Stochastic Privacy


Adish Singla


Eric Horvitz


Ece Kamar


Ryen White

AAAI'14 | July 30, 2014

Stochastic Privacy

Bounded, small level of *privacy risk*—the likelihood of data being accessed and harnessed in a service.

Procedures (and proofs) work within bound to acquire data from population needed to provide quality of service for all users.

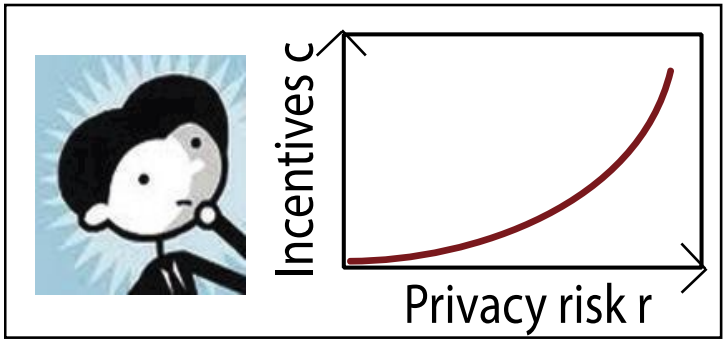
Stochastic Privacy

Probabilistic approach to sharing data

- Guaranteed upper bound on likelihood data is accessed
- Users offered/choose small *privacy risk* (*e.g*, $r < 0.000001$)
- Small probabilities of sharing may be tolerable to users
  - e.g., We all live with possibility of a lightning strike ☺
- Different formulations of events/data shared
  - e.g., Over time or volume of user data

Encoding user preferences

- Users choose *r* in some formulations
- System may offer incentives
  - e.g. Discount on software, premium service



Background: Online Services & Data

Accessing User Data

- Online services often rely on user data
  - Search engines, recommender systems, social networks
  - Click logs, personal data, e.g., demographics and location
- Personalization of services, increased revenue

User Consent

- Request permission or declare policy at sign-up / real time
- Typically: Binary response sought (yes/no)
  - *Terms of services* with *opt-in* or *opt-out*
  - Sets of questions: “May I access location to enhance services?”

    Response often: “Hmm…sounds good. Yes, sure!” ☺

Privacy Concerns and Approaches

Privacy Concerns and Data Leakage

- Sharing or selling to third parties or malicious attacks
- Unusual scenarios, e.g., AOL data release (2005)
- Privacy advocates and government institutions
  - FTC charges against major online companies (2011, 2012)

Approaches

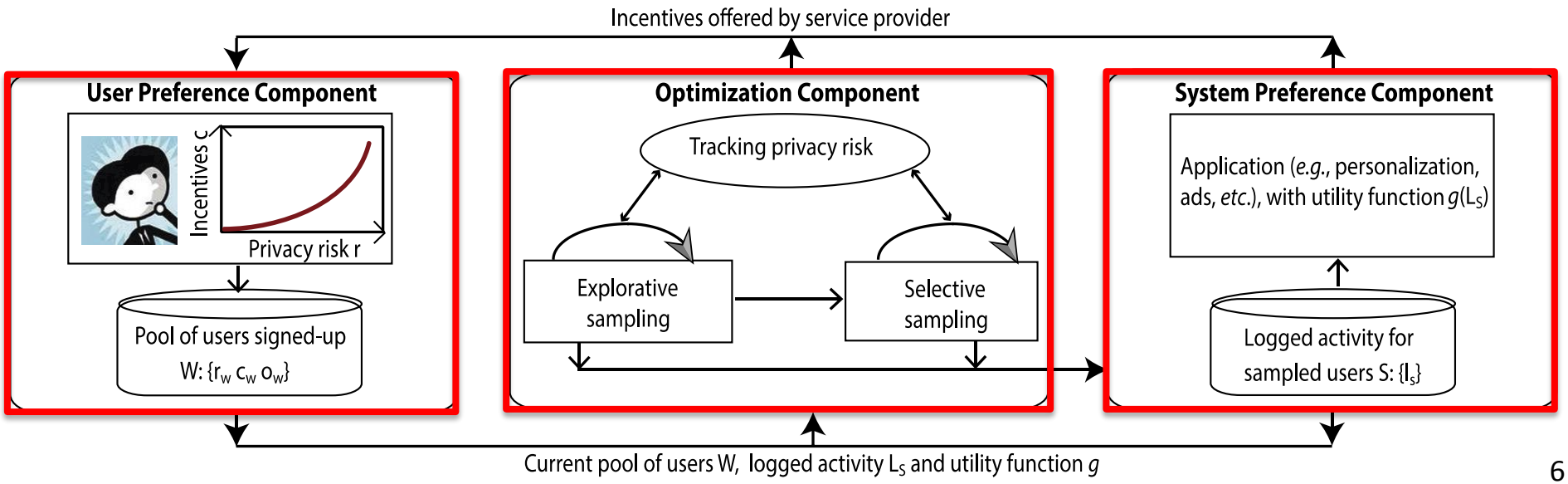
- Understand and design for user preferences as top priority
  - Type of data that can be logged *(Olson, Grudin, and Horvitz 2005)*
- Privacy—utility tradeoff *(Krause and Horvitz 2008)*
- Granularity and level of identifiability — *k-anonymity, differential-privacy*

