Privacy-Preserving Enhanced Collaborative Tagging

Javier Parra-Arnau, Andrea Perego, Elena Ferrari, *Fellow*, *IEEE*, Jordi Forné, and David Rebollo-Monedero

Abstract—Collaborative tagging is one of the most popular services available online, and it allows end user to loosely classify either online or offline resources based on their feedback, expressed in the form of free-text labels (i.e., tags). Although tags may not be *per se* sensitive information, the wide use of collaborative tagging services increases the risk of cross referencing, thereby seriously compromising user privacy. In this paper, we make a first contribution toward the development of a privacy-preserving collaborative tagging service, by showing how a specific privacy-enhancing technology, namely *tag suppression*, can be used to protect end-user privacy. Moreover, we analyze how our approach can affect the effectiveness of a policy-based collaborative tagging system that supports enhanced web access functionalities, like content filtering and discovery, based on preferences specified by end users.

Index Terms—Policy-based collaborative tagging, social bookmarking, tag suppression, privacy-enhancing technology, Shannon's entropy, privacy-utility tradeoff

1 Introduction

Collaborative tagging is one of the most diffused and popular services available online. First provided by social bookmarking sites only—for example, Delicious (http://delicious.com), Digg (http://digg.com), StumbleUpon (http://stumbleupon.com)—it is currently supported by nearly any type of social web application, and it is used to annotate any kind of online and offline resources (e.g., webpages, images, videos, movies, music, and even blog posts).

The main purpose of collaborative tagging is to loosely classify resources based on end-user's feedback, expressed in the form of free-text labels (i.e., tags). The novelty of such an approach to content/resource categorization has been seen, in recent years, as a challenging research topic. In fact, collaborative tagging may be the basis for a semantic network connecting online resources based on their characteristics, and not only their URIs. At the same time, the undefined semantics of tags, which are per se ambiguous and expressed in multiple languages, makes it difficult to enforce semantic interoperability and to grant a reasonable level of accuracy when determining the "meaning" of a tag.

- *J. Parra-Arnau, J. Forné, and D. Rebollo-Monedero are with the Department of Telematics Engineering, Universitat Politècnica de Catalunya, Campus Nord, C./ Jordi Girona 1-3, 08034 Barcelona, Spain. E-mail: {javier.parra, jforne, david.rebollo}@entel.upc.edu.*
- *A. Perego is with the European Commission - Joint Research Centre of the European Commission, Institute for Environment & Sustainability, Unit H06 - Digital Earth & Reference Data, Via E. Fermi, 2749 - TP 262, 21027 Ispra VA, Italy. E-mail: andrea.perego@jrc.ec.europa.eu.*
- *E. Ferrari is with the Department of Theoretical and Applied Science, University of Insubria, Via Mazzini 5, I-21100 Varese, Italy. E-mail: elena.ferrari@uninsubria.it.*

Manuscript received 14 Nov. 2011; revised 8 May 2012; accepted 7 Nov. 2012; published online 28 Dec. 2012.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2011-11-0695. Digital Object Identifier no. 10.1109/TKDE.2012.248.

Based on such considerations, most research work has investigated how to effectively reuse tag collections (referred to as *folksonomies*) in the semantic Web framework (see, e.g., [1], [2], [3]), and analyzed collaborative tagging practices to enforce strategies addressing the semantic ambiguity issue (e.g., as in [4]), by statistically analyzing tag collections to infer, whenever possible, a semantic alignment of at least a subset of tags.

Although collaborative tagging is mainly used to support tag-based resource discovery and browsing, it could also be exploited for other purposes. As an example, the tags collected by social bookmarking services can be exploited to enforce enhanced web access functionalities, like content filtering and discovery, based on preferences specified by the end user. However, to achieve this enhanced use, the current architecture of collaborative tagging services must be extended by including a *policy layer*. The aim of this layer will be to enforce user preferences, intensionally denoting resources on the basis of the set of tags associated with them, and, possibly, other parameters concerning their trustworthiness (the percentage of users who have added a given tag, the social relationships and characteristics of those users, etc.). This is a new research topic, and, to the best of our knowledge, the only work addressing this issue is reported in [5], where a multilayer policy-based collaborative tagging system is described.

However, besides the support to policy enforcement, enhanced collaborative tagging requires another layer which addresses an issue so far not deeply investigated, i.e., *privacy protection*. Although the collection of end-users' private information stored by social services, like Facebook, is now recognized as a privacy threat [6], [7], it is worth noting that the public availability of user-generated data (as tags are) could be used to extract an accurate snapshot of users' interests or *user profiles*, containing sensitive information, such as health-related information, political preferences,

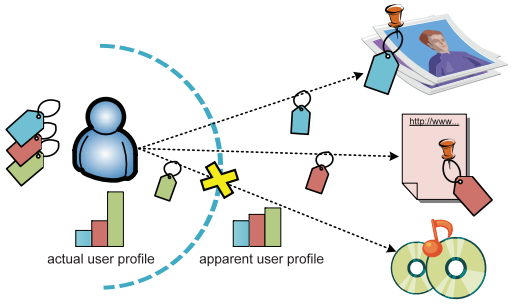


Fig. 1. Conceptually speaking, our tag suppression technique enables a user to protect his/her privacy by refraining from tagging some resources. In doing so, the actual user profile, that is, the profile capturing the user genuine interests, is observed from the outside as a perturbed profile. We refer to it as the apparent user profile.

salary or religion. Actually, the huge number of users using collaborative tagging services, and the fact that collaborative tagging is a service supported virtually by any social online application, increases the risk of cross referencing, thereby seriously compromising user privacy. Indeed, it could be possible to correlate the account of a user with other accounts he/she may have at different services, which would imply gaining far more precise information about the user profile.

Consequently, collaborative tagging requires the enforcement of mechanisms that enable users to protect their privacy by allowing them to hide certain user-generated contents (unless they desire otherwise), without making them useless for the purposes they have been provided in a given online service. This means that privacy-preserving mechanisms must not negatively affect the service accuracy and effectiveness (e.g., tag-based browsing, filtering, or personalization).

In this paper, we make a first contribution in this direction by showing how a specific privacy-enhancing technology (PET), namely *tag suppression* [8], can be used to protect end-user privacy; and second, we analyze how our approach can affect the effectiveness of policy-based collaborative tagging systems. Tag suppression is a technique that has the purpose of preventing privacy attackers from profiling users' interests on the basis of the tags they specify (see Fig. 1). Conceptually, our approach protects user privacy to a certain extent, by dropping those tags that make a user profile show bias toward certain categories of interest. Mathematically, we model a user profile as a histogram of relative frequencies of tags across categories of interest, and quantify the degree of privacy attained by the modified profile as its Shannon entropy. Building on these premises, we extend our preliminary work [8], without empirical evaluation, and conduct a thorough experimental analysis to assess the impact of tag suppression on one of the most popular social bookmarking services, Delicious.

More precisely, we illustrate an architecture, built on top of Delicious, consisting of two additional services. The former enables users to specify policies both to block undesired web content and to denote resources of interest. The latter implements tag suppression. The combination of these two services allows us to broaden the functionality of collaborative tagging systems and, at the same time, to provide users with a mechanism to preserve their privacy while tagging. Moreover, we carry out an extensive performance evaluation of this architecture, showing its effectiveness in terms of privacy guarantees, data utility and filtering capabilities for two key scenarios, for example, parental control and resource recommendation. Since we are not aware of similar experimental studies, we believe that what reported in this paper can be useful to evaluate further future developments in the area.

The remainder of this paper is organized as follows: Section 2 discusses related work, whereas Section 3 describes the proposed approach. Section 4 illustrates our tag suppression mechanism. Section 5 introduces the two reference scenarios on which the tag suppression approach has been tested, whereas performance results are reported and discussed in Section 6. Section 7 concludes the paper and outlines future research directions. Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2012.248, details the methodology we apply for tag categorization.

2 Related Work

Social/collaborative tagging has been early recognized as a challenging research topic [4], [9]. From 2005, several papers have studied its specific characteristics, the similarities, and differences with "traditional" annotation techniques, as well as how tags' collections evolve over time.

For instance, recent work has proposed interesting approaches to *tag recommendation* and *prediction*. Tag prediction concerns the possibility of identifying the most probable tags to be associated with a nontagged resource, whereas tag recommendation is meant to suggest to users the tags to be used to describe resources they are bookmarking. In both cases, existing approaches apply techniques usually enforced in recommendation systems [10]. For instance, in [11], [12], tags are predicted based on resource's content and its similarity with already tagged resources. By contrast, in [13], [14], [15] tag recommendation is enforced by computing tag-based user profiles, and by suggesting tags specified on a given resource by users having similar characteristics/interests, whereas in [16] the authors use a rating-based approach, as in reputation systems [17]. However, so far, very few works have investigated how social tagging can be used to enhance users access to web resources. Actually, research on this issue mainly focused on how and/or whether social tagging can improve web search (see, e.g., [18], [19]).

Another interesting issue concerns the exploitation of the "explicit" relationships between users (i.e., the actual social network underlying a folksonomy) to address issues like annotation relevance and/or trustworthiness. Such topic has been thoroughly studied in social media focused on offline resources (e.g., movies and music), where the relationships existing between users are used to weigh the relevance of the collected ratings and/or tags for specific users (see, e.g., [20]). However, such issue has not been yet enough investigated in social bookmarking services.

Note that all the works discussed so far consider only the relationships that can be inferred based on profile similarity, whereas the lists of contacts created by registered users are totally ignored. "Explicit" user preferences are another under investigated topic in the field of social bookmarking.

Although, no collaborative tagging service allows its members to specify preferences, these could be exploited to enhance web access by supporting features like resource quality assessment and personalization. This means investigating how social bookmarking services can be extended to support and enforce user-specified policies.

