Stochastic Privacy

Adish Singla*
ETH Zurich
adish.singla@inf.ethz.ch

Eric Horvitz
Microsoft Research
horvitz@microsoft.com

Ece Kamar
Microsoft Research
eckamar@microsoft.com

Ryen White
Microsoft Research
ryen.white@microsoft.com

Abstract

Online services such as web search and e-commerce applications typically rely on the collection of data about users, including details of their activities on the web. Such personal data is used to maximize revenues via targeting of advertisements and longer engagements of users, and to enhance the quality of service via personalization of content. To date, service providers have largely followed the approach of either requiring or requesting consent for collecting user data. Users may be willing to share private information in return for incentives, enhanced services, or assurances about the nature and extent of the logged data. We introduce *stochastic privacy*, an approach to privacy centering on the simple concept of providing people with a guarantee that the probability that their personal data will be shared does not exceed a given bound. Such a probability, which we refer to as the *privacy risk*, can be given by users as a preference or communicated as a policy by a service provider. Service providers can work to personalize and to optimize revenues in accordance with preferences about privacy risk. We present procedures, proofs, and an overall system for maximizing the quality of services, while respecting bounds on privacy risk. We demonstrate the methodology with a case study and evaluation of the procedures applied to web search personalization. We show how we can achieve near-optimal utility of accessing information with provable guarantees on the probability of sharing data.

Introduction

Online services such as web search, recommendation engines, social networks, and e-commerce applications typically rely on the collection of data about activities (*e.g.*, click logs, queries, and browsing information) and personal information (*e.g.*, location and demographics) of users. The availability of such data enables providers to personalize services to individuals and also to learn how to enhance the service for all users (*e.g.*, improved search results relevance). User data is also important to providers for optimizing revenues via better targeted advertising, extended user engagement and popularity, and even the selling of user data to third party companies. Permissions are typically obtained via broad consent agreements that request user permission to share their data through system dialogs or via complex *Terms of Service*. Such notices are typically difficult to understand and are often ignored (Technet 2012). In other cases, a plethora of requests for information, such as attempts to gain access to users' locations, may be shown in system dialogs at run time or installation time. Beyond the normal channels for sharing data, potential breaches of information are possible via attacks by malicious third parties and malware, and through surprising situations such as the AOL data release (Arrington 2006; Adar 2007) and de-anonymization of released Netflix logs (Narayanan and Shmatikov 2008). The charges by the Federal Trade Commission against Facebook (FTC 2011) and Google (FTC 2012) highlight increasing concerns by privacy advocates and government institutions about the large-scale recording of personal data.

Ideal approaches to privacy in online services would enable users to benefit from machine learning over data from populations of users, yet consider users' preferences as a top priority. Prior research in this realm has focused on designing privacy-preserving methodologies that can provide for control of a privacy-utility tradeoff (Adar 2007; Krause and Horvitz 2008). Research has also explored the feasibility of incorporating user preferences over what type of data can be logged (Xu et al. 2007; Cooper 2008; Olson, Grudin, and Horvitz 2005; Krause and Horvitz 2008).

We introduce a new approach to privacy that we refer to as *stochastic privacy*. Stochastic privacy centers on providing a guarantee to users about the likelihood that their data will be accessed and used by a service provider. We refer to this measure as the assessed or communicated *privacy risk*, which may be increased in return for increases in the quality of service or other incentives. Very small probabilities of sharing data may be tolerated by individuals (just as lightning strikes are tolerated as a rare event), yet offer service providers sufficient information to optimize over a large population of users. Stochastic privacy depends critically on harnessing inference and decision making to make choices about data collection within the constraints of a guaranteed privacy risk.

We explore procedures that can be employed by service providers when preferences or constraints about the shar-

*Adish Singla performed this research during an internship at Microsoft Research.

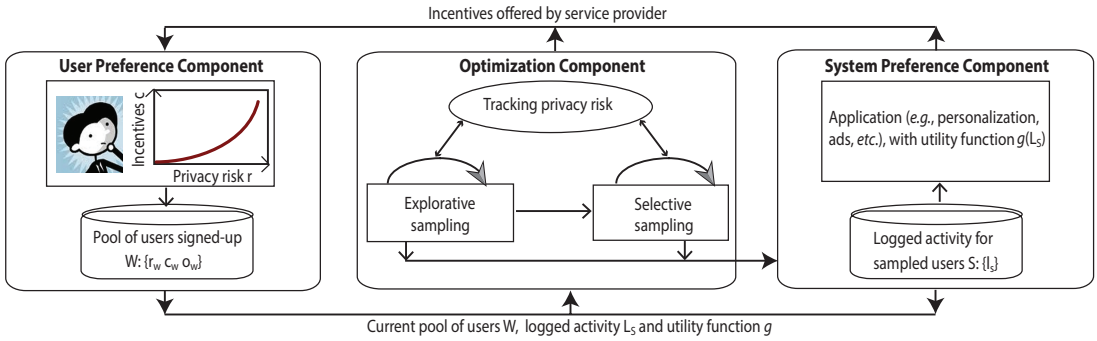


Figure 1: Overview of stochastic privacy.

