A Utility-Theoretic Approach to Privacy and Personalization

Andreas Krause[*]
Carnegie Mellon University

Eric Horvitz
Microsoft Research

Abstract

Online services such as web search, news portals, and e-commerce applications face the challenge of providing high-quality experiences to a large, heterogeneous user base. Recent efforts have highlighted the potential to improve performance by personalizing services based on special knowledge about users. For example, a user's location, demographics, and search and browsing history may be useful in enhancing the results offered in response to web search queries. However, reasonable concerns about privacy by both users, providers, and government agencies acting on behalf of citizens, may limit access to such information. We introduce and explore an *economics of privacy* in personalization, where people can opt to share personal information in return for enhancements in the quality of an online service. We focus on the example of web search and formulate realistic objective functions for search efficacy and privacy. We demonstrate how we can identify a near-optimal solution to the utility-privacy tradeoff. We evaluate the methodology on data drawn from a log of the search activity of volunteer participants. We separately assess users' preferences about privacy and utility via a large-scale survey, aimed at eliciting preferences about peoples' willingness to trade the sharing of personal data in returns for gains in search efficiency. We show that a significant level of personalization can be achieved using only a small amount of information about users.

Introduction

Information about the preferences, activities, and demographic attributes of people using online applications can be leveraged to personalize the services for individuals and groups of users. For example, knowledge about the locations of users performing web searches can help identify their informational goals. Researchers and organizations have pursued explicit and implicit methods for personalizing online services. For web search, explicit personalization methods rely on users indicating sets of topics of interest that are stored on a server or client. Implicit methods make use of information collected in the absence of user effort and awareness. Data collected implicitly in web search can include users' locations and search activities, capturing such information as how people specify and reformulate queries and click, dwell, and navigate on results. Beyond web search, data collected about users in an implicit manner can be used to custom-tailor the behaviors of a broad spectrum of online applications from informational services like news summarizers to e-commerce services that provide access to online shopping, and that seek to maximize sales with targeted advertising.

The potential value of harnessing data about people to enhance online services coupled with the growing ubiquity of online services raises reasonable concerns about privacy. Both users and the hosts of online applications may benefit from the custom-tailoring of services. However, both may be uncomfortable with the access and use of personal information. There has been increasing discussion about incursions into the privacy of users implied by the general logging and storing of online data (Adar 2007). Beyond general anxieties with sharing personal information, people may more specifically have concerns about becoming increasingly *identifiable*; as increasing amounts of personal data are acquired, users become members of increasingly smaller groups of people associated with the same attributes.

Most work to date on personalizing online services has either ignored the challenges of privacy and focused efforts solely on maximizing utility (*c.f.*, Sugiyama, Hatano, and Ikoma 2004) or has completely bypassed the use of personal data. One vein of research has explored the feasibility of personalizing services with methods that restrict the collection and analysis of personal data to users' own computing devices (Horvitz 2006). Research in this realm includes efforts to personalize web search by making use of the content stored on local machines, as captured within the index of a desktop search service (Teevan, Dumais, and Horvitz 2005; Xu, Zhang, and Wang 2007).

Rather than cutting off opportunities to make personal data available for enhancing online services or limit personalization to client-side analyses, we introduce and study utility-theoretic methods that balance the costs of sharing of personal data with online services in return for the benefits of personalization. Such a decision-theoretic perspective on privacy can allow systems to weigh the benefits of enhancements that come with adaptation with the costs of sensing and storage according to users' preferences.

[*] Andreas Krause performed this research during an internship at Microsoft Research.

We characterize the utility of sharing attributes of private data via value-of-information analyses that take into consideration the preferences to users about the sharing of personal information. We explicitly quantify preferences about utility and privacy and then solve an optimization problem to find the best trade. Our approach is based on two fundamental observations. The first is that, for practical applications, the utility gained with sharing of personal data may often have a diminishing returns property; acquiring more information about a user adds decreasing amounts to utility given what is already known about the user's needs or intentions. On the contrary, the more information that is acquired about a user, the more concerning the breach of privacy becomes. For example, a set of individually non-identifying pieces of information may, when combined, hone down the user to membership in a small group, or even identify an individual. We map the properties of diminishing returns on utility and the concomitant accelerating costs of revelation to the combinatorial concepts of *submodularity* and *supermodularity*, respectively.

Although the economic perspective on privacy is relevant to a wide spectrum of applications, and to studies of the foundations of privacy more broadly, we shall illustrate the concepts in application to personalizing web search. We employ a probabilistic model to predict the website that a searcher is going to visit given the search query and attributes describing the user. We define the utility of a set of personal attributes by the focusing power of the information gained with respect to the prediction task. Similarly, we use the same probabilistic model to quantify the risk of identifying users given a set of personal attributes. We then combine the utility and cost functions into a single objective function, which we use to find a small set of attributes which maximally increases the likelihood of predicting the target website, while making identification of the user as difficult as possible.

The challenges of this optimization are in learning the benefits and costs and grappling with its computational hardness. Solving for the best set of attributes for users to reveal (and hence for the optimal setting of the utility-privacy tradeoff) is an NP-hard search problem, and thus intractable in general for large sets of attributes. We shall demonstrate how we can use the submodularity of the utility and supermodularity of privacy in order to find a *near-optimal* tradeoff efficiently. To our knowledge, no existing approach (such as LeFevre, DeWitt, and Ramakrishnan 2006, Chen, LeFevre, and Ramakrishnan 2007, Hore and R. Jammalamadaka 2007) provides such theoretical guarantees. We evaluate our approach on real-world search log data, as well as from data collected from a user study with over 1400 participants focused on the elicitation of preferences about sharing sensitive information. Our results indicate the existence of prominent "sweet spots" in the utility-privacy tradeoff curve, at which most of the utility can be achieved with the sharing of a minimal amount of private information.