To the best of our knowledge, exploitation of explicit relationships and user preferences has been studied only in [5], where a multilayer architecture is proposed integrating a basic social tagging service with trust relationships and user preferences. One of the notable characteristics of such framework is the support of a rule layer, which can be used to express and enforce user preferences. Such preferences are coded into policies explicitly specifying the set of trustworthy tags by denoting their creators in terms of their relationships and/or characteristics. Also, they state which action must be performed by the system when accessing a resource associated with a given set of tags (mark it as trustworthy or not, as un/safe, etc.).

Privacy protection in social tagging services is another issue that has not been thoroughly investigated. Nevertheless, personalized web access and content filtering may be ultimately considered as content recommendation services. In the context of recommendation systems, numerous approaches have been proposed to protect user privacy. These approaches basically suggest perturbing the information provided by users. For instance, [21] proposes that users add random values to their ratings and then submit these perturbed ratings to the recommender. After receiving these ratings, the system executes an algorithm and sends the users some information that allows them to compute the prediction. When the number of participating users is sufficiently large, the authors find that user privacy is protected to a certain extent and the system reaches a decent level of accuracy. In this line, [22] applies the same perturbative technique to collaborative filtering algorithms based on singular-value decomposition. However, even though a user disguises all his/her ratings, it is evident that the items themselves may uncover sensitive information. In other words, the simple fact of showing interest in a certain item may be more revealing than the ratings assigned to that item. Apart from this critique, other works [23], [24] stress that the use of *randomized* data distortion techniques might not be able to preserve privacy.

3 Overview of the Proposed Approach

As we discussed in Section 1, social bookmarking services are among the most used social services, and, thanks to their support to collaborative tagging, they can be currently considered as the most valuable knowledge acquisition tools, as far as online resources are concerned.

We have also pointed out that collaborative tagging is not exploited to its full potential, since it is typically used just to support tag-based resource browsing and search, despite the fact that collaborative tagging systems can be easily enhanced without modifying their core architecture, because they provide access to the collected information via APIs, which can be easily exploited by external applications. One of the reasons is that the size of the collected data sets is too big to allow the enforcement of even simple mechanisms, concerning, for example, personalization, content filtering, and quality assessment.

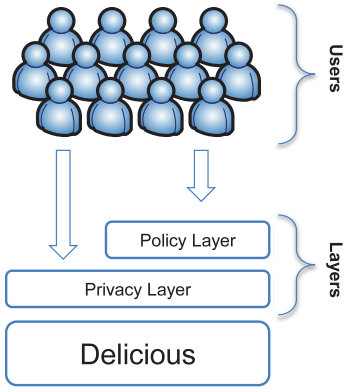


Fig. 2. Architecture of the proposed enhanced social tagging service.

For these reasons, in this paper, we describe an enhanced collaborative tagging system that consists of a "traditional" bookmarking service, such as Delicious, and two main additional services built on top of it (see Fig. 2). Such services address two main issues. The former allows end users to specify policies that can be used either to explicitly denote resources of interests or to enforce blocking conditions on the browsed data. The latter features a specific PET, namely, *tag suppression*, to preserve the privacy of registered users by hiding the specific characteristics of their profiles. Such an architecture is a specific implementation of the multilayer framework presented in [5], with the relevant difference that in [5] the privacy layer is missing. Lastly, we would also like to emphasize that our approach is not limited to the specific bookmarking application here contemplated, i.e., Delicious. As a matter of fact, it could be built on top of any collaborative tagging system, such as BibSonomy, CiteULike or Flickr.

But which is the purpose of combining a policy layer and a privacy layer? As discussed in Section 1, privacy is usually considered an issue for those social services that collect end-users' sensible information (e.g., personal data, opinions, photos, and videos). Social bookmarking services do not fall in this category, because they do not require the user to specify personal data (with the exception of the users' name and email) and they do not collect user-generated contents. Due to this, social bookmarking services do not provide data protection mechanisms—even those available, for example, in Facebook, which are not enough to prevent the disclosure of private data. As an example, Delicious allows registered users just to flag a bookmark as public (default option) or private. When a user marks a bookmark as private, this bookmark and its associated tags are hidden to other Delicious users. Note, however that, even if a user flags all his/her bookmarks and tags as private, the Delicious server still records this information.

Nevertheless, if tags were not sensible information per se, they could easily be exploited to infer users' personal information, such as personal interests, preferences, and opinions. This is even easier when it is possible to statistically analyze huge collections of tags as those made publicly available by social bookmarking services, thus obtaining accurate tag-based user profiles. In this field, privacy-preserving techniques should guarantee, at the

Fig. 3. The tags posted by a user in a collaborative tagging service are frequently depicted as a tag cloud. This representation is equivalent to the model of user profile assumed in Section 4.3. The tag cloud shown here is also represented as a PMF in Fig. 4.

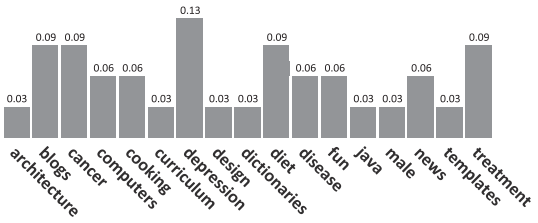


Fig. 4. The example of user profile here depicted shows the equivalence between the tag cloud illustrated in Fig. 3 and a histogram of relative frequencies of tags.

same time, 1) privacy protection and 2) the correctness of the results obtained by analyzing the data set.

The problem here is not only to find the correct tradeoff between these two issues. In fact, since collaborative tagging is used to find/browse resources based on the associated tags, suppressing tags might decrease accuracy, and increase the number of false positives/negatives. Moreover, if tags are used for more sensible purposes (parental control, quality assessment, etc.), this might have even worse consequences. For these reasons, the support to privacy-preserving techniques is a key requirement when we come to enhanced policy-based uses of collaborative tagging. In fact, in such cases, users may tend to annotate resources by using tags which can be reused for specific purposes—for example, parental control. Such tags are then even more sensitive than the ones collected by traditional collaborative tagging services. Our aim is to verify whether and how tag suppression can be effectively applied also in an enhanced collaborative tagging service, as the one illustrated in this paper.

Next, we first illustrate our tag suppression technique, and then we describe the reference scenarios we have used to carry out our experiments.

4 Tag Suppression

In our scenario of collaborative tagging, users tag resources on the web, for example, music, pictures, videos or bookmarks, according to their personal preferences. Users therefore contribute to describe and classify those resources, but this is *inevitably* at the expense of revealing their profile. To avoid being accurately profiled by tagging systems, or in general by any attacker able to collect such information, users may adopt a privacy-enhancing technology based on *data perturbation*.

The data-perturbative technology considered in this work is *tag suppression*, a technique that allows a user to refrain from tagging certain resources in such a manner that the profile resulting from this perturbation does not capture their interests so precisely. Our conceptually simple technique protects user privacy to a certain degree, but at the cost of the semantic loss incurred by suppressing tags. Other approaches based on data perturbation include the submission of false tags. For example, a user wishing to tag the webpage www.mentalhelp.net with "depression" could use the tag "sports" instead, to conceal their interest for this resource. In doing so, the user distorts their actual profile, although at the expense of a far greater impact on semantic functionality than suppression does—resources are assigned tags that do not describe, in principle, the actual content of such resources. A more intelligent form of tag perturbation consists in replacing (specific) user tags with (general) tag categories. In conceptual terms, and resorting to the example above, the user would use the tag "health" instead of "depression." In this manner, the user would hide, to a certain extent, their genuine interest in that resource, but clearly at the cost of some vagueness or inaccuracy in the description of that webpage.

Among these approaches, we consider tag suppression as a suitable strategy for the enhancement of user privacy in the scenario of collaborative tagging, not only because of its simplicity in terms of implementation costs, but also because of its lower impact on semantic functionality. Lastly, we would like to emphasize the synergic effect of our approach in combination with other strategies based on data perturbation.

4.1 User Profile Model

In the scenario of social bookmarking, a user browses the web bookmarks pages and assigns tags to them according to his/her profile of interests. As in our previous work on tag suppression [8], we consider n tag categories, indexed by $1, \ldots, n$, and model the profile of a user as a probability mass function (PMF) $q = (q_1, \ldots, q_n)$, that is, a histogram of relative frequencies of tags across these categories. Our model of user profile is equivalent to the *tag clouds* that numerous collaborative tagging services use to visualize the tags posted by users; a tag cloud is a visual representation in which tags are weighted according to their frequency of use. This equivalence is illustrated in Fig. 3 where we show an example of user profile modeled as a tag cloud. The font size of these tags is then used in Fig. 4 to represent the same user profile as a PMF.

Under this model, a privacy attacker, possibly the social bookmarking provider itself, supposedly observes a perturbed version of this profile, according to a tag suppression strategy, and is unaware or ignores the fact that the observed user profile, also in the form of a histogram, does not reflect the actual profile of interests of the user in question. By refraining from tagging on certain categories, the *actual* profile of interests q is perceived from the outside as the *apparent* PMF $s = \frac{q-r}{1-\sigma}$. In this expression, $\sigma \in [0, 1)$ denotes the *suppression rate*, that is, the total fraction of tags a user is willing to eliminate, and r the *suppression strategy*, which models the specific frequencies of categories suppressed, with $0 \leq r_i \leq q_i$ and $\sum r_i = \sigma$. In a nutshell, the apparent profile may be interpreted intuitively as the result of the suppression of some tags from the actual profile, and the posterior normalization by $\frac{1}{1-\sigma}$ so that $\sum_{i=1}^{n} s_i = 1$.

