Privacy, Personalization, and the Web: A Utility-theoretic Approach

Andreas Krause[1]
SCS, CMU

Eric Horvitz
Microsoft Research

Abstract

Online offerings such as web search face the challenge of providing high-quality service to a large, heterogeneous user base. Recent efforts have highighted the potential to improve performance by introducing methods to personalize services based on special knowledge about users. For example, a user's location, demographics, and past search and browsing may be useful in enhancing the efficiency and accuracy of web search. However, reasonable concerns about privacy by both users and providers limit access by services to such information. We explore the rich space of possibility where people can opt to share, in a standing or a real-time manner, personal information in return for expected enhancements in the quality of an online service. We present methods and studies on addressing such tradeoffs between privacy and utility in online services. We introduce concrete and realistic objective functions for efficacy and privacy and demonstrate how we can efficiently find a provably near-optimal optimization of the utility-privacy tradeoff. We evaluate our methodology on data drawn from a large-scale web search log of people who volunteered to have their logs explored so as to contribute to enhancing search performance. In order to incorporate personal preferences about privacy and utility, and the willingness to trade off revealing some quantity of personal data to a search system in returns for gains in efficiency, we performed a user study with 1400 participants. Employing utility and preferences estimated from the real-world data, we show that a significant level of personalization can be achieved using only a small amount of information about users.

1 Introduction

Information about people searching the web can be used to enhance web search. For example, knowing a searcher's location can help identify their informational goals when confronted with queries like "sports" "movies," or "pizza." Researchers and organizations have pursued explicit and implicit methods for personalizing search. Explicit personalization procedures include such methods as storing sets of topics of interest on a server or client. Richer implicitly mined data can also be employed. Studies have demonstrated how personal data about individual users, such as information captured by the index of desktop search services, and used only in privately-held, local analyses, can be used in order to provide personalization of web search [23, 25]. These techniques have demonstrated the potential of greatly improving the relevance of displayed search results by disambiguating queries based on personal information about the users. On the implicit side, Web search services have relied on the logging of data in order to enhance and audit their performance. Search services have access to great amounts of data about people, both individually and in aggregate, including such attributes as how people specify and reformulate queries and click, dwell, and navigate on results and links over time, and the coarse location of people (available via IP lookup).

Information about people can enhance the accuracy of search engines, but the sensing and storage of such information may also conflict with personal preferences about privacy. Indeed, there has been increasing discussion about potential privacy concerns implied by the general logging and storing of such data by online services [2]. 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 more and more identifiable as they are in increasingly smaller sets of others associated with the same attributes. A fundamental utility-privacy tradeoff exists where the more information that is acquired, the higher the utility via, e.g., personalization, but, at the same time, the greater the privacy concerns.

[1] Work performed during an Internship at Microsoft Research

Previous work either has ignored privacy problems and focused efforts on maximizing utility [21], or has tried to avoid privacy incursion by using no personal data or only using data available on the local machine [23, 25].

We shall explicitly examine the promise of methods that allow for a smooth tradeoff between privacy and the utility of enhanced personalization of online services by taking a decision-theoretic perspective. 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, subsequently, 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 intent. On the contrary, privacy behaves the opposite way; typically, 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 can bring together the properties of diminishing-returns on utility and the concommitant accelerating costs of revelation via the combinatorial concepts of *submodularity* and *supermodularity*, respectively.

We shall apply these concepts for the case of personalized search. We employ a probabilistic model to predict the website that a searcher is going to visit given the search query and the 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 predict, given a set of personal attributes, the users who matches the same attributes. Our privacy objective is chosen to favor sets of attributes that make the prediction of the users as difficult as possible. We then combine our 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.

Unfortunately, solving for the best set of attributes (and hence for the optimal setting of the utility-privacy tradeoff) is NP-hard, and hence an intractable computation for large sets of attributes. We demonstrate how we can use the submodularity of the utility and supermodularity of privacy in order to find a *near-optimal* tradeoff efficiently. The approximation is guaranteed to be close to the optimal solution. To our knowledge no existing approach (such as [16, 12]) has such strong approximation guarantees. We shall evaluate our approach on real-world search log data, and demonstrate 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.

