.63	71.79	**85.34**	85.29	81.59	77.70

Table 10. Comparison between different classification models.

A first finding from this experiment relates to the superior performance of logistic regression and neural networks. When used as a single model, LR and NNets outperform the other classifiers (Knn and LSVM), which exhibit rather poor performance. A second finding pertains to the effect of ensemble models. While all predictive models seem to benefit from ensemble learning, LR has only marginal or no improvements, while neural networks and LSVM classifiers exhibit the greatest gains. These findings indicate that logistic regression is a reasonable choice as a high-accuracy and computationally simple classifier, yet other classification models may prove competitive when used in ensembles.

Reliability of predictions: One of the key requirements for the inference mechanism to be deployed in the system is the capability to produce confidence scores with respect to the conducted inferences. To achieve this, we devised the ensemble scheme and the reliability score described in the previous section (please note that some types of classifiers offer alternative mechanisms for confidence estimation as well). We first used the FS-95 feature selection technique (cf. Table 7) and then used $B = 25$ bags to create an equal number of LR models and tested different values of α between 20% and 100% (the latter is equivalent to not using bags). The obtained results are reported in Table 11.

A remarkable observation is that even for variables that can be predicted with high accuracy, there is a sizeable percent of users that are classified to one of the values with very low confidence. For instance, in the case of the gay/straight label and for α=80%, there are 11.8% of the test users that are classified to one of the two variables with a reliability score below 0.5. As expected, for those users the classification accuracy drops considerably

(56.6% compared to the overall accuracy of 83%). This is even more surprising, given the fact that in the case of α=80%, the 25 classification models were built using many common training samples (since each of those is an 80% random subset of the same set). For lower values of α, the percentage of users who are classified with low confidence is even higher, e.g. 32.2% for the variable liberal/conservative and α=20%.

Variable	α	AUC	U_{HC} (%)	AUC_{HC}	U_{LC} (%)	AUC_{LC}
G/S	80	83.0	70.5	90.6	11.8	56.6
	40	82.9	46.5	94.6	22.9	60.6
	20	82.3	28.7	97.6	35.1	65.1
S/M	80	69.8	72.6	75.0	11.1	53.4
	40	69.8	52.2	78.5	20.6	52.1
	20	70.1	33.1	80.1	31.2	55.5
L/C	80	77.3	77.3	82.9	8.6	53.8
	40	77.7	53.6	88.8	19.3	57.1
	20	77.7	31.0	94.0	32.2	59.3
C/M	80	85.1	64.5	94.7	14.7	53.9
	40	84.1	40.9	97.2	23.6	58.8
	20	84.4	27.4	95.7	29.3	67.1

Table 11. Reliability of attribute inferences. AUC_{HC} and AUC_{LC} stand for the AUC scores for high- and low-confidence classifications respectively. A high-confidence classification for user x occurs when S_x=1, while a low-confidence one occurs when S_x≤0.5.

Hence, given the fact that such highly unreliable classifications are often used for targeting users, and sometimes have serious real-world consequences (e.g. not being selected for a job), one should be really cautious against inferred user profiles, and could raise serious ethical concerns with respect to the overall practice of mining user profiles from observed data. Yet, being able to quantify the reliability of the performed inferences based on the developed methodology could be at least used as a measure to mitigate the risk of erroneous profiling (by deciding to not make inferences at all for those users for which the *S* scores are low). This in fact strongly justifies the decision to include a confidence score in our scoring framework.

Regarding the deployment of the predictive module on the system, the last two experiments make clear that using ensembles of classifiers is beneficial in two ways: a) improving the prediction accuracy of the inferences, b) deriving a reliability score, which strongly correlates with the accuracy of the prediction, and can thus be used as the confidence score of inferences defined by the USEMP privacy scoring model.

5. Privacy Scores using Latent Topics

In the previous Chapter, we examined a method for inferring personal information using the set of pages that the user has liked. This was found to be a versatile and effective inference approach; however, it has some limitations. The most important is that many pages will have received likes only from one or very few users. This will make inference based on these likes highly unreliable (as shown in the feature selection experiments of the previous Chapter). Nevertheless, it is clear that it should be possible to infer something about a user that has liked pages that have received very few likes, based on the content of the liked pages.

In this Chapter, we propose to generate an alternative content-based representation for the set of pages that the user has liked, leading to an approach that does not suffer from the aforementioned limitation. This stems from the fact that, in a variety of prediction tasks, the extraction of appropriate features or latent variables to be used as inputs can improve performance significantly. In general, there are two ways in which such features may be extracted. The first is using some prior knowledge about the nature of the features to be extracted. The second is to use some purely statistical method to extract the features in an unsupervised manner. When prior knowledge is available, the first is to be preferred. In the case of profile attribute prediction, such a set of features is the set of topics that a person is interested in. Intuitively, the distribution of the topics that a user is interested in will reflect some things about him/her, at least for some attributes. For instance, younger people are more likely to be interested in topics related to education than older people. Also, extrovert people are more likely to express a wider distribution of interests and so on. Thus, in the following, we present an approach for personal information prediction, utilizing as features the set of topics that the user is interested in. As will be discussed, this approach is applicable both to the scenario of processing sets of liked pages, as well as to the scenario of processing other types of content posted by the user.

Note that the approach described here is complementary to the Explicit Semantic Analysis (ESA) approach presented in D5.1, in that ESA relies on an explicit semantic model (concept space) such as Wikipedia, whereas the method described here makes use of a latent topic space that is “learned” from large corpora of Facebook pages. We expect that ESA will be a robust and generally applicable approach for mapping users to specific concepts/attributes of interest, while the approach described here would become more accurate as soon as a set of appropriately selected Facebook pages are collected in order to generate an appropriate latent topic space (i.e. one that covers well the topics of interest for the test users).

5.1. Approach overview

The main idea of the proposed approach is to extract a set of topics, reflected in the set of pages that the user has liked, and use them as features in a set of classifiers that predict various personal attributes. The mapping from sets of liked pages to sets of topics of interest can be done by performing an unsupervised topic detection technique on the description or the posts of the liked pages. The basic idea is illustrated in Figure 4.

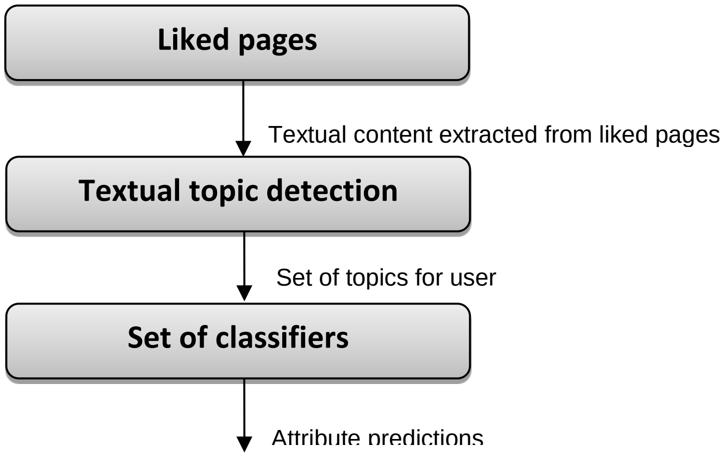


Figure 4. Topic-based attribute inference.

It should be noted here that the same SVD transformation that was used in the previous chapter essentially generates appropriate features, this time from the text content of liked pages, to be used for classification. In some sense SVD performs a kind of topic detection. More particularly, the well-known topic detection technique of Latent Semantic Analysis (LSA) performs SVD on a term-document matrix, resulting in a set of topics, each of which is related to a set of terms. The approach described in the previous chapter on the other hand, performs SVD directly on a like-user matrix, resulting again in a set of topics, each of which is related to a set of pages, rather than to a set of terms. The advantage of having topics represented by sets (or distributions) of terms is that it will be possible to extract topics even for pages never seen before, based on the content that appears on these pages (assuming that the respective terms are included in the learned topic model). Moreover, as mentioned, such topics are applicable to text data coming from other sources, not only from liked pages, resulting in an additional channel of data that can be used for conducting inferences.

In order to perform this procedure, it is necessary to obtain a large set of possible topics from a large and representative corpus (the topic corpus). Then, the text from new (and potentially unknown) pages would be used to associate the page with the topics of the topic corpus. The reason that this is done, rather than performing topic detection from scratch for the text documents of each user, is to keep the semantics of the extracted features constant. That is, if we extracted a new set of topics for the liked pages of some users, then we would have to map them in a common representation so that they can be used as inputs to a classifier. If on the other hand we perform topic assignment on a set of already known topics, then we make sure that a consistent feature representation is used.

In short, the proposed procedure is the following:

1. Create a topic model
    - Collect a large set of documents representing a very wide range of possible topics
    - Perform topic modelling with the use of Latent Dirichlet Allocation (LDA). Other topic detection and modelling are possible to use in case they are found to be more effective.
2. Extract topic-based representation for users with known labels
3. Train attribute predictors and apply them to “new” users

The last step in the above procedure, involves a supervised learning task; therefore training data for the personal attributes that will be predicted are required. That is, it will be necessary to have the set of liked pages (and/or other textual content) of a number of users in order to build their topical profiles, but we will also need their values for the considered attributes so that we can train a classifier.

As mentioned, this is applicable to data other than liked pages. Actually, any type of textual content that expresses the interests of a user may be used. For instance:

- Posts
- Self-description in profile page
- Descriptions of joined groups

In the case of the latter (descriptions of Facebook groups), care should be taken, since the description are not authored by the user him/herself, but by group administrators and they can be modified after the user has subscribed. Hence, the decision on whether such descriptions should be used for performing inferences should be carefully evaluated.

5.2. Experimental study

The data we used for this study are also part of the myPersonality dataset. In particular, the myPersonality dataset includes the distribution of (fully anonymized) topics related to a set of users. These topics have been derived applying LDA on the descriptions of the liked pages. Thus, we used these data together with the available labels to train and test our classifiers. Importantly though, after the research pilot study, and provided that sufficient data are available, we will train new classifiers that will be used to the deployed system with our own data. Eventually, the classifiers that we will build will target the particular privacy attributes that we focus on in USEMP.

In the following, for comparability purposes, we focus on the binary classification tasks that we also tested in the previous chapter. In particular, we examine the following pairs of labels: a) gay/straight (G/S), b) single/married (S/M), c) liberal/conservative (L/C), and d) christian/muslim (C/M). We used an SVM classifier and our performance measure was again AUC. The results can be seen in Table 12. For comparison, we also list here the best results achieved for each classification task when different predictive models are used (from Table 10) and when different feature selection mechanisms are used (from Table 6-Table 9).