4.2 Measuring the Privacy of a User Profile

In this section, we shall present an information-theoretic criterion to quantify the privacy of user profiles. To make the presentation of our privacy criterion suited to a wider audience, next we shall review two fundamental quantities of information theory, namely Shannon's entropy and Kullback-Leibler (KL) divergence.

Recall [25] that the Shannon entropy $\mathrm{H}(s)$ of a PMF s is defined as $\mathrm{H}(s) = -\sum_{i=1}^{n} s_i \log_2 s_i$, that is, as a measure of the uncertainty of the outcome of a random variable distributed according to such PMF, and that it is maximized, among all distributions on $\{1, \ldots, n\}$, by the uniform distribution $u_i = 1/n$ for all i, for which $\mathrm{H}(u) = \log_2 n$. Recall also that, given two probability distributions s and p over the same alphabet, KL divergence $\mathrm{D}(s \parallel p)$ is defined as $\mathrm{D}(s \parallel p) = \sum_{i=1}^{n} s_i \log_2 s_i/p_i$. The KL divergence is often referred to as *relative entropy*, as it may be considered as a generalization of the entropy of a distribution, relative to another. Conversely, entropy may be regarded as a special case of KL divergence when p is the uniform distribution:

$$\mathrm{D}(s \parallel u) = \log_2 n - \mathrm{H}(s). \tag{1}$$

Recently, these two information-theoretic quantities have been used as measures of the privacy of user profiles. Some examples include [26], [27], which quantify privacy as the entropy of the apparent distribution s, and [8], [28], which measure privacy risk as the KL divergence between this apparent profile s and the average population's profile p. In this work, we rule out the use of divergence as a privacy criterion because of the unrealistic assumption that the population's tag distribution p is available to users. In the absence of this information, as it is the case of Delicious and many other collaborative tagging systems, it is then reasonable to assume the uniform distribution, i.e., $p = u$. Under this assumption, note that both measures of privacy are essentially equivalent, owing to (1). Motivated by all this, in this paper we measure the level of privacy attained by the apparent profile s as its Shannon entropy $\mathrm{H}(s)$.

An intuitive justification of our privacy metric stems from the observation that an attacker will have actually gained some information about a user provided that his/her interests are significantly concentrated on a subset of categories. Put differently, an apparent user profile s close to u makes the user go unnoticed in the sense that his/her interests appear to be equally distributed across all categories.

A richer justification of Shannon's entropy as a privacy criterion may be found in [29], where it is interpreted under the perspective of Jaynes' rationale behind entropy-maximization methods [30], the method of types and large deviation theory. The leading idea is that the method of types from information theory establishes an approximate monotonic relationship between the likelihood of a PMF in a stochastic system and its entropy. Loosely speaking and in our context, the higher the entropy of a profile, the more likely it is, the more users behave similarly. Under this interpretation, entropy is, more precisely, a measure of anonymity, *not* in the sense that the user's identity remains unknown, but only in the sense that higher likelihood of an apparent tag profile, believed by an external observer to be the actual profile, makes that profile more common, hopefully helping the user go unnoticed, less interesting to an attacker whose objective is to target peculiar users.

The use of entropy as a measure of privacy, in the widest sense of the term, is by no means new. As a matter of fact, Shannon's work in the fifties introduced the concept of *equivocation* as the conditional entropy of a private message given an observed cryptogram [31], later used in the formulation of the problem of the wiretap channel [32] as a measure of confidentiality. More recent studies [33], [34] rescue the suitable applicability of the concept of entropy as a measure of privacy, by proposing to measure the degree of anonymity observable by an attacker as the entropy of the probability distribution of possible senders of a given message.

In addition, we would like to point out that the coexistence of a number of users benefiting from our mechanism has a synergic effect in terms of the privacy attained. The reason is that as each apparent profile of a population of such users approaches the uniform distribution, their apparent profiles also approach each other.

4.3 Optimization of the Privacy-Suppression Tradeoff

Equipped with a quantitative measure of privacy, now we are interested in choosing a suppression strategy r so that s maximizes $\mathrm{H}(s)$ for a given σ. Formally speaking, we would like to solve the multiobjective optimization problem given by the *privacy-suppression* function:

$$\mathcal{P}(\sigma) = \max_{\substack{r \\ 0 \le r_i \le q_i \\ \sum r_i = \sigma}} \mathrm{H}\Big(\frac{q - r}{1 - \sigma}\Big) \tag{2}$$

which characterizes the optimal tradeoff between privacy and tag suppression rate. Although this optimization will be carried out for suppression rate as a measure of utility, which makes the problem tractable, the remainder of our work is devoted to assess the loss in data utility and accuracy due to tag suppression in terms of certain percentages regarding missing tags on bookmarks, on the one hand, and on the other, in terms of false positives and negatives.

As stressed in Section 4.2, our formulation is built upon the premise that the population's tag distribution p is unknown to users, what leads us to assume $p = u$. Under this assumption, and on account of (1), entropy maximization is a special case of divergence minimization. Note, however, that if p was available to users, it would be preferable to use KL divergence as a measure of privacy (risk). This is because divergence minimization may reduce the degradation in utility compared to entropy maximization, which strives to make the apparent profile close to u, ignoring the fact that certain categories may be more popular than others.

In light of formulation (2), we would also like to remark two important advantages in modeling the privacy of a user profile as an entropy. First, the mathematical tractability demonstrated in previous, related work [28]. Second, the privacy-suppression function is defined in terms of an optimization problem, whose objective function is concave, subject to affine constraints. As a consequence, this problem

belongs to the extensively studied class of convex optimization problems [35] and may be solved numerically, using a number of extremely efficient methods, such as interior-point methods.

Another important aspect that follows directly from our formulation is the intuitive fact that there must exist a tag suppression rate beyond which the privacy-suppression function achieves its maximum value or *critical privacy* $\mathcal{P}_{\text{crit}} = \text{H}(u) = \log_2 n$. We refer to this suppression rate as the *critical suppression rate* and define it formally as

$$\sigma_{\text{crit}} = \min\{\sigma : \mathcal{P}(\sigma) = \mathcal{P}_{\text{crit}}\}. \tag{3}$$

Interestingly, it can be shown that $\sigma_{\text{crit}} = 1 - n \min_i q_i$, which implies that critical privacy is never attained for $\sigma < 1$, provided that q has at least one zero component. To see this, next we sketch a proof. Let r^* be the solution to the maximization problem $\mathcal{P}(\sigma)$. Assume that σ is such that $\mathcal{P}(\sigma) = \mathcal{P}_{\text{crit}}$, or equivalently, that $\text{H}(\frac{q-r^*}{1-\sigma}) = \log_2 n$. Note that this assumption is satisfied if, and only if, the optimal suppression strategy is $r_i^* = q_i - \frac{1-\sigma}{n}$ for $i = 1, \ldots, n$. However, r^* must also satisfy the constraints of the optimization problem, $\sum r_i^* = \sigma$ and $0 \leq r_i^* \leq q_i$. Check that the equality constraint holds and, by virtue of $\sigma < 1$, that the right-hand inequality is verified. Finally, observe that the left-hand inequality is equivalent to $q_i - \frac{1-\sigma}{n} \geq 0$ for all i, and ultimately to $\sigma \geq \max_i 1 - n\, q_i$, which leads to the aforementioned expression of σ_{crit}.

The importance of this result lies in the fact that a user not tagging across all categories will not achieve an apparent user profile close to u for any meaningful tag suppression rate lower than 100 percent. Put differently, no suppression strategy fulfilling the constraints in (2) can lead to the uniform distribution whenever the genuine profile vanishes at some components. This fundamental property about our tag suppression mechanism will be later used to justify some of the results shown in Section 6.

5 Reference Scenarios

As most PETs, tag suppression must address two main issues: protecting user privacy and granting that the perturbed data set can be effectively used. Specifically, we must verify whether the semantic loss incurred by tag suppression to protect private data can be acceptable. Clearly, the acceptable semantic loss threshold may highly depend on the purpose for which social bookmarking is used. Depending on it, we may require different levels of semantic accuracy, and we may have a higher or lower error tolerance (may such errors concern false positives and/or negatives).

As an example, we can figure out two different scenarios, which are both examples of enhanced uses of social bookmarking, and share the notion of "user-defined policy," i.e., a tag-based intensional definition of resource classes, explicitly expressed by end users. Such classes, depending on the purpose for which policies are specified, may denote an assessment of the quality, safety, relevance, and so on, of tagged resources. In the former scenario end users specify policies to inform the bookmarking service about the resources they consider relevant. Based on them, the social bookmarking service regularly updates end users, for example, by using web feeds, about the resources denoted by the policies. It can be considered as a subscription service that makes use of a recommendation system relying primarily on the explicit preferences expressed by end users—and not, as traditionally, on the "implicit" preferences that can be inferred by users' past behavior. The latter scenario concerns parental control. Here, policies denote which resources are un/safe. Whenever a user requests access to a resource, such policies are then used to determine whether access to that resource can be granted or should be denied.

Note that the parental control scenario has very low tolerance of false negatives; we refer to *false negatives* as those resources classified as safe, but that are actually unsafe. More precisely, in this scenario, granting access to an unsafe resource is not acceptable at all. In contrast, in the former scenario we can tolerate a higher threshold of false negatives, because recommending a not relevant resource would not compromise the safety of end users.

We introduce here the general definition of policy which can be applied to both scenarios.