In addition to identifiability considerations, people often have preferences about revealing different attributes of personal data. For example, knowledge of whether a person is interested in adult websites poses very low risk to identifiability, but may nevertheless be considered highly sensitive information by a searcher. On the other hand, knowledge about the searcher's country may be far more identifying, but potentially less sensitive . In order to elicit sensitivity of different demographic attributes and other personal information relevant to web search, we conducted a user study including over 1400 participants. Beyond providing a rich view into preferences about the sharing of private information, the data allowed us to calibrate the utility-privacy tradeoff based on perceived sensitivities.

2 Privacy-aware personalization

We consider the challenge of personalization as diagnosis under uncertainty. We seek to predict a searcher's information goals, given such clues as query terms and potentially additional attributes that describe users and their interests and activities. We frame the problem probabilistically (as done, e.g., by [5, 6] in the search context), 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 $\mathcal{V} = \{V_1, V_2, \ldots, V_m\}$ 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 concrete attributes used in this work for the web search context in Section 5.2. Additional examples can be found in [6] or [23]. 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 private data. Upon receiving a new request Q, and given a subset $\mathcal{A} \subseteq \mathcal{V}$ of the attributes, we can use the probabilistic model to predict the 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 (*c.f.*, [4]) $H(X \mid Q, \mathcal{A}) = -\sum_{x,q,\mathbf{a}} P(x, q, \mathbf{a}) \log_2 P(x \mid q, \mathbf{a})$. 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 been previously studied by [5].

Cost of private data. There is a large amount of work on mathematically modeling privacy (*c.f.*, [1, 22, 17, 7]). Our cost function is motivated by the consideration that sets of attributes $\mathcal{A} \subseteq \mathcal{V}$ should be preferred, which make identification of an individual user as difficult as possible. 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. L is chosen in a way, such that if there exists a user y such that $P(Y = y)$ is close to 1, then the loss $L(P(Y))$ is very large. If $P(Y)$ is the uniform distribution, then $L(P(Y))$ is close to 0. In our experiments we use the *maxprob* loss, $L_m(P(Y)) = \max_y P(y)$. Other losses, e.g., based on k-anonymity [22] are possible as well. 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 $\mathcal{A} = \mathbf{a}$.

In addition to identifiability, we introduce an additional additive cost component $S(\mathcal{A}) = \sum_{a \in \mathcal{A}} s(a)$, where $s(a) \geq 0$ is a nonnegative quantity modeling 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., $C(\mathcal{A}) = \rho I(\mathcal{A}) + (1-\rho) S(\mathcal{A})$.

2.1 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 $\mathcal{A}$, for which $U(\mathcal{A})$ is as large as possible, while keeping $C(\mathcal{A})$ as small as possible. In order to solve for this tradeoff, we use scalarization [3], by defining a new, scalar objective $F_\lambda(\mathcal{A}) = U(\mathcal{A}) - \lambda C(\mathcal{A})$. Hereby, λ can be considered a Langrangean multiplier which controls the privacy-to-utility conversion factor. The goal is to solve the following optimization problem:

$$\mathcal{A}^*_\lambda = \operatorname*{argmax}_{\mathcal{A}} F_\lambda(\mathcal{A}) \tag{2.1}$$

By varying λ, we can find different solutions $\mathcal{A}^*_\lambda$. If we choose a very small λ, we find solutions with higher utility and higher cost; large values of λ will lead to lower utility, but also lower privacy cost.

If the set of attributes $\mathcal{V}$ is large, then (2.1) is a difficult (NP-hard) search problem, as the number of subsets $\mathcal{A}$ grows exponentially in the size of $\mathcal{V}$. Given the complexity, we cannot expect to efficiently find an optimal solution $\mathcal{A}^*$. 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.

3 Theoretical 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. 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* [18], 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 $G(\mathcal{A}) \leq G(\mathcal{B})$.