Variable	**TB**	**LP**	**FS**
G/S	80.01	77.50	**86.39**
S/M	68.19	67.94	**70.92**
L/C	77.02	72.85	**83.01**
C/M	**89.88**	85.34	87.27

Table 12. Comparison between results achieved using topic-based features (TB) and results achieved using the approach of the previous chapter with the original features, i.e. liked pages (LP) and using Feature Selection (FS)

Interestingly, the topic-based approach performs much better than the like-based approach of Chapter 4 in the absence of feature selection. However, it can be seen that when feature selection is performed, the performance for most of the variables is even higher with the exception of the last attribute, where the topic-based approach performs significantly better.

An advantage of the topic-based method over the feature selection one is that it is applicable to all users (even those with a single liked page in their history), while the approach using feature selection is only applicable for users whose features are retained by the selection process.

In fact, the above observations indicate that indeed, the topic-based approach can improve the performance and that it is a necessary complement of the like-based approach. Additionally, the above results also support the hypothesis that the performance may be further improved when feature selection is performed; therefore, testing the feature selection mechanisms also on the topic-based approach appears to offer a promising extension that should be explored in the future. Moreover, these results, just as those from the previous chapter, indicate that some attributes are easier to predict than others.

Additionally, despite the fact that labels to train the classifiers for the specific attributes that USEMP will focus on will be available after the start of the pilots, we have already experimented with the creation of relevant topic models. In particular, we collected the basic page information for a large number of public Facebook pages (~600,000) using the Facebook Graph API. These were collected using as query very popular English words, such as “a”, “the”, “very”, etc. in order to collect a very wide range of topics. For each of the pages we created a document containing the following fields: about, description, category. We used this set of documents to perform topic detection using LDA. Table 13 lists some example topics that were produced. Each topic is represented with a list of keywords. A similar approach will be repeated to derive topics in the languages of interest for USEMP (French, Dutch, and Swedish).

#	**Keywords**
1	image, artist, photo, capture, video, picture, photographer, moments, camera
2	song, music, band, track, chart, record, album, release, video
3	career, education, work, job, recruitment, position, job, interview
4	makeup, product, beauty, natural, color, nail, skin, health, hair, body
5	sports, yoga, health, weight, exercise, training, body, fit, nutrition, workout, fitness, gym
6	country, law, political, court, people, government, police, public, rights, organization, legal, national, justice, state, human, party

Table 13. Example topics produced by performing LDA on the descriptions of 600K Facebook pages

After the pre-pilots, provided we have collected a sufficient amount of training data, we will perform the next steps of the procedure and will conduct more thorough evaluations.

6. Privacy Scoring based on User Links

In this chapter, we examine another mechanism for inferring personal information, alternative to the ones developed in WP5 and to those of the previous two chapters. This time we look at a quite different source of information. More particularly, we look at what the network around a user can reveal about the user. In fact, when OSN users connect and interact with each other, they implicitly (and most of the times without intention) disclose information about each other. It is therefore important to study the implications stemming from a user's links and interactions with other OSN users on their private information, and to be able to quantify the level of privacy risk they are exposed to when interacting online with each other.

To this end, we devised a novel graph-based supervised learning approach that produces inferences about user attributes based on a number of users with known attributes who operate in the same network. In particular, the proposed approach extracts a set of graph structure features for each user in a social network, and then uses these features and a set of known users to generate a predictive model that relies only on the structure of the network. Hence, even when a user does not disclose personal information about him/her, it is still possible to draw inferences about them on the basis of the user's connections. To study the effectiveness of our approach we test it on a publicly available Twitter dataset, including the interactions between users (by means of mentions) and reference user attributes (political opinions, religion, and location) that are constructed semi-automatically. The approach, however, is generic in nature, and could be applied on any online social network assuming that access to information about the network structure is available.

This study is of utmost relevance to the USEMP scope and objectives; hence, as part of the deliverable, we make available a prototype implementation of the approach and an associated anonymized dataset to demonstrate its potential for conducting inferences solely based on linkage information. However, there has been a decision not to make use of it in the pilot studies: this decision was necessitated by legal and ethical constraints, since utilizing data coming from a user's connections without their explicit consent was deemed as an unacceptable practice even when conducted purely in the interest of scientific research.

6.1. Approach overview

Similar to the previous Chapter, user attribute inference is formulated as a machine learning problem, and more specifically as a semi-supervised learning problem: given a set of interactions among the OSN users of interest, we first build an interaction graph (or network) $G=(V,E)$, where V is the set of users (graph nodes) and E is the set of the connections between them (graph edges). The connections can be either explicit links, e.g. friendship or followership relations, or interactions, e.g. one commenting on another's post (in Facebook), one mentioning the other (on Twitter), etc. Once the graph is constructed, we extract graph-based features X, i.e. a k-dimensional feature representation for each user. In the final step, using the known values for the attributes of a small number of seed nodes (users) for training, we build a predictive model that can be used to infer these attributes for the rest of the nodes. Although the last step could be implemented using standard supervised learning methods, the overall framework is semi-supervised due to the fact that the k-dimensional features are constructed based on the full user graph, i.e. the graph comprising both "known" (training) users and "unknown" (test) ones. The framework is illustrated in Figure 5.

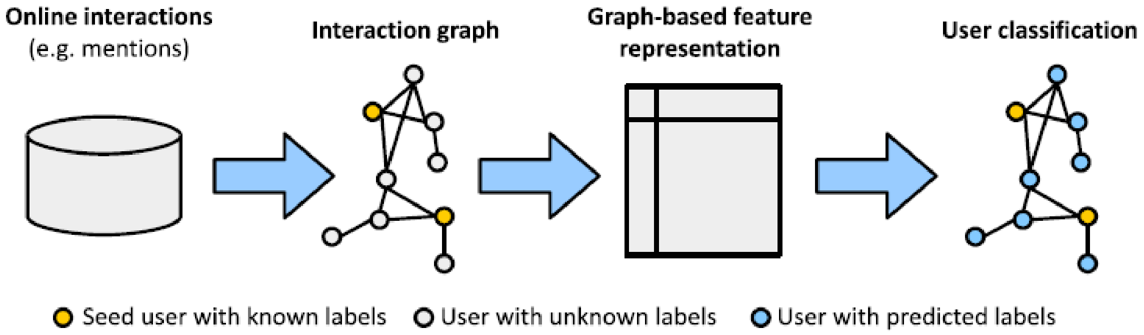


Figure 5. Overview of user attribute inference framework based on user links to other OSN users.

The proposed framework, which we term Absorbing Regularized Commute-Time Embedding (ARCTE), is grounded on the principle of *homophily* (McPherson et al., 2001), according to which people sharing the same beliefs and interests tend to connect to each other and as such we expect them to form denser than average communities. There are existing approaches for graph-based user classification that involve *either embedding the graph* onto some latent Euclidean space such that similar users are proximal to each other or alternatively using *community detection* methods (Papadopoulos et al., 2012) to find dense communities of similar users that capture common latent variables and employing them as features for a classifier to infer user interests.

The first group of existing methods, based on graph embedding, attempt to project a network onto a latent space such that its structure is retained in some sense. Such techniques reduce the dimensionality of the problem, providing compact features in the form of continuous coordinates of data points. For example, Laplacian Eigenmaps (Belkin & Niyogi, 2003) aim to preserve local structure. Probabilistic modeling of networks has resulted in the Latent Space Model (Hoff et al., 2002), which for now is limited to smaller scale networks. Another recently popular family of similar methods is Neighbor Embedding (Hinton & Roweis, 2002), in which the latent space distances denote the probability that nodes are neighbors, however, their optimization is computationally costly and prone to local optima.

The second group of existing methods, based on community detection, considers that membership of nodes to communities can act as an effective feature representation for user classification. A number of scalable methods have been proposed that include Louvain (Blondel et al., 2008), OSLOM (Lancichinetti et al., 2011) and BigClam (Yang & Leskovec, 2013). Furthermore, there exist methods explicitly designed for the problem of user classification: MROC (Wang et al., 2013) and EdgeCluster (Tang & Liu, 2009). Of interest are local methods that exhibit scalability, such as the random walk-based Nibble algorithm (Spielman & Teng, 2008) and its improvement PageRank-Nibble (Andersen et al., 2006) that have led to improved speed and cuts of good conductance.

The proposed approach, ARCTE, builds upon the concept of a local PageRank search to generate a feature representation for an input node (user) consisting of other users in the network that should be similar (or proximal) to it. Here, similarity or proximity is defined on the basis of the underlying link structure. To identify this set of similar users, ARCTE looks beyond the immediate neighbors of the input user and may include additional users in the feature representation due to their common membership in a community.

6.2. Studied extensions and contributions

ARCTE is divided into two steps to get two sets of communities: X = embed($C_1 \cup C_2$). The first set of output communities is the set of *base communities*. Each base community around node j is defined as its immediate neighbourhood plus the node itself: $c_j = N(j) \cup j, \forall j \in N$. As for the second set, we want to identify sets of nodes around each user by searching locally for highly probable random walk destinations. Specifically, we first compute the Regularized Commute-Time kernel (Fouss et al., 2012), a variant of the well-known PageRank, then sort nodes according to the degree-normalized ranking, and select the fewest possible highest ranking nodes as the local community such that they comprise a strict superset of the corresponding base community.

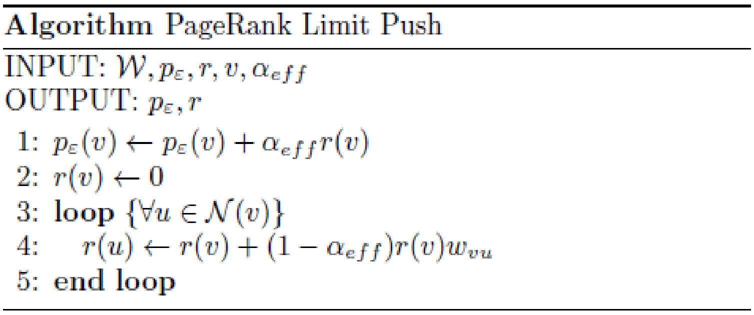
A key contribution of ARCTE involves a much more efficient means of computing the Regularized Commute-Time (RCT) kernel (i.e. pairwise similarities between nodes of the graph), by avoiding self-transitions of probability, in order to identify the nodes belonging to community C_2. This makes it applicable to massive interaction networks. To build the RCT kernel, i.e. the similarity matrix that encodes the pairwise similarities between nodes, ARCTE makes use of Random Walks with Restart (RWR). Each row i from the RWR matrix is equal to the PageRank (Page et al., 1999) when the probability is concentrated on the i-th node in the personalized distribution s (Haveliwala, 2002). We call this distribution a *node-centric PageRank*. The equation describing this process is the following:

