Privacy2.0: Towards Collaborative Data-Privacy Protection

Erik Buchmann, Klemens Böhm, Oliver Raabe

Abstract Data protection is challenging in scenarios where numerous devices collect personal data without drawing attention from the individual concerned, e.g., in Ubiquitous Computing applications, Sensor Networks or Radio Frequency Identification installations. Technical mechanisms for data protection force the individual to keep track of his personal data and require a thorough understanding of technology. Regulatory approaches cannot keep pace with the advent of new privacy threats. This paper proposes and describes a new multidisciplinary research direction for data protection: The idea is to use Web2.0 mechanisms which let users share their experiences, observations and recommendations regarding the privacy practices of service providers in an intuitive and flexible manner. We define an innovative framework at the logical level, i.e., identify the components of the architecture. The core of the framework is a folksonomy of tagged geo-locations, physical items and Internet addresses that might have an impact on privacy. Our framework envisioned helps the user to decide if a data collector handles personal information compliant with legal regulations and according to the user preferences. We find out which current technologies can be adapted to implement our framework, and we discuss design alternatives and new research directions.

1 Introduction

It has never been as simple as today to collect large volumes of personal data. In the near future, advances in the areas of Ubiquitous Computing [21], RFID [28] or Sensor Networks [14] will bridge the gap between the online and offline world and challenge data protection significantly [19]. Using current and novel information technologies in everyday life will shape the society of the future. If we do not find practical approaches for privacy protection, it might become as simple as using a search engine to assemble comprehensive personality profiles of individuals.

Erik Buchmann, Klemens Böhm
Institute for Program Structures and Data Organization, Universität Karlsruhe (TH), Germany,
e-mail: {buchmann | boehm}@ipd.uka.de

Oliver Raabe
Center for Applied Jurisprudence, Universität Karlsruhe (TH), Germany,
e-mail: raabe@ipd.uka.de

Current solutions for data protection divide into (1) legal norms and regulations and (2) technical mechanisms. Throughout the European Union, directives establish data protection as a fundamental human right [10]. However, the regulatory approach often results in a daunting number of norms that is unmanageable both for the persons concerned and for the data collectors. Further, there is a lack of legal certainty for new technologies which the legislator has not considered yet. Ensuring that regulations are met is particularly challenging in scenarios where the collection and processing of private data is *intransparent* to the individuals concerned [19]. This is the rule in most Ubiquitous Computing-, RFID- and Sensor Network scenarios. From a technical perspective, all current privacy-related approaches are *technology-centered*, *isolated* implementations and require a thorough understanding of the technology. For example, an entry 'X-No-Archive: Yes' in the header of Usenet messages prevents them from being filed. But it is hard to explain to persons without a technical background how to make use of such features. We expect that the situation will become worse with more sophisticated technology.

Intuitively speaking, one is neither interested in going through lawsuits nor in implementing technical mechanisms to enforce privacy. Instead of bothering with the details of technologies and regulations, the individual concerned simply wants to know: "*Can I trust a certain service provider to handle my personal information compliant with regulations and according to my preferences?*". Since existing research does not directly address this demand, it is necessary to investigate new issues and directions for future research and development.

The core idea behind this paper is the deployment of Web2.0 technologies to support individuals in protecting their privacy. Web2.0 mechanisms like folksonomies, blogs and social network communities have become popular in the recent past. Our vision is to leverage them for data protection. This paper is the first step in the direction of a holistic Privacy2.0-framework that lets the individuals concerned share their experiences, observations and recommendations about privacy practices of service providers in an intuitive and flexible manner. We make the following contributions:

- We explain from an interdisciplinary point of view why existing solutions for data protection are not sufficient in current and future scenarios. We say why we think that social software mechanisms from the Web2.0 has the potential to overcome these limitations.
- We introduce a Privacy2.0 framework on the logical level based on social software mechanisms. It is flexible enough to meet individual privacy needs, and provides a holistic view on privacy threats coming from a broad range of technologies in the online and offline world, e.g., search engines, web shops, sensor networks or RFID-tagged products. We identify the components of our framework at the logical level.
- We review current technologies in order to find out which approaches can be adapted for an implementation of our framework, and we say which functionality is missing and requires further development. In addition, we discuss open issues for further research in multiple disciplines.

The remainder of this paper is organized as follows: The next section reviews technologies with an impact on privacy and solutions for data protection. Section 3 introduces the components of our framework, and in Section 4 we sketch an anonymous variant. Section 5 discusses applicability issues and outlines future work, followed by the conclusion in Section 6.

2 Privacy Threats

According to a recent survey of the IEEE Spectrum [15], more than 60% of 700 scientists interviewed expect that intelligent, interconnected devices performing individual-related services will have penetrated our daily lives in the next 10 years. In the following we will briefly overview prominent technologies that are relevant in this context, and we will explain how they affect privacy.