Privacy-Aware Personalization

We consider the challenge of web search personalization as diagnosis under uncertainty. We seek to predict the searcher's information goals, given such noisy clues as query terms and potentially additional attributes that describe users and their interests and activities. We frame the problem probabilistically (*e.g.*, as done by Dou, Song, and Wen 2007 and Downey, Dumais, and Horvitz 2007) by modeling a joint distribution P over random variables, which comprise the target intention X, some request-specific attributes (*e.g.*, the query term) Q, the identity of the user Y, and several attributes $\mathcal{V} = \{V_1, V_2, \ldots, V_n\}$ containing private information. Such attributes include user-specific variables (such as demographic information, search history, word frequencies on the local machine, etc.) and request-specific variables (such as the period of time since an identical query was submitted). We describe the attributes used in this work for the web search context in Table 1. We use statistical techniques to learn such a model P from training data for frequent queries. Then, we present methods for trading off utility and privacy in the context of this probabilistic model.

Utility of accessing private data. Upon receiving a new query Q, and given a subset $\mathcal{A} \subseteq \mathcal{V}$ of the attributes, we can use the probabilistic model to predict the user's target intention by performing inference, computing the conditional distribution $P(X \mid Q, \mathcal{A})$. Then, we use this distribution to inform the decision of, *e.g.*, which search results to present to the user. The hope in personalization is that additional knowledge about the user (*i.e.*, the observed set of attributes $\mathcal{A}$) will help to simplify the prediction task, via reducing the uncertainty in $P(X \mid Q, \mathcal{A})$. Based on this intuition, we quantify the uncertainty in our prediction using the conditional Shannon entropy $H(X \mid Q, \mathcal{A}) = -\sum_{x,q,\mathbf{a}} P(x, q, \mathbf{a}) \log_2 P(x \mid q, \mathbf{a})$. Hence, for any subset $\mathcal{A} \subseteq \mathcal{V}$, we define its utility $U(\mathcal{A})$ to be the *information gain*, i.e., expected entropy reduction achieved by observing $\mathcal{A}$: $U(\mathcal{A}) = H(X \mid Q) - H(X \mid Q, \mathcal{A})$. Click entropy has previously been found effective by Dou, Song, and Wen (2007).

Cost of sharing private data. Several different models of privacy have been proposed in prior work (*c.f.*, Sweeney 2002, Machanavajjhala et al. 2006, Dwork 2006). Our cost function is motivated by the consideration that, *ceteris paribus*, people prefer sets of attributes $\mathcal{A} \subseteq \mathcal{V}$ which make identification of an individual user more difficult. We can consider the observed attributes $\mathcal{A}$ as noisy observations of the (unobserved) identity $Y = y$ of the user. Intuitively, we want to associate high cost $C(\mathcal{A})$ with sets $\mathcal{A}$ which allow accurate prediction of Y given $\mathcal{A}$, and low cost for sets $\mathcal{A}$ for which the conditional distributions $P(Y \mid \mathcal{A})$ are highly uncertain. For a distribution $P(Y)$ over users, we hence define an *identifiability loss* function $L(P(Y))$ which maps probability distributions over users Y to the real numbers. We will choose L such that if there exists a user y with

$P(Y = y)$ close to 1, then the loss $L(P(Y))$ is high. If $P(Y)$ is the uniform distribution, then $L(P(Y))$ should be low.

One possible choice for the loss L is the negative entropy of the distribution $P(Y)$, as used to quantify utility. However, in the context of privacy, this choice is rather poor: Consider a case where we want to quantify the identifiability cost of learning the searcher's gender. Assuming an equal distribution of the gender, learning the gender of the searcher would roughly halve the space of possible searchers, hence increasing the entropy loss by roughly 1. However, this increase is *independent* of whether we start with (A) one billion or (B) only two searchers. In contrast to the influence on utility, where halving the search space of pages to consider is a very large gain independently of how many pages we start with (Dou, Song, and Wen 2007), this diminishment of privacy is enormous: In case (A), an adversary trying to identify the searcher based on knowing their gender has almost no chance of success, whereas in case (B) they would always identify the person. Motivated by this consideration, we represent the privacy cost in our experiments as the *maxprob* loss (Chen, LeFevre, and Ramakrishnan 2007), $L_m(P(Y)) = \max_y P(y)$. Other losses, *e.g.*, based on k-anonymity (Sweeney 2002) are possible as well (see Lebanon et al. 2006 for a justification of using decision theory to quantify privacy and how inferential attacks through side information can be handled). Based on the loss function, we define the *identifiability cost* $I(\mathcal{A})$ as the expected loss of the conditional distributions $P(Y \mid \mathcal{A} = \mathbf{a})$, where the expectation is taken over the observations $\mathcal{A} = \mathbf{a}$.

We also introduce an additive cost component $S(\mathcal{A}) = \sum_{a \in \mathcal{A}} s(a)$, where $s(a) \geq 0$ models the subjective *sensitivity* of attribute a, and other additive costs, such as data acquisition cost, etc. The final cost function $C(\mathcal{A})$ is a convex combination of the identifiability cost $I(\mathcal{A})$ and sensitivity $S(\mathcal{A})$, *i.e.*, $C(\mathcal{A}) = \rho I(\mathcal{A}) + (1 - \rho) S(\mathcal{A})$.