$$p(a,s) = \alpha \cdot s + (1-\alpha)\, p(\alpha, s \cdot W)$$

where α is the restart probability and W is the random walk probability transition matrix. We denote by w_{uv} the probability of transitioning from node u to node v. An ε-exact algorithm that scales as $O(d_{ave}/\alpha\varepsilon)$, where d_{ave} is the average degree of the graph, was first described in (Andersen et al., 2006). Here, we describe two innovations with respect to this algorithm, one pertaining to time efficiency and the other to the calculation of a similarity slice with greater predictive potential. The algorithm of (Andersen et al., 2006) involves retaining two distributions, the approximate PageRank p_ε and the residual distribution r, such that

$$p_\varepsilon + p(\alpha,r) = p(\alpha,s)$$

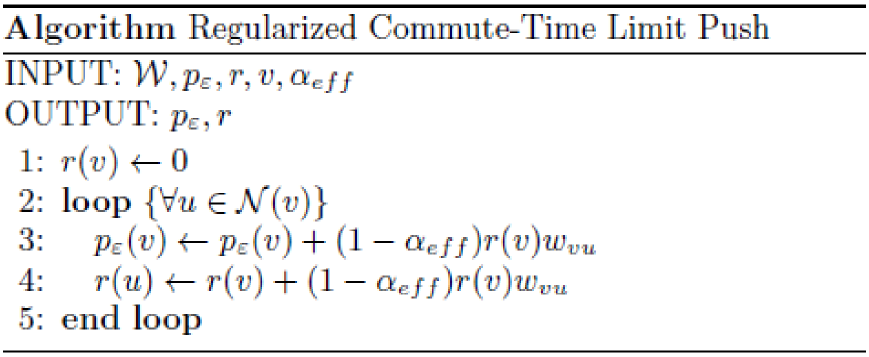
If $r(u)/d(u) \geq \varepsilon$, where $d(u)$ denotes the degree of node u, probability mass proportional to α is pushed from $r(u)$ to $p_\varepsilon(u)$ and the rest is allocated to the node's neighbours as if a random step were applied on it. The aforementioned push step, as described in (Andersen et al., 2006) is carried out in an iterative manner, which is sub-optimal from a computational point of view. To this end, in ARCTE we devised the PageRank Limit Push algorithm that converges to the same solution without the need to perform iterations to account for self-transitions.



```
Algorithm PageRank Limit Push
INPUT: $\mathcal{W}, p_\varepsilon, r, v, \alpha_{eff}$
OUTPUT: $p_\varepsilon, r$
$p_\varepsilon(v) \leftarrow p_\varepsilon(v) + \alpha_{eff} r(v)$
$r(v) \leftarrow 0$
loop $\{\forall u \in \mathcal{N}(v)\}$
    $r(u) \leftarrow r(v) + (1 - \alpha_{eff}) r(v) w_{vu}$
end loop
```

Figure 6. PageRank Limit Push algorithm

Our second improvement is an adaptation of the pairwise similarity step, such that it calculates similarities with better predictive power, with the same number of push operations. Consider a random walk process that with probability *α* the walker is absorbed, and with (1-*α*) proceeds as normal. The cumulative probability of being at each node is the similarity value we calculate. This is further specified in the following Regularized Commute-Time Limit Push algorithm of Figure 7.



```
Algorithm Regularized Commute-Time Limit Push
INPUT: $\mathcal{W}, p_\varepsilon, r, v, \alpha_{eff}$
OUTPUT: $p_\varepsilon, r$
$r(v) \leftarrow 0$
loop $\{\forall u \in \mathcal{N}(v)\}$
    $p_\varepsilon(v) \leftarrow p_\varepsilon(v) + (1 - \alpha_{eff}) r(v) w_{vu}$
    $r(u) \leftarrow r(v) + (1 - \alpha_{eff}) r(v) w_{vu}$
end loop
```

Figure 7. Regularized Commute-Time Limit Push algorithm

6.3. Experimental study

We tested ARCTE on the interaction graph (comprising replies/mentions) extracted from the publicly available SNOW dataset (Papadopoulos et al., 2014). The largest weakly connected component of this graph consists of 533,874 users and 965,821 mention/reply edges. This graph was used to generate the feature-based representations *X* for all nodes. User labels were then generated for 13,000 users of this graph, by computing the PageRank on the mention graph and selecting the top ranking users. For each of those, we used the Twitter API to collect up to 500 Twitter lists that these users belong to. The list names and descriptions were then tokenized, stop words were removed, and the remaining tokens were lemmatized. With manual inspection, some lemmas were removed and others were merged into a final list of labels, which is illustrated in Table 14.

Dimensions	**Attributes**
Political Opinion	democrat, libertarian, conservative, feminist
Religious Beliefs	islamic, christian, atheism, judaic
Location	china, france, malaysia, london, sweden, south_west, greece, leicester, italy, colombia, spain, usa, chile, newzealand, atlanta, england, iraq, germany, korea, iran, newyork, east_midlands, pakistan, argentina, manchester, thailand, afghanistan, belgium, jersey, north_west, india, birmingham, indonesia, canada, egypt, texas, wales, libya, yorkshire, liverpool, east_of_england, russia, saudiarabia, israel, scotland, japan, south_east, west_midlands, florida, turkey, jordan, ireland, brazil, california, mexico, seattle, lebanon, bristol, glastonbury, illinois, nigeria, australia, hongkong

Table 14. Private information attributes used to label the SNOW dataset Twitter accounts (note that the tags are automatically produced by the labelling process, hence some of them may not be linguistically and grammatically optimal).

Using TF-IDF scoring, we computed a user-label matrix and then selected as “correct” labels, those that were associated with a user with a score equal to, or higher than the 75th percentile of normalized frequencies. We then used between 1% and 30% of nodes for training and the remaining for testing, using an SVM to learn the feature-based representation. Classification accuracy was measured in terms of the micro- and macro-average F-score. Macro-averaging means that the F-score is calculated independently for each different label and then averaged, while micro-averaging implies summing up the individual true positives, false positives, and false negatives of the methods over all labels. To test the effectiveness of our proposed representation, we compare it with the F-score achieved using previously proposed representations, namely the Laplacian Eigenmaps - LEs (Belkin & Niyogi, 2003), MROC (Wang et al., 2013), and the base communities (BaseCom) defined above.

		1%	**2%**	**3%**	**4%**	**5%**	**10%**	**15%**	**20%**	**25%**	**30%**
F1macro	ARCTE	12.48	13.28	13.65	15.25	15.52	18.14	**20.5**	**22.31**	**23.60**	**24.57**
	LE	**16.53**	**17.94**	**18.81**	**19.40**	**19.55**	**20.65**	20.33	20.67	20.67	21.05
	MROC	11.34	11.85	12.14	12.77	12.79	14.54	15.71	17.28	18.94	19.73
	BaseCom	11.22	11.50	11.79	12.15	12.03	12.95	13.73	14.24	15.27	16.17
F1micro	ARCTE	**47.34**	**47.57**	**47.74**	**48.71**	**48.71**	**50.69**	**51.29**	**51.93**	**52.75**	**53.48**
	LE	33.07	38.04	37.63	40.76	41.52	42.73	43.58	43.85	43.98	44.10
	MROC	47.08	47.37	47.77	48.04	48.02	49.18	49.42	50.24	51.05	51.65
	BaseCom	46.97	47.20	47.56	47.58	47.49	48.09	47.92	48.19	48.72	49.19

Table 15. Comparative classification accuracy in terms of F1-macro and F1-micro scores between the proposed method (ARCTE) and other three state-of-the-art methods: Laplacian Eigenmaps (LE), MROC and Base Communities (BaseCom). Classification is reported for different training sizes. The accuracy is measured with respect to the Political Opinion and Religious Belief labels of Table 14.

		1%	**2%**	**3%**	**4%**	**5%**	**10%**	**15%**	**20%**	**25%**	**30%**
F1macro	ARCTE	5.05	5.83	7.55	8.24	9.00	**11.13**	**12.97**	**14.12**	**14.99**	**15.38**
	LE	**6.68**	**7.44**	**8.27**	**8.77**	**9.52**	10.34	10.86	11.22	11.63	11.73
	MROC	3.25	3.39	4.54	5.28	5.65	7.74	9.68	10.78	11.94	12.76
	BaseCom	2.83	3.01	3.60	3.94	4.27	6.15	7.62	8.94	9.82	10.60
F1micro	ARCTE	**25.35**	**26.64**	**28.48**	**29.60**	**30.21**	**32.34**	**33.81**	**34.56**	**35.06**	**35.61**
	LE	19.38	24.12	26.34	27.59	29.13	30.91	31.54	31.90	32.38	32.60
	MROC	23.73	24.24	25.13	26.15	26.39	28.75	30.91	31.80	32.58	33.30
	BaseCom	23.32	23.73	24.08	24.27	24.28	25.74	27.13	28.07	28.60	29.49

Table 16. Comparative classification accuracy using the same setup of Table 15. The accuracy is measured with respect to the Location labels of Table 14.

Table 15 and Table 16 present the obtained results. According to those, it appears that the proposed method, ARCTE, performs consistently the best compared to all alternatives in terms of micro-averaged F-score, while in terms of macro-averaged F-score it performs best when more than 10% of the training samples are available. For lower numbers of training samples, LE appears to perform better (in terms of macro-averaged F-score). However, it is noteworthy that the macro-averaged score is particularly sensitive to labels that are very sparse (i.e. for which very few users are associated with them). Furthermore, ARCTE is the fastest feature representation to compute compared to the rest (with the exception of the BaseCom features, which are trivial to extract). Overall, it appears that ARCTE offers an effective and at the same time scalable method to perform multi-label user classification.

The fact that the feature representation of users in these experiments was generated based on their online interactions has two profound issues. First, online interactions appear in a streaming fashion, hence the structure of the corresponding interaction networks and ultimately the values of the extracted features will depend on the period and time of observation. In the case of the SNOW dataset, for instance, the interactions were recorded over the course of 24 hours. One could therefore doubt about the validity or stability of labels for those users. To address this concern, longer observational studies of a longitudinal nature would be required. Second, several of the observed interactions are not motivated by interest on a topic, but are the result of other social mechanisms (e.g. casual chat) or even spam (e.g. attempts to maximize one's reach). To mitigate the confounding impact of such interactions on the proposed analysis, appropriate filtering mechanisms should be devised. A related issue appears in the context of interactions between users that “agree” on a certain label, but disagree on others, necessitating the development of appropriate label-dependent homophily characterization mechanisms.

In conclusion, the proposed method and associated experimental study demonstrated the potential of graph-based content-agnostic methods for generating inferences about private attributes of OSN users based on the network of their interactions with other OSN users. The implications of this work are profound for the analysis of institutional privacy concerns. In particular, given that OSN service operators and third parties have access to rich network-based data in the form of user interactions and connections, and that a considerable part of their users choose to disclose several pieces of private information about them, it is possible for them (the OSN providers) to considerably enrich the set of inferences they make for the rest of the users, not only based on the content of their posts (e.g. using techniques described in Deliverables D5.1 and D5.2) and their behavioural data (e.g. using techniques described in the previous chapters), but simply on the basis of their interactions and links with other OSN users.