2.1 Future Technologies with an Impact on Privacy

A **Sensor Network** [17] consists of many sensor nodes equipped with sensing devices, radio transmitters and limited computational resources. By using self-organization technologies for sensing and network formation, sensor networks are able to fulfill complex measurement tasks in the fields of surveillance, border control or facility monitoring. The toll collect network, which operates approximately 2800 nodes[1], can be seen as a first large-scale sensor network in public spaces. Toll collect identifies trucks on German highways with an average capture rate of 99.75%[2]. Although toll collect's objective is to charge trucks for the use of highways, some parties have already demanded access this data for other purposes, e.g., civilian law enforcement.

The idea of **Radio Frequency Identification** (RFID) is to assign a globally unique identification number to physical objects for applications like object tracking or stock management [3]. The objects are labeled with a RFID tag containing a radio transmitter as well as very limited computing and storage capabilities. The tags can be read over a distance without requiring a straight line of sight. One of the most prominent RFID applications is to replace barcode-based processes, in particular at the points of sale of retailers [13]. Thus, RFID technology is about to enter public spaces at a large scale. But while barcodes contain only information about the product group and have to be visibly mounted and scanned, RFID labels can be embedded into products invisibly and be read without the knowledge of the individual concerned.

Ubiquitous Computing (Ubicomp) means equipping objects of everyday life (refrigerators, microwave ovens, etc.) with "intelligence" in order to ease repetitive tasks in household, business or medical care [29]. For instance, consider an Ubi-

[1] Bundesanstalt für Straßenwesen (BASt), 01/01/2008, http://www.mauttabelle.de

[2] Press release 14/12/07, "Truck toll in Germany: Three years", http://www.toll-collect.de

comp system that tracks the position of a certain user and reminds him of important appointments either via the handsfree set in his car, the hi-fi system in his living room or any other interconnected device with audio output in his vicinity. Ubicomp installations can monitor the behavior of their users around the clock.

As a cross-cutting service, the **Internet** enables a broad variety of devices and services to interact in order to drive highly personal, interconnected applications. When looking at today's Internet, privacy threats arise from an unmanageably large number of different technical protocols, services and service providers. On the Web, privacy-sensitive information can be gathered by using cookies, iframes, web-bugs, affiliation programs, services that require a personal login etc. The situation is expected to become even more unclear in the future [15].

Technologies like Sensor Networks, RFID or Ubicomp have a large impact on privacy, for the following reasons:

- They bridge the gap between the online and the offline world. Thus, the situation is not as simple as switching off the computer to leave the respective data-privacy issues behind.
- The technologies use networked devices to collect, transfer and process personal data in the background and without the assistance and the knowledge of the individual concerned. Thus, it is virtually impossible for each individual to keep track of all service providers which have his personal data.
- As the level of detail of the collected data is comprehensive, and personal data from multiple sources can be easily linked, the potential of any misuse of this information is huge.
- The applications outlined yield a clear benefit for their users. Thus, it is not an option to strictly avoid their use.

2.2 Solutions for Data Protection

In this subsection we will briefly outline the range of data-protection approaches.

Laws and Regulations. The European Union harmonizes the data-protection law of its members by issuing directives in sectors like e-commerce [12] or electronic communication [11]. However, the debates on transposing these directives into national law show that regulatory approaches have fundamental limitations. The legislator cannot predict new technologies. This involves periods of time without legal certainty, until regulations have been adopted for new privacy threats. Further, new technologies often result in a flood of new regulations. For example, German law contains approximately 1500 norms for data protection. But, to give an example, there still is no regulation for Peer-to-Peer Overlay Networks [25] where each peer can be a service provider which handles personal data. The regulatory approach is often ineffective in ensuring data protection: Service providers cannot find out which particular norms apply among a daunting but incomplete amount of regulations. Authorities are overloaded with an increasing number of regulations, and enforcing them requires a fundamental understanding of the technical background.

In consequence, it is *intransparent* for the individual if a service provider handles personal data with care or not, even if appropriate regulations exist.

Technical Mechanisms. The number of privacy enhancement technologies available is large. For example, epic.org lists approximately 200 privacy tools, and vunet.com finds about 100 commercial privacy suites. Nevertheless, existing technical mechanisms cannot ensure data protection for the majority of people, for various reasons. For example, P3P-enabled web servers [5] inform the web browser of the user about the privacy policy of the service provider, and let the browser reject cookies which pose a privacy threat according to the user preferences. But understanding the impact of cookies on privacy and therefore setting the preferences accordingly requires a thorough understanding of the Internet protocols. Finally, P3P cannot express all details required from EU privacy regulations, and addresses only a tiny fraction of privacy threats in the Internet (cf. [9]). Other technical mechanisms face similar problems. k-anonymity [27] handles the problem that anonymized microdata, e.g., from surveys or field studies, can be linked with public data to obtain personal profiles. However, it turned out that it is even challenging for experts to anonymize personal data properly [24]. This does not mean that it is not important to develop such mechanisms. The finding simply indicates that it cannot be left to the individual to use a large number of specific privacy techniques efficiently.