ing of data are represented as privacy risk. The goal is to maximize the utility of service using data extracted from a population of users, while abiding by the agreement reached with users on privacy risk. We show that optimal selection of users under these constraints is NP-hard and thus intractable, given the massive size of the online systems. As solutions, we propose two procedures, RANDGREEDY and SPGREEDY, that combine greedy value of information analysis with obfuscation to offer mechanisms for tractable optimization, while satisfying stochastic privacy guarantees. We present performance bounds for the expected utility achievable by these procedures compared to the optimal solution. Our contributions can be summarized as follows:

- Introduction of stochastic privacy, an approach that represents preferences about the probability that data will be shared, and methods for trading off privacy risk, incentives, and quality of service.
- A tractable end-to-end system for implementing a version of stochastic privacy in online services.
- RANDGREEDY and SPGREEDY procedures for sampling users under the constraints of stochastic privacy, with theoretical guarantees on the acquired utility.
- Evaluation to demonstrate the effectiveness of the proposed procedures on a case study of user selection for personalization in web search.

Stochastic Privacy Overview

Figure 1 provides an overview of stochastic privacy in the context of a particular design of a system that implements the methodology. The design is composed of three main components: (i) a user preference component, (ii) a system preference component, and (iii) an optimization component for guiding the system's data collection. We now provide details about each of the components and then formally specify the optimization problem for the *selective sampling* module.

User Preference Component

The user preference component interacts with users (*e.g.*, during signup) and establishes an agreement between a user and service provider on a tolerated probability that the user's data will be shared in return for better quality of service or incentives. Representing users' tolerated privacy risk allows for the design of controls that provide options for sharing data. The incentives offered to users can be personalized based on general information available about a user (*e.g.*, general location information inferred from a previously shared IP address) and can vary from guarantees of improved service (Krause and Horvitz 2010) to complementary software and entries in a lottery to win cash prizes (as done by the comScore service (Wikipedia-comScore 2006)).

Formally, let W be the population of users signed up for a service. Each user $w \in W$ is represented with the tuple $\{r_w, c_w, o_w\}$, where o_w is the metadata information (*e.g.*, IP address) available for user w prior to selecting and logging finer-grained data about the user. r_w is the privacy risk assessed by the user, and c_w is the corresponding incentive provided in return for the user assuming the risk. The elements of this tuple can be updated through interactions between the system and the user. For simplicity of analysis, we shall assume that the pool W and user preferences are static.

System Preference Component

The goal of the service provider is to optimize the quality of service. For example, a provider may wish to personalize web search and to improve the targeting of advertising for maximization of revenue. The provider may record the activities of a subset of users (*e.g.*, sets of queries issued, sites browsed, *etc.*) and use this data to provide better service globally or to a specific cohort of users. We model the private data of activity logs of user w by variable $l_w \in 2^L$, where L represents the web-scale space of activities (*e.g.*, set of queries issued, sites browsed, *etc.*) . However, l_w is observed by the system only after w is selected and the data from w is logged. We model the system's uncertain belief of l_w by a random variable Y_w, with l_w being its realization distributed according to conditional probability distribution $P(Y_w = l_w|o_w)$. In order to make an informed decision about user selection, the distribution $P(Y_w = l_w|o_w)$ is learned by the system using data available from the user and recorded logs of other users. We quantify the utility of application by logging activities L_S from selected users S through function $g : 2^L \rightarrow \mathbb{R}$, given by $g(\bigcup_{s \in S} l_s)$. The expected value of the utility that the system can expect to gain by selecting users S with observed attributes O_S is characterized by distribution P and utility function g as: $\tilde{g}(S) \equiv \mathbb{E}_{Y_S}\left[g(\bigcup_{s \in S} l_s)\right] = \sum_{L_S \in 2^L \times S} \left(P(Y_S = L_S|O_S) \cdot g(\bigcup_{s \in S} l_s)\right)$. However, the application itself may be using the logs L_S in a complex manner (such as training a

ranker (Bennett et al. 2011)) and evaluating this on complex user metrics (Hassan and White 2013). Hence, the system uses a surrogate utility function $f(S) \approx \tilde{g}(S)$ to capture the utility through a simple metric, for example, coverage of query-clicks obtained from the sampled users (Singla and White 2010) or reduction in uncertainty of click phenomena (Krause and Horvitz 2008).