7. Personal Data Value Scoring

7.1. Overview of personal data value scoring

In modern societies personal data is an asset for those that generate them as well as for those organizations that collect, store, analyze and use them. Similar to every asset in modern societies, the valuation of personal data is necessary to identify which information is the most important and the most valuable. The value extracted from European consumers' personal data was worth €315bn in 2011 and has the potential to grow to nearly €1tn annually in 2020, according to recent research conducted by the Boston Consulting Group (Rose et al., 2012).

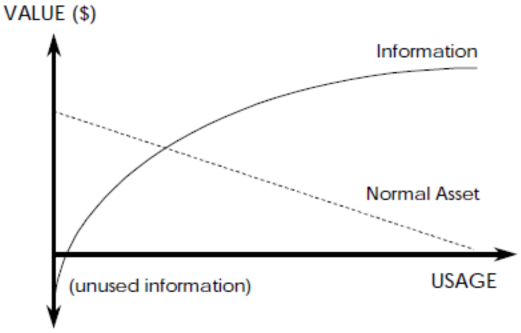


Figure 8. The value of information increases with its use.

Information actually increases in value the more it is used (Figure 8), while the major cost of information is in its capturing, storage and maintenance (Moody & Walsh, 2002). The measurement of the value of personal data is a complex and difficult task. There is no commonly accepted methodology for estimating the value of personal data. Existing approaches rely on:

a) market valuations of personal data, or other related market measures
b) individual perceptions of value of personal data and privacy.

For the first approach the market cap/revenues/net income per data record, market prices for data, cost of a data breach, data prices in illegal markets are some of the proposed methods to estimate the value of personal data. Alternatively, surveys and economic experiments can be used to estimate the individual valuation of personal data and the individual valuation of privacy. However, this is a complex and context-dependent task (OECD, 2013), and in most cases somehow related to a monetary value.

In the context of the USEMP project, feedback to users on the value of personal data is expected to raise their awareness concerning the value of personal information that they share through OSNs by addressing the research question of which shared information is most important or most valuable to a user's OSN social graph (alternatively referred to as user's audience).

For that purpose, it is necessary to model the process of the personal data valuation performed by OSNs. In this context, the USEMP project investigates a new composite approach by collecting and computing indicators of scores related to the audience in a network (data producers, data consumers), and the usage of data. The audience describes the number of OSN users that have an action (e.g., share, like, comment) on an object that another OSN user creates, while the usage of data describes the type of actions taking into account the time duration of these actions. Figure 9 depicts the introduced framework for the valuation of personal data. The above-mentioned indicators that are calculated for each OSN user and the created objects are provided as an input to the utility function for the estimation of the value of personal data. The calculated value is provided to the end user in order to increase his/her awareness.

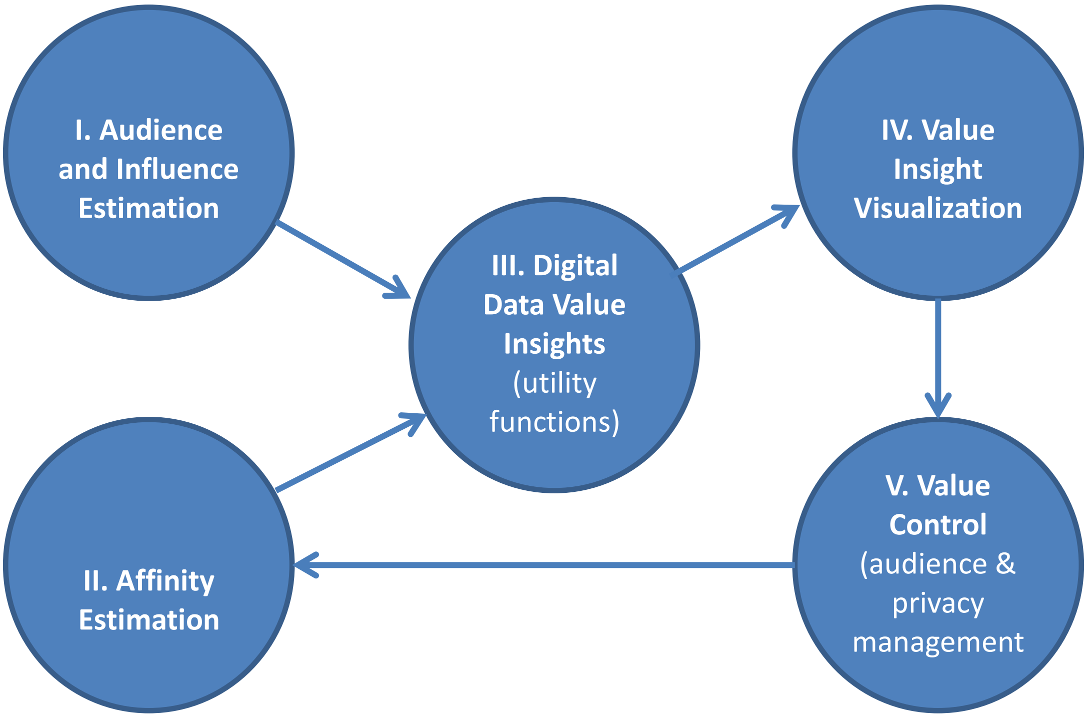

I. Audience and Influence Estimation
IV. Value Insight Visualization
III. Digital Data Value Insights (utility functions)
II. Affinity Estimation
V. Value Control (audience & privacy management

Figure 9. USEMP framework for Personal Data Value

7.2. Personal data value metrics

In the following paragraphs we describe the first set of metrics that are proposed to be evaluated for the estimation of user's shared data influence and importance in the context of an OSN. These metrics that can be estimated in an actual OSN environment (such as Facebook) are proposed to be evaluated along with additional affinity measures and digital value insights (like data privacy utility functions estimation) in additional theoretical and experimentation work in WP6.

The value of personal data, as introduced in Section 3.3, is denoted by *V* and it is calculated based on the influence of the person that creates an object to the OSN, denoted by *I* as well as on the importance of the object (picture/video/post) that is posted to the OSN, denoted by *M*. Both values range between 0 and 1.

The *Influence score* of a specific person is estimated based on the history of the objects that the specific person has created, while taking into consideration the

- number of connections comparing to the total number of users of the network;
- the types of actions (share, like, comment) of the first and the second hop friends on the objects that the corresponding person has uploaded/created to the OSN.

$$I = \frac{\sum_{i=1}^{n}\left(\frac{\alpha \frac{\sum_{j=1}^{e} A_j^i}{f_1} + \beta \frac{\sum_{k=1}^{g} A_k^i}{f_2}}{2}\right)}{n}$$

For the calculation of user influence I the following parameters are introduced:

- *n:* number of objects (i.e., picture/video/post) that a user has created
- *i:* index of an object that a user creates
- α: number of first-hop friends comparing to the total number of users of the OSN
- β: number of second-hop friends comparing to the total number of users of the OSN
- j: index of first-hop friends that had an action on an object of the corresponding user
- *k*: index of the second-hop friends that had an action on an object of the corresponding user
- e: total number of first hop friends that had an action on the object *i*
- g: total number of second hop friends that had an action on the object *i*
- A_j^i: type of action (i.e., share, like, comment) of user *j* on the object *i*
- A_k^i: type of action (i.e., share, like, comment) of user *k* on the object *i*
- f_1: total number of first-hop friends
- f_2: total number of second-hop friends

Parameters *α* and *β* provide an indication of the position of the user in the OSN in terms of the number of first- and second-hop friends, respectively. For the calculation of *α*, the total number of the first-hop friends is associated with the average number of first-hop friends of all OSN users, denoted by $\bar{f_1}$, and the average number of first-hop friends of the most popular OSN users (top 10%), denoted by $\overline{f_1^{10\%}}$:

$$\alpha = \begin{cases} w_1 \frac{f_1}{\bar{f_1}} + (1-w_1)\frac{f_1}{\overline{f_1^{10\%}}}, & f_1 < \bar{f_1} \; and \; f_1 < \overline{f_1^{10\%}} \\ w_1 + (1-w_1)\frac{f_1}{\overline{f_1^{10\%}}}, & f_1 > \bar{f_1} \; and \; f_1 < \overline{f_1^{10\%}} \\ w_1 \frac{f_1}{\bar{f_1}} + (1-w_1), & f_1 < \bar{f_1} \; and \; f_1 > \overline{f_1^{10\%}} \\ w_1 + (1-w_1), & f_1 > \bar{f_1} \; and \; f_1 > \overline{f_1^{10\%}} \end{cases}$$

where w_1 reflects the importance of each of the two terms in the above equation.

$$\beta = \begin{cases} w_2 \frac{f_2}{\bar{f_2}} + (1-w_2)\frac{f_2}{\overline{f_2^{10\%}}}, & f_2 < \bar{f_2} \; and \; f_2 < \overline{f_2^{10\%}} \\ w_2 + (1-w_2)\frac{f_2}{\overline{f_2^{10\%}}}, & f_2 > \bar{f_2} \; and \; f_2 < \overline{f_2^{10\%}} \\ w_2 \frac{f_2}{\bar{f_2}} + (1-w_2), & f_2 < \bar{f_2} \; and \; f_2 > \overline{f_2^{10\%}} \\ w_2 + (1-w_2), & f_2 > \bar{f_2} \; and \; f_2 > \overline{f_2^{10\%}} \end{cases}$$

For the calculation of *β,* the total number of second hop-friends is associated with the average number of second-hop friends of all OSN users, denoted by $\overline{f_2}$, and the average number of second-hop friends of the most popular OSN users (top 10%), denoted by $\overline{f_2^{10\%}}$. w_2 reflects the importance of each of the two terms in the calculation of *β*.