Optimizing the Utility-Privacy Tradeoff

Previously, we described how we can quantify the utility $U(\mathcal{A})$ for any given set of attributes $\mathcal{A}$, and its associated privacy cost $C(\mathcal{A})$. Our goal is to find a set $\mathcal{A}$, that maximizes $U(\mathcal{A})$ while keeping $C(\mathcal{A})$ as small as possible. In order to solve for this tradeoff, we use scalarization (*c.f.*, Boyd and Vandenberghe 2004), by defining a new, scalar objective $F_\lambda(\mathcal{A}) = U(\mathcal{A}) - \lambda C(\mathcal{A})$. Hereby, λ can be considered a privacy-to-utility conversion factor. The goal is to solve the following optimization problem:

$$\mathcal{A}^*_\lambda = \operatorname*{argmax}_{\mathcal{A}} F_\lambda(\mathcal{A}) \tag{1}$$

By varying λ, we can find different solutions $\mathcal{A}^*_\lambda$. Choosing a small λ, leads to solutions with higher utility and higher cost, while selecting large values of λ will lead to solutions with lower utility, but also lower privacy cost. If the set of attributes $\mathcal{V}$ is large, then solving (1) poses a difficult search problem; the number of subsets $\mathcal{A}$ grows exponentially in the size of $\mathcal{V}$. It can be shown that the solution to this problem is even hard to approximate:

Theorem 1. *If there is a constant $\alpha > (1 - 1/e)$ and there exists an algorithm which is guaranteed to find a set $\mathcal{A}'$ such that $F_1(\mathcal{A}') \geq \alpha \max_{\mathcal{A}} F_1(\mathcal{A})$, then $P = NP$.*

The proofs of all theorems are presented in (Krause and Horvitz 2007). Given the complexity, we cannot expect to find a solution $\mathcal{A}^*$ efficiently which achieves even slightly more than $(1 - 1/e) \approx 63\%$ of the optimal score. However, as we show in the following, we can find a solution which is guaranteed to achieve at least $1/3$ of the optimal score.

Properties of the Utility-Privacy Tradeoff

As mentioned above, we would expect intuitively that the more information we already have about a user (*i.e.*, the larger $|\mathcal{A}|$), the less the observation of a new, previously unobserved, attribute would help with enhancing a service. The combinatorial notion of *submodularity* formalizes this intuition. A set function $G : 2^{\mathcal{V}} \to \mathbb{R}$ mapping subsets $\mathcal{A} \subseteq \mathcal{V}$ into the real numbers is called *submodular* (Nemhauser, Wolsey, and Fisher 1978), if for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$, and $V' \in \mathcal{V} \setminus \mathcal{B}$, it holds that $G(\mathcal{A} \cup \{V'\}) - G(\mathcal{A}) \geq G(\mathcal{B} \cup \{V'\} - G(\mathcal{B})$, i.e., adding V' to a set $\mathcal{A}$ increases G more than adding V' to a superset $\mathcal{B}$ of $\mathcal{A}$. G is called *nondecreasing*, if for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ it holds that $G(\mathcal{A}) \leq G(\mathcal{B})$.

A result described in Krause and Guestrin (2005) shows that, under certain conditional independence conditions, the click entropy reduction is submodular and nondecreasing:

Theorem 2 (Krause and Guestrin 2005)**.** *Assume, the attributes $\mathcal{A}$ are conditionally independent given X. Then $U(\mathcal{A})$ is submodular in $\mathcal{A}$.*

We discussed earlier how we expect the privacy cost to behave differently than the utility with the addition of attributes: Adding a new attribute would likely make a stronger incursion into personal privacy when we already know a great deal about a user, and less if we know little. This "accelerating costs" property naturally corresponds to the combinatorial notion of *supermodularity*: A set function $G : 2^{\mathcal{V}} \to \mathbb{R}$ is called *supermodular* (Nemhauser, Wolsey, and Fisher 1978), if for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$, and $V' \in \mathcal{V} \setminus V$, it holds that $G(\mathcal{A} \cup \{V'\}) - G(\mathcal{A}) \leq G(\mathcal{B} \cup \{V'\} - G(\mathcal{B})$, *i.e.*, adding V' to a large set $\mathcal{B}$ increases G more than adding V' to a subset $\mathcal{A}$ of $\mathcal{B}$.

Theorem 3. *Assume that the attributes $\mathcal{V}$ are marginally independent and that the user Y is completely characterized by the attributes,* i.e., *$Y = (\mathcal{V})$. Then the* maxprob *loss $I(\mathcal{A})$ is supermodular in $\mathcal{A}$.*

Note that the attribute sensitivity $S(\mathcal{A})$ is additive per definition and hence supermodular as well. Thus, as a positive linear combination of supermodular functions, $C(\mathcal{A}) = \rho I(\mathcal{A}) + (1 - \rho) S(\mathcal{A})$ is supermodular in $\mathcal{A}$. In our empirical evaluation, we verify the submodularity of $U(\mathcal{A})$ and supermodularity $C(\mathcal{A})$ even without the assumptions made by Theorem 2 and Theorem 3.

Motivated by the above insights about the combinatorial properties of utility and privacy, in the following we present a general approach for trading off utility and privacy. We assume only that the utility $U(\mathcal{A})$ is a submodular set function, whereas $C(\mathcal{A})$ is a supermodular set function. We define the general utility-privacy tradeoff problem as follows:

Problem 4. *Given a set $\mathcal{V}$ of possible attributes to select, a nondecreasing submodular utility function $U(\mathcal{A})$, a nondecreasing supermodular cost function $C(\mathcal{A})$, and a constant $\lambda \geq 0$, our goal is to find a set $\mathcal{A}^*$ such that*

$$\mathcal{A}^* = \operatorname*{argmax}_{\mathcal{A}} F_\lambda(\mathcal{A}) = \operatorname*{argmax}_{\mathcal{A}} U(\mathcal{A}) - \lambda C(\mathcal{A}) \qquad (2)$$

Since $C(\mathcal{A})$ is supermodular if and only if $-C(\mathcal{A})$ is submodular, and since nonnegative linear combinations of submodular set functions are submodular as well, the scalarized objective $F_\lambda(\mathcal{A}) = U(\mathcal{A}) - \lambda C(\mathcal{A})$ is submodular as well. Hence, problem (2) requires the maximization of a submodular set function.