We require the set function f to be *non-negative*, *monotone* (*i.e.*, whenever $A \subseteq A' \subseteq W$, it holds that $f(A) \leq f(A')$) and *submodular*. Submodularity is an intuitive notion of diminishing returns, stating that, for any sets $A \subseteq A' \subseteq W$, and any given user $a \notin A'$, it holds that $f(A \cup \{a\}) - f(A) \geq f(A' \cup \{a\}) - f(A')$. These conditions are general, and are satisfied by many realistic, as well as complex utility functions (Krause and Guestrin 2007), such as reduction in click entropy (Krause and Horvitz 2008). As a concrete example, consider the setting where attributes O represent geo-coordinates of the users and $D : O \times O \to \mathbb{R}$ computes the geographical distance between any two users. The goal of the service is to provide location-based personalization of web search. For such an application, click information from local users provides valuable signals for personalizing search (Bennett et al. 2011). The system's goal is to select a set of users S, and to leverage data from these users to enhance the service for the larger population of users. For search queries originating from any other user w, it uses the click data from the nearest user in S, given by $\arg\min_{s \in S} D(o_s, o_w)$. One approach for finding such a set S is solving the *k-medoid* problem which aims to minimize the sum of pairwise distances between selected set and the remaining population (Mirzasoleiman et al. 2013; Kaufman and Rousseeuw 2009). Concretely, this can be captured by the following submodular utility function:

$$f(S) = \frac{1}{|W|} \sum_{w \in W} \Big(\min_{x \in X} D(o_x, o_w) - \min_{s \in S \cup X} D(o_s, o_w) \Big) \tag{1}$$

Here, X is any one (or a set of) fixed reference location(s), for example, simply representing origin coordinates and is used ensure that function f is non-negative and monotone. Lemma 1 formally states the properties of this function.

Optimization Component

To make informed decisions about data access, the system computes the expected value of information (VOI) of logging the activities of a particular user, *i.e.*, the marginal utility that the application can expect by logging the activity of this user (Krause and Horvitz 2008). In the absence of sufficient information about user attributes, the VOI may be small, and hence needs to be learned from the data. The system can randomly sample a small set of users from the population that can be used to learn and improve the models of VOI computation (*explorative sampling* in Figure 1). For example, for optimizing the service for a user cohort speaking a specific language, the system may choose to collect logs from a subset of users to learn how languages spoken by users map to geography. If preferences about privacy risk are overlooked, VOI can be used to select users to log with a goal of maximizing the utility for the service provider (*selective sampling* in Figure 1). Given that the utility function of the system is submodular, a greedy selection rule makes near-optimal decisions about data access (Krause and Guestrin 2007). However, this simple approach could violate guarantees on privacy risk. To act in accordance with the guarantee, we design selective sampling procedures that couple obfuscation with VOI analysis to select the set of users to provide data.

The system needs to ensure that both the explorative and selective sampling approaches respect the privacy guarantees, *i.e.*, the likelihood of sampling any user w throughout the execution of the system must be less than the privacy risk factor r_w. The system tracks the sampling risk (likelihood of sampling) that user w faces during phases of the execution of explorative sampling, denoted r_w^{ES}, and selective sampling, denoted r_w^{SS}. The privacy guarantee for a user is preserved as long as: $r_w - \big(1 - (1 - r_w^{ES}) \cdot (1 - r_w^{SS})\big) \geq 0$.

Optimization Problem for Selective Sampling

We now focus primarily on the selective sampling module and formally introduce the optimization problem. The goal is to design a sampling procedure M that abides by guarantees of stochastic privacy, yet optimizes the utility of the application in decisions about accessing user data. Given a budget constraint B, the goal is to select users S^M:

$$S^M = \arg\max_{S \subseteq W} f(S) \tag{2}$$
$$\text{subject to } \sum_{s \in S} c_s \leq B \text{ and } r_w - r_w^M \geq 0 \; \forall w \in W.$$

Here, r_w^M is the likelihood of selecting $w \in \mathcal{W}$ by procedure M and hence $r_w - r_w^M \geq 0$ captures the constraint of stochastic privacy guarantee for w. Note that we can interchangeably write utility acquired by procedure as $f(M)$ to denote $f(S^M)$ where S^M is the set of users selected by running M. We shall now consider a simpler setting of constant privacy risk rate r for all users and unit cost per user (thus reducing the budget constraint to a simpler cardinal constraint, given by $|S| \leq B$). These assumptions lead to defining $B \leq W \cdot r$, as that is the maximum possible set size that can be sampled by any procedure for Problem 2.