In the notation used in the formula above we refer to the following quantity as object estimated influence for an object (I_o) for a given OSN user object (piece of shared info).

$$\left(\frac{\alpha \frac{\sum_{j=1}^{e} A_j^i}{f_1} + \beta \frac{\sum_{k=1}^{g} A_k^i}{f_2}}{2} \right)$$

The importance of an object (picture/video/post) that is posted to the OSN (*M*) is calculated taking into consideration the type of action on the specific object of the first- and second-hop friends of the object creator. The calculation of *M* is following the rationale of the equation for the estimation of *I;* the number of connections of the OSN user that creates the object as well as the actions of other OSN users on the object are parameters that characterize the importance of the object. However, another parameter that should be taken into account is the duration of activity in terms of actions on the uploaded object. The longer is this duration the higher is the importance of the respective object. For that reason parameter *d* is introduced, which denotes the duration of activity, in terms of days, on the respective object (e.g., *d*=3 days), while d_{max} denotes the maximum number of days that we monitor the activity on an object after its creation (e.g., d_{max}=7 days).

$$M = \frac{d}{d_{max}} \left(\frac{\alpha \frac{\sum_{j=1}^{e} A_j^i}{f_1} + \beta \frac{\sum_{k=1}^{g} A_k^i}{f_2}}{2} \right)$$

The higher is the influence of a person the larger are the possibilities to increase the value of the personal data. The latter also increases proportional to the importance of the object. Hence, taking into account the above analysis, the personal data value combines these two factors (*I*, *M*) and it is calculated as follows:

$$V = I \cdot M$$

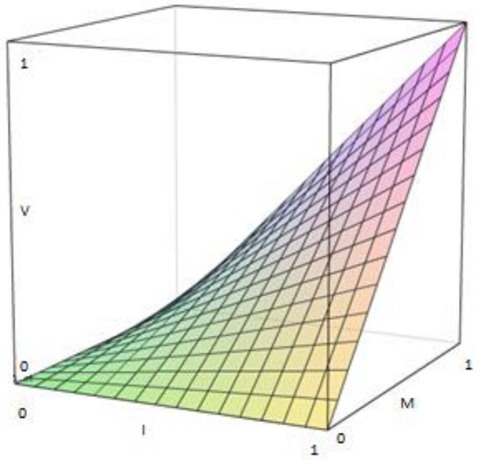


Figure 10. Personal Data Value (V)

This initial set of defined value indicators (data value *V*, user influence score *I*, object importance *M*) is defined so that it can be computed from actual OSN data (like Facebook). As part of future work in WP6 these value indicators will be evaluated with actual data collected from the pilots on top of Facebook and with simulated data from theoretical models. Additionally this set of value indicators will be extended with additional indicators of value based on measures of affinity between users/shared data and evaluations of utility functions for data sharing (the general framework that will be used to integrate the current and future research in the USEMP platform is described in D7.1, section 2.5.2.2).

7.3. Working example

In the example below we consider a specific OSN user that has created five objects (for example, he/she updated his status, posted a message, uploaded a video, liked a web page or uploaded an image) with some assumptions about the OSN social graph this specific user belongs to. The assumptions we make for the average user of our proposed OSN social graph are described in Table 17.

Parameter	Definition	Assumed Value
$\overline{f_1}$	Average number of first-hop friends in the OSN	175
$\overline{f_1^{10\%}}$	Average number of first-hop friends in the OSN (top 10%)	1500
$\overline{f_2}$	Average number of first-hop friends in the OSN	500
$\overline{f_2^{10\%}}$	Average number of first-hop friends in the OSN (top 10%)	3000

Table 17. OSN characteristics

Let us consider the example of a specific user in that OSN social graph that has 200 first hop friends (f_1= 200) and 430 second hop friends (f_2= 430).

Also assume that three types of actions are into account to estimate the proposed metrics:

- A1: Like on Object *i*.
- A2: Share of Object *i*.
- A3: Comment on Object *i*.

In this example, we consider that all actions have the same weight. For this specific user the number for each type of action (A1, A2, A3) s/he has performed at a certain point in time for each of the five created objects (O) is described in Table 18, along with the duration of activities considered for each object.

		First-hop friends				**Second-hop friends**			
O	**d**	**|A1|**	**|A2|**	**|A3|**	**Sum$_1$**	**|A1|**	**|A2|**	**|A3|**	**Sum$_2$**
1	1	40	3	10	53	50	4	5	59
2	3	150	10	10	170	200	15	25	240
3	2	100	5	15	120	140	8	20	168
4	1	80	3	10	93	50	4	5	59
5	1	40	10	20	70	200	4	5	209

Table 18. Number of actions for each object by the OSN friends

Using the proposed equations, the value of each object (i.e., personal data) has been calculated and the results are presented in Table 19. The influence of the OSN user that has created the objects is *I* = 0.092, which is calculated based on the impact of each object that the user has created.

O	**$\mathrm{Sum}_1 + \mathrm{Sum}_2$**	***d***	**I_O**	***M***	***V***
1	112	1	0.019	0.004	0.00034
2	410	3	0.213	0.128	0.01172
3	288	2	0.106	0.042	0.00387
4	152	1	0.051	0.010	0.00092
5	279	1	0.069	0.014	0.00126

Table 19. Calculated I_O, M and V values for each piece of shared personal data (object) compared with total number of actions and activity period.

From the numerical example above, it is demonstrated that for the proposed score *V*, for a given time period and for a specific user (i.e., where the influence *I of* a specific user is the same for all objects), the value score of a given user item depends on the importance *M* of the item. The Importance *M* of the item grows larger as the number of actions from other OSN members on the item grows larger. Additionally, the longer activities are performed on an item compared to a monitoring period (in terms of days) the higher is the importance of the respective object and consequently its value (e.g., Object 2).

As part of future work in WP6 these value indicators will be evaluated with actual data collected from the pilots on top of Facebook and with simulated data from theoretical models. Additionally this set of value indicators will be extended with additional indicators of value based on measures of affinity between users, shared data and evaluations of utility functions for data sharing.

8. Conclusions and Next Steps

The core part of this deliverable is the design and development of the first version of the USEMP privacy scoring framework. The first step towards the development of this framework was the definition of privacy dimensions. This is essentially a hierarchical organization of personal attributes and their possible values. On top of that schema, a privacy scoring framework was developed, which enriches the set of nodes at each level of the hierarchy with a number of scores, each capturing a different aspect of privacy. Moreover, we have described an approach for aggregating the evidence coming from OSN presence data to fill the scores at all levels of the hierarchy.

Furthermore, a number of inference mechanisms were developed for inferring personal information. These mechanisms operate in a complementary way to the modules developed within WP5, leveraging different types of input data. In particular, whereas WP5 focuses on single pieces of media content (text messages, images), here we have investigated collective inference approaches that take into account the behavioural observations (e.g. likes) and the links/interactions of multiple OSN users to generate additional inferences about them. A number of experimental studies on publicly available datasets were presented and useful insights and conclusions were drawn. In particular:

- It is indeed feasible to infer with satisfactory accuracy a number of personal traits and attributes of an OSN user based on their observed behaviour (like histories), using SVD and standard Machine Learning (ML) techniques. Yet, there are a significant number of OSN users, for which the conducted inferences are highly unreliable. Bagging multiple ML models was shown to be effective both for increasing the accuracy of conducted inferences and to derive a measure of confidence in them.
- Topic modelling approaches may be used as an alternative inference mechanism that addresses some of the limitations of the previous approach: it provides an easy to interpret representation of a user profile and is directly linked with the content that a user interacts with in the context of an OSN service.
- Graph-based approaches offer a powerful means of conducting inferences about OSN users solely on the basis of their links/interactions with other users. Using a public dataset of Twitter interactions, our newly proposed approach was shown to outperform existing ones in terms of inference accuracy.

Finally, a preliminary approach was proposed for deriving relative value estimates for the personal data of OSN users. The proposed approach models the value of personal data as a product of two main factors: a) the online audience of an OSN user, and b) the influence (in terms of reactions and interactions) that a user's OSN posts have on their audience.

Research and development in the upcoming period will be carried out at two levels. The first involves integration and development. As the pilots are approaching, we will complete the preparation of modules that are responsible for the realization of the privacy scoring framework and of modules that implement various inference mechanisms. Moreover, the presented formulations of both the scoring framework and the inference mechanisms are expected to be adapted according to the results of the first pilots.

At a second level, work will focus on open research questions and alternative approaches:

- New approaches, e.g. probabilistic, for aggregating the results of different inference mechanisms and results based on different types of data.
- New ways for topic modelling and assignment with the goal of increasing the performance and flexibility of topic-modelling approaches.
- New mechanisms for building the network structure, on which graph-based methods (such as the proposed ARCTE method) operate.

9. Annex I: Privacy Dimensions & Attributes

Here, we provide the list of identified attributes for each dimension of the USMEP privacy scoring framework. The Demographics table is the same as the one of Chapter 3.

#	Attribute	Description	Example values and range
A.1	Age	Rather than using the absolute number of years, it is typical to use age groups.	The following is a preliminary set of age groups: 6-12, 12-18, 15-25, 25-35, 35-45, 45-55, 55-65, 65-75, older than 75 years
A.2	Gender	The gender of the user	Male, Female
A.3	Nationality	Nationality of the user	French, Belgian, Greek
A.4	Racial origin	The racial background of the user.	Asian, African, Caucasian, Latino/Hispanic, Other
A.5	Ethnicity	The ethnic origin of the user that could relate to a combination of racial background, language and religion.	List of target ethnicities. e.g. Arabic, Easter-European, etc.
A.6	Literacy level	Literacy level will be represented by the highest degree or level of school attended.	None, Nursery school, High school, Bachelor's degree, Master's degree, Ph.D., Other
A.7	Employment status	The employment status of the user	Employed, Unemployed, Retired, Other
A.8	Income level	In the questionnaire, this is represented as a perceived deviation from the average national income; however, this will internally be transformed to an absolute number that will be mapped to a specific range of values	Qualitative ranges of monthly income, e.g. a 5-scale range from low to high.
A.9	Family status	Marital status of the user.	Single, in a relationship, married, separated, other

Table 20. Demographic attributes

#	Name	Description	Example values and range
B.1	Emotional stability	This describes the overall reaction of a person to negative emotions.	• Sensitive / nervous • Secure / confident
B.2	Agreeableness	A measure of one's trustful and helping nature.	• Friendly / compassionate • Analytic / detached
B.3	Extraversion	Quite generally, this describes the tendency to seek social interaction.	• Outgoing / energetic • Solitary / reserved
B.4	Conscientiousness	Planned rather than spontaneous behaviour.	• Efficient / organized • Easy going / careless
B.5	Openness	Reflects the degree of intellectual curiosity.	• Inventive / curious • Consistent / cautious

Table 21. Psychological traits

#	Name	Description	Example values and range
C.1	Sexual preference	Sexual preference.	• Heterosexual • Homosexual • Bisexual

Table 22. Sexual profile

#	Name	Description	Example values and range
D.1	Parties	We will have separate lists per country.	Part of list for Belgium: • CD&V • Groen! • N-VA • Open VLD • etc. Part of list for Sweden: • Centerpartiet • Vansterpartiet • Folkpartiet liberalerna • etc.
D.2	Political ideology	This is clearly correlated to the above ; however, we included it in order to take into account the case that the specific party that the user supports is not listed.	• Communist • Socialist • Green • Liberal • Christian democratic • Conservative • Right-wing extremist

Table 23. Political attitudes

#	Name	Description	Example values and range
E.1	Supported religion	A finer division may eventually be used.	• Atheist • Agnostic • Christian • Muslim • Hinduist • Buddhist • Other