Definition 1 (Policy). *A policy pol is a pair (CC, sign), where: 1) CC is a conjunction of category constraints* $(cc_1 \wedge \ldots \wedge cc_n)$*, and 2)* $\text{sign} \in \{+, -\}$*. Each category constraint is a triple* (c, op, t)*, where c is a tag category,* $t \in [0, 1]$*, and op is a comparison operator.*

A category constraint intensionally denotes the set of resources associated with a percentage of tags in the category c which is greater than (less than, equal to, etc., depending on op) the value denoted by t. For example, category constraint $(c, >, 0.5)$ denotes those resources associated with a percentage of tags in category c which is greater than 50 percent. In contrast, the semantics of the $sign$ component depends on the scenario. More precisely, in the resource recommendation scenario it denotes whether the resources matching the category constraint CC are relevant $(+)$ or not $(-)$, whereas in the parental control scenario it denotes whether they are safe $(+)$ or unsafe $(-)$.

Since the support for both positive and negative policies may raise conflicts (i.e., we may have a resource covered by both positive and negative policies), a conflict resolution mechanism must be enforced. The scientific literature provides several examples of approaches which can be adopted (see, e.g., [36] for a survey on this topic). Here, for simplicity we adopt the one according to which negative policies are prevailing, since this approach is the one giving stronger guarantees with regard to the risk of accessing not appropriate contents. However, other conflict resolution policies can be easily adopted as well.

Following, we provide examples of policies for the reference scenarios introduced above. In the examples, we will refer to some of the tag categories we have obtained from our experimental data set (see Section 6 for more details). For brevity, we will denote the relevant tag categories by $c_1, \ldots, c_n$. Also, for simplicity and clarity, in the examples we will keep using the policy formal notation introduced in Definition 1. We would like to note, however, that such notation, describing how policies are actually implemented in the system, is supposed to be made transparent in the front end both to improve usability and

to help users specify policies reflecting as much as possible their intentions. Several strategies can be devised for this purpose, for example, the use of textual labels instead of numeric values and comparison operators. However, a discussion on such issues is out of the scope of this paper.

Example 1 (Policies for resource recommendation). Suppose that Carol (C) is interested in literature, but not in resources concerning science-fiction. C realizes that the relevant tag categories are c_1 ("books") and c_2 ("literary criticism"), and she decides that the resources she is interested in are those associated with not less than 40 percent of the tags in either c_1 or c_2. In contrast, C finds out that the tag category that corresponds to the resources she is not interested in is c_3 ("science-fiction, fantasy"), and she decides to discard all the resources associated with not less than 20 percent of the tags in c_3. Consequently, C specifies the following policies:

- $pol_1 = (\{(c_1, \geq, 0.4)\}, +)$,
- $pol_2 = (\{(c_2, \geq, 0.4)\}, +)$,
- $pol_3 = (\{(c_3, \geq, 0.2)\}, -)$.

Suppose now that there exists a resource R_1, which satisfies content constraints $(c_1, \geq, 0.4)$, $(c_2, \geq, 0.4)$, and $(c_3, \geq, 0.2)$. In such a case, we have a conflict, since all policies pol_1, pol_2, and pol_3 apply. According to our conflict resolution mechanism, policy pol_3 prevails over policies pol_1 and pol_2, since the latter are positive policies. Consequently, resource R_1 is marked as irrelevant to C.

Example 2 (Policies for parental control). Suppose that Alice (A) would like to enable a web filter for her son Bob (B) by granting him access only to contents specifically tailored for children. By checking the available tag categories, she realizes that the suitable one is c_4: "entertainment for children." She then decides that resources suitable to children are those associated with not less than 60 percent of the tags from category c_4. Moreover, just to be sure that no harmful content is accessed, she also would like to prevent access to "entertainment" resources that may include any content for adults. To achieve this, A specifies the following policies:

- $pol_4 = (\{(c_4, \geq, 0.6)\}, +)$;
- $pol_5 = (\{(c_5, \geq, 0.1)\}, -)$, where c_5 is the tag category corresponding to "entertainment for adults."

Suppose now that B requests access to a resource R_2, which satisfies both content constraints $(c_1, \geq, 0.6)$ and $(c_2, \geq, 0.1)$. In such a case, we have a conflict, since both policies pol_4 and pol_5 apply. According to our conflict resolution mechanism, policy pol_5 prevails over pol_4, because pol_5 is a negative policy. Consequently, Bob is denied access to resource R_2.

In the following section, we report the results of a series of experiments which have been carried on for the parental control scenario. The reason of this choice is that such scenario is the most demanding as far as error tolerance is concerned. Therefore, if good results are obtained for this more demanding scenario, they can be extended to the other one as well.

6 Experimental Analysis

In this section, we delve into the impact that tag suppression may have on an enhanced collaborative tagging system based on Delicious. With this aim, Section 6.1 first examines the data set that we used to conduct the experimental evaluation. To make user profiles tractable, Section 6.2 summarizes the methodology that we followed for mapping tags into a small set of meaningful categories of interest. Finally, Section 6.3 shows a comprehensive analysis of the degradation in data utility and accuracy, incurred by the application of our privacy-protecting technique.

6.1 Data Set

In our experiments, we used the Delicious data set retrieved by the Distributed Artificial Intelligence Laboratory (DAI-Labor), at Technische Universität Berlin [37]. This data set includes those bookmarks and tags marked as public by approximately 950,000 users. The information is organized in the form of triples (*username*, *bookmark*, *tag*), each one modeling the action of a user associating a bookmark with a tag. The data set contains 420 millions of these triples, which were posted from September 2003 to December 2007. It is worth mentioning that no preprocessing has been done, though *usernames* have been anonymized by applying a hash function.

The data set that we considered in our analysis is a subset of the entire data set described above. Concretely, we selected out a subset covering approximately one year and including 1,241,029 triples. We decided to choose this subset because, on the one hand, it spanned a significant period of time, and, on the other, it did not involve the processing of millions of triples that would overload our experiments. Our data subset therefore contains 9,588 users, 390,008 resources and 59,505 tags.

6.2 Tag Categorization

The representation of a user profile as a normalized histogram across these 59,505 tags would be certainly unfeasible from various practical perspectives, mainly concerning the unavailability of data to reliably, accurately measure interests across such fine-grained categorization, and, should the data be available, its overwhelming computational intractability. Further, in our experiments but also in data mining procedures, a coarser categorization makes it easier to have a quick overview of the user interests. For example, for users posting the tags "welfare," "Dubya" and "Katrina" it would be preferable to have a higher level of abstraction that enables us to conclude, directly from the inspection of the user profile, that these users are interested in politics.

Motivated by this, we decided to categorize the tags in our data set into a coarser representation with just a few high-level tag categories. Specifically, we used Lloyd's algorithm [38] to group tags into 20 categories; and then, for each of those categories, we clustered its tags into 10 subcategories. The result of this hierarchical clustering yielded a total of 200 subcategories. In the end, we sorted the tags in each subcategory in decreasing order of proximity to the centroid. Fig. 5 shows the subcategories used in the example of policies for the parental control

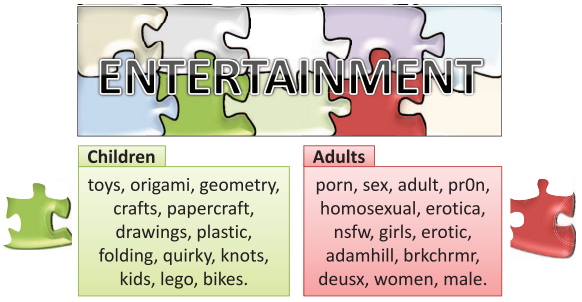


Fig. 5. According to our hierarchical clustering, each category is composed of 10 subcategories. In this example, we represent two subcategories belonging to the "entertainment" category. In particular, we show the tags falling into the subcategory 62 "entertainment for children" and 68 "entertainment for adults," which are used in the specification of policies for the parental control scenario described in Section 5. The two examples of subcategories shown here also illustrate a key result of the categorization process—tags in each subcategory are sorted in decreasing order of proximity to the centroid, which in practice means that those tags at the top of the list are the most representative tags of the subcategory they belong to.

scenario. Appendix A, available in the online supplemental material, provides a complete description of the categorization process summarized here, and Appendix B, available in the online supplemental material, describes the aforementioned Lloyd's algorithm.

6.3 Results

This section presents a number of experimental results that will allow us to evaluate our architecture in terms of privacy protection, utility loss, and filtering accuracy. Specifically, Section 6.3.1 analyses the privacy gain as a result of the application of our tag suppression technique, Section 6.3.2 evaluates the utility loss, whereas Section 6.3.3 provides insight into the loss in filtering accuracy for the parental control scenario.

6.3.1 Privacy

In our architecture, a user specifies a suppression rate indicating the fraction of tags he/she is disposed to eliminate. Based on this suppression rate and the user profile across the $n = 200$ subcategories, our approach numerically solves the optimization problem (2)—see Section 4.3. The numerical method chosen is the interior-point algorithm [39], [40], [41] implemented by the Matlab R2011a function `fmincon`. The algorithm in question makes use of the so-called barrier functions and has a polynomial-time complexity with respect to the number of subcategories [35]. On a side note, we would like to emphasize that this same computational complexity holds for the more general optimization problem consisting in the minimization of the privacy risk measured as the KL divergence between the apparent user distribution and the population's.