Other Mechanisms. Considering that both legal and technical mechanisms are not sufficient to ensure data protection, recent political debates suggest the legislator to focus on self-regulation, education and training [4]. Privacy seals are one approach for self-regulation. They certify that service providers follow specific privacy guidelines [1]. Thus, privacy seals signalize trust in the data-handling practices of the provider audited. However, as long as the prerequisites to obtain a seal are unclear for the most of the public, the significance of privacy seals is limited.

3 A Collaborative Framework for Data Protection

We have shown that all available technical solutions for data protection are technology-centered, isolated mechanisms which require thorough knowledge from the individuals, and we have explained the fundamental problems of regulatory approaches. Before proposing our novel *collaborative* framework for data protection, we will introduce the requirements for data-protection mechanisms in current and future scenarios, as we see them. The requirements rely on the assumption that the user does not want to deal with technical and legal details; instead, he simply wants to know which service he can commit personal information to.

(R1) One Mechanism for All Data-Privacy Issues. From the user perspective it is not sufficient to develop isolated technical mechanisms tailored to specific privacy threats. Instead, the user requires a *holistic* view of all possible threats. It should indicate if it is safe to entrust personal data to a particular service. It should be independent from the technologies and protocols used by the service.

Obviously, one typically cannot foresee new data-privacy threats. This calls for flexible and adaptive mechanisms.

(R2) Privacy Self-Protection. Currently, the authorities are overloaded with the enforcement of a daunting number of regulations, and new technologies often result in a lack of legal certainty. Furthermore, most existing technical privacy mechanisms have to be implemented at the data collector, and it is impossible for the majority of people to assess their effectiveness. Thus, from the perspective of the individuals concerned, data protection currently is a matter of trust in the willingness and ability of the data collectors to care for the privacy of their customers. But, according to our perception, individuals want powerful tools which put data protection in their hands.

(R3) Intuitive Use. The widespread use of information technology in public spaces makes data protection a concern for broad parts of the society. In consequence, it is of utmost importance that a privacy measure is applicable without requiring a special training or in-depth knowledge from its user. Otherwise, the result could be a two-tier society where the educated part of the population is able to keep its privacy while the other one is not.

(R4) Individual Preferences. As [7, 6] have shown, the desired level of privacy varies at large scale. While some persons are willing to provide private information to a significant extent just for comfort or gaming, others request a considerable compensation for their private data. Thus, it is important to support an individual not only according to existing norms and regulations, but also by representing his preferences. In line with the requirement ’Intuitive Use’, this calls for a set of basic preference templates, e.g., “discreet”, “standard” and “communicative”. Individuals then may adapt these templates.

3.1 Overview

The vision behind the framework proposed is to use modern Web2.0 techniques for data protection. The framework lets individuals share their experiences with the data-handling practices of service providers, and it provides a warning before someone reveals personal information against his preferences. As observations by privacy activists and committees for data protection show[3], attentive persons can detect many privacy violations, and they are willing to communicate their findings. Sharing data-privacy issues would put pressure on services which violate the privacy of their consumers and lets society enforce social standards for data protection. Thus, a privacy framework based on Web2.0 allows the individuals concerned to supervise the data collectors and to put data protection in their own hands.

We describe our logical framework as it could be implemented at a *trusted third party*. The framework stores the pseudonyms of its users, which could have an impact on their privacy. In Section 4 we will sketch an alternative Peer-to-Peer realization that provides complete anonymity.

[3] See http://www.trust-us.ch, http://www.ccc.de/club/archiv, http://netzpolitik.org for examples.