Desirable Characteristics

- User preferences as top priority
- Controls, clarity for users
- Allow larger systems to optimize under the privacy preferences of users

Overall Architecture

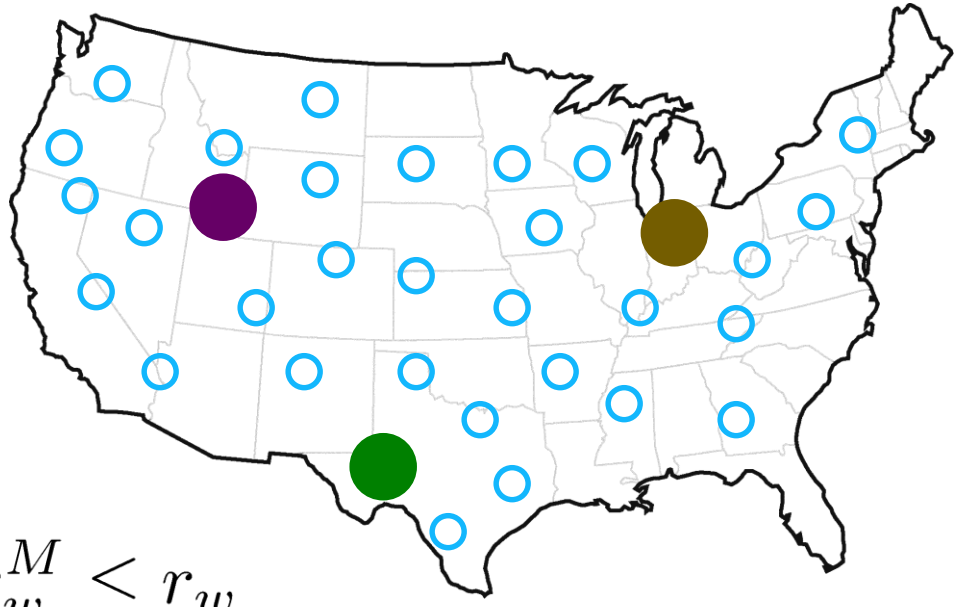
- **User preferences** – choosing risk *r*, offer incentives
- **System preferences** – application utility
- **Optimization** – user sampling while managing privacy risk
  - *Explorative sampling* (*e.g*, learning user's utilities)
  - *Selective sampling* (*e.g*, utility-driven user selection)
  - *Engagement* (optional) on incentive offers (*e.g*, engage users with option/incentive to take on higher privacy risk)



Optimization: Selective Sampling

Example: Location based personalization

- Population W
  - Privacy risk r_w, cost c_w
  - Observed attributes o_w
- Under budget or cardinality B
- Select set $S \subseteq W$
- Utility $f(S)$
- Risk of sampling by mechanism $r_w^M \leq r_w$



Class of Utility Functions

- Consider submodular utility functions f
- Capture notion of diminishing returns
  - Suitable for various applications such as maximizing click coverage
- Near-optimal polynomial-time solutions *(Nemhauser'78)*