Table 24. Religious beliefs

#	Name	Description	Example variables and range
F.1	Smoking	We may eventually use finer classes that will also reflect the frequency of smoking	• Smoker • Non-smoker
F.2	Drinking (alcohol)	Alcohol consumption.	• Non-drinker • Social drinker • Drinker • Alcoholic
F.3	Drug use	This includes all types of substances that could be classified as drugs	• Yes • No
F.4	Chronic diseases	The list on the right column is only indicative. Can take multiple values simultaneously.	• Diabetes • Epilepsy • Cardiomiopathy • Hypertension • etc.
F.5	Disabilities	Different kinds of disabilities (physical, mental, sensory, etc.). The list on the right column is only indicative.	• Lower limb loss • Vision impairment • Balance disorder • etc.
F.6	Other health factors	Additional health factors. Likely to be extended. Can take multiple values simultaneously.	• Exercise (yes / no) • Late night shifts (yes / no) • Staying up late

Table 25. Health Factors and Condition

#	Name	Description	Example values and range
G.1	Home	The place where the user resides. Different levels of accuracy are possible.	City, address, country, GPS
G.2	Work	The place where the user works. Different levels of accuracy are possible.	City, address, country, GPS
G.3	Favourite places	Favourite places of the user. Different levels of accuracy are possible. Can take multiple values simultaneously.	City, address, country, GPS
G.4	Visited places	Places that the user has visited. Different levels of accuracy are possible. Can take multiple values simultaneously.	City, address, country, GPS

Table 26. Location

#	Name	Description	Example values / range
H.1	Brand attitude	Brand + stance (favourable, non-favourable). Can take multiple values simultaneously.	• List of pairs, each of which has a brand name and either the value “favourable” or the value “non-favourable”
H.3	Hobbies	Hobbies. Can take multiple values simultaneously.	A list of hobbies
H.4	Devices	Used electronic devices. Can take multiple values simultaneously.	• Smartphone • Tablet • PC • etc.

Table 27. Consumer Profile

10. Annex II: Implementation Details

In the following we present some details about the implementation of the scoring framework. In particular, we first present the JSON schema for the scoring framework. In addition, we sketch the computational procedures by which the inference modules will update the scores for a user during system execution. That is, we identify specific events during which specific processing events will be triggered.

10.1. JSON serialization

Two alternative JSON schemas are actually considered and will be used in different settings. In the first, there is a distinct record type for values, attributes, dimensions and users and each node of the hierarchy will be stored as a separate entry. This formulation is clearly more modular and allows for very fast access and updating of scores. However, it is also handy to have all data about a user in a single record that more directly represents the hierarchical structure of the framework.

Let us first examine the modular schema. As mentioned, the modular schema uses one record type for the values, one for the attributes, one for the dimensions and one for the users. However, there are two additional record types, The first contains a single support element for some particular value. This structure essentially links privacy values to OSN presence data and consists of four fields: inference mechanism, an identifier for the OSN data that was used to perform the inference, confidence and level of control. This structure may appear only as part of a value record. Please note that a particular value may be supported by multiple inferences, based on different data or different inference mechanisms, therefore, multiple support records may appear as part of a single value record. The JSON schema for the Support record is illustrated in Table 28. The second additional record type is related to the personal data value score and contains the importance of data items posted by the user. This structure can only appear as part of a user record (under both formulations). The JSON schema for the Support record is illustrated in Table 29Table 28.

At the next higher level is the Value record. This is a rich structure with a number of fields attached to it: confidence, sensitivity, visibility, overall privacy score, level of control, declared/inferred and support. The schema for the value record is listed below. Please note that the fields “user” and “value” should uniquely identify the value node for a specific user, i.e. there should be no two value records with the same “user” and “value” fields. Also, please note that the support field is an array with elements of type Support. The JSON schema for the Value record is included in Table 30.

One level above the Value record is the Attribute record. This is quite simpler than the value record as it only has an overall privacy score, a visibility field (apart from the user and attribute fields that uniquely identify the record for a particular user) and level of control. Its schema is presented in Table 31.

One level further up, we have the Dimension record, which has a schema very similar to that of Attribute (Table 32).

Finally, at the top level we have a User record, which again has a similar structure to that of the Attribute and the Dimension records, and is documented in Table 33.

```
{
  "type":"object",
  "$schema":"http://json-schema.org/draft-03/schema",
  "name":"support_record",
  "description":"This is a support record. It represents a performed inference that supports a value. This record
                 cannot appear by itself, it will always appear as part of a value record",
  "id":"http://jsonschema.net",
  "required":false,
  "properties":{
    "support_confidence":{
      "type":"number",
      "description":"This is the confidence provided for each support record for that particular value",
      "minimum":"0",
      "maximum":"1",
      "id":"http://jsonschema.net/support_confidence",
      "required":true
    },
    "data_pointer":{
      "type":"string",
      "description":"This points to the data that has been used to perform the inference, e.g. some picture, some
                     post or the network around the user",
      "id":"http://jsonschema.net/data_pointer",
      "required":true
    },
    "support_level_of_control":{
      "type":"number",
      "description":"This represents the ability of the user to control the disclosure of the data based on which this
                     inference is based",
      "id":"http://jsonschema.net/support_level_of_control",
      "required":true
    },
    "inference_mechanism":{
      "type":"string",
      "description":"This is the specific inference mechanism for that particular support value",
      "id":"http://jsonschema.net/inference_mechanism",
      "required":true
    }
  }
}
```

Table 28. Support record JSON schema

```
{
  "type":"object",
  "$schema":"http://json-schema.org/draft-03/schema",
  "name":"item_importance_record",
  "description":"This is an item importance record. It represents the importance for some particular piece of data
                 that the user has posted. This record cannot appear by itself, it will always appear as part of a
                 user record.",
  "id":"http://jsonschema.net",
  "required":false,
  "properties":{
    "data_pointer":{
      "type":"string",
      "description":"This points to a particular piece of data",
      "id":"http://jsonschema.net/data_pointer",
      "required":true
    },
    "importance":{
      "type":"number",
      "description":"This is the actual importance score of a piece of data posted by the user. This is denoted as M
                     in Chapter 7 of the deliverable and is used in combination with the influence score of the user
                     (I) to compute the overall personal data value score.",
      "id":"http://jsonschema.net/item_importance",
      "required":true
    }
  }
}
```

Table 29. Item importance record JSON schema

```
{
  "type":"object",
  "$schema":"http://json-schema.org/draft-03/schema",
  "description":"This is a record that represents a value in the privacy dimensions framework. It can either be
                 stored by itself or may be stored as part of an attribute record",
  "id":"http://jsonschema.net",
  "required":false,
  "properties":{
    "user":{
      "type":"string",
      "$ref":"support_record",
      "description":"This is the user identifier to which the data for this value record apply. In the case that this
                     record becomes an item in the list of a variable record, then it is missing. ",
      "id":"http://jsonschema.net/user",
      "required":true
    },
    "value":{
      "type":"string",
      "description":"This is the name of the value that is represented by this record.",
      "id":"http://jsonschema.net/value",
      "required":true
    },
    "value_confidence":{
      "type":"number",
      "description":"This is the aggregated confidence score for the specified value.",
      "id":"http://jsonschema.net/value_confidence",
      "required":true
    },
    "sensitivity":{
      "type":"string",
      "description":"This is the sensitivity score for this particular value. It is either provided by the user or is
                     computed from prior knowledge.",
      "id":"http://jsonschema.net/sensitivity",
      "required":false
    },
    "support":{
      "type":"array",
      "$ref":"support_record",
      "description":"This is an array that includes support records. Support records can only appear as parts of
                     this array.",
      "id":"http://jsonschema.net/support",
      "required":true
    },
    "value_visibility_overall":{
      "type":"number",
      "description":"This is the overall visibility score for this particular value. It will depend on the user's privacy
                     settings on the content that supports the value.",
      "minimum":"0",
      "maximum":"1",
      "id":"http://jsonschema.net/value_visibility",
      "required":true
    },
    "value_visibility_label":{
      "type":"string",
      "description":"This is qualitative visibility label that represents the widest group of audience to which
                     information about this value  is accessible",
      "id":"http://jsonschema.net/value_visibility_label",
      "required":true
    },
    "value_visibility_actual_audience":{
      "type":"number",
      "description":"This is an estimate of the actual audience that has access to information related to this value",
       "minimum":"0",
       "id":"http://jsonschema.net/value_visibility_actual_audience",
      "required":true
    },
    "declared_inferred":{
      "type":"boolean",
      "description":"This is a binary field  that defines if the value has been declared by the user or has been
                        inferred.",
      "id":"http://jsonschema.net/declared_inferred",
      "required":true
    },
    "level_of_control":{
      "type":"number",
      "description":"This represents the ability of the user to control the disclosure of information related to this
                        value",
      "id":"http://jsonschema.net/value_level_of_control",
      "required":true
    },
    "value_privacy_score":{
      "type":"string",
      "description":"This is the overall privacy score for that particular value, which is a function of the sensitivity,
                        (aggregated)  confidence and (aggregated) visibility scores.",
      "id":"http://jsonschema.net/value_privacy_score",
      "required":true
    }
  }
}
```

Table 30. Value record JSON schema

```
{
  "type":"object",
  "$schema":"http://json-schema.org/draft-03/schema",
  "description":"This is a record that represents an attribute in the privacy dimensions  framework. It can either be
                      stored by itself or may be stored as part of a dimensions record",
  "id":"http://jsonschema.net",
  "properties":{
    "user":{
      "type":"string",
      "description":"This is the user identifier to which the data for this attribute record apply. In the case that this
                        record becomes an item in the list of a dimension record, then it is missing. ",
      "id":"http://jsonschema.net/user",
      "required":true
    },
    "attribute":{
      "type":"string",
      "description":"This is the name of the attribute that is represented by this record",
      "id":"http://jsonschema.net/attribute",
      "required":true
    },
    "attribute_level_of_control":{
      "type":"number",
      "description":"This represents the ability of the user to control the disclosure of information related to this
                      attribute",
      "id":"http://jsonschema.net/attribute_level_of_control",
      "required":true
    },
    "attribute_privacy_score":{
      "type":"number",
      "description":"This is the overall privacy score for that particular attribute. It is aggregated from the privacy
                      scores of the values under this attribute",
      "minimum":"0",
      "maximum":"1",
      "id":"http://jsonschema.net/privacy_score",
      "required":true
    },
    "attribute_visibility_overall":{
      "type":"number",
      "description":"This is the overall visibility score for this particular attribute. It depends on the     visibility
                      scores of the values under it.",
      "minimum":"0",
      "maximum":"1",
      "id":"http://jsonschema.net/visibility",
      "required":true
    },
    "attribute_visibility_label":{
      "type":"string",
      "description":"This is qualitative visibility label that represents the widest group of audience to which
                      information about this dimension  is accessible",
      "id":"http://jsonschema.net/attribute_visibility_label",
      "required":true
    },
    "attribute_visibility_actual_audience":{
      "type":"number",
      "description":"This is an estimate of the actual audience that is aware of information related to this
                     dimension",
      "minimum":"0",
      "id":"http://jsonschema.net/attribute_visibility_actual_audience",
      "required":true
    }
  }
}
```