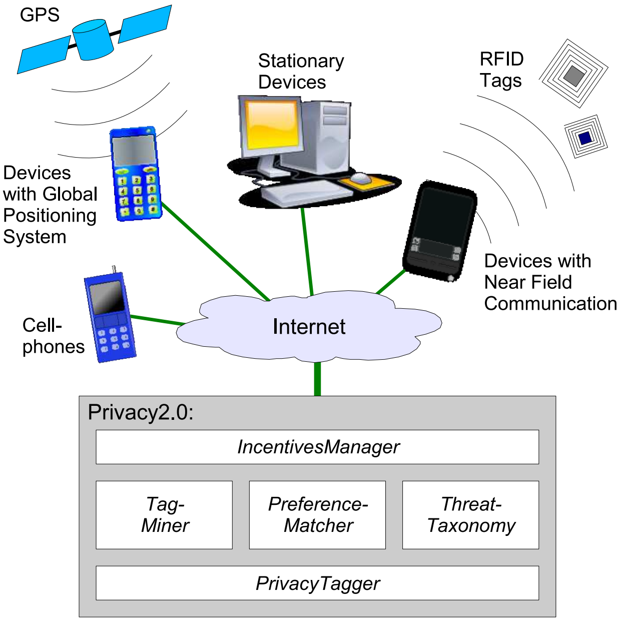

Fig. 1 Overview of the Privacy2.0 Framework

Figure 1 provides an overview of our logical framework. Its core component is the **PrivacyTagger**, a user-managed taxonomy of data-privacy issues. It stores tuples of the form (*privacy threat*, *user pseudonym*, *label*), and it lets the users intuitively label privacy-threatening Internet-service providers, geographical locations or objects referenced by RFID tags. For example, one could tag a certain geographical position with the labels “video”, “surveillance” and “no privacy policy”. The **TagMiner** component extracts the semantics of the labels provided, i.e., it tries to identify labels with the same meaning. For example, the labels “worse policy” and “spammer” provide the same negative assessment. In order to find out if the privacy threats identified by one particular user are relevant for another one, the **PreferenceMatcher** computes the similarity of two users, based on the tagged objects and the tags they have created. The **ThreatTaxonomy** component determines similar services, locations and objects. Finally, social software systems require a minimum number of active users to become operational. Thus, the **IncentivesManager** motivates the users to create useful and reliable labels, e.g., by providing better service for dependable users.

In the following, we will explain the functionality of each component, and we will provide a discussion of design alternatives, implementation issues and technology available.

G	Geographic locations
P	Physical objects
I	Internet-services
O	Privacy issues referenced by the tags
U	Pseudonyms of users who provided the tags
T	Tags

Table 1 Symbols used to describe the framework.

3.2 The PrivacyTagger Component

The PrivacyTagger lets the users gather information on the data-privacy practices of a wide range of service providers in the online and offline world collaboratively. We propose a *folksonomy* [22] ("folk taxonomy") as the basis infrastructure for this component. *Social tagging* means that the users can label privacy threats with schema-free *tags*, e.g., "spammer" or "good privacy policy". Technorati and Flickr[4] are prominent examples of social tagging. The tags can be arbitrarily chosen; the users do not need to agree on global standards. Since tags consist of only a few letters each, it is possible to support a wide range of end-user devices. Even the 160 letters allowed in an SMS would be sufficient to generate tags en route with a cellphone. More systematically, folksonomies are in line with our requirements: They make no restrictions regarding the tags and objects tagged (R1), help the users protecting their privacy by making privacy violations transparent (R2), and their use is intuitive (R3). Note that R4 is addressed by other components.

A folksonomy typically is a tripartite network [20] described by (*object*, *user*, *tag*)-tuples. Let O be the set of objects that can be referenced, e.g., the URL of a web shop or the location of a surveillance camera which might have an impact on privacy. U is the set of users who provide the tags. T stands for the set of all tags (Table 1 contains all symbols used). Thus, the PrivacyTagger component stores the records $(O,U,T) = \{(o_1,u_1,t_1),\cdots,(o_n,u_n,t_n)\}$ where $o \in O$, $u \in U$, $t \in T$. Since Requirement R1 calls for holistic mechanisms, we have to ensure that our framework can address a broad range of privacy threats. For this reason, we propose to let the users tag geographic locations (G), physical items (P) and Internet services (I). The range of taggable objects is $dom(O) = dom(G) \cup dom(P) \cup dom(I)$. $dom(G) = (latitude, longitude, height)$ describes geographic locations. Internet services are referenced by its uniform resource locator (URL), and RFID labels[5] can identify physical objects. Extending the framework for other privacy threats simply requires extending $dom(O)$.

Due to its simple structure, the implementation of a folksonomy is straightforward. Alternatively, one can adapt open-source implementations like Scuttle[6].

[4] http://www.technorati.com, http://www.flickr.com/tags

[5] Each RFID label stores a globally unique EPC-ID, cf. http://www.epcid.com.

[6] http://sourceforge.net/projects/scuttle

3.3 The ThreatTaxonomy Component

The ThreatTaxonomy component computes the similarity of privacy threats o_a, o_b which are represented differently in the PrivacyTagger component. There are many reasons why similar threats can have different representations. For example, the accuracy of civilian Global Positioning System (GPS) receivers used to tag geographic locations is typically about 15 meters. Further, one company might be represented in the Internet, e.g., with a web shop, at certain geographic locations, e.g., outlets in shopping malls, and with RFID-tagged objects, e.g., bonus cards. In order to provide a holistic view on all privacy issues (Requirement R1), our framework has to identify similar privacy threats. Thus, the ThreatTaxonomy implements the function $tt(o_a, o_b) = x$ with $x \in [0,1]$, where $x = 1$ refers to exactly the same and $x = 0$ to completely different privacy threats.

The similarity between differently represented objects can be computed in numerous ways; it has to be investigated which method is adequate for an implementation of our logical framework. For example, an inspection of the commercial register provides a taxonomy of the corporate structure and interconnected companies. Based on the accuracy of GPS receivers, two locations could be distinguished as follows:

$$tt(g_a, g_b) = \begin{cases} 1 & \text{if } distance(g_a, g_b) \leq 2 \cdot 15m \\ 0 & \text{otherwise} \end{cases}$$

It is also possible to infer similar privacy threats from the PrivacyTagger data. The records $\{(o_1, u_1, t_1), \cdots, (o_n, u_n, t_n)\}$ can be regarded as a graph structure $S = (V, E)$ where the set of vertices is $V = O \cup U \cup T$, and each record (o, u, t) constitutes three undirected edges $\{(o,u), (u,t), (t,o)\} \in E$. Now a wide range of well-known graph-based measures can be used to determine the similarity of two privacy threats o_a, o_b, e.g., the number of disjoint paths or the structural equivalence. See [26] for a comparison of graph-based distance measures that could be applied in our context.