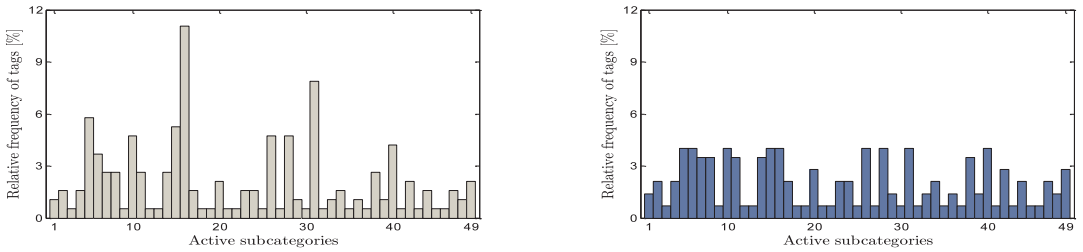
The result of this optimization is a suppression strategy r, that is, an n-tuple containing the percentage of tags that a user should eliminate in each subcategory. With this information, our suppression algorithm proceeds to select which particular tags should be dropped. At this point, recall that, as a result of the categorization process described in Section 6.2, all tags within each subcategory were sorted in decreasing order of proximity to the centroid. Leveraging on this, our tag elimination algorithm operates as follows—since the closer tags to the centroid are the most representative tags of the subcategory they belong to, those tags are given priority in elimination. As an example, suppose that a particular user posted the tags "sex," "erotica," and "women," all of them corresponding to the subcategory "entertainment for adults," and assume that the suppression strategy r recommends eliminating one of those three tags. Since the tag "sex" is closer to the centroid than the other two tags, as shown in Fig. 5, our algorithm would suggest that the user should eliminate this tag. The strategy adopted by this algorithm favors privacy in detriment of data utility, because the unique purpose of our tag suppression technique is to enhance user privacy, given a constraint on utility. However, we would like to note that other strategies putting more emphasis on utility are also possible. An example of this would be the suppression of those tags that are furthest away from the centroid.

In Section 4.3, we mentioned that the critical suppression (3) beyond which critical privacy is attained is given by $\sigma_{\text{crit}} = 1 - n \min_i q_i$. A consequence of this fact is that, in the case when a user does not tag across all subcategories, the critical privacy $\mathcal{P}_{\text{crit}} = \log_2 n$ is not attained for any $\sigma < 1$. This is precisely what happens in our data set, that is, no user has tagged across all subcategories, which in practice means that these users will not achieve an apparent user profile close to u. However, without loss of generality we may consider the subset of subcategories that have been tagged by a particular user. We denote these categories as the *active* subcategories of that user, and the cardinality of this subset as n_{act}. In terms of these subcategories, we may assume the existence of an equivalent critical suppression σ'_{crit} and an equivalent critical privacy $\mathcal{P}'_{\text{crit}}$, in the sense that, beyond this suppression rate, s becomes the uniform distribution across the active subcategories and $\mathcal{P}'_{\text{crit}} = \mathrm{H}(s) = \log_2 n_{\text{act}}$. This interesting property is illustrated in Fig. 6, where we plot the apparent profile of a specific user.[1]

The figure in question shows the user's apparent profile just for the active subcategories. For convenience, we rearranged these subcategories and indexed them from 1 to 49. Clearly, when no suppression is applied, the apparent profile is in fact the actual user profile q. On the other hand, when $\sigma = 0.25$ we observe that the subcategories affected by suppression are those with a percentage of tags furthest away from u. In the special case when the user consents to eliminate a fraction of tags $\sigma \geq \sigma'_{\text{crit}} \approx 0.74$, s becomes the uniform distribution across the active subcategories and, hence, $\mathrm{H}(s)$ attains its maximum value, $\log_2 49$. This effect is also highlighted in Fig. 7, where we represent the privacy-suppression tradeoff of this particular user. In short, the results shown in these two figures confirm our assumption on the existence of an (equivalent) critical suppression rate beyond which the privacy-suppression function achieves its maximum value. Incidentally, we observe that the tradeoff between privacy and tag suppression rate is concave.

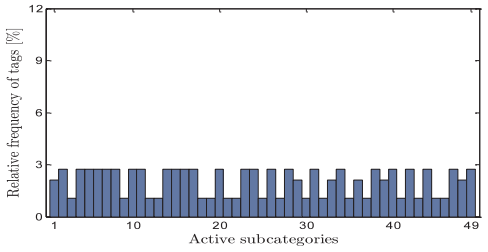
In addition, we plot in Fig. 8 an example of suppression strategy in the case when $\sigma = \sigma'_{\text{crit}}$. In this figure, we

1. This particular user is identified by the string `674f779ba3b445937fd9876054a6e` in [37].

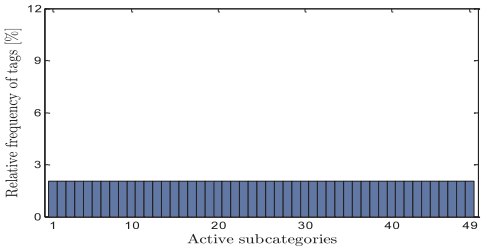


(a) $\sigma = 0.00$, $\mathrm{H}(s) \simeq 5.0254$.

(b) $\sigma = 0.25$, $\mathrm{H}(s) \simeq 5.3182$.



(c) $\sigma = 0.50$, $\mathrm{H}(s) \simeq 5.5000$.



(d) $\sigma = 0.75$, $\mathrm{H}(s) \simeq 5.6147$.

Fig. 6. We represent the apparent profile of a particular user, that is, the perturbed profile resulting from the suppression of tags and observed from the outside. We only show the active subcategories of this profile, i.e., those subcategories tagged by the user. In this particular case, the user posted 190 tags belonging to 49 subcategories. As expected, we observe that as σ increases, s approaches u and $\mathrm{H}(s)$ tends to $\log_2 49 \simeq 5.6147$. When $\sigma = 0$, the apparent profile is plotted in gray to emphasize that this profile is actually the genuine profile. This is consistent with Fig. 8.

superimpose the optimal suppression strategy on the genuine user profile q, to reflect the proportion of tags that the user should eliminate from each subcategory of q to become the uniform distribution. Noteworthy is the fact that $r_i = q_i - \min_j q_j$ for any active subcategory i. Lastly,

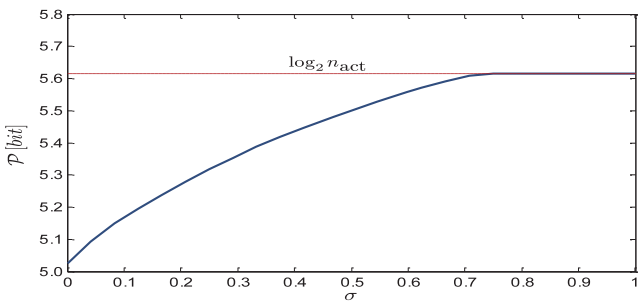


Fig. 7. Our PET poses a tradeoff between privacy and tag suppression rate. This is illustrated here, where we plot function (2) for the particular user considered in Section 6.3.1. Consistently with Section 4.3, we observe that when $\sigma \geq \sigma'_{\mathrm{crit}} \approx 0.74$, the function achieves its maximum value $\mathcal{P}'_{\mathrm{crit}}$, which is given by the number of active subcategories n_{act}.

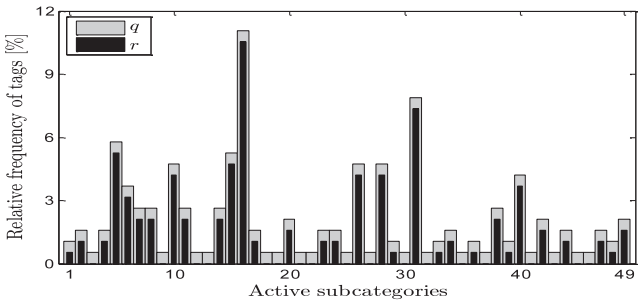


Fig. 8. In this figure, we represent the genuine user profile q of the particular user considered in Section 6.3.1. In addition, we plot the suppression strategy r solving the optimization problem (2) in the special case when $\sigma = \sigma'_{\mathrm{crit}} \approx 0.74$. As explained in Section 4.3, this suppression strategy r gives us the fraction of tags that the user should refrain from tagging in each subcategory to approach u.

Fig. 9 shows the privacy protection that users achieve as a result of the suppression of tags. More accurately, we consider the case when all users in our data set have adhered to tag suppression and use the same suppression rate. Under these assumptions, we plot the percentile curves (10th, 50th, and 90th) of relative privacy gain.

In closing, the results shown in this section illustrate how our mechanism perturbs the user profile observed from the outside and how this perturbation enables users to protect their privacy to a certain degree.

6.3.2 Data Utility

As we have just seen, our approach helps users protect their privacy. Nevertheless, as in any perturbative mechanism, this protection comes at the expense of a loss in data utility. In this section, we assess quantitatively the degradation in data utility caused by our privacy-protecting mechanism.

In our previous work on tag suppression [8], we used a preliminary, simplified measure of loss in data utility, namely the tag suppression rate. In this work, we do evaluate the impact that suppression has on utility, by

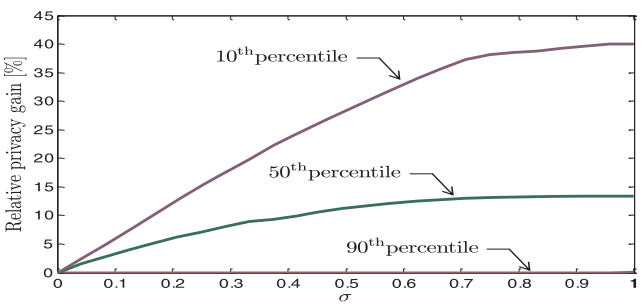


Fig. 9. As described in Section 6.3.1, we consider the case when all users in our data set protect their privacy by using a common tag suppression rate. Built on this premise, we then plot some percentiles curves of privacy gain against this common suppression rate.