In [14], it was shown that, under certain conditional independence conditions, the click entropy reduction is submodular and nondecreasing:

Theorem 3.1 ([14]). *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: Adding a new attribute would likely make a stronger incursion into personal privacy when we know a great deal about a user, and less if we know little. This "increasing costs" property naturally corresponds to the combinatorial notion of *supermodularity*: A set function $G : 2^{\mathcal{V}} \to \mathbb{R}$ is called *supermodular* [18], 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 3.2. *Assume, the attributes $\mathcal{V}$ are marginally independent, and the user Y is completely characterized by the attributes, i.e., $Y = (\mathcal{V})$. Then the* maxprob *loss $I_m(\mathcal{A})$ is supermodular in $\mathcal{A}$.*

All proofs are available in [15]. Note that the attribute sensitivity $S(\mathcal{A})$ is per definition additive 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 $\mathcal{A}$, for both choices of $I_i(\mathcal{A})$ or $I_m(\mathcal{A})$. In our empirical evaluation, we verify the submodularity of $U(\mathcal{A})$ and supermodularity $C(\mathcal{A})$ even without the assumptions made by Theorem 3.1 and Theorem 3.2.

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 only assume 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 3.3. *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*

$$\mathcal{A}^* = \operatorname*{argmax}_{\mathcal{A}} F_\lambda(\mathcal{A}) = \operatorname*{argmax}_{\mathcal{A}} U(\mathcal{A}) - \lambda C(\mathcal{A}) \tag{3.1}$$

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 (3.1) requires the maximization of a submodular set function.

4 Optimization Algorithms

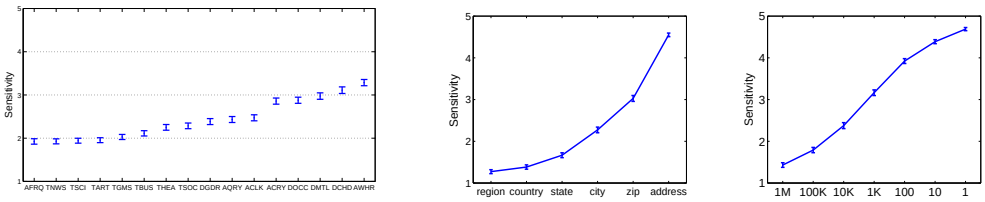
As the number of subsets $\mathcal{A} \subseteq \mathcal{V}$ grows exponentially with the size of $\mathcal{V}$, and because of the NP-hardness of Problem (2.1), we cannot expect to find the optimal solution $\mathcal{A}^*$ efficiently. A fundamental result by Nemhauser et.al. [18] characterized the 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 $1 - 1/e$ away from the optimal solution. While this result would allow to, e.g., select a near-optimal set of k private attributes maximizing the utility $U(\mathcal{A})$ (which satisfies the conditions of the result from [18]), it unfortunately does not apply in our 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 [8]. A local search algorithm, named LS, was proved to guarantee a near-optimal solution $\mathcal{A}_{LS}$, if F is an nonnegative[2] (but not necessarily nondecreasing) submodular function:

1. Let $V^* \leftarrow \operatorname{argmax}_{V' \in \mathcal{V}} F(\{V'\})$ and init. $\mathcal{A} \leftarrow \{V^*\}$
2. If there exists an element $V' \in \mathcal{V} \setminus \mathcal{A}$ such that $F(\mathcal{A} \cup \{V'\}) > (1 + \frac{\varepsilon}{n^2})F(\mathcal{A})$, then let $\mathcal{A} \leftarrow \mathcal{A} \cup \{V'\}$, and repeat step 2.
3. If there exists an element $V' \in \mathcal{A}$ such that $F(\mathcal{A} \setminus \{V'\}) > (1 + \frac{\varepsilon}{n^2})F(\mathcal{A})$, then let $\mathcal{A} \leftarrow \mathcal{A} \setminus \{V'\}$, and go back to step 2.
4. Return $\mathcal{A}_{LS} \leftarrow \operatorname{argmax}\{F(\mathcal{A}), F(\mathcal{V} \setminus \mathcal{A})\}$.

In [8], it is proven that the local search is a polynomial time algorithm using at most $\mathcal{O}(\frac{1}{\varepsilon} n^3 \log n)$ function evaluations, and, for the solution $\mathcal{A}_{LS}$ returned by LS, it holds that

[2]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.



(a) *Attributes shared.* (b) *Location* (c) *Discriminability*

Figure 1: (a) Sensitivity of individual attributes (with 95% confidence intervals). (b) Sensitivity of sharing location under different levels of discretization. (c) Sensitivity of k-discriminability levels (right). Plots show 95% confidence intervals.