Selective Sampling with Stochastic Privacy

We now present desiderata of the selection procedures, discuss the hardness of the problem, and review several different tractable approaches, as summarized in Table 1.

Desirable Properties of Sampling Procedures

The problem defined by Equation 2 requires solving an NP-hard discrete optimization problem, even when the stochastic privacy constraint is removed. The algorithm for finding the optimal solution of this problem without the privacy constraint, referred as OPT, is intractable (Feige 1998). We address this intractability by exploiting the submodular structure of the utility function f and offer procedures providing provable near-optimal solutions in polynomial time. We aim at designing procedures that satisfy the following desirable properties: *(i)* provides competitive utility w.r.t. OPT with

Procedure	Competitive utility	Privacy guarantees	Polynomial runtime
OPT	✓	✗	✗ $\mathcal{O}(\lvert W\rvert^B)$
GREEDY	✓	✗	✓ $\mathcal{O}(B \cdot \lvert W\rvert)$
RANDOM	✗	✓	✓ $\mathcal{O}(B)$
RANDGREEDY	✓	✓	✓ $\mathcal{O}(B \cdot \lvert W\rvert \cdot r)$
SPGREEDY	✓	✓	✓ $\mathcal{O}(B \cdot \lvert W\rvert \cdot log(1/r))$

Table 1: Properties of different procedures. RANDGREEDY and SPGREEDY satisfy all of the desired properties.

provable guarantees, *(ii)* preserves stochastic privacy guarantees, and *(iii)* runs in polynomial time.

Random Sampling: RANDOM

RANDOM samples the users at random, without any consideration of cost and utility. The likelihood of any user w to be selected by the algorithm is $r_w^{\text{RANDOM}} = B/W$ and hence privacy risk guarantees are trivially satisfied since $B \leq W \cdot r$ as defined in Problem 2. In general, RANDOM can perform arbitrarily poorly in terms of acquired utility, specifically for applications targeting particular user cohorts.

Greedy Selection: GREEDY

Next, we explore a greedy sampling strategy that maximizes the expected marginal utility at each iteration to guide decisions about selecting a next user to log. Formally, GREEDY starts with empty set $S = \emptyset$. At an iteration i, it greedily selects a user $s_i^* = \arg\max_{w \subseteq W \setminus S} \big(f(S \cup w) - f(S)\big)$ and adds the user to the current selection of users $S = S \cup \{s_i^*\}$. The procedure halts when $|S| = B$.

A fundamental result by Nemhauser, Wolsey, and Fisher (1978) states that the utility obtained by this greedy selection strategy is guaranteed to be at least $(1 - 1/e)$ $(= 0.63)$ times that obtained by OPT. This result is tight under reasonable complexity assumptions ($P \neq NP$) (Feige 1998). However, such a greedy selection clearly violates the stochastic privacy constraint in Problem 2. Consider the user w^* with highest marginal value: $w^* = \arg\max_{w \subseteq W} f(w)$. The likelihood that this user will be selected by the algorithm $r_{w^*}^{\text{GREEDY}} = 1$, regardless of the promised privacy risk r_{w^*}.

Sampling and Greedy Selection: RANDGREEDY

We combine the ideas of RANDOM and GREEDY to design procedure RANDGREEDY which provides guarantees on stochastic privacy and competitive utility. RANDGREEDY is an iterative procedure that samples a small batch of users $\psi(s)$ at each iteration, then greedily selects $s^* \in \psi(s)$ and removes the entire set $\psi(s)$ for further consideration. By keeping the batch size $\psi(s) \leq W \cdot r/B$, the procedure ensures that the privacy guarantees are satisfied. As our user pool W is static, to reduce complexity, we consider a simpler version of RANDGREEDY that defers the greedy selection. Formally, this is equivalent to first sampling the users from W at rate r to create a subset $\widetilde{W}$ such that $|\widetilde{W}| = |W| \cdot r$, and then running the GREEDY algorithm on $\widetilde{W}$ to greedily select a set of users of size B.