Optimization Algorithms

As the number of subsets $\mathcal{A} \subseteq \mathcal{V}$ grows exponentially with the size of $\mathcal{V}$, and because of the NP-hardness of Problem (1), we cannot expect to find the optimal solution $\mathcal{A}^*$ efficiently. A fundamental result by Nemhauser, Wolsey, and Fisher (1978) characterizes the performance of the simple greedy algorithm, which starts with the empty set $\mathcal{A} = \emptyset$ and greedily adds the attribute which increases the score the most, i.e., $\mathcal{A} \leftarrow \mathcal{A} \cup \operatorname{argmax}_{V'} F(\mathcal{A} \cup \{V'\})$, until k elements have been selected (where k is a specified constant). It was shown that, if F is nondecreasing, submodular and $F(\emptyset) = 0$, then the greedy solution $\mathcal{A}_G$ satisfies $F(\mathcal{A}_G) \geq (1 - 1/e) \max_{|\mathcal{A}|=k} F(\mathcal{A})$, i.e., the greedy solution is at most a factor of $1 - 1/e$ off from the optimal solution. Although this result would allow us to select a near-optimal set of k private attributes maximizing the utility $U(\mathcal{A})$ (which satisfies the conditions of the result from Nemhauser, Wolsey, and Fisher 1978), it does not apply in the more general case, where our objective $F_\lambda(\mathcal{A})$ is *not* nondecreasing.

The problem of maximizing such *non-monotone* submodular functions has been resolved recently by Feige, Mirrokni, and Vondrak (2007). A local search algorithm, named LS, was proved to guarantee a near-optimal solution $\mathcal{A}_{LS}$, if F is a nonnegative[1] (but not necessarily nondecreasing) submodular function. More formally, for the solution $\mathcal{A}_{LS}$ returned by LS, it holds that $F(\mathcal{A}_{LS}) \geq \left(\frac{1}{3} - \frac{\varepsilon}{n}\right) \max_{\mathcal{A}} F(\mathcal{A})$.

Evaluating utility and cost. To run LS, we need to be able to efficiently evaluate the utility $U(\mathcal{A})$ and cost $C(\mathcal{A})$. In principle, we can compute the objective functions from the empirical distribution of the training data, by explicitly evaluating the sums defining $U(\mathcal{A})$ and $C(\mathcal{A})$. However, this approach is very inefficient—$\Omega(N^2)$ where N is the number of training examples. Instead, we can estimate $U(\mathcal{A})$ and $C(\mathcal{A})$ by Monte Carlo sampling.

The following Lemma by Krause & Guestrin (2005) provides sample complexity bounds for estimating the click entropy reduction $U(\mathcal{A})$.

Lemma 5 (Krause & Guestrin 2005)**.** *For any $\varepsilon > 0$ and $\delta > 0$, we need $\left\lceil \frac{1}{2} \left(\frac{\log_2 |\operatorname{dom}(X)|}{\varepsilon} \right)^2 \log \frac{1}{\delta} \right\rceil$ samples in order to estimate $U(\mathcal{A})$ to absolute error ε with confidence at least $1 - \delta$.*

Hereby, $|\operatorname{dom}(X)|$ is the number of different values X can assume. Using a similar argument, we can also prove sample complexity bounds for calculating the cost $C(\mathcal{A})$ of sets of attributes:

Lemma 6. *For any $\varepsilon > 0$ and $\delta > 0$, we need $\left\lceil \frac{1}{2\varepsilon^2} \log \frac{1}{\delta} \right\rceil$ samples in order to estimate $C(\mathcal{A})$ to absolute error ε with confidence at least $1 - \delta$.*

Lemmas 5 and 6 bound the number of samples required to approximate $U(\mathcal{A})$ and $C(\mathcal{A})$ to arbitrary precision ε, with high probability $1 - \delta$.

We also show, how we can generalize the result from Feige, Mirrokni, and Vondrak (2007) to also hold in the case where utility and cost are estimated only up to small constant error ε. The following theorem summarizes our analysis:

Theorem 7. *If λ such that $F_\lambda(\mathcal{V}) \geq 0$, then* LS*, using sampling to estimate $C(\mathcal{A})$ and $U(\mathcal{A})$, computes a solution $\mathcal{A}_{ELS}$ such that $F_\lambda(\mathcal{A}_{ELS}) \geq \left(\frac{1}{3} - \frac{\varepsilon}{n}\right) \max_{\mathcal{A}} F_\lambda(\mathcal{A}) - n\varepsilon_S$, with probability at least $1 - \delta$. The algorithm uses at most $\mathcal{O}\left(\frac{1}{\varepsilon} n^3 \log n \left(\frac{\log_2(|\operatorname{dom}(X)|)}{\varepsilon_S} \right)^2 \log \frac{1}{\delta n^3} \right)$ samples.*

Finding the optimal solution. While LS allows us to find a near-optimal solution in polynomial time, submodularity of F_λ can also be exploited to find an *optimal* solution in a more informed way, allowing us to bypass an exhaustive search through all exponentially many subsets $\mathcal{A}$. Existing algorithms for optimizing submodular functions include branch and bound search, *e.g.*, in the *data-correcting algorithm* (Goldengorin et al. 1999), as well as mixed-integer programming (Nemhauser and Wolsey 1981). These techniques do not require any assumptions about the positiveness of F_λ.