$F(\mathcal{A}_{LS}) \geq \left(\frac{1}{3} - \frac{\varepsilon}{n}\right) \max_{\mathcal{A}} F(\mathcal{A})$. In [15], we show how LS can be significantly sped up by using a lazy evaluation technique.

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 $C(\mathcal{A})$ (*c.f.*, Section 2). 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 sampling. In [15], we show, how we can use Hoeffding's inequality [11] in order to 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 [8] 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 4.1. *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 $\mathcal{O}\left(\frac{1}{\varepsilon} n^3 \log n \left(\frac{\log_2(\#intents)}{\varepsilon_S}\right)^2 \log \frac{1}{\delta n^3}\right)$ 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* [9], as well as mixed-integer programming [19].

5 Experimental results

5.1 Survey on Privacy Preferences

Although identifiability is an important part of privacy, people have different preferences about sharing individual attributes [20]. Related work has explored elicitation of private information (*c.f.*, [13, 24, 10]). 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 (in this case, in terms of enhanced search efficiency). As previous studies show [20], willingness to share information greatly depends 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 tried to be as specific as possible, by specifying a low-risk situation, in which the "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 within 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.

Questions about individual attributes. We first asked the participants to classify the sensitivity of the attributes on a Likert scale from 1 (not very sensitive) to 5 (highly sensitive). The order of the questions was randomized. Figure 1(a) presents the results. As might be expected, frequency of search engine usage (AFRQ), as well as very general topic interests, e.g., in news pages (TNWS), are considered to be of low sensitivity. Interestingly, we found that there are significant differences among participants even for sharing with a service interests in different topics; 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 . We found that participants were most sensitive to sharing whether they are at work while performing a query (AWHR).

Questions about identifiability. We also asked questions in order to elicit sensitivity about different levels of data aggregation and identifiability. First, we sought to identify how different levels of detail, or *granularity* (and hence, risk of being identified) affect sensitivity. More specifically, we asked, how sensitive the participant was to sharing their location at the region, country, state, city, zip code or address level. Figure 1(b) 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). Here, we varied k among 1, 10, 100, 1,000, 10,000, 100,000 and 1 million. Figure 1(c) presents the results of this experiment. The experiment shows that study participants have strong preferences about the granularity of the shared information. Moreover, as explained below in Section 5.4, we can use the information obtained from this experiment to explicitly take into account peoples' preferences when trading off privacy and utility.

Questions about utility. In addition to assessing the sensitivity of sharing different kinds of personal information, we asked the participants, what kind of improvement they would require in order to share attributes of a given sensitivity level. 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 1/2/...". 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 associated with a speedup by a factor of 4). We also allowed the participant the option of selecting to opt for never sharing information at the specified sensitivity level. Using the responses of this experiment, in addition to the sensitivity assessments, we can establish sensitivity as a common currency of utility and cost.

5.2 Search log data and attributes

Our experiments are based on a total of 247,684 queries performed by 9,523 users from 14 months between December 2005 and January 2007. The data was obtained from users who had volunteered to participate in a data sharing program that would make use of information about their search activities to enhance search. Our data contains only frequent queries which have been performed by at least 30 different users, resulting in a total of 914 different queries. From the demographic information and the search logs, we compute 28 different user / query specific attributes. In selecting our attributes, we chose very coarse-granular discretization. No attribute is represented by more than 2 bits, and most attributes are binary.

For demographic information, only location was available in the search log data (by inverse IP lookup). We discretized the location into four broad regions (DREG).

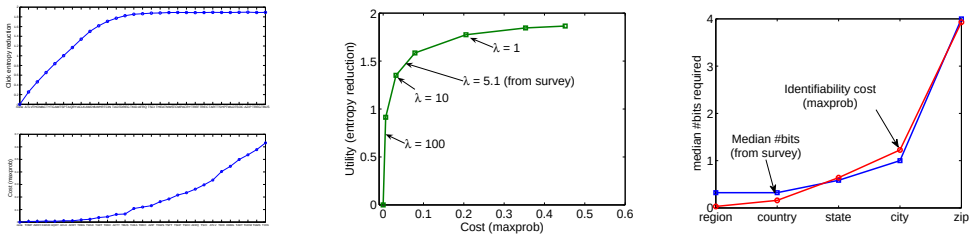
The next set of attributes contains features extracted from search history data. For each query, we determine whether the same query has been performed before (AQRY), as well as whether the searcher has visited the same webpage (ACLK) before. The attribute AFRQ describes whether the user performed at 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 ever performs queries from at least 2 different zip codes (AZIP), cities (ACTY) and countries (ACRY), by performing reverse DNS lookup of the query IP addresses. For each query, we also store whether the query was performed during working hours (AWHR; between 7 am and 6 pm) and during workdays (AWDY; Mon-Fri) or weekend (Sat, Sun).