The initial random sampling ensures a guarantee on the privacy risk for users during the execution of the procedure. In fact, for any user $w \in W$, the likelihood of w being sampled and included in subset $\widetilde{W}$ is $r_w^{\text{RANDGREEDY}} \leq r$. We further analyze the utility obtained by this procedure in the next section and show that, under reasonable assumptions, the approach can provide competitive utility compared to OPT.

Greedy Selection with Obfuscation: SPGREEDY

SPGREEDY uses an inverse approach of mixing RANDOM and GREEDY: it does greedy selection, followed by obfuscation, as illustrated in Procedure 1. It assumes an underlying distance metric $D : W \times W \to \mathbb{R}$ which captures the notion of distance or dissimilarity among users. As in GREEDY, it operates in iterations and selects the user s^* with maximum marginal utility at each iteration. However, to ensure stochastic privacy, it obfuscates s^* with nearest $1/r$ number of users using distance metric D to create a set $\psi(s^*)$. Then, it samples one user randomly from $\psi(s^*)$ and removes the entire set $\psi(s^*)$ from further consideration.

The guarantees on privacy risk hold by the following arguments: During the execution of the algorithm, any user w becomes a possible candidate of being selected if the user is part of $\psi(s^*)$ in some iteration (*e.g.*, iteration i). Given that $|\psi(s^*)| \geq 1/r$ and algorithm randomly sample $v \in \psi(s^*)$, the likelihood of w being selected in iteration i is at most r. The fact that set $\psi(s^*)$ is removed from available pool $\widetilde{W}$ at the end of the iteration ensures that w can become a possible candidate of selection only once.

Procedure 1: SPGREEDY

Input: *users* W; *cardinality constraint* B; *privacy risk* r; *distance metric* $D : W \times W \to \mathbb{R}$;
Initialize:
- **Outputs**: *selected users* $S \leftarrow \emptyset$;
- **Variables**: *remaining users* $W' \leftarrow W$;

begin
  while $|S| \leq B$ **do**
    $s^* \leftarrow \arg\max_{w \in W'} f(S \cup w) - f(S)$;
    Set $\psi(s^*) \leftarrow s^*$;
    while $|\psi(s^*)| < 1/r$ **do**
      $v \leftarrow \arg\min_{w \in W' \setminus \psi(s^*)} D(w, s^*)$;
      $\psi(s^*) \leftarrow \psi(s^*) \cup \{v\}$;
    Randomly select $\widetilde{s^*} \in \psi(s^*)$;
    $S \leftarrow S \cup \{\widetilde{s^*}\}$;
    $W' \leftarrow W' \setminus \psi(s^*)$;
Output: S

Performance Analysis

We now analyze the performance of the proposed procedures in terms of the utility acquired compared to that of OPT as a baseline. We first analyze the problem in a general setting and then under a set of practical assumptions on the structure of underlying utility function f and population of users W. The proofs of all the results are available in the extended version (Singla et al. 2014).

General Case

In the general setting, we show that one cannot do better than $r \cdot f(OPT)$ in the worst case. Consider a population of users W where only one user w^* has utility value of 1, and rest of the users $W \setminus w^*$ have utility of 0. OPT achieves a utility of 1 by selecting $S^{\text{OPT}} = \{w^*\}$. Consider any procedure M that has to respect the guarantees on privacy risk. If the privacy rate of w^* is r, then M can select w^* with only a maximum probability of r. Hence, the maximum expected utility that any procedure M for Problem 2 can achieve is r.

On a positive note, a trivial algorithm can always achieve a utility of $(1 - 1/e) \cdot r \cdot f(OPT)$ in expectation. This result can be reached by running GREEDY to select a set S^{GREEDY} and then choosing the final solution to be S^{GREEDY} with probability r, and otherwise output an empty set. Theorem 1 formally states these results for the general problem setting.

Theorem 1. *Consider the Problem 2 of optimizing a submodular function f under cardinality constraint B and privacy risk rate r. For any distribution of marginal utilities of population W, a trivial procedure can achieve an expected utility of at least $(1 - 1/e) \cdot r \cdot f(OPT)$. In contrast, there exists an underlying distribution for which no procedure can have expected utility of more than $r \cdot f(OPT)$.*

Smoothness and Diversification Assumptions

In practice, we can hope to do much better than the worst-case results described in Theorem 1 by exploiting the underlying structure of users' attributes and utility function. We start with the assumption that there exists a distance metric $D : W \times W \to \mathbb{R}$ which captures the notion of distance or dissimilarity among users. For any given $w \in W$, let us define its α-neighborhood to be the set of users within a distance α from w (*i.e.*, α-close to w): $N_\alpha(w) = \{v : D(v, w) \leq \alpha\}$. We assume that population of users is large and that the number of users in the $N_\alpha(w)$ is large. We capture these requirements formally in Theorems 2,3.