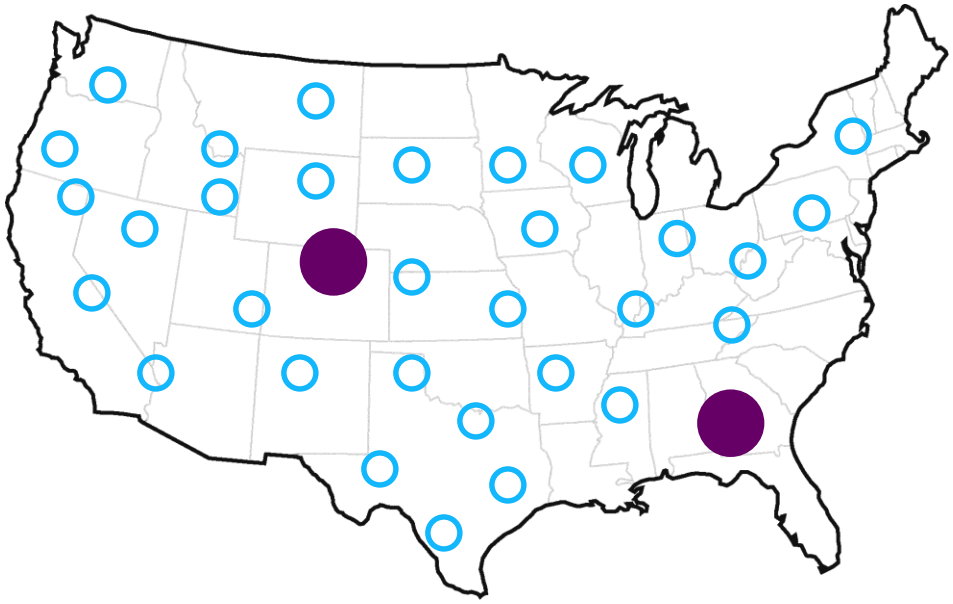
Sampling Procedures

- **Desirable Properties:**
  - Competitive utility
  - Privacy guarantees
  - Polynomial runtime
- **OPT:** Even ignoring privacy risk constraints: solution is intractable *(Feige'98)*
- **GREEDY:** Selection based on maximal marginal utility by cost ratio *(Nemhauser'78)*
- **RANDOM:** Selecting users randomly

	Competitive Utility	Privacy Guarantees	Polynomial Runtime
OPT	✓	X	X
GREEDY	✓	X	✓
RANDOM	X	✓	✓
RANDGREEDY	✓	✓	✓
SPGREEDY	✓	✓	✓

Procedure II: SPGREEDY

- Greedy selection followed by obfuscation in batches
- Iteratively builds the solution as follows:
  - Greedily selects s^*
  - Obfuscates s^* with "*similar*" users to create set $\psi(s^*)$
  - Randomly samples $\widetilde{s^*} \in \psi(s^*)$
  - Removes the entire set $\psi(s^*)$ for further consideration



Analysis: General Case

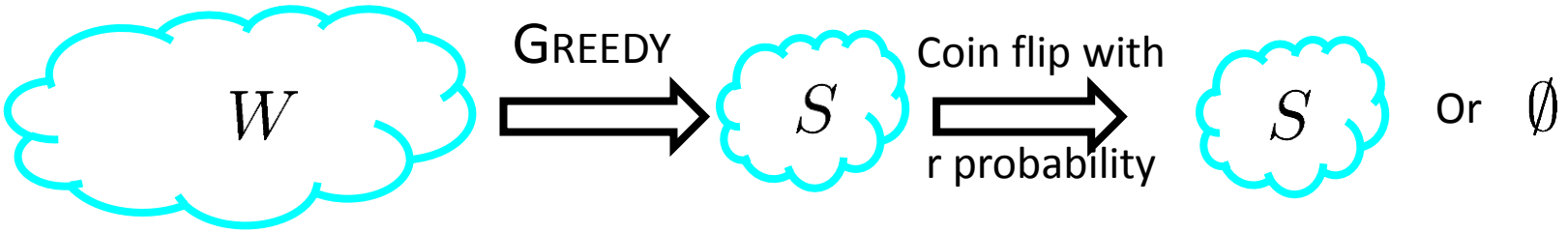
Worst-case upper bound

Theorem. *There exists an underlying distribution of marginal utilities for which no procedure can have expected utility of more than* $r \cdot f(OPT)$.