We also 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 use a binary attribute (acronyms start with T) indicating whether the user has ever visited a website in the category.

5.3 Computing Utility and Cost

We evaluate 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 independent samples in order to estimate $U(\mathcal{A})$ and $I(\mathcal{A})$.

We first used the greedy algorithm to select an increasing number of attributes, maximizing the utility and ignoring the cost. Figure 2(a) presents the greedy ordering and the achieved entropy reductions. The greedy algorithm selects the attributes ATLV, THOM, ACTY, TGAM, TSPT, AQRY, ACLK,



(a) *Utility & Cost* (b) *Tradeoff-curve* (c) *Calibration*

Figure 2: Utility (a, top) and cost (a, bottom) for an increasing number of greedily chosen attributes. (b) Tradeoff-curve for varying λ. (c) Calibrating the tradeoff.

AWDY, AWHR, TCIN, TADT, DREG, TKID, AFRQ in this order. After selecting these attributes, the utility does not increase significantly anymore. The entropy reduction levels off at roughly 1.92 bits. Figure 2(a) clearly indicates the diminishing-returns property of click entropy reduction.

Similarly, we generate a greedy ordering of the attributes, in order of minimum incremental cost. Figure 2(a) presents the results of this experiment, using the maxprob cost metric. As expected, the curve looks convex (apart from small variations due to the sampling process). The cost initially increases very slowly, and the growth increases as more attributes are selected. This behavior empirically corroborates the supermodularity assumption for the cost metric.

5.4 Calibrating the tradeoff with the survey

In Section 5.3 we optimized utility and cost separately. We now use the scalarization (3.1) in order to trade off utility and cost. In order to do that, 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 2(b) 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 of the utility lead of big increases in cost, and vice versa. Hence, at this knee, one achieves near-maximal utility at near-minimum cost, which is very encouraging.

In order to take into account people's preferences in determining the tradeoff, we performed the following *calibration* procedure. From the search log data, we determined, how increasing details about a person's location increase the privacy cost. As levels of detail, we vary the location granularity from *region* (coarsest) to *zip code* (finest). For example, we computed the values $I_m(\{\text{zip code}\})$, $I_m(\{\text{city}\})$, etc. from data. We compared these values with responses from the survey as follows. As explained in Section 5.1, we asked the subjects to assess the sensitivity of sharing the different location granularities. Similarly, we asked, which improvement in search performance would be required in order to share attributes of a given sensitivity. With each level of improvement, we associated a number of bits: A speedup 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 now perform 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 2(c), 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 2(b).

5.5 Optimizing the utility-privacy tradeoff

Based on the calibration described in Section 5.4, our goal is to find a set of attributes $\mathcal{A}$ maximizing the calibrated objective $F_\lambda(\mathcal{A})$ according to (3.1). We use 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 2 bytes of the IP address. Figure 3 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 ≈ 1.5. The search statistics $\mathcal{A}_{search}$ performs second best, with a score of 0.57, but achieving a

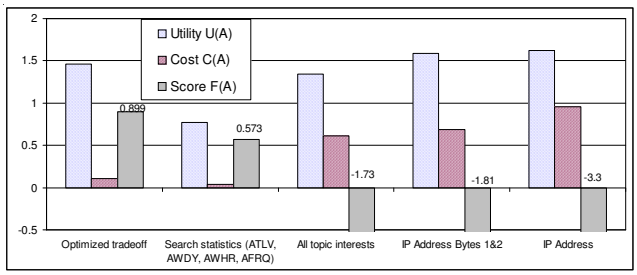


Figure 3: Comparison with heuristics

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. The reason for this is that knowledge of the exact topic interest profile frequently suffices to uniquely identify a searcher. As expected, the IP address (even the first 2 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 the optimized solution outperforms heuristic choices.