First, we consider utility functions that change gracefully with changes in inputs, similar to the notion of λ-*Lipschitz* set functions used in Mirzasoleiman et al. (2013). We formalize the notion of smoothness in the utility function f w.r.t metric D as follows:

Definition 1. *For any given set of users S, let us consider a set $\widetilde{S}_\alpha$ obtained by replacing every $s \in S$ with any $w \in N_\alpha(s)$. Then, $|f(S) - f(\widetilde{S}_\alpha)| \leq \lambda_f \cdot \alpha \cdot |S|$, where parameter λ_f captures the notion of smoothness of function f.*

Secondly, we consider utility functions that favor diversity or dissimilarity of users in the subset selection w.r.t the distance metric D. We formalize this notion of diversification in the utility function as follows:

Definition 2. *Let us consider any given set of users $S \subseteq W$ and a user $w \in W$. Let $\alpha = \min_{s \in S} D(s, w)$. Then, $f(S \cup w) - f(S) \leq \Upsilon_f \cdot \alpha$, where parameter Υ_f captures the notion of diversification of function f.*

The utility function f introduced in Equation 1 satisfies both of the above assumptions as formally stated below.

Lemma 1. *Consider the utility function f in Equation 1. f is submodular, and satisfies the properties of smoothness and diversification, i.e. has bounded λ_f and Υ_f.*

We note that for the functions with unbounded λ and Υ (*i.e.*, $\lambda_f \to \infty$ and $\Upsilon_f \to \infty$), it would lead to the general problem settings (equivalent to no assumptions) and hence results of Theorem 1 apply.

Performance Bounds

Under the assumption of smoothness (*i.e.*, bounded λ_f), we can show the following bound on utility of RANDGREEDY:

Theorem 2. *Consider the Problem 2 for function f with bounded λ_f. Let S^{OPT} be the set returned by* OPT *for Problem 2 without the privacy constraints. For a desired $\epsilon < 1$, let $\alpha_{rg} = \arg\min_\alpha \{\alpha : |N_\alpha(s)| \geq 1/r \cdot \log(B/\epsilon)\ \forall s \in S^{\text{OPT}}$ and $N_\alpha(s) \cap N_\alpha(s') = \emptyset\ \forall s, s' \in S^{\text{OPT}}\}$. Then, with probability at least $(1 - \epsilon)$,*

$$\mathbb{E}[f(\text{RANDGREEDY})] \geq (1 - 1/e) \cdot \big(f(\text{OPT}) - \alpha_{rg} \cdot \lambda_f \cdot B\big)$$

Under the assumption of smoothness and diversification (*i.e.*, bounded λ_f and Υ_f), we can show the following bound on utility of SPGREEDY:

Theorem 3. *Consider the Problem 2 for function f with bounded λ_f and Υ_f. Let S^{GREEDY} be the set returned by* GREEDY *for Problem 2 without the privacy constraints. Let $\alpha_{spg} = \arg\min_\alpha \{\alpha : |N_\alpha(s)| \geq 1/r\ \forall s \in S^{\text{GREEDY}}\}$. Then,*

$$\mathbb{E}[f(\text{SPGREEDY})] \geq (1 - 1/e) \cdot f(\text{OPT}) - 2 \cdot (\lambda_f + \Upsilon_f) \cdot \alpha_{spg} \cdot B$$

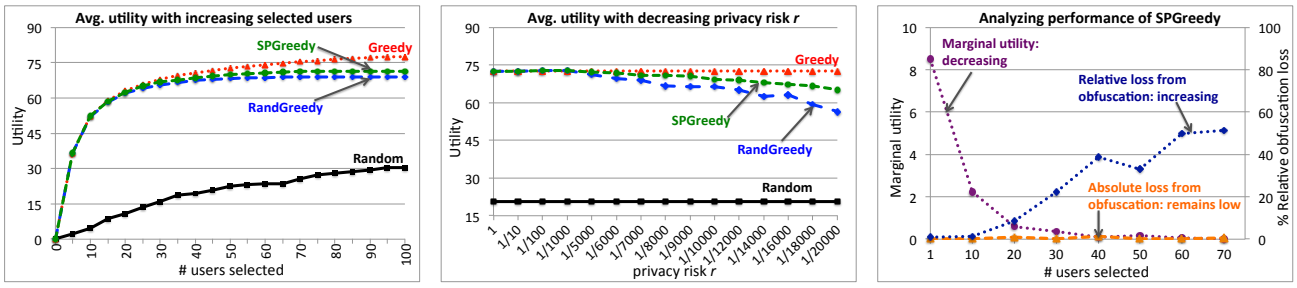
Intuitively, these results imply that both RANDGREEDY and SPGREEDY achieve competitive utility w.r.t OPT, and that the performance degrades smoothly as the privacy risk r is decreased or the bounds on smoothness and diversification for function f increase.