Table 31. Attribute record JSON schema

```
{
  "type":"object",
  "$schema":"http://json-schema.org/draft-03/schema",
  "description":"This is a record that represents a privacy dimension in the privacy dimensions framework. It can
                 either be stored by itself or may be stored as part of a dimensions record",
  "id":"http://jsonschema.net",
  "properties":{
    "user":{
      "type":"string",
      "description":"This is the user identifier to which the data for this dimension record apply. In the case that
                     this record becomes an  item in the list of a user record, then it is missing.",
      "id":"http://jsonschema.net/user",
      "required":true
    },
    "dimension":{
      "type":"string",
      "description":"This is the name of the attribute that is represented by this record",
      "id":"http://jsonschema.net/variable",
      "required":true
    },
     "dimension_level_of_control":{
      "type":"number",
      "description":"This represents the ability of the user to control the disclosure of information related to this
                    dimension",
      "id":"http://jsonschema.net/dimension_level_of_control",
      "required":true
    },
   "dimension_privacy_score":{
      "type":"number",
      "description":"This is the overall privacy score for that particular dimension. It is aggregated from the privacy
                    scores of the attributes under this dimension",
      "minimum":"0",
      "maximum":"1",
      "id":"http://jsonschema.net/privacy_score",
      "required":true
    },
    "dimension_visibility_overall":{
      "type":"number",
      "description":"This is the overall visibility score for this particular privacy dimension. It depends on the
                    visibility scores of the values under it.",
      "minimum":"0",
      "maximum":"1",
      "id":"http://jsonschema.net/visibility",
      "required":true
    },
    "dimension_visibility_label":{
        "type":"string",
        "description":"This is qualitative visibility label that represents the widest group of audience to which
                       information about this dimension is accessible",
        "id":"http://jsonschema.net/dimension_visibility_label",
        "required":true
      },
      "dimension_visibility_actual_audience":{
        "type":"number",
        "description":"This is an estimate of the actual audience that is aware of information related to this
                       dimension",
        "minimum":"0",
        "id":"http://jsonschema.net/dimension_visibility_actual_audience",
        "required":true
      }
    }
}
```

Table 32. Dimension record JSON schema

```
{
    "type":"object",
    "$schema":"http://json-schema.org/draft-03/schema",
    "description":"This is a record that represents an individual user, it acts a container for the scores. ",
    "id":"http://jsonschema.net",
    "properties":{
      "user":{
        "type":"string",
        "description":"This is the user identifier to which the data for this user record apply.",
        "id":"http://jsonschema.net/user",
        "required":true
      },
      "user_level_of_control":{
        "type":"number",
        "description":"This represents the ability of the user to control the disclosure of information",
        "id":"http://jsonschema.net/user_level_of_control",
        "required":true
      },
      "user_privacy_score":{
        "type":"number",
        "description":"This is the overall privacy score for that particular user. It is aggregated from the privacy
                       scores of the dimensions under this user.",
        "minimum":"0",
        "maximum":"1",
        "id":"http://jsonschema.net/privacy_score",
        "required":true
      },
      "user_visibility_overall":{
        "type":"number",
        "description":"This is the overall visibility score for this particular user. It depends on the visibility scores of
                       the privacy dimensions under it.",
        "minimum":"0", "maximum":"1",
        "id":"http://jsonschema.net/visibility",
        "required":true
      },
      "user_visibility_label":{
        "type":"string",
        "description":"This is qualitative visibility label that represents the widest group of audience to  which
                       information about this user  is accessible",
        "id":"http://jsonschema.net/user_visibility_label",
        "required":true
      },
      "user_visibility_actual_audience":{
        "type":"number",
        "description":"This is an estimate of the actual audience that is aware of information related to this user",
        "minimum":"0",
        "id":"http://jsonschema.net/user_visibility_actual_audience",
        "required":true
      }
      "overall_personal_data_value":{
        "type":"number",
        "description":"This is an estimate of the user's personal data value",
        "minimum":"0",
        "id":"http://jsonschema.net/personal_data_value",
        "required":true
      }
      "user_influence":{
        "type":"number",
        "description":"This is the overall user influence (denoted as I in Chapter 7) that is used for computing the
                          personal data value",
        "minimum":"0",
        "id":"http://jsonschema.net/personal_data_value",
        "required":true
      }
      "personal_data_value_per_item":{
        "type":"array",
        "$ref":"item_importance_record",
        "description":"This is an array that includes item importance records. Support records can only appear as
                       parts of this array. In combination with the user influence score, they are used to compute the
                       overall personal data value score",
        "id":"http://jsonschema.net/personal_data_value_per_item",
        "required":true
      }
    }
}
```

Table 33. User record JSON schema

Let us now present the alternative schema, in which a single record represents the scoring hierarchy for a single user. This is a much more complicated schema and for presentation reasons we present a simplified version. In particular, the full schema would include the complete array of dimensions, under each dimension all its attributes would be listed and under each attribute all its values would be listed. However, instead of this, we simply refer to arrays of dimensions, attributes and values and we stop the description at the attributes level. This alternative schema is illustrated in Table 34.

```
{
  "type":"object",
  "$schema":"http://json-schema.org/draft-03/schema",
  "description":"This is a record that represents the complete scoring framework for a specific user.",
  "id":"http://jsonschema.net",
  "properties":{
    "user":{
      "type":"string",
      "description":"This is the user identifier to which the data for this user record apply.",
      "id":"http://jsonschema.net/user",
      "required":true
    },
    "user_privacy_score":{
      "type":"number",
      "description":"This is the overall privacy score for that particular user. It is aggregated from the privacy
                    scores of the dimensions under this user.",
      "minimum":"0",  "maximum":"1",
      "id":"http://jsonschema.net/privacy_score",
      "required":true
    },
    "user_visibility_overall":{
      "type":"number",
      "description":"This is the overall visibility score for this particular user. It depends on the visibility scores of
                    the privacy dimensions under it.",
      "minimum":"0",   "maximum":"1",
      "id":"http://jsonschema.net/visibility",
      "required":true
    },
    "user_visibility_label":{
      "type":"string",
      "description":"This is qualitative visibility label that represents the widest group of audience to which
                    information about this user is accessible",
      "id":"http://jsonschema.net/user_visibility_label",
      "required":true
    },
    "user_visibility_actual_audience":{
      "type":"number",
      "description":"This is an estimate of the actual audience that is aware of information related to this user",
      "minimum":"0",
      "id":"http://jsonschema.net/user_visibility_actual_audience",
      "required":true
    },
     "overall_personal_data_value":{
      "type":"number",
      "description":"This is an estimate of the user's personal data value",
      "minimum":"0",
      "id":"http://jsonschema.net/personal_data_value",
      "required":true
    }
    "user_influence":{
      "type":"number",
      "description":"This is the overall user influence (denoted as I in Chapter 7) that is used for computing the
                    personal data value",
      "minimum":"0",
      "id":"http://jsonschema.net/personal_data_value",
      "required":true
    }
    "personal_data_value_per_item":{
      "type":"array",
      "$ref":"item_importance_record",
      "description":"This is an array that includes item importance records. Support records can only appear as
                    parts of this array. In combination with the user influence score, they are used to compute the
                    overall personal data value score",
      "id":"http://jsonschema.net/personal_data_value_per_item",
      "required":true
    }
   "dimensions":{
      "type":"array",
      "description":"This is an array that holds pointers for the first Level of the hierarchy.",
      "required":true,
      "items":{
        "title":"dimension",
        "type":"object",
        "description":"This is almost the same as the dimensions record defined before, most details are omitted
                      for brevity.",
        "properties":{
          "dimension":{
            "type":"string"
          },
          "dimensions_privacy_score":{
            "type":"number"
          },
          "dimension_visibility_overall":{
            "type":"number"
          },
          "dimension_visibility_label":{
            "type":"string"
          },
          "dimension_visibility_actual_audience":{
            "type":"number"
          },
          "dimension_level_of_control":{
            "type":"number"
          },
          "attributes":{
            "type":"array",
            "description":"This is an array that holds pointers for the second level of the hierarchy",
            "required":true,
            "items":{
              "title":"dimension",
              "type":"object",
              "properties":{
                "attribute":{
                  "type":"string"
                },
                "attribute_privacy_score":{
                  "type":"number"
                },
                "attribute_visibility_overall":{
                  "type":"number"
                },
                "attribute_visibility_label":{
                  "type":"string"
                },
                "attribute_visibility_actual_audience":{
                  "type":"number"
                },
                "attribute_level_of_control":{
                  "type":"number"
                },
                "values":{   "_comment":"”:"For brevity we do not list here the final level of the hierarchy.It will be an
                                  array of entries, but this time it will have the fields for the values level."
                }
              }
            }
          }
        }
      }
    }
  }
}
```

Table 34. Alternative User JSON schema

10.2. Computational process for privacy scoring

When the user first registers with the USEMP platform, a new set of records that represent his scores are generated and inserted into the database. At first, all privacy scores are set to 0 and fields like support will be empty. The platform will then fetch the historical OSN presence data of the user and will process them using the inference mechanisms described in WP5 and in this deliverable. Different inference mechanisms will examine different subsets of the data. Each inference mechanism may update the scores for different sets of values. Once all inference mechanisms are executed and the relevant scores have been filled at the value level, then the bottom-up aggregation procedure that was described in section 3.2 will be executed.

Additionally, during the operation of the system, the user is going to post new content, become linked with other OSN users, delete content, etc. In any of these cases, execution of the relevant inference mechanisms that may operate on the involved type of data will be triggered. A list of branches of the hierarchy that are affected is kept and once execution of all relevant inference mechanisms has been completed, the scores for these branches will be propagated upwards.

11. Annex III: Prototype Implementations

This document is accompanied by a number of prototype implementations. More details about each of them and short guidelines about how to run them are provided in the following.

11.1. Privacy scoring using digital trails

The first prototype implementation is about the likes-based inference method that has been presented in Chapter 4. Both code and fully anonymized test data are provided in order to reproduce the experiments that have been presented. In particular, test data are provided for the following four classification tasks: a) Gay/Straight (G/S), b) Married/Single (M/S), c) Conservative, Liberal (C/L), d) Christian/Muslim (C/M).

The provided data are part of the myPersonality dataset. Please note that only the subset of the data that is necessary to reproduce the presented experiments is provided with this package. Also, the data is only meant for restricted access for review purposes and will not be made publicly available.