Experimental Results

We now describe the real-world grounding of the utility-theoretic methods via the acquisition of preferences about privacy and evaluation with a log of search activity.

Survey on Privacy Preferences

Although identifiability is an important part of privacy, people may have different preferences about sharing individual attributes (Olson, Grudin, and Horvitz 2005). We set out to assess preferences about cost and benefits of sharing personal data. Related work has explored the elicitation of private information (*c.f.*, Huberman, Adar, and Fine 2005, Wattal et al. 2005, Hann et al. 2002). We are not familiar with a similar study for the context of web search. Our survey was designed specifically to probe preferences about revealing different attributes of private data in return for increases in the utility of a service. Prior research (Olson, Grudin, and Horvitz 2005) has demonstrated that peoples'

[1] If F takes negative values, then it can be normalized by considering $F'(\mathcal{A}) = F(\mathcal{A}) - F(\mathcal{V})$, which however can impact the approximation guarantees.

willingness to share information depends greatly on the type of information being shared, with whom the information is shared, and how the information is going to be used. In designing the survey, we assessed preferences for sharing in a low-risk situation, where "personal information would be shared and used only with respect to a single specified query, and discarded immediately thereafter." Our survey contained questions both on the sensitivity of individual attributes and on concerns about identifiability. The survey was distributed to multiple divisions within the Microsoft Corporation via an online survey tool. We motivated people to take the survey by giving participants a chance to win a media player via a random drawing. The survey was open to worldwide entries, and we received a total of 1451 responses.

Label	bits	Description
DGDR	1	Gender (*)
DAGE	2	Age group (<18, 18-50, >50) (*)
DOCC	3	Occupation (6 groups of related jobs) (*)
DREG	2	Region (4 geographic regions)
DMTL	1	Marital status (*)
DCHD	1	Whether the searcher has children or not (*)
AQRY	1	Performed same query before
ACLK	1	Visited same website before
AFRQ	1	User performs $\geq$ 1 query/day on average
AZIP	1	User queried from $\geq$ 2 different zip codes
ACTY	1	User queried from $\geq$ 2 different cities
ACRY	1	User queried from $\geq$ 2 different countries
AWHR	1	Current query during working hours
AWDY	1	Current query during workday / weekend
ATLV	2	Top-level of query IP (.com, .net, .org, .edu)
TART	1	User prev. visit. arts rel. webpage
TADT	1	User prev. visit. webpage with adult content
TBUS	1	User prev. visit. business rel. webpage
TCMP	1	User prev. visit. compute rel. webpage
TGMS	1	User prev. visit. games rel. webpage
THEA	1	User prev. visit. health rel. webpage
THOM	1	User prev. visit. home rel. webpage
TKID	1	User prev. visit. kids / teens rel. webpage
TNWS	1	User prev. visit. news rel. webpage
TREC	1	User prev. visit. recreation rel. webpage
TREF	1	User prev. visit. reference rel. webpage
TREG	1	User prev. visit. webpage w. regional content
TSCI	1	User prev. visit. science rel. webpage
TSHP	1	User prev. visit. shopping rel. webpage
TCIN	1	User prev. visit. consumer inform. webpage
TSOC	1	User prev. visit. society rel. webpage
TSPT	1	User prev. visit. sports rel. webpage
TWLD	1	User prev. visit. world rel. webpage

Table 1: 33 Attributes used in our experiments. First letter indicates (D)emographic, (A)ctivity or (T)opic related attributes. Attributes marked (*) were not available in search log data. Total number of bits = 38.

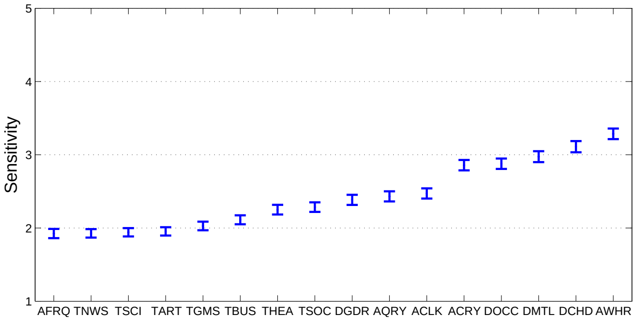
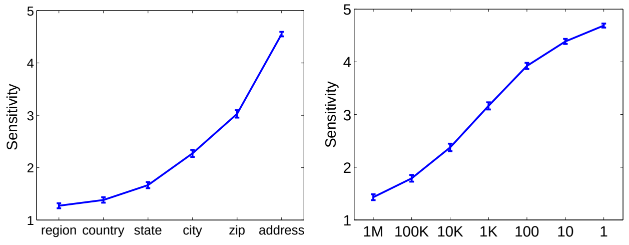


Figure 1: Sensitivity of individual attributes (with 95% confidence intervals).



(a) *Location* (b) *Discriminability*

Figure 2: (a) Sensitivity of sharing location under different levels of discretization. (b) Sensitivity of k-discriminability levels (right). Plots show 95% confidence intervals.

Questions about individual attributes. We first asked the participants to classify the sensitivity of a set of attributes on a Likert scale from 1 (not very sensitive) to 5 (highly sensitive). The order of the questions was randomized. Figure 1 presents the results (see Table 1 for used acronyms). As might be expected, the frequency of search engine usage (AFRQ), as well as very general topic interests, *e.g.*, news pages (TNWS), are considered to be of low sensitivity. We found that preferences for sharing topical interests with a service depended on the topic; participants showed significantly greater sensitivity to sharing interest in health or society related websites (THEA, TSOC) than in news or science-related pages (TNWS, TSCI). The biggest "jump" in sensitivity occurs between attributes ACLK, referring to sharing a repeated visit to same website, and ACRY, referring to having recently traveled internationally. Participants were most sensitive to sharing whether they are at work when performing a query (AWHR).