Experimental Evaluation

We now report on experiments aimed at providing insights on the performance of the stochastic privacy procedures with a case study of the selective collection of user data in support of the personalization of web search.

Benchmarks and Metrics

We compare the performance of the RANDGREEDY and SPGREEDY procedures against the baselines of RANDOM and GREEDY. While RANDOM provides a trivial lower benchmark for any procedure, GREEDY is a natural upper bound on the utility, given that OPT itself is intractable. To analyze the robustness of the procedures, we vary the level of privacy risk r. We further carried out experiments to understand the loss incurred from the obfuscation phase during the execution of SPGREEDY.



(a) Vary budget B (b) Vary privacy risk r (c) Analyze SPGREEDY

Figure 2: Fig. 2(a) shows increases in the average utility of proposed procedures and GREEDY with increases in the budget B on the number of selected users, at a constant privacy risk of $r = 1/10000$. Fig. 2(b) displays smooth decreases in utility as the level of privacy risk r for the population is reduced for applying RANDGREEDY and SPGREEDY with a fixed budget $B = 50$. Fig. 2(c) shows small losses at each step incurred by SPGREEDY via obfuscation.

Experimental Setup

We consider the application of providing location-based personalization for queries issued for the business domain (*e.g.*, real-estate, financial services, *etc.*). The goal is to select a set of users S who are experts at web search in this domain. We seek to leverage click data from these users to improve the relevance of search results shown to the broader population of users searching for local businesses. The experiments are based on using a surrogate utility function as introduced in Equation 1. As we study the domain of business-related queries, we modify the utility function in Equation 1 by restricting S to users who are experts in the domain, as further described below. The acquired utility can be interpreted as the average reduction in the distance for any user w in the population to the nearest expert $s \in S$.

The primary source of data for the study is obtained from interaction logs on a major web search engine. We consider a fraction of users who issued at least one query in the month of October 2013, restricted to queries coming from IP addresses located within ten neighboring states in the western region of the United States. This results in a pool W of seven million users. We consider a setting where the system has access to metadata information of geo-coordinates of the users, as well as a probe of the last 20 search-result clicks for each user, which together constitute the observed attributes of user denoted as o_w. Each of these clicks are then classified into a topical hierarchy from a popular web directory named the Open Directory Project (ODP) (dmoz.org), using automated techniques (Bennett, Svore, and Dumais 2010). With a similar objective to White, Dumais, and Teevan (2009), the system then uses this classification to identify users who are expert in the business domain. We used the simple rule of classifying a user as an expert if at least one click was issued in the domain of interest. With this, the system marks a set of users $W' \subseteq W$ as experts, and the set S in Equation 1 is restricted to W'. We note that the specific thresholds or variable choices do not influence the overall results below.

Results

We now review results from the experiments.

Varying the budget B: In the first set of experiments, we vary the budget B of the number of users selected, and measure the utility acquired by different procedures. We fix the privacy risk $r = 1/10000$. Figure 2(a) illustrates that both RANDGREEDY and SPGREEDY are competitive w.r.t GREEDY and outperform the naive RANDOM baseline.

Varying the privacy risk r: We then vary the level of privacy risk, for a fixed budget $B = 50$, to measure the robustness of the RANDGREEDY and SPGREEDY. The results in Figure 2(b) demonstrate that the performance of RANDGREEDY and SPGREEDY degrades smoothly, as per the performance analysis in Theorems 2 and 3.

Analyzing performance of SPGREEDY: Last, we perform experiments to understand the execution of SPGREEDY and the loss incurred from the obfuscation step. SPGREEDY removes $1/r$ users from the pool at every iteration. As a result, for a small privacy risk r, the relative loss from obfuscation (*i.e.*, relative % difference in marginal utility acquired by a user chosen by greedy selection as compared to a user picked following obfuscation) can increase over the execution of the procedure. Such an increase is illustrated in Figure 2(a), which displays results computed using a moving average of window size 10. However, the diminishing returns property ensures that SPGREEDY incurs low absolute loss in marginal utility from obfuscation at each step.