$f(W \setminus w^*) = 0$ $W \setminus \{w^*\}$ w^* $f(w^*) = 1$

Lower bound for any distribution

Theorem. *For any distribution of marginal utilities, a trivial procedure can achieve expected utility of at least* $(1 - 1/e) \cdot r \cdot f(OPT)$.



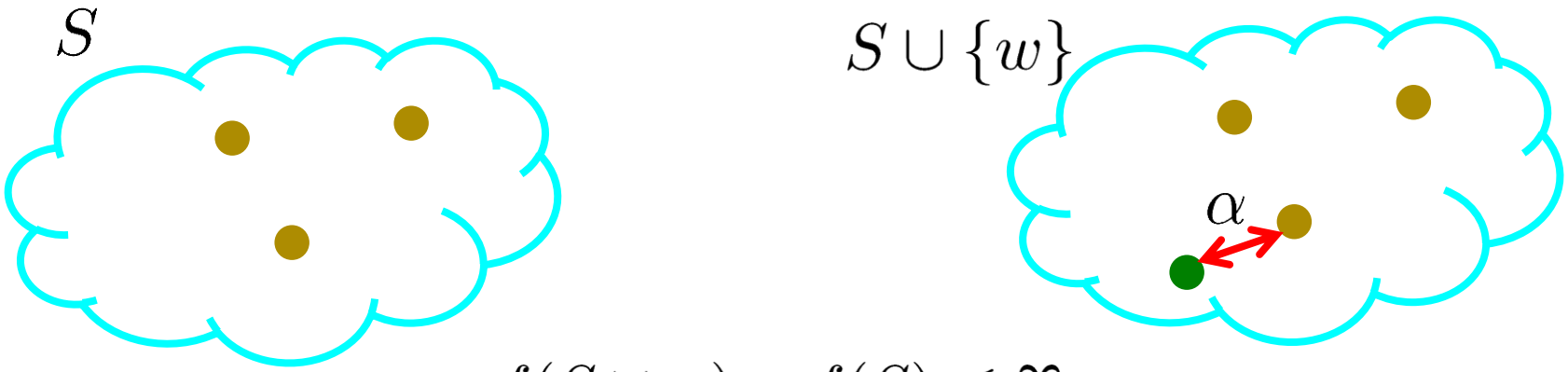
Can we achieve better performance bounds for realistic scenarios?

Additional Structural Properties

Smoothness λ_f

$$|f(S) - f(S_\alpha)| \leq \lambda_f \cdot \alpha \cdot |S|$$

Diversification Υ_f



$$f(S \cup w) - f(S) \leq \Upsilon_f \cdot \alpha$$

Main Theoretical Results

RandGreedy

Theorem. *Consider the utility function f with bounded λ_f. Let S^{Opt} be the set returned by* Opt *after relaxing privacy constraints. For a desired $\epsilon < 1$, let $\alpha_{rg} = \arg\min_\alpha\{\alpha : |N_\alpha(s)| \geq {}^1/_r \cdot \log({}^B/_\epsilon)\ \forall s \in S^{\text{Opt}},$ where $N_\alpha(s_i) \cap N_\alpha(s_j) = \emptyset\ \forall s_i, s_j \in S^{\text{Opt}}\}$. Then, with probability at least $(1-\epsilon)$,*

$$\mathbb{E}[f(\text{RandGreedy})] \geq (1 - {}^1/_e) \cdot \big(f(\text{Opt}) - \alpha_{rg} \cdot \lambda_f \cdot B\big).$$

SPGreedy

Theorem. *Consider the utility function f with bounded λ_f and Υ_f. Let S^{Greedy} be the set returned by* Greedy. *Let $\alpha_{spg} = \arg\min_\alpha\{\alpha : |N_\alpha(s)| \geq {}^1/_r\ \forall s \in S^{\text{Greedy}}\}$. Then,*

$$f(\text{SPGreedy}) \geq (1 - {}^1/_e) \cdot f(\text{Opt}) - 2 \cdot (\lambda_f + \Upsilon_f) \cdot \alpha_{spg} \cdot B.$$