Running the experiments is a two-step process. In the first, one must fill in the parameters in the MATLAB script PrepareData.m that is located in the code directory and run it. The parameters that need to be set include the location of the data in the local file system, the number of cross validation folds, etc. Please look at the comments in the script file for more details. Actual learning and testing is carried out in the ML.m script. Before running it, please fill in again the parameters in it. These parameters define the classifier that will be used in the experiments as well as its parameters. Again, please see the comments in the script ML.m for more details.

11.2. Privacy scoring using latent topics

The second prototype implementation is about the latent topics-based inference method that has been presented in Chapter 5. Both code and fully anonymized test data are provided in order to reproduce the four different classification tasks that were examined in the respective chapter. These are the same classification tasks as those for the previous prototype.

For each of the examined classes, we have the topics distribution (for 600 identified topics) for a number of individuals that belong to that class. The required test data are part of the myPersonality dataset. Please note that again only the subset of the data that is necessary to reproduce the presented experiments is provided with this package and is only meant for use for review purposes.

The code directory includes a Java project that performs N-fold cross-validation and reports the results. To run it, the appropriate parameters must first be set in the file Parameters.txt which is located in the subdirectory:

```
Code\TopicBasedPrediction\target\TopicBasedPrediction
```

The parameters that need to be set are the locations of the files that contain the data for the two classes, the parameters of the SVM classifier and the number of folds for the cross-validation procedure. Please note that the SVM parameters must be in the form required by the Weka library. For more details please see the relevant Weka documentation:

http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMO.html

To run the code, please run the following command from the directory where the parameters file is located:

```
java -jar TopicBasedPrediction.jar
```

11.3. Privacy scoring based on user links

This is a prototype implementation of the network-based inference method that has been presented in Chapter 6. This package requires the Python language as well as the Python packages numpy, scipy and scikit-learn. A handy way to obtain an installation with these dependencies is to install the Anaconda Python distribution.

To reproduce the results please do the following:

1. Open the .../sensitive_data/src/configure_experiment.py file and set the output of the get_package_path() function to the location of the folder 'sensitive_data' (this folder).
2. Open the .../sensitive_data/bin/classify_user.py file and set the desired experiment parameters on the top of the file.
3. Run .../sensitive_data/bin/classify_user.py. Select the appropriate method and label set by uncommenting the corresponding lines.

11.4. JSON schema

Finally, we provide in separate JSON files the schemas for the scoring framework that were presented in the previous Annex. According to it, we defined two alternative JSON schemas.

The set of schemas for the first formulation can be found in the files:

- value.schema.json
- attribute.schema.json
- dimension.schema.json
- user.schema.json
- support.schema.json
- item_importance.schema.json

whereas the schema for the second can be found in the file user_alternative.schema.json.

12. References

A. Acquisti, C. M. Fong. (2012). An Experiment in Hiring Discrimination Via Online Social Networks. Social Science Research Network Working Paper Series, Apr. 2012

Adl, R. K., Askari, M., Barker, K. and Safavi-Naini, R. (2012). Privacy consensus in anonymization systems via game theory. In DBSec 2012, author's copy

Andersen, R., Chung, F. And Lang, K. (2006). Local graph partitioning using PageRank vectors. Proceedings of 47th Annual IEEE Symposium on Foundations of Computer Science, FOCS'06, pp. 475-486

Aperjis, C. and Huberman, B. A. A Market for Unbiased Private Data. HP Technical Report

Backstrom, L. and Kleinberg, J. (2014). Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook. In CSCW'14

Backstrom, L. and Leskovec, J. (2011). Supervised random walks: Predicting and recommending links in social networks. In WSDM '11. 635 - 644

Belkin, M. and Niyogi, P.. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 15(6), pp. 1373-1396

Bernstein, M., Bakshy, E., Burke, M. and Karrer, B. (2013). Quantifying the Invisible Audience in Social Networks. In CHI 2013

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E.. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008, 2008

Brandimarte, L., Acquisti, A. and Loewenstein, G. (2010). Misplaced Confidences: Privacy and the Control Paradox. In WEIS 2010

L. Breiman. (1996). "Bagging predictors". Journal of Machine Learning 24(2), pp. 123-140

Carrascal, J. P., Riederer, C., Erramilli, V., Cherubini, M. and de Oliveira, R. (2013). Your Browsing Behavior for a Big Mac: Economics of Personal Information Online. In WWW 2013

Hakravarthy, S. L. and Kumari, V. V. (2011). Preserving Data Privacy Using Coalitional Game Theory. In ECML PKDD 2011, author's copy

Conover, M.D., Goncalves, B., Ratkiewicz, J., Flammini, A. and Menczer, F. (2011). Predicting the Political Alignment of Twitter Users. In SocialCom 2011

Conti, M., Hasani, A. and Crispo B. (2011). Virtual Private Social Networks. In CODASPY'11

Domingo-Ferrer, J. (2010). Rational privacy disclosure in social networks. Modeling decisions for artificial intelligence. LNCS. Volume 6408

Fouss, F., Francoisse, K., Yen, L., Pirotte, A. and Saerens, M. (2012). An experimental investigation of kernels on graphs for collaborative recommendation and semi-supervised classification. Neural Networks 31, pp. 53-72, July 2012

Haveliwala, T.. (2002). Topic-sensitive PageRank. Proceedings of the 11th international conference on World Wide Web, pp. 517-526. ACM

Haron, H., Yusof, F.B.M. (2010). Cyber stalking: The social impact of social networking technology. International Conference on Education and Management Technology. pp. 237-241.

He, X., Machanavajjhala, A. and Ding, B. (2014). Blowfish Privacy: Tuning Privacy-Utility Trade-offs using Policies. In SIGMOD'14 (http://arxiv.org/abs/1312.3913)

Hinton, G. and Roweis, S.. (2002). Stochastic neighbor embedding. Proceedings of NIPS volume 2, pp. 833-840

Hoff, P., Raftery, A. and Handcock, M.. (2002). Latent space approaches to social network analysis. Journal of the American Statistical association, 97(460), pp. 1090-1098

Hong, J. I., Ng, J. D., Lederer, S., Landay, J. A. (2004). Privacy risk models for designing privacy-sensitive ubiquitous computing systems. In Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques (pp. 91-100). ACM.

James Fontanella-Khan, "Personal data value could reach €1tn", Financial Times, 2012

Jernigan, C. and Mistree, B. (2009) Gaydar: Facebook friendships expose sexual orientation. First Monday, Sep 2009, http://firstmonday.org/article/view/2611/2302

Kosinski, M., Stillwell, D. and Graepel, T. (2013). Private Traits and Attributes are Predictable from Digital Records of Human Behavior. PNAS 110 5802 - 5805.

Lancichinetti, A., Radicchi, F., Ramasco, J.J. and Fortunato, S. (2001). Finding statistically significant communities in networks. PloS one, 6(4):e18961, 2011

Liu, K. and Terzi, E. (2009). A framework for computing the privacy scores of users in online social networks. In ICMD 2009

Liu, K. and Terzi, E. (2010). A framework for computing the privacy scores of users in online social networks. In ACM Transactions on Knowledge Discovery from Data. Vol. 5. No. 1. Article 6

McPherson, M., Smith-Lovin, L. and Cook, J. M.. (2001). Birds of a feather: Homophily in social networks. Annual review of sociology, pp. 415-444

Moody, D.L., Walsh, P.A. (2002). Measuring The Value Of Information: An Asset Valuation Approach. Guidelines for Implementing Data Resource Management (4th Edition), B. Morgan and C. Nolan, (Eds.), DAMA International Press, Seattle, USA

Nepali, R.K. and Wang (2013). Sonet: A social network model for privacy monitoring and ranking. ICDCS 2013

OECD (2013). Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value. OECD Digital Economy Papers, No. 220, OECD Publishing. http://dx.doi.org/10.1787/5k486qtxldmq-en

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford Technical Report.

Papadopoulos, S., Kompatsiaris, Y., Vakali, A. and Spyridonos, P. (2012). Community detection in social media. Data Mining and Knowledge Discovery, 24(3), pp. 515-554

Papadopoulos, S., Corney, D. and Aiello, L. (2014). SNOW 2014 Data Challenge: Assessing the Performance of News Topic Detection Methods in Social Media. Proceedings of the SNOW 2014 Data Challenge co-located with (WWW 2014), pp. 1-8

Peng, H., Long, F. and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), pp. 1226-1238

Pennacchiotti, M. and Popescu, A.-M. (2011). Democrats, republicans and starbucks afficionados: User classification in twitter. In KDD'11. 430 - 438

Popescu, A., and Grefenstette, G. (2010). Mining User Home Location and Gender from Flickr Tags. In ICWSM 2010.

Raman, A. S., Barloon, J. L., Welch, D. M. (2012). Social media: Emerging fair lending issues. The Review of Banking and Financial Services, 28(7), July 2012

Rao, D., Yarowsky, D., Shreevats, A. and Gupta, M. (2010). Classifying Latent User Attributes in Twitter. In SMUC 2010

Rose, J., O. Rehse, B. Rober. (2012). The Value of our Digital Identity. Report published online by Liberty Global Inc.

Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P. and Ungar, L.H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE 9 e73791

Schneier, B., A Taxonomy of Social Networking Data, Security & Privacy, IEEE , vol.8, no.4, pp.88,88, July-Aug. 2010

Spielman, D. and Teng, S.-H. (2008). A local clustering algorithm for massive graphs and its application to nearly-linear time graph partitioning. arXiv preprint:0809.3232, 2008

Sramka, M. (2015). Evaluating privacy risks in social networks from the user's perspective. Advanced Research in Data privacy. Studies in computational intelligence 517

Srivastava, A. and Geethakumari, G. (2013). Measuring privacy leaks in online social networks. ICACCI 2013.

Stutzman, F., Gross, R. and Acquisti, A. (2012). Silent Listeners: The Evolution of Privacy and Disclosure on Facebook. Journal of Privacy and Confidentiality 4 7 - 41

Tang, L. and Liu, H. (2009). Scalable learning of collective behavior based on sparse social dimensions. In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 1107-1116. ACM

Wagner, C., Asur, S. and Hailpern J. (2013). Religious Politicians and Creative Photographers: Automatic User Categorization in Twitter. In SocialCom 2013.

Wang, X., Tang, L., Liu, H. and Wang, L.. (2013). Learning with multi-resolution overlapping communities. Knowledge and Information Systems, 36(2), pp. 517-535

Wang, Y., Nepali, R. and Nicolai, J. (2014). Social network privacy measurement and simulation. In ICNC 2014.

Yang, J. and Leskovec, J. (2013). Overlapping community detection at scale: a nonnegative matrix factorization approach. Proceedings of the sixth ACM international conference on Web search and data mining, pp. 587-596. ACM

Yassine, A. and Shirmohammadi, S. (2008). Privacy and the Market for Private Data: A Negotiation Model to Capitalize on Private Data. In AICCSA 2008, author's copy

Zheleva, E. and Getoor., L. (2009). To Join or not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles. In WWW 2009