Summary and Future Directions

We introduced *stochastic privacy*, a new approach to managing privacy that centers on service providers abiding by guarantees about not exceeding a specified likelihood of accessing users' data, and maximizing information collection in accordance with these guarantees. We presented procedures and an overall system design for maximizing the quality of services while respecting an assessed or communicated privacy risk. We showed bounds on the performance of the RANDGREEDY and SPGREEDY procedures, as compared to the optimal, NP-Hard solution and evaluated the algorithms on a web personalization application.

Research directions ahead on stochastic privacy include studies of user preferences about the probability of sharing data. We are interested in understanding how people in different settings may trade increases in privacy risk for en-

hanced service and monetary incentives. We seek an understanding of preferences, policies, and corresponding analyses that consider the sharing of data as a privacy risk *rate* over time. We are also interested in exploring different overall designs for the operation of a large-scale system, spanning study of different ways that users might be engaged. In one design, a provider might simply publish a universal policy on privacy risk or privacy risk rate. In another approach, users might additionally be notified when they are selected to share data and can decide at that time whether to accept and receive a gratuity or to decline the request for data. Inferences about the preferences of subpopulations about privacy risk and incentives could be folded into the selection procedures, and systems could learn to recognize and counter informational biases that might be associated with data accessed from these subgroups. We are excited about the promise of stochastic privacy to provide understandable approaches to enhancing privacy while enabling rich, personalized online services.

References

Adar, E. 2007. User 4xxxxx9: Anonymizing query logs. In *Workshop on Query Log Analysis at WWW'07*.

Arrington, M. 2006. Aol proudly releases massive amounts of private data. http://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data.

Bennett, P. N.; Radlinski, F.; White, R. W.; and Yilmaz, E. 2011. Inferring and using location metadata to personalize web search. In *Proc. of SIGIR*, 135–144.

Bennett, P. N.; Svore, K.; and Dumais, S. T. 2010. Classification-enhanced ranking. In *Proc. of WWW*, 111–120.

Cooper, A. 2008. A survey of query log privacy-enhancing techniques from a policy perspective. *ACM Trans. Web* 2(4):19:1–19:27.

Feige, U. 1998. A threshold of ln n for approximating set cover. *Journal of the ACM* 45:314–318.

FTC. 2011. FTC charges against Facebook. http://www.ftc.gov/opa/2011/11/privacysettlement.shtm.

FTC. 2012. FTC charges against Google. http://www.ftc.gov/opa/2012/08/google.shtm.

Hassan, A., and White, R. W. 2013. Personalized models of search satisfaction. In *Proc. of CIKM*, 2009–2018.

Kaufman, L., and Rousseeuw, P. J. 2009. *Finding groups in data: an introduction to cluster analysis*, volume 344. Wiley.com.

Krause, A., and Guestrin, C. 2007. Near-optimal observation selection using submodular functions. In *Proc. of AAAI, Nectar track*.

Krause, A., and Horvitz, E. 2008. A utility-theoretic approach to privacy and personalization. In *Proc. of AAAI*.

Krause, A., and Horvitz, E. 2010. A utility-theoretic approach to privacy in online services. *Journal of Artificial Intelligence Research (JAIR)* 39:633–662.

Mirzasoleiman, B.; Karbasi, A.; Sarkar, R.; and Krause, A. 2013. Distributed submodular maximization: Identifying representative elements in massive data. In *Proc. of NIPS*.

Narayanan, A., and Shmatikov, V. 2008. Robust de-anonymization of large sparse datasets. In *Proc. of the IEEE Symposium on Security and Privacy*, 111–125.

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

Singla, A., and White, R. W. 2010. Sampling high-quality clicks from noisy click data. In *Proc. of WWW*, 1187–1188.

Singla, A.; Horvitz, E.; Kamar, E.; and White, R. 2014. Stochastic privacy (extended version). http://research.microsoft.com/~horvitz/StochasticPrivacy-extended.pdf.

Technet. 2012. Privacy and technology in balance. http://blogs.technet.com/b/microsoft_on_the_issues/archive/2012/10/26/privacy-and-technology-in-balance.aspx.

White, R. W.; Dumais, S. T.; and Teevan, J. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proc. of WSDM*, 132–141.

Wikipedia-comScore. 2006. ComScore-#Data_collection_and_reporting. http://en.wikipedia.org/wiki/ComScore\#Data_collection_and_reporting.

Xu, Y.; Wang, K.; Zhang, B.; and Chen, Z. 2007. Privacy-enhancing personalized web search. In *Proc. of WWW*, 591–600. ACM.