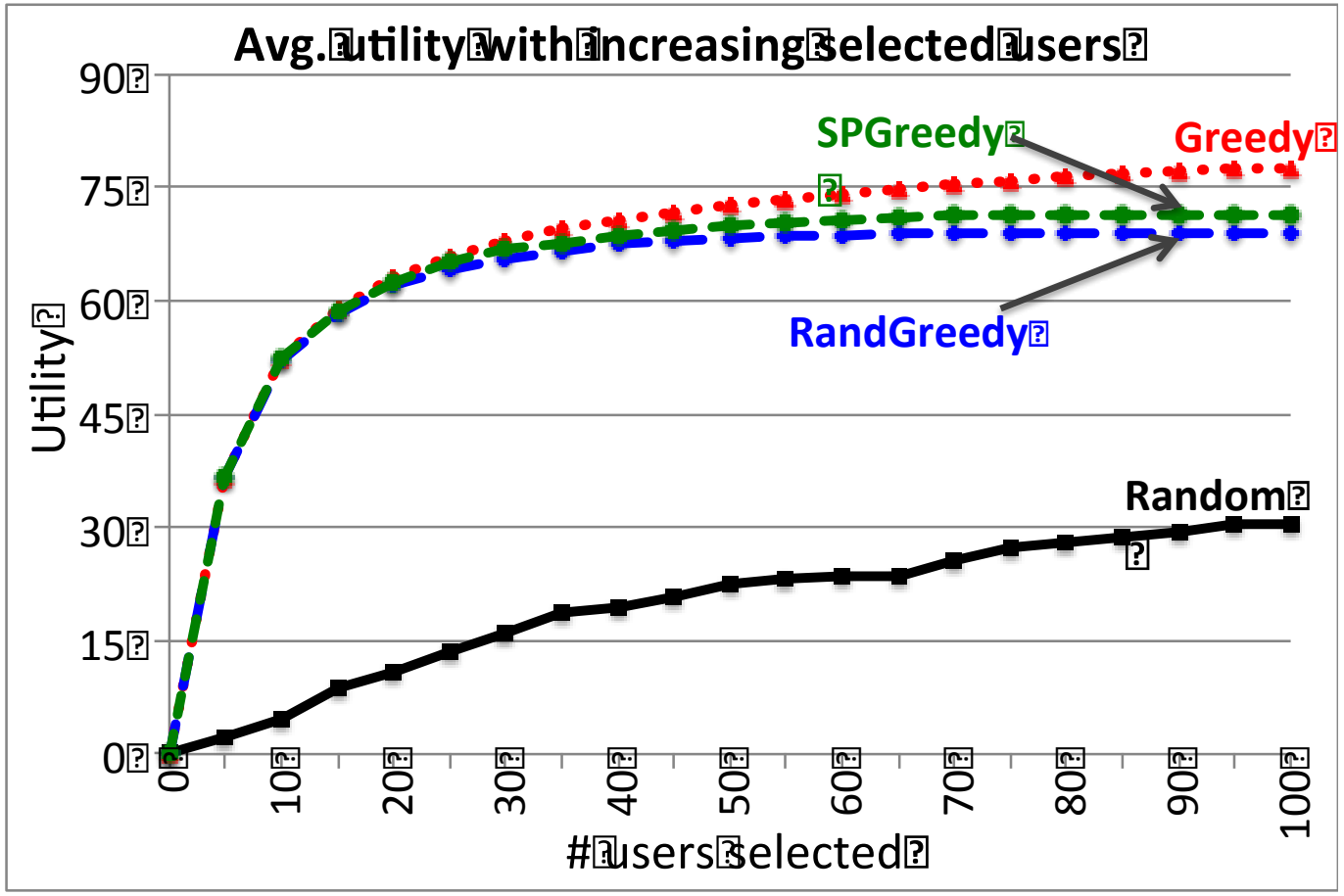
RandGreedy vs. SPGreedy

- RandGreedy doesn't requires additional assumption of diversification
- SPGreedy bounds always hold, whereas RandGreedy bounds are probabilistic
- SPGreedy has smaller constants in the bounds α_{spg} compared to α_{rg}