Finally, it would be feasible to employ a second folksonomy to let the users create a taxonomy of similar privacy threats.

3.4 The TagMiner Component

Folksonomies can be intuitively used (cf. Requirement R3), but this might result in less accurate tags. According to [16], the tags provided by typical users can be, amongst other challenges:

- ambiguous, imprecise or inexact,
- misspelled or language dependent,
- homonymous, synonymous or consisting of singular and plural forms,
- consisting of compound tags and separators.

The TagMiner component extracts the meaning from the tags provided. In the context of this framework it is sufficient to learn if a certain user finds a particular privacy issue either threatening or not. Thus, TagMiner implements the function $tm(t) = x$ with $x \in \{positive, negative\}$.

A realization of the TagMiner component could use information-retrieval and knowledge-discovery mechanisms, e.g., rely on existing work in the area of opinion extraction [8, 18]. Opinion extraction aims to find out if customers like a certain product or not from, say, textual product reviews provided at shopping portals like Ciao.com. In order to extract the opinion from a text, complex preprocessing steps are required to identify the "opinion words" and to consider the order of words in a sentence. Opinion extraction on folksonomies should be slightly easier, because the tags are always descriptive [16] and do not require such preprocessing. However, the tags are not necessarily expressed in natural language but contain special characters and separators, which might stress existing opinion-extraction techniques.

3.5 The PreferenceMatcher Component

The preferences of different users regarding the desired level of privacy vary significantly. This has an influence on the tags the users generate. For example, while one user labels a web shop with "bad privacy policy", a less suspicious user might come up with "acceptable" for the same issue (cf. Requirement R4). Because it is important to know if the tags from one user can serve as a recommendation for another one, the PreferenceMatcher determines the similarity between two users.

For this reason, the PreferenceMatcher stores a set of preferences of all users $\{\pi_1, \cdots, \pi_n\}$. Based on the preferences π_a, π_b of two users a, b, the PreferenceMatcher computes the function $pm(\pi_a, \pi_b) = x$ with $x \in [0,1]$. $x = 0$ means that two users have complementary preferences, while $x = 1$ stands for users with equal attitudes regarding data protection.

It remains to be discussed how the preferences π_u of user u should be represented. Identifying the *real* preferences of individuals is challenging: Since privacy is a highly emotional topic, observations show that persons are rarely able to estimate their desired level of privacy exactly [6]. In order to approach this problem, we define the preference of each user u as a set of $(object, tag)$-pairs, i.e., $\pi_u = \{(o_1, t_1), \cdots, (o_m, t_m)\}$. As a starting point, each new user chooses a set of popular objects and tags π^t with $\forall (o,t) \in \pi^t : o \in O \wedge t \in T$. As the user subsequently labels new objects, these join his preferences, i.e., $\pi_u = \pi^t \cup \{(\hat{o}, \hat{t}) \mid (\hat{o}, u, \hat{t}) \in (O, U, T)\}$. Since the choice of tags and tagged threats represent the opinion of the user, newly created tags refine the preferences towards his real objectives.

The folksonomy can be represented as a graph structure $S = (V, E)$ where the set of vertices is $V = O \cup U \cup T$, and each record (o, u, t) constitutes three undirected edges $\{(o,u), (u,t), (t,o)\} \in E$ (cf. Subsection 3.3). The preferences of each user u form a subgraph of $S' = (V', E')$, i.e., $\forall (o,t) \in \pi_u : o \in V' \wedge \{(o,u), (u,t), (t,o)\} \in E'$ and $S' \subseteq S$. Thus, the PreferenceMatcher can determine the similarity of two users by finding overlapping subgraphs [30] or the distance between subgraphs [26] in the graph structure.

3.6 The IncentivesManager Component

People participate in social software systems like folksonomies for various reasons: for personal use, to express their opinion, to attract attention, for self presentation issues, or to play and compete with others (cf. [22] for an exhaustive description). The purpose of the IncentivesManager component is to achieve *incentive compatibility* between the individuals participating and the global objectives of the framework envisioned. The IncentivesManager has to motivate the users to:

- **participate.** A social framework for data protection becomes effective only if a *sufficient number of attentive users* observes and labels privacy threats.
- **tag the right objects.** As research has pointed out, tags in a folksonomy usually follow a power-law distribution [16]: Few popular objects are tagged frequently, while most objects are labeled with a few tags at most. However, Requirement R2 targets at a sufficient number of tags on less popular privacy threats, too.
- **provide high-quality tags.** The description of the TagMiner component lists a number of issues like ambiguous or misspelled tags the framework has to deal with. In order to facilitate this, the IncentivesManager should motivate the users to *create tags carefully* right from the start.