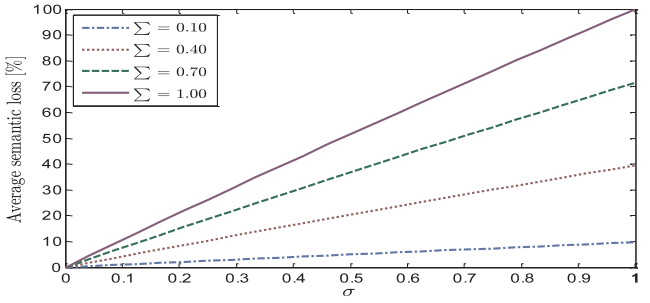


Fig. 10. We plot the loss in semantic functionality as a function of the tag suppression rate. Interestingly, we observe that, regardless of the fraction of users eliminating tags, the semantic loss exhibits a linear behavior with the suppression rate.

considering a more sophisticated metric—the percentage of tags that each bookmark loses as a result of the elimination of tags. Throughout this section, we shall also refer to data utility as *semantic functionality* to highlight that tags make bookmarks meaningful.

In our experiments, the set of tags that users assign to a particular bookmark is referred to as the *bookmark profile* and is modeled exactly as we do with user profiles, that is, as a normalized histogram of these tags across the $n = 200$ subcategories mentioned in Section 6.2. In addition to this characterization, we contemplate a fraction Σ of the user population suppressing tags with a common suppression rate, and assume that the remaining users do not eliminate their tags. To calculate the utility loss experienced by every bookmark in our data set, we solve numerically the optimization problem (2) for every user suppressing tags.[2] Then, the resulting suppression strategies are applied to the specific bookmarks tagged by these users. Fig. 10 shows the semantic loss averaged for all bookmarks. The results clearly indicate that the average semantic loss is roughly linear with the common suppression rate. Specifically, we appreciate that our measure of utility is given approximately by the multiplication of σ and Σ.

Fig. 11 provides more extensive results with regard to semantic loss, but in the form of histograms of relative frequencies. In particular, this figure depicts the percentage of bookmarks affected by a given semantic loss, for $\sigma = 0.25, 0.50, 0.75, 0.99$ and the worst-case scenario where all users adhered to tag suppression, i.e., $\Sigma = 1.00$. Additionally, Fig. 12 plots the curves of semantic loss for different values of Σ. More accurately, we depict a curve for the percentage of bookmarks with a 10 percent loss in the number of tags with respect to the case without suppression, and similarly for $20, 30, \ldots, 100$ percent, where 100 percent refers to completely untagged bookmarks.

In summary, this section has examined the extent to which the application of tag suppression affects data utility, in terms of percentages of missing tags on bookmarks, depending on the fraction of users suppressing tags and a common suppression rate. Remarkably enough, the results show that the average semantic loss is approximately linear with the common suppression rate.

2. This section evaluates our approach by using a data set with 8,882 profiles across 200 subcategories. The computation of the optimal suppression strategy for each of those profiles took a maximum of 0.681 seconds on Windows 7 SP1, on an Intel Xeon 5650 CPU at 2.66 GHz.

6.3.3 Accuracy in Content Filtering

In this section, we quantitatively evaluate the degradation in the classification of web content due to the suppression of tags. Specifically, this section measures the loss in accuracy in the parental control scenario described in Section 5. Throughout this section, we shall resort to the example of web filter referred to as "Example 2," which classifies resources on the web into two states, "granted" or "denied."

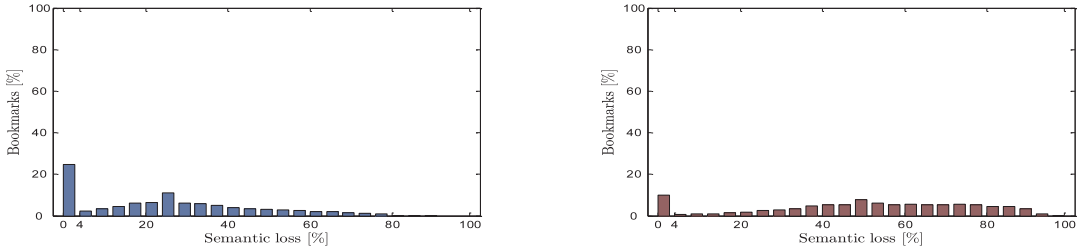
Recall that our web filter first retrieves the profile of the webpage to be accessed, which we model as a normalized histogram of tags across the set of subcategories described in Section 6.2, and second checks whether certain subcategories of this profile exceed a particular threshold. The subcategories of our example are "entertainment for children" and "entertainment for adults," identified, after the categorization process, as the subcategories 62 and 68, respectively.[3] The threshold values for these subcategories are $t_{62} = 60\%$ and $t_{68} = 10\%$. That said, suppose w is the profile of a webpage and that w_{62} and w_{68} are the components of this profile, corresponding to the aforementioned subcategories. According to Section 5, the operation of the filter is as follows: if $w_{68} < t_{68}$ and $w_{62} \geq t_{62}$, then that resource is classified as granted; otherwise, the access to the webpage is denied.

Having reviewed how the parental control filter works, in this series of experiments we shall assume that this filter is installed by default in the users' web browser. In other words, we shall suppose that all users specify the same policies for parental control, which may describe a fairly realistic scenario, as most users do not change default settings [42]. Moreover, we shall assume that the filter works perfectly when tag suppression is not applied. When users skip tagging some resources, however, this filter may classify them incorrectly. In this regard, we shall refer to the *initial* state and the *final* state of a resource as the states before and after the suppression of tags, respectively.

To quantify the loss in the accuracy of this filter, we contemplate the following measures of utility: the number of false negatives and false positives, precision, and recall. In our scenario, a *false negative* is defined as a resource that changes from the initial state denied to the final state granted, as a consequence of tag suppression. To illustrate this case, consider Alice enables our web filter for her son Bob. Suppose that, at some point, Bob wishes to access a webpage with profile w, and components $w_{62} = 50\%$ and $w_{68} = 10\%$. According to the operation of the filter, the access to this resource would be blocked. Nevertheless, after the suppression of tags by other users, it could be possible that this webpage experienced a reduction in the percentage of tags such that $w_{68} < t_{68}$. Due to the fact that we are dealing with relative frequencies, this reduction could cause that $w_{62} \geq t_{62}$, and therefore, Bob would be able to access said resource. Should this be the case, we would classify this webpage as a false negative.

Having described the case of a false negative, next we contemplate the other three possible combinations for the al and final states. Specifically, we define a *true negative* as a resource whose access is granted before and after the suppression of tags. Similarly, a *false positive* denotes a

3. The list with the 200 subcategories resulting from our hierarchical clustering may be downloaded at http://hdl.handle.net/2117/16623.



(a) $\sigma = 0.25$, $\Sigma = 1.00$. (b) $\sigma = 0.50$, $\Sigma = 1.00$.

(c) $\sigma = 0.75$, $\Sigma = 1.00$. (d) $\sigma = 0.99$, $\Sigma = 1.00$.

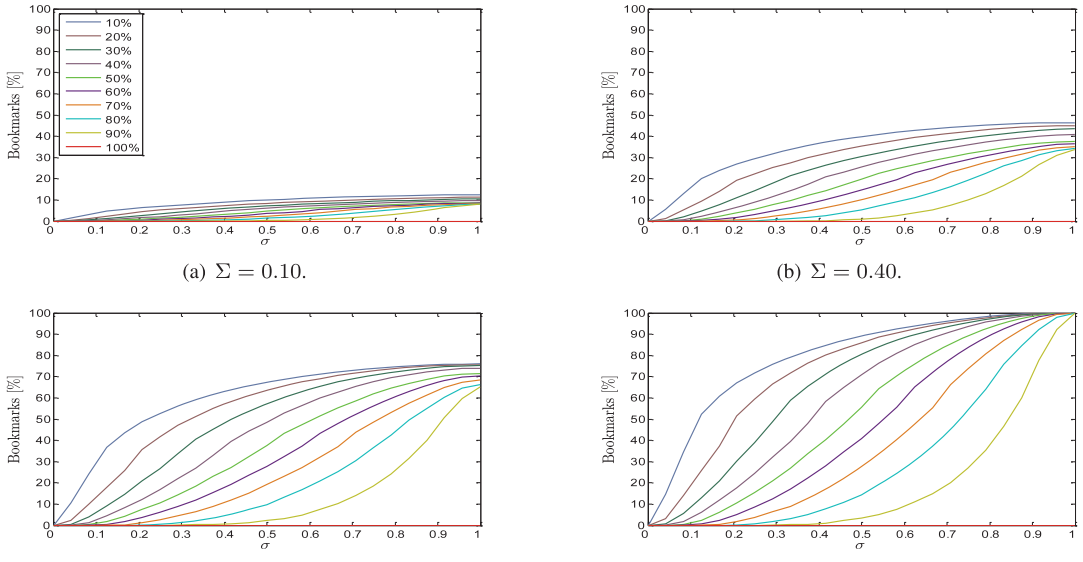
Fig. 11. Loss in semantic functionality in the case when all users apply tag suppression. Not entirely unexpectedly, when these users eliminate almost all their tags, we observe that nearly all resources are affected by a semantic loss between 96 and 100 percent.

resource passing from the initial state granted to the final state denied. And finally, a *true positive* corresponds to a resource that is blocked before and after tag suppression.

Note that all bookmarks in our data set belong to one of these four cases—every resource is classified as denied or granted before (initial state) and after (final state) our technique is applied, which means they necessarily fall into one of the cases mentioned above. However, among these cases, false negatives are clearly the most sensitive in the scenario of parental control, as described in our example. On the other hand, false positives are less critical, even if important, because they represent resources that should be granted but are blocked due to tag suppression. Thus, false positives could be considered as an availability problem rather than a disclosure of potentially dangerous content.