Questions about identifiability. We also elicited preferences about sharing personal data at different degrees of precision and with different levels of identifiability. First, we sought to identify changes in the sensitivity associated with sharing personal data at increasingly higher resolution. More specifically, we asked participants to assess how sensitive they are to sharing their location at the region, country, state, city, zip code, or address level of precision. Figure 2(a) presents the mean sensitivity with 95% confidence intervals for this experiment. We also asked the participants about how sensitive they would be if, in spite of sharing the information, they would be guaranteed to remain indistinguishable from at least k other people (thereby eliciting preferences about k of k-anonymity). We varied k among 1, 10, 100, 1,000, 10,000, 100,000 and 1

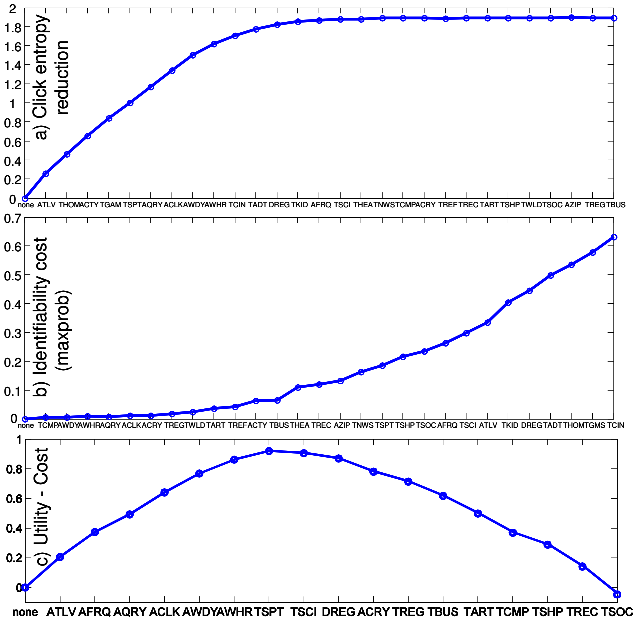


Figure 3: Utility (a), cost (b), and net benefit (c) for an increasing number of attributes selected in a greedy manner.

million. Figure 2(b) presents the results of this experiment. The survey data revealed that the sensitivities of study participants to sharing personal data depend significantly on the granularity of the shared information.

Questions about utility. In addition to assessing the sensitivity of sharing different kinds of personal information, we asked the participants to assess the improvements they would require in a service so as to be willing to share attributes of different sensitivity levels. More specifically, we asked: "How much would a search engine have to improve its performance, such that you would be willing to share information you consider to be of sensitivity x". Hereby, "improvement" was measured in the decrease of time required to obtain the desired result (*e.g.*, 100% improvement means getting results twice as fast). As response options, we offered average improvements by 25%, 50%, 100%, as well as immediately presenting the desired page 95% of the time (which we valued as a speedup by a factor of 4). We also allowed the participant the option of indicating that they would never share information of a specified sensitivity level. These responses, in conjunction with the earlier sensitivity assessments, allowed us to establish sensitivity as a *common currency* of utility and cost. We used this currency in exploring the privacy and utility tradeoff with a large-scale log of web search activity.

Search Log Data and Attributes

Using the preference data collected from the survey, we performed experiments with a log containing 247,684 queries performed by 9,523 users over a period of 14 months between December 2005 and January 2007. The data was obtained from users who had volunteered to participate in a public Microsoft data sharing program that would use information about their search activities to enhance search. The data was filtered to include only those queries which had been performed by at least 30 different users, resulting in a total of 914 different queries. From the search logs, we computed 28 different user / query specific attributes (Table 1). In selecting our attributes, we chose coarse discretizations; no attribute is represented by more than 2 bits, and most attributes are binary.

The first set of attributes contains features extracted from the search history data. For each query, we determine whether the same query has been performed before (AQRY), and whether the searcher has visited the same webpage (ACLK) before. The attribute AFRQ describes whether the user performed at least one query each day. We also log the top-level domain (ATLV), determined by reverse DNS lookup of the query IP address, and used only the domains .net, .com, .org and .edu. In addition, we determined if a user had ever performed queries from at least two different zip codes (AZIP), cities (ACTY) and countries (ACRY), by performing reverse DNS lookup of the query IP addresses. For each query, we also stored whether the query was performed during working hours (AWHR; between 7 am and 6 pm) and during workdays (AWDY; Mon-Fri) or during a weekend (Sat, Sun).

We looked up all websites visited by the user during 2006 in the 16 top-level category of the Open Directory Project directory (www.dmoz.org). For each category, we used a binary attribute indicating whether the user had ever visited a website in the category (acronyms for topics are indicated with prefix T).

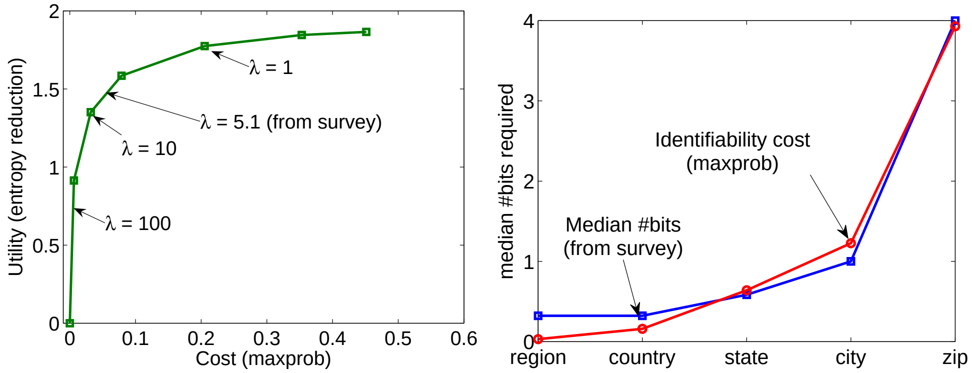
For demographic information, only location was available in the search log data, accessible by via a reverse IP lookup. We discretized the location obtained in this manner into four broad regions (DREG).