Case study of web search personalization

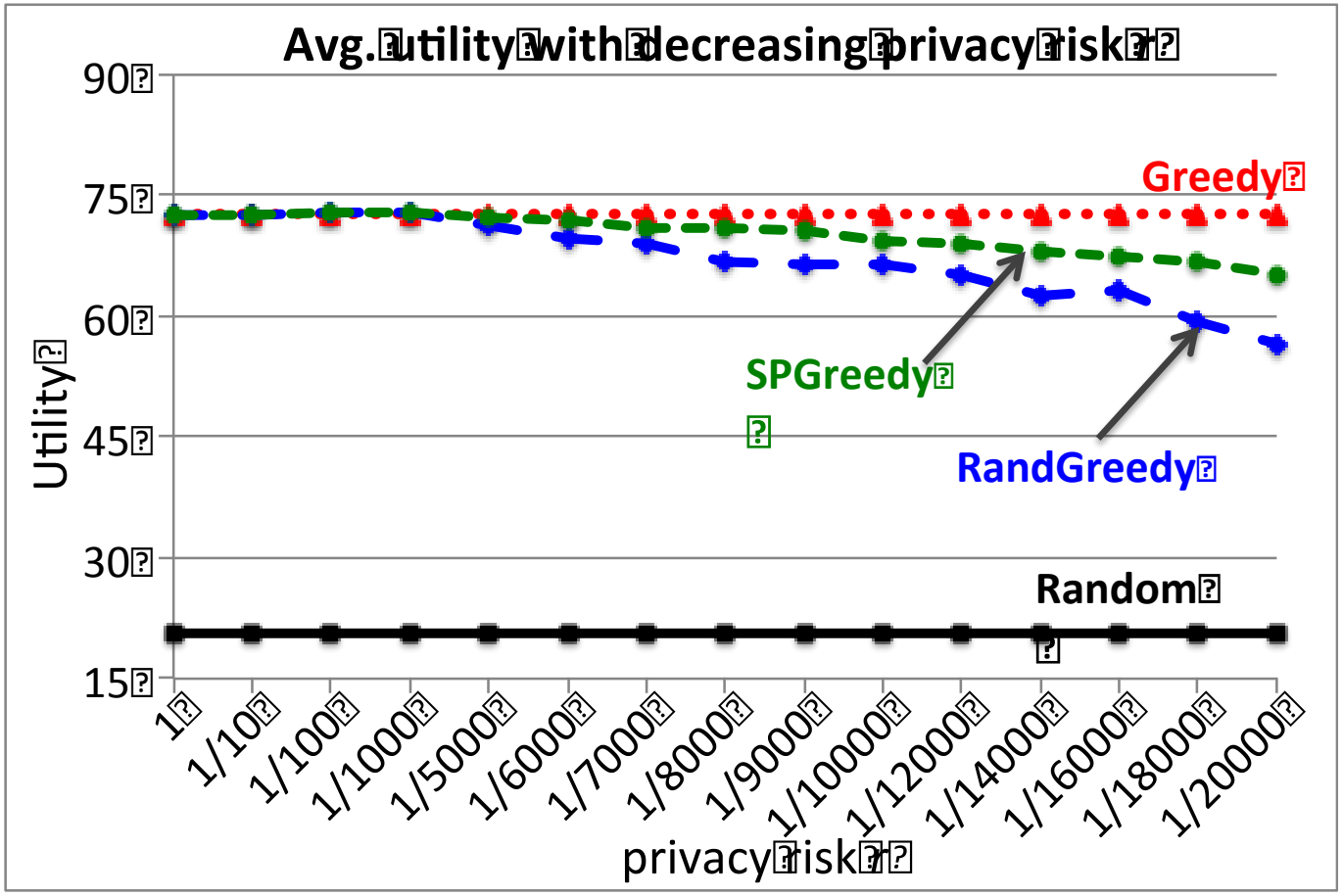
- Location based personalization for queries issued for specific domain
  - Business *(e.g.* real-estate, financial services)
- Search logs from Oct'2013, restricted to 10 US states (7 million users)
- Observed attributes of users prior to sampling
  - Meta-data about geo-coordinates
  - Meta-data about last 20 search-result clicks (to infer expert profile)
- ODP (Open Directory Project) used to assign expert profile *(White, Dumais, and Teevan 2009)*

Results: Varying Budget