We shall refer to fn, tn, fp, and tp as the number of false negatives, true negatives, false positives, and true positives. According to this notation, *precision* may be defined as $\frac{tp}{tp+fp}$ and *recall* as $\frac{tp}{tp+fn}$. These two measures may be interpreted in probabilistic terms—precision may be regarded as the probability that a resource with final state denied has been classified correctly; and recall as the probability that a resource is classified correctly, given that its initial state is denied.

The experimental results are shown in Figs. 13, 14, 15, and 16, in the special case when all users eliminate tags, i.e., $\Sigma = 1.00$. In these figures, we test the web filter described at the beginning of this section, specified more formally in Section 5. However, to enrich our analysis, we also include two slight variations of this (original) filter. Particularly, we



(a) $\Sigma = 0.10$. (b) $\Sigma = 0.40$.

(c) $\Sigma = 0.70$. (d) $\Sigma = 1.00$.

Fig. 12. Curves of semantic loss showing the percentages of bookmarks that experienced a $10, 20, \ldots, 100$ percent loss in the number of tags, for distinct fractions of the population suppressing tags Σ. The 100 percent curve of semantic loss refers to those bookmarks that lost all their tags.

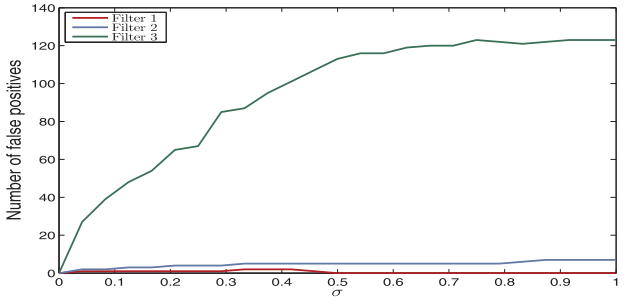


Fig. 13. A false positive represents a resource that changes from the initial state granted to the final state denied, due to the suppression of tags. In this figure, we observe that the most permissive filter (filter 3) exhibits much more false positives than the other two filters.

contemplate different values for the thresholds t_{62} and t_{68}. Accordingly, in our experiments we refer to the original filter as *filter 2*. A more restrictive version of this filter is *filter 1*, whereas *filter 3* is more permissive. Next, we summarize the set of filters used in our evaluation:

- *filter 1*, with $t_{62} = 75\%$ and $t_{68} = 5\%$,
- *filter 2*, with $t_{62} = 60\%$ and $t_{68} = 10\%$,
- *filter 3*, with $t_{62} = 45\%$ and $t_{68} = 15\%$.

Fig. 13 shows the number of false positives. As can be observed, the maximum number of cases is around 120 for the least restrictive filter. Since the total number of resources is 310,923, the number of false positives only represents 0.04 percent of all cases. The differences in terms of false positives between filter 3 on the one hand and filters 1 and 2 on the other are due to the nature of the resources granted by those filters. In particular, before the suppression of tags, 99 percent of the resources classified as granted by filter 1 have a distribution of tags such that *all* tags are concentrated on the subcategory "entertainment for children." In other words, the profile of each of those webpages has only one positive component, namely the component 62. As a consequence of this fact, the profile of those resources will remain exactly the same no matter which suppression rate is applied. Recall that profiles are *relative* histograms of tags and that our suppression approach simply subtracts tags from positive components. Therefore, after the suppression of tags, almost none of those resources will be blocked and, consequently, they will not be considered as false positives. This is the reason why the number of false positives is so low in the case of filter 3.

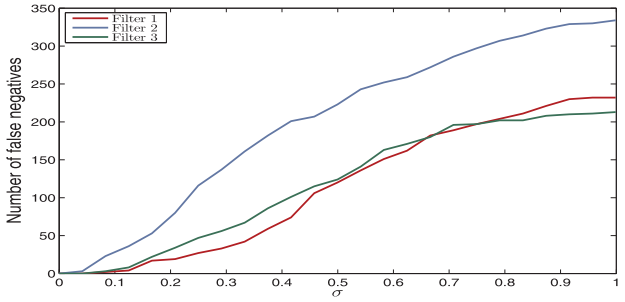


Fig. 14. A false negative refers to a web resource whose access is denied before tag suppression, but after the elimination of tags, the access to this resource is granted.

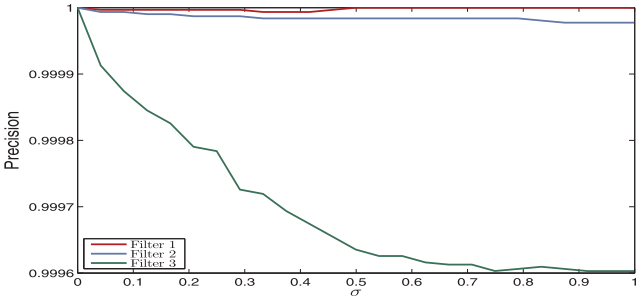


Fig. 15. Precision may be interpreted as the probability that a resource has been classified correctly, given the fact that it was considered denied after the suppression of tags. While at first glance it may seem that there is a great difference, in terms of precision, between filters 1 and 2 on the one hand, and filter 3 on the other, it should be noted that suppression has a negligible effect on the precision of any of the three filters.

The above reasoning also applies to filter 2, where, before tag suppression, 94 percent of the resources granted have a profile with 100 percent of their tags in the subcategory 62. But this is not the case of filter 3—this particular distribution of tags is only observed in 54 percent of the resources classified as granted. As a result, we notice a greater number of resources blocked after the suppression of tags, and therefore, a larger number of false positives, as shown in Fig. 13.

The number of false negatives is plotted in Fig. 14. Here we observe that the maximum number of cases is around 340, which accounts for 0.11 percent of all cases. In Fig. 15, we appreciate that precision is practically unaffected by the suppression of tags. The differences between filter 3 on the one hand, and filters 1 and 2 on the other, are essentially due to the larger number of false positives observed in filter 3, an effect that we examined above. Similarly, Fig. 16 shows that recall is reduced only by a 0.11 percent in the worst-case scenario, corresponding to filter 2.

In closing, these results indicate that tag suppression does not have a significant impact on the accuracy of a parental control filter. Further, because the scenario of resource recommendation described in Section 5 is more tolerant to false negatives than the scenario analyzed in these experiments, we may extend the above results to the former scenario and then assert that our technique would have a similar impact on the accuracy of the recommendations.

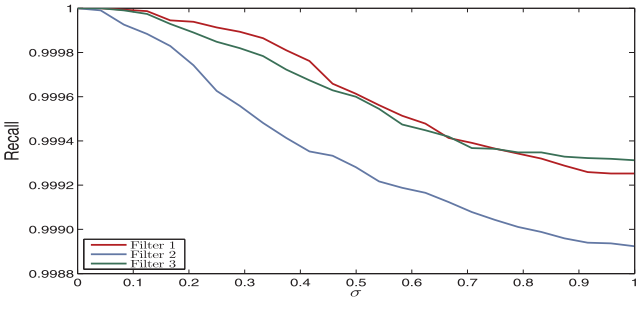


Fig. 16. Recall lends itself to be interpreted in probabilistic terms. In particular, it may be regarded as the probability that a resource with initial state denied has been classified correctly. In this figure, we observe how tag suppression decreases this probability, but only to an insignificant extent.

7 Conclusions and Future Work

Collaborative tagging is currently an extremely popular online service. Although nowadays it is basically used to support resource search and browsing, its potential is still to be exploited. One of these potential applications is the provision of web access functionalities such as content filtering and discovery. For this to become a reality, however, it would be necessary to extend the architecture of current collaborative tagging services so as to include a policy layer that supports the enforcement of user preferences.

On the other hand, as collaborative tagging has been gaining popularity, it has become more evident the need for privacy protection; not only because tags are sensitive information per se, but also because of the risk of cross referencing. In a nutshell, collaborative tagging would also benefit from a layer helping users protect their privacy.

Motivated by all this, our first contribution is an architecture that incorporates two layers on support of enhanced and private collaborative tagging. More specifically, the proposed architecture consists of a bookmarking service and two additional services built on it. The former service enables users to specify policies both to block undesired web content and to denote resources of interest. The latter implements tag suppression, a privacy-preserving technology based on data perturbation. The combination of these two services allows us then to broaden the functionality of collaborative tagging systems and, at the same time, provide users with a mechanism to preserve their privacy while tagging. However, the fact that our PET comes at the cost of data utility poses a tradeoff between privacy on the one hand, and on the other hand, the effectiveness of the enhanced collaborative tagging services enabled by said policy layer.

Our second contribution is an extensive performance evaluation of this architecture, showing its effectiveness in terms of privacy guarantees, data utility, and filtering capabilities for two key scenarios, for example, parental control and resource recommendation. Since we are not aware of similar experimental studies, we believe that what reported in this paper can be useful to evaluate further future developments in the area. Future work includes the development of a full prototype for the experimented system and its testing and use in further scenarios.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments. This work was partly supported by the Spanish Government through projects Consolider Ingenio 2010 CSD2007-00004 "ARES," TEC2010-20572-C02-02 "Consequence" and by the Government of Catalonia under grant 2009 SGR 1362.

References

[1] P. Mika, "Ontologies Are Us: A Unified Model of Social Networks and Semantics," *Proc. Int'l Semantic Web Conf. (ISWC '05)*, Y. Gil, E. Motta, V. Benjamins, and M. Musen, eds., pp. 522-536, 2005.

[2] X. Wu, L. Zhang, and Y. Yu, "Exploring Social Annotations for the Semantic Web," *Proc. 15th Int'l World Wide Web Conf. (WWW)*, pp. 417-426, 2006.

[3] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and S. Gerd, "Evaluating Similarity Measures for Emergent Semantics of Social Tagging," *Proc. 18th Int'l Conf. World Wide Web (WWW)*, pp. 641-650, 2009.