Computing Utility and Cost

We evaluated utility and cost based on the empirical distribution of the data. In order to avoid overfitting with sparse data, we applied Dirichlet smoothing. In our experiments, we used 1000 samples in order to estimate $U(\mathcal{A})$ and $I(\mathcal{A})$.

We first used the greedy algorithm to select increasing numbers of attributes, maximizing the utility and ignoring the cost. Figure 3(a) shows the greedy ordering (attributes ATLV, THOM, ACTY, TGAM, TSPT, AQRY, ACLK, AWDY, AWHR, TCIN, TADT, DREG, TKID, AFRQ) and the achieved entropy reductions. The entropy reduction levels off at roughly 1.9 bits. Figure 3 clearly indicates the diminishing-returns property of click entropy reduction.

We also generated a greedy ordering of the attributes, in order of minimum incremental cost. Figure 3(b) presents the results of this experiment, using the maxprob cost metric. As expected, the curve appears convex (apart from small variations based in the sampling process). The cost initially increases very slowly, but the growth accelerates as more attributes are selected. This behavior empirically corroborates the supermodularity assumption for the cost metric.



(a) *Tradeoff-curve* (b) *Calibration*

Figure 4: (a) Tradeoff curve for different λ. (b) Calibrating the tradeoff based on survey data. Best alignment of survey and log data leads to a choice of λ mapping into the sweet spot of (a).

Calibrating the Tradeoff with the Survey

We now employ scalarization as in Eq. (2) to trade off utility of personalization with the cost associated with the sharing of personal data. In order to do this, we need to choose a particular tradeoff parameter λ. Instead of committing to a single value of λ, we use LS to generate solutions for increasing values of λ, and plot their utility and cost. Figure 4(a) shows the tradeoff curve obtained from this experiment. We can see that this curve exhibits a prominent "knee;" for values $1 \leq \lambda \leq 10$, small increases in the utility lead to big increases in cost, and vice versa. Hence, at this knee, one achieves near-maximal utility at near-minimum cost, a finding we found to be encouraging.

To integrate peoples' preferences in the analysis of the tradeoff, we performed the following *calibration* procedure: From the search log data, we determined how increasing the resolution in a peoples' locations would increase the privacy cost. We varied the location granularity from *region* (coarsest) to *zip code* (finest). For example, we computed the values $I(\{\text{zip code}\})$, $I(\{\text{city}\})$, etc. from the data. We compared these values with responses from the survey. As explained above, we had assessed sensitivities of participants about sharing their locations at different levels of precision. Similarly, we asked them to assess the degree of search performance enhancement required for them to share attributes of a given sensitivity. With each level of improvement, we associated a number of bits: A speed up by a factor of x would require $\log_2 x$ bits (*i.e.*, doubling the search performance would require 1 bit, etc.). We then concatenated the mappings from location granularity to sensitivity, and from sensitivity to utility (bits), and computed the median number of bits required for sharing each location granularity.

We performed a linear regression analysis to align the identifiability cost curve estimated from data with the curve obtained from the survey. The least-squares alignment is presented in Figure 4(b), and obtained for a value of $\lambda \approx 5.12$. Note that this value of λ maps exactly into the sweet spot $1 \leq \lambda \leq 10$ of the tradeoff curve of Figure 4(a). Figure 3(c) presents the scalarized net benefit F_λ for greedily chosen attributes after calibration.

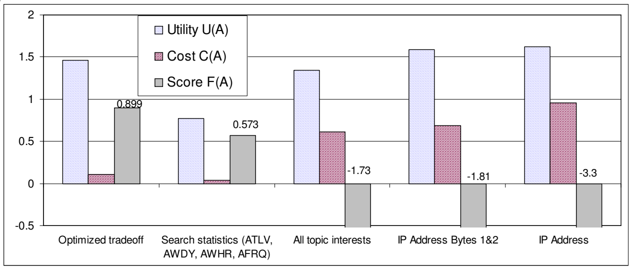


Figure 5: Comparison with heuristics

Optimizing the Utility-Privacy Tradeoff

Based on the calibration, our goal was to find a set of attributes $\mathcal{A}$ maximizing the calibrated objective $F_\lambda(\mathcal{A})$ according to (2). We used LS to approximately solve this optimization problem. The algorithm returns the solution TSPT, AQRY, ATLV, AWHR, AFRQ, AWDY, TGMS, ACLK.

We also compared the optimized solution $\mathcal{A}_{opt}$ to various heuristic solutions. For example, we compared it to the candidate solution $\mathcal{A}_{topic}$ where we select all topic interest attributes (starting with T); $\mathcal{A}_{search}$ including all search statistics (ATLV, AWDY, AWHR, AFRQ); $\mathcal{A}_{IP}$, the entire IP address or $\mathcal{A}_{IP2}$, the first two bytes of the IP address. Figure 5 presents the results of this comparison. The optimized solution $\mathcal{A}_{opt}$ obtains the best score of 0.90, achieving a click entropy reduction of $\approx\ 1.5$. The search statistics $\mathcal{A}_{search}$ performs second best, with a score of 0.57, but achieving a drastically lower utility of only 0.8. Perhaps surprisingly, the collection of topic interests, $\mathcal{A}_{topic}$ results in a negative total score of -1.73, achieving less utility than the optimized solution. We believe that this is because knowledge of the exact topic interest profile frequently suffices to uniquely identify a searcher. As expected, the IP address (even the first two bytes) is quite identifying in this data set, and hence has very high cost. This experiment shows that the optimization problem is non-trivial, and that the optimized solution outperforms the choices made by the heuristic policies.