6 Conclusions

We presented an approach for explicitly optimizing the utility-privacy tradeoff in personalized services such as web search. We showed that utility functions like click entropy reduction satisfy submodularity, an intuitive diminishing returns property. In contrast, privacy concerns behave supermodularly; the more private information we combine, the higher sensitivity and risk of identifiability. Based on these 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 real-world web search data. We demonstrated how the quantitative tradeoff can be calibrated according to personal preferences, obtained from a user study with 1400 participants. Overall, we found that significant personalization can be achieved using only a small amount of information about users.

References

[1] N. R. Adam and J. C. Wortmann, *Security-control methods for statistical databases: A comparative study*, ACM Computing Surveys **21** (1989), no. 4, 515–556.
[2] Eytan Adar, *User 4xxxxx9: Anonymizing query logs*, Query Logs Workshop, WWW, 2007.
[3] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge UP, March 2004.
[4] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley Interscience, 1991.
[5] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen, *A large-scale evaluation and analysis of personalized search strategies*, WWW, 2007.
[6] D. Downey, S. Dumais, and E. Horvitz, *Models of searching and browsing: Languages, studies, and applications*, IJCAI, 2007.
[7] Cynthia Dwork, *Differential privacy*, ICALP, 2006.
[8] U. Feige, V. Mirrokni, and J. Vondrak, *Maximizing non-monotone submodular functions*, FOCS, 2007.
[9] Boris Goldengorin, Gerard Sierksma, Gert A. Tijssen, and Michael Tso, *The data-correcting algorithm for the minimization of supermodular functions*, Management Science **45** (1999), no. 11, 1539–1551.
[10] I. Hann, K. Hui, T. Lee, and I. Png, *Online-information privacy: Measuring the cost-benefit tradeoff*, International Conference on Information Systems, 2002.
[11] Wassily Hoeffding, *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association **58** (1963), no. 301, 13–30.
[12] B. Hore and S. Mehrotra R. Jammalamadaka, *Flexible anonymization for privacy preserving data publishing: A systematic search based approach*, SDM, 2007.
[13] Bernardo A. Huberman, Eytan Adar, and Leslie R. Fine, *Valuating privacy*, IEEE Security & Privacy **3** (2005), no. 5, 22–25.
[14] A. Krause and C. Guestrin, *Near-optimal nonmyopic value of information in graphical models*, UAI, 2005.
[15] Andreas Krause and Eric Horvitz, *Privacy, personalization, and the web: A utility-theoretic approach*, Tech. Report MSR-TR-2007-135, Microsoft Research, October 2007.
[16] K. LeFevre, D. DeWitt, and R. Ramakrishnan, *Mondrian multidimensional k-anonymity*, ICDE, 2006.
[17] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam, *L-diversity: Privacy beyond k-anonymity*, ICDE, 2006.
[18] G. Nemhauser, L. Wolsey, and M. Fisher, *An analysis of the approximations for maximizing submodular set functions*, Mathematical Programming **14** (1978), 265–294.
[19] G. L. Nemhauser and L. A. Wolsey, *Studies on graphs and discrete programming*, ch. Maximizing Submodular Set Functions: Formulations and Analysis of Algorithms, pp. 279–301, North-Holland, 1981.
[20] Judith S. Olson, Jonathan Grudin, and Eric Horvitz, *A study of preferences for sharing and privacy*, CHI, 2005.
[21] K. Sugiyama, K. Hatano, and T. Ikoma, *Adaptive web search based on user profile constructed without any effort from users*, WWW, 2004.
[22] L. Sweeney, *k-anonymity: a model for protecting privacy*, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems **10** (2002), no. 5, 557–570.
[23] J. Teevan, S. T. Dumais, and E. Horvitz, *Personalizing search via automated analysis of interests and activities*, SIGIR, 2005.
[24] S. Wattal, R. Telang, T. Mukhopadhyay, and P. Boatwright, *Examining the personalization-privacy tradeoff an empirical investigation with email advertisements*, Management Science (2005).
[25] Y. Xu, B. Zhang, and K. Wang, *Privacy-enhancing personalized web search*, WWW, 2007.