Research in the area of social networking provides a number of incentives mechanisms possible. Examples are a ranking of the most active users or comfort features for users who have provided reliable information. As another example, [2] introduces a social psychology approach to motivate the members of a film community (MovieLens) to review particular movies. The participants received emails emphasizing that the receiver has been chosen because he provides better reviews than others, and that his review would be useful for many people. [2] shows that this approach motivates to participate in a way an operator has devised a priori. At this point, we do not impose any restriction on the incentive mechanisms, except that they have to follow the Requirements R1–R4.

3.7 The Privacy2.0 Framework

Having introduced the components of our framework, we can now specify how the framework decides if it is safe for a certain user to commit his personal data to a service provider. The provider is identified by a particular physical object, geographic location or Internet address. Formally, the framework requires a possible privacy threat q with $dom(q) \in dom(o), o \in O$ and the requesting user u as input. Based on the folksonomy (O,U,T) and the user preferences $\{\pi_1 \cdots \pi_n\}$, the framework computes $f(q,u) = x$ with $x \in \{true, false, unknown\}$. The value *unknown* is returned if the folksonomy does not contain tags related to q. A return value $true$ means that q identifies a service provider in the folksonomy, and the privacy practices of the service provider match the preferences of user u. Otherwise, $false$ is returned.

To ease the presentation, we divide the computation into three steps. First the TagMiner partitions the (O,U,T)-tuples stored in the PrivacyTagger component into two sets of $(object, tag)$-pairs with positive (M^{pos}) and negative (M^{neg}) tags:

$$M^{pos} = \left\{(\hat{o},\hat{u}) \mid (\hat{o},\hat{u},\hat{t}) \in (O,U,T) \ \wedge \ tm(\hat{t}) = positive\right\}$$
$$M^{neg} = \left\{(\hat{o},\hat{u}) \mid (\hat{o},\hat{u},\hat{t}) \in (O,U,T) \ \wedge \ tm(\hat{t}) = negative\right\}$$

Second, we find out if q refers to a privacy issue that is threatening for the user or not. Therefore, we compute a score over each pair $(\hat{o},\hat{u}) \in M^{pos}, M^{neg}$. The ThreatTaxonomy component computes a measure that quantifies the similarity between the threat in question q and a threat $\hat{o}$ that has been tagged before. The PreferenceMatcher provides a measure for the similarity of the current user u and the user $\hat{u}$ who provided the tag. One way to compute the score is to sum the products of these values:

$$score = \sum_{(\hat{o},\hat{u})\in M^{pos}} tt(\hat{o},q) \cdot pm(\pi_u,\pi_{\hat{u}}) - \sum_{(\hat{o},\hat{u})\in M^{neg}} tt(\hat{o},q) \cdot pm(\pi_u,\pi_{\hat{u}})$$

Finally, the framework returns *unknown* to the user if there is no privacy threat similar to q in the folksonomy. It returns *true* if the score is positive and $false$ otherwise.

$$f(q,u) = \begin{cases} unknown & \text{if } \forall o \in O : tt(q,o) = 0 \\ true & \text{if } score > 0 \\ false & \text{otherwise} \end{cases}$$

Note that we have kept the framework simple on purpose to ease presentation. It has to be investigated if more elaborate methods result in an increased utility. For example, the score could outweigh negative tags to reflect that false positives are more problematic than false negatives. Further, the score could weigh the tags of the questioner higher than the tags provided by others. Another extension could provide an aging mechanism to remove outdated privacy issues. It requires a prototypical implementation and user studies to investigate the effect of these extensions.

4 Anonymity vs. Pseudonymity

The framework proposed so far requires unique identifiers for the users. Although pseudonyms are well-suited as identifiers, knowledge of all (o,u,t) records of a certain user actually is a privacy threat. In the following we discuss how the framework must change in order to provide full anonymity.

In its anonymity-preserving variant, the PrivacyTagger component stores $(object, tag)$ records, i.e, $(O,T) = \{(o_1,t_1),\cdots,(o_n,t_n)\}$. Because it is possible to infer the user identity based on his IP address and the queries he issues, PrivacyTagger should store these records in a Peer-to-Peer data structure (see [23] for an overview), where the records are distributed among many peers, and communication can be encrypted. Without knowing the set of users U and their preferences π_u, the framework is restricted to compute $f(q)$ based on the similarity between the threat q and the set of already tagged threats O:

$$M^{pos} = \{o \mid (o,t) \in (O,T) \ \wedge \ tm(t) = positive\}$$

$$M^{neg} = \{o \mid (o,t) \in (O,T) \ \wedge \ tm(t) = negative\}$$

$$f(q) = \begin{cases} unknown & \text{if } \ \forall o \in O : tt(q,o) = 0 \\ true & \text{if } \sum\limits_{o \in M^{pos}} tt(o,q) - \sum\limits_{o \in M^{neg}} tt(o,q) > 0 \\ false & \text{otherwise} \end{cases}$$

After having sketched an anonymous privacy framework, we compare it with the pseudonymous variant. We see two reasons why the framework should be implemented using pseudonyms:

- Not knowing which user has assigned a tag to a privacy issue would degrade the service quality. It is impossible to provide incentives for the users, i.e., amount and quality of the anonymous tags would be worse in comparison to a pseudonymous implementation. Further, it is not possible to compute the similarity between the current user and the user who provided a certain tag. Thus, the decision of the anonymous framework relies solely on the accumulated opinions and the experiences of the majority of users. While the framework would represent the joint privacy standards of the society, it cannot consider individual preferences.
- The framework can be used to assess itself, i.e., the users are free to generate tags on the URL of a service that implements the framework. To provide an extreme example, immediately after a new user has specified his preferences, a warning could appear that he should delete his profile and log off. However, the framework does not require to disclose any personal information. Given that a responsible and dependable provider implements the framework, such extreme cases should not happen.

It is up to future research to find out if the concerns regarding a pseudonymous privacy mechanism outweigh its increased usefulness, as compared to an anonymous variant.

5 Discussion

The benefits of the Privacy2.0 framework envisioned are broad. It promises to produce a new level of transparency on privacy violations. While it is hard for the individual to estimate if a certain provider meets a daunting number of regulations and follows acceptable privacy standards, our framework profits from the observations and experiences of a community of users. Due to the flexibility of the underlying folksonomy infrastructure, the framework addresses a wide range of privacy threats and user preferences without forcing the users to master the details of a large number of isolated data-protection techniques. By providing transparency for the data handling practices of service providers, the framework fits into the EU directive [10] and could be a cornerstone of self-regulation approaches for data protection.

Legal Aspects of our Framework. Because our framework does not restrict the tags generated, users are free to provide slander, gossip or negative opinions without any reason. This raises legal challenges for a trusted third party which operates an implementation of the framework. As an instance for the situation in the EU, we will briefly summarize recent developments in Germany.

Reviews on the eBay platform[7] are similar in length and nature to the tags in our framework. Thus, the legal risks of the framework can be derived from judicial decisions on eBay lawsuits. On April 3, 2006, the Higher Regional Court of Oldenburg has decided on the legitimacy of negative reviews (file reference 13 U 71/05). Specifically, the court has defined under which premises a few words constitute an untrue claim of fact which violates the personal rights of the individual concerned. Another relevant decision comes from the Higher Regional Court of Koblenz at July 12, 2007 (file reference 2 U 862/06). The court has specified the borderline between claims of fact that can be verified, subjective expressions of opinion, value judgments and illegal abusive criticism. However, an in-depth investigation of the legal issues cannot be performed solely on the framework, but requires a concrete implementation of the components and of the score function.

Directions for Future Research. The framework presented tries to establish a new research direction in the field of data protection with a focus on society. From this point of view, both technical and legal measures for data protection are of utmost importance: Since it cannot be guaranteed that users detect all privacy threats, the framework proposed complements existing legal and technical methods, but does not want to replace them. In that sense our framework uses collaborative mechanisms to observe and communicate if a service provider implements appropriate and effective techniques.

As a consequence, research on the framework calls for multidisciplinarity. Computer scientists would explore efficient and effective technologies for the infrastructure. Sociologists would investigate how interactions between people of different cultures, social classes, gender etc. influence the use of the framework proposed. Since both privacy threats and privacy techniques affect the behavior of the users, research in technology assessment is needed to estimate the impact of an implementation of the framework on society. There are open questions regarding data-privacy legislation, liability for misuse, copyright etc. which require the attention of jurists. Finally, economists have to investigate the relationships between pricing and the handling of private data in the presence of a mechanism that makes privacy violations public.

An evaluation of the framework must reflect this multidisciplinary alignment and therefore consider different perspectives. The technical point of view requires to prove that the basic infrastructure is effective and scales well in a global setting

[7] http://www.ebay.com

where many people issue large numbers of queries in parallel. One option to evaluate the social and legal aspects is to initiate and supervise a public discourse in collaboration with privacy activists and media partners. Field tests with a prototypical implementation in a supervised environment can provide insight in the behavior of the users and of individuals concerned, the quality of the tags and the applicability of the framework components in isolation.

6 Conclusion

The integration of current and future technologies in the everyday life will shape the society of the future not only because of its benefits, but also due to significant new challenges with regard to data protection. Current solutions for data protection require time and effort from the user. In the presence of many networked devices to collect, transfer and process personal data, these solutions cannot ensure privacy.

The objective of this paper is to propose investigating the deployment of Web2.0 technologies to support individuals in protecting their privacy. To this end, we have specified a framework based on a folksonomy of tagged geo-locations, physical items and Internet addresses with an impact on privacy. It lets the users share observations and experiences on data-privacy issues. By comparing possible privacy threats to the user preferences, the framework helps the user to decide if a particular data collector handles personal information with care and according to his preferences.

A broad variety of technology needed to implement the framework is currently being researched or developed. Future research issues are multidisciplinary and involve computer science, social science, economic science and jurisprudence.