[4] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, to Read," *Proc. 17th Conf. Hypertext and Hypermedia (HYPERTEXT)*, pp. 31-40, 2006.

[5] B. Carminati, E. Ferrari, and A. Perego, "Combining Social Networks and Semantic Web Technologies for Personalizing Web Access," *Proc. Fourth Int'l Conf. Collaborative Computing: Networking, Applications and Worksharing*, pp. 126-144, 2008.

[6] R. Gross and A. Acquisti, "Information Revelation and Privacy in Online Social Networks," *Proc. ACM Workshop Privacy Electronic Soc. (WPES)*, pp. 71-80, 2005.

[7] S.B. Barnes, "A Privacy Paradox: Social Networking in the United States," *First Monday*, vol. 11, no. 9, Sept. 2006.

[8] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "A Privacy-Preserving Architecture for the Semantic Web Based on Tag Suppression," *Proc. Seventh Int'l Conf. Trust, Privacy, Security, Digital Business (TrustBus)*, pp. 58-68, Aug. 2010.

[9] J. Voß, "Tagging, Folksonomy & Co - Renaissance of Manual Indexing?" *Computer Research Repository*, vol. abs/cs/0701072, 2007.

[10] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Trans. Knowledge Data Eng.*, vol. 17, no. 6, pp. 734-749, June 2005.

[11] P. Heymann, D. Ramage, and H. Garcia-Molina, "Social Tag Prediction," *Proc. 31st Ann. Int'l ACM SIGIR Conf. Research Development Information Retrieval*, pp. 531-538, 2008.

[12] E. Frías-Martinez, M. Cebrián, and A. Jaimes, "A Study on the Granularity of User Modeling for Tag Prediction," *Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence Intelligent Agent Technology (WIIAT)*, pp. 828-831, 2008.

[13] Z. Yun and F. Boqin, "Tag-Based User Modeling Using Formal Concept Analysis," *Proc. IEEE Eighth Int'l Conf. Computer Information Technology (CIT)*, pp. 485-490, 2008.

[14] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke, "Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering," *Proc. ACM Conf. Recommender Systems (RecSys)*, pp. 259-266, 2008.

[15] M. Bundschus, S. Yu, V. Tresp, A. Rettinger, M. Dejori, and H.-P. Kriegel, "Hierarchical Bayesian Models for Collaborative Tagging Systems," *Proc. IEEE Int'l Conf. Data Mining (ICDM)*, pp. 728-733, 2009.

[16] X. Li, C.G.M. Snoek, and M. Worring, "Learning Social Tag Relevance by Neighbor Voting," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1310-1322, Nov. 2009.

[17] S. Marti and H. Garcia-Molina, "Taxonomy of Trust: Categorizing P2P Reputation Systems," *Computer Networks*, vol. 50, pp. 472-484, Mar. 2006.

[18] K. Bischoff, C.S. Firan, W. Nejdl, and R. Paiu, "Can All Tags Be Used for Search?" *Proc. 17th ACM Conf. Information and Knowledge Management (CIKM)*, pp. 193-202, 2008.

[19] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Can Social Bookmarking Improve Web Search?" *Proc. Int'l Conf. Web Search Data Mining (WSDM)*, pp. 195-206, 2008.

[20] J. Golbeck, "Combining Provenance with Trust in Social Networks for Semantic Web Content Filtering," *Proc. Int'l Conf. Provenance and Annotation of Data*, pp. 101-108, 2006.

[21] H. Polat and W. Du, "Privacy-Preserving Collaborative Filtering Using Randomized Perturbation Techniques," *Proc. SIAM Int'l Conf. Data Mining (SDM)*, 2003.

[22] H. Polat and W. Du, "SVD-Based Collaborative Filtering with Privacy," *Proc. ACM Int'l Symp. Applied Computing (SASC)*, pp. 791-795, 2005.

[23] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," *Proc. IEEE Int'l Conf. Data Mining (ICDM)*, pp. 99-106, 2003.

[24] Z. Huang, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," *Proc. ACM SIGMOD Int'l Conf. Management Data*, pp. 37-48, 2005.

[25] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, second ed. Wiley, 2006.

[26] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, "Coprivate Query Profile Obfuscation by Means of Optimal Query Exchange between Users," *IEEE Trans. Dependable and Secure Computing,* vol. 9, no. 5, pp. 641-654, Sept.-Oct. 2012.
[27] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "A Privacy-Protecting Architecture for Collaborative Filtering via Forgery and Suppression of Ratings," *Proc. Int'l Workshop Data Privacy Management, Autonomous Spontaneus Security (DPM),* pp. 42-57, Sept. 2011.
[28] D. Rebollo-Monedero and J. Forné, "Optimal Query Forgery for Private Information Retrieval," *IEEE Trans. Information Theory,* vol. 56, no. 9, pp. 4631-4642, Sept. 2010.
[29] D. Rebollo-Monedero, J. Parra-Arnau, and J. Forné, "An Information-Theoretic Privacy Criterion for Query Forgery in Information Retrieval," *Proc. Int'l Conf. Security Technology (SecTech),* pp. 146-154, Dec. 2011.
[30] E.T. Jaynes, "On the Rationale of Maximum-Entropy Methods," *Proc. IEEE,* vol. 70, no. 9, pp. 939-952, Sept. 1982.
[31] C.E. Shannon, "Communication Theory of Secrecy Systems," *Bell Systems Technical J.,* vol. 28, pp. 656-715, 1949.
[32] A. Wyner, "The Wiretap Channel," *Bell Systems Technical J.,* vol. 54, pp. 1355-1367, 1975.
[33] C. Díaz, S. Seys, J. Claessens, and B. Preneel, "Towards Measuring Anonymity," *Proc. Workshop Priv. Enhanc. Technology (PET),* pp. 54-68, Apr. 2002.
[34] C. Díaz, "Anonymity and Privacy in Electronic Services," PhD dissertation, Katholieke Univ. Leuven, Dec. 2005.
[35] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge Univ. Press, 2004.
[36] E. Ferrari and B. Thuraisingham, "Secure Database Systems," *Advanced Database Technology and Design,* M. Piattini and O. Diaz, eds., ch. 11, pp. 353-403, Artech House, Inc., 2000.
[37] http://www.dai-labor.de/en/competence_centers/irml/datasets/. 2013.
[38] S.P. Lloyd, "Least Squares Quantization in PCM," *IEEE Trans. Information Theory,* vol. IT-28, no. 2, pp. 129-137, Mar. 1982.
[39] R.H. Byrd, J.C. Gilbert, and J. Nocedal, "A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming," *Math. Programming,* vol. 89, no. 1, pp. 149-185, 2000.
[40] R.H. Byrd, M.E. Hribar, and J. Nocedal, "An Interior Point Algorithm for Large-Scale Nonlinear Programming," *SIAM J. Optimization,* vol. 9, no. 4, pp. 877-900, 1999.
[41] R.A. Waltz, J.L. Morales, J. Nocedal, and D. Orban, "An Interior Algorithm for Nonlinear Optimization that Combines Line Search and Trust Region Steps," *Math. Programming,* vol. 107, no. 3, pp. 391-408, 2006.
[42] W.E. Mackay, "Triggers and Barriers to Customizing Software," *Proc. SIGCHI Conf. Human Factor Computing Systems,* pp. 153-160, 1991.
[43] M. Grahl, A. Hotho, and G. Stumme, "Conceptual Clustering of Social Bookmarking Sites," *Proc. Int'l Conf. Knowledge Management (I-KNOW),* pp. 356-364, Sept. 2007.
[44] L. Specia and E. Motta, "Integrating Folksonomies with the Semantic Web," *Proc. Int'l Semantic Web Conf.,* pp. 624-639, 2007.
[45] D.S. Hochbaum and D.B. Shmoys, "A Best Possible Heuristic for the k-Center Problem," *Math. Operations Research,* vol. 10, no. 2, pp. 180-184, 1985.
[46] G. Hamerly and C. Elkan, "Alternatives to the k-Means Algorithm that Find Better Clusterings," *Proc. 11th Int'l Conf. Information Knowledge Management (CIKM),* pp. 600-607, 2002.

Javier Parra-Arnau received the MS degree in telecommunications engineering from the Universitat Politècnica de Catalunya (UPC) in 2004. Since 2009, he has been working toward the PhD degree at the Information Security Group of the UPC, where he investigates mathematical models dealing with the tradeoff between privacy and data utility in information systems.

Andrea Perego is a researcher at the Joint Research Centre of the European Commission, located at Ispra (Italy). His research interests include access control, privacy protection, and personalization for the Social Semantic Web, cross-domain and cross-language semantic interoperability, linked open data, ontology design, and Semantic Web technologies for eGovernment data and services. Results of his research work have been published in international journals and conference proceedings.

Elena Ferrari is a full professor of computer science at the University of Insubria, Italy, and a scientific director of the K&SM Research Center. Her research activities include various aspects of data management, including access control, privacy and trust in social networks and the social web, secure cloud computing and emergency management, and secure Semantic Web. In 2009, she received the IEEE Computer Society's Technical Achievement Award for "outstanding and innovative contributions to secure data management." She is a fellow of the IEEE.

Jordi Forné received the MS degree in telecommunications engineering from UPC in 1992, and the PhD degree in 1997. Currently, he is an associate professor at the Telecommunications Engineering School of Barcelona (ETSETB), and is with the Information Security Group.

David Rebollo-Monedero received the MS and PhD degrees in electrical engineering from Stanford University, in 2003 and 2007, respectively. He is a postdoctoral researcher with the Information Security Group of the Department of Telematics Engineering at UPC, where he investigates the application of data compression formalisms to privacy in information systems.