Summary and Conclusions

We presented a utility-theoretic approach to privacy in online services that takes user preferences into consideration. We focused on the use of the methods to optimize the utility-privacy tradeoff in web search. We showed that utility functions that quantify enhancements of search with the use of personal data satisfy submodularity, a property captured intuitively as diminishing returns with access to increasing quantities of personal data. In contrast, we found that privacy concerns can behave in a supermodular manner; the sensitivity and the risk of identifiability accelerate with additional data. Based on the submodular utility and supermodular cost functions, we demonstrated how we can efficiently find a provably near-optimal utility-privacy tradeoff. We evaluated our methodology on a log of over twelve months of web search data, calibrated with preferences assessed from a survey of over 1400 people. We found that a significant value of personalizing web search can be achieved using only a small amount of information about users.

The insights and methods can be harnessed in a variety of different ways in real-world systems, from their use to guide overall designs of privacy policies to creating specific personalization machinery that executes in real-time in online services. The former includes the use offline analyses to support decisions about the best ways to limit the logging of users' online activities. The latter includes methods that make use of standing profiles of personal data that users are comfortable with sharing and interactive systems that engage users with session-based requests for personal data that promises to best enhance the user's acute experiences, given goals and needs that are inferred in real time.

We believe that the principles and methods employed in the utility-theoretic analysis of tradeoffs for web search have applicability to the personalization of a broad variety of online services. The results underscore the value of taking a decision-theoretic approach to privacy, where we seek to jointly understand the utility of personalization that can be achieved via access to information about users, and the preferences of users about the costs and benefits of selectively sharing their personal data with online services.

References

Adar, E. 2007. User 4xxxxx9: Anonymizing query logs. In *Query Logs Workshop, WWW*.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge UP.

Chen, B.-C.; LeFevre, K.; and Ramakrishnan, R. 2007. Privacy skyline: Privacy with multidimensional adversarial knowledge. In *VLDB*.

Dou, Z.; Song, R.; and Wen, J.-R. 2007. A large-scale evaluation and analysis of personalized search strategies. In *WWW*.

Downey, D.; Dumais, S.; and Horvitz, E. 2007. Models of searching and browsing: Languages, studies, and applications. In *IJCAI*.

Dwork, C. 2006. Differential privacy. In *ICALP*.

Feige, U.; Mirrokni, V.; and Vondrak, J. 2007. Maximizing non-monotone submodular functions. In *FOCS*.

Goldengorin, B.; Sierksma, G.; Tijssen, G. A.; and Tso, M. 1999. The data-correcting algorithm for the minimization of supermodular functions. *Mgmt Sci* 45(11):1539–1551.

Hann, I.; Hui, K.; Lee, T.; and Png, I. 2002. Online-information privacy: Measuring the cost-benefit tradeoff. In *ICIS*.

Hore, B., and R. Jammalamadaka, S. M. 2007. Flexible anonymization for privacy preserving data publishing: A systematic search based approach. In *SDM*.

Horvitz, E. 2006. Machine learning, reasoning, and intelligence in daily life: Directions and challenges. Technical Report TR-2006-185, Microsoft Research.

Huberman, B. A.; Adar, E.; and Fine, L. R. 2005. Valuating privacy. *IEEE Security & Privacy* 3(5):22–25.

Krause, A., and Guestrin, C. 2005. Near-optimal nonmyopic value of information in graphical models. In *UAI*.

Krause, A., and Horvitz, E. 2007. A utility-theoretic approach to handling privacy in online personalization. Technical Report MSR-TR-2007-135, Microsoft Research.

Lebanon, G.; Scannapieco, M.; Fouad, M. R.; and Bertino, E. 2006. Beyond k-anonymity: A decision theoretic framework for assessing privacy risk. *Springer Lecture Notes in Computer Science* 4302:217–232.

LeFevre, K.; DeWitt, D.; and Ramakrishnan, R. 2006. Mondrian multidimensional k-anonymity. In *ICDE*.

Machanavajjhala, A.; Kifer, D.; Gehrke, J.; and Venkitasubramaniam, M. 2006. L-diversity: Privacy beyond k-anonymity. In *ICDE*.

Nemhauser, G. L., and Wolsey, L. A. 1981. *Studies on Graphs and Discrete Programming*. North-Holland. chapter Maximizing Submodular Set Functions: Formulations and Analysis of Algorithms, 279–301.

Nemhauser, G.; Wolsey, L.; and Fisher, M. 1978. An analysis of the approximations for maximizing submodular set functions. *Math. Prog.* 14:265–294.

Olson, J. S.; Grudin, J.; and Horvitz, E. 2005. A study of preferences for sharing and privacy. In *CHI*.

Sugiyama, K.; Hatano, K.; and Ikoma, T. 2004. Adaptive web search based on user profile constructed without any effort from users. In *WWW*.

Sweeney, L. 2002. k-anonymity: a model for protecting privacy. *Int. Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5):557–570.

Teevan, J.; Dumais, S. T.; and Horvitz, E. 2005. Personalizing search via automated analysis of interests and activities. In *SIGIR*.

Wattal, S.; Telang, R.; Mukhopadhyay, T.; and Boatwright, P. 2005. Examining the personalization-privacy tradeoff an empirical investigation with email advertisements. *Mgmt Sci*.

Xu, Y.; Zhang, B.; and Wang, K. 2007. Privacy-enhancing personalized web search. In *WWW*.