References

1. Beatty, P., Reay, I., Dick, S., Miller, J.: P3P Adoption on E-Commerce Web sites: A Survey and Analysis. IEEE Internet Computing (IC) **11**(2), 65–71 (2007)
2. Beenen, G., et al.: Using Social Psychology to Motivate Contributions to Online Communities. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW'04) (2004)
3. Chawathe, S.S., et al.: Managing RFID Data. In: Proceedings of the 30st International Conference on Very Large Data Bases (VLDB'04) (2004)
4. Concil of the European Union: European Policy Outlook RFID (draft version). Working document for the expert conference "RFID: Towards the Internet of Things" (2007)
5. Cranor, L., et al.: The Platform for Privacy Preferences 1.0 (p3p1.0). W3C Recommendation, Available at http://www.w3.org/TR/P3P/ (2002)
6. Cvrcek, D., Kumpost, M., Matyas, V., Danezis, G.: A study on the value of location privacy. In: Proceedings of the 5th Workshop on Privacy in the Electronic Society (WPES'06) (2006)
7. Danezis, G., Lewis, S., Anderson, R.: How Much is Location Privacy Worth? In: Proceedings of the 4th Workshop on Economics of Information Security (WEIS'05) (2005)

8. Dave, K., Lawrence, S., Pennock, D.M.: Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In: Proceedings of the 12th International World Wide Web Conference (WWW'03) (2003)
9. Electronic Privacy Information Center: Pretty Poor Privacy: An Assessment of P3P and Internet Privacy. Available at http://www.epic.org/reports/prettypoorprivacy.html (2000)
10. European Parliament and the Council of the European Union: Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L 281, 11/23/1995, p.31. (1995)
11. European Parliament and the Council of the European Union: Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market. Official Journal L 178 , 07/17/2000 p.1–16 (2000)
12. European Parliament and the Council of the European Union: Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector. Official Journal L 201 , 31/07/2002 p.37–47 (2002)
13. Gaukler, G.M., Seifert, R.W., Hausman, W.H.: Item-Level RFID in the Retail Supply Chain. Production and Operations Management **16**, 65–76 (2007)
14. Gehrke, J., Madden, S.: Query Processing in Sensor Networks. IEEE Pervasive Computing **03**(1), 46–55 (2004)
15. Gorbis, M., Pescovitz, D.: Bursting Tech Bubbles Before They Balloon. In: IEEE Spektrum (2006)
16. Guy, M., Tonkin, E.: Folksonomies: Tidying Up Tags? D-Lib Magazine **12**(1) (2006)
17. Haenselmann, T.: An FDL'ed Textbook on Sensor Networks. Published under GNU FDL at http://www.informatik.uni-mannheim.de/~haensel/sn_book (2005)
18. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: Proceedings of the 10th Conference on Knowledge Discovery and Data Mining (KDD'04) (2004)
19. Klüver, L., et al.: ICT and Privacy in Europe – A Report on Different Aspects of Privacy Based on Studies Made by EPTA Members in 7 European Countries. Available at DOI: http://dx.doi.org/10.1553/ITA-pb-a44s (2006)
20. Lambiotte, R., Ausloos, M.: Collaborative Tagging as a Tripartite Network. Available at http://arxiv.org/abs/cs.DS/0512090 (2005)
21. Langheinrich, M.: A Privacy Awareness System for Ubiquitous Computing Environments. In: Proceedings of the 4th International Conference on Ubiquitous Computing (UbiComp'02), pp. 237–245 (2002)
22. Marlow, C., et al.: HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. In: Proceedings of the 17th ACM Conference on Hypertext and Hypermedia (Hypertext'06) (2006)
23. Milojicic, D.S., Kalogeraki, V., Lukose, R., Nagaraja, K., Pruyne, J., Richard, B., Rollins, S., Xu, Z.: Peer-to-Peer Computing. Tech. Rep. HPL-2002-57, HP Labs (2002).
24. Ninghui, L., Tiancheng, L., Venkatasubramanian, S.: t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. Proceedings of the 23rd International Conference on Data Engineering (ICDE'07) (2007)
25. Raabe, O., Dinger, J., Hartenstein, H.: Telekommunikationsdienste in Next-Generation-Networks am Beispiel von Peer-to-Peer-Overlay-Systemen. In: Kommunikation und Recht (2007)
26. Schenker, A., et al.: Comparison of Distance Measures for Graph-Based Clustering of Documents. In: Proceedings of the 4th International Workshop on Graph Based Representations in Pattern Recognition (GbRPR'03) (2003)
27. Sweeney, L.: k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(5), 557–570 (2002)
28. Wang, F., Liu, P.: Temporal Management of RFID Data. In: Proceedings of the 31st International Conference on Very Large Data Bases (VLDB'05), pp. 1128–1139 (2005)
29. Weiser, M.: The Computer for the 21st Century. ACM SIGMOBILE Mobile Computing and Communications Review **3**(3) (1999)
30. Yan, X., Yu, P.S., Han, J.: Substructure Similarity Search in Graph Databases. In: Proceedings of the 24th International Conference on Management of Data (SIGMOD'05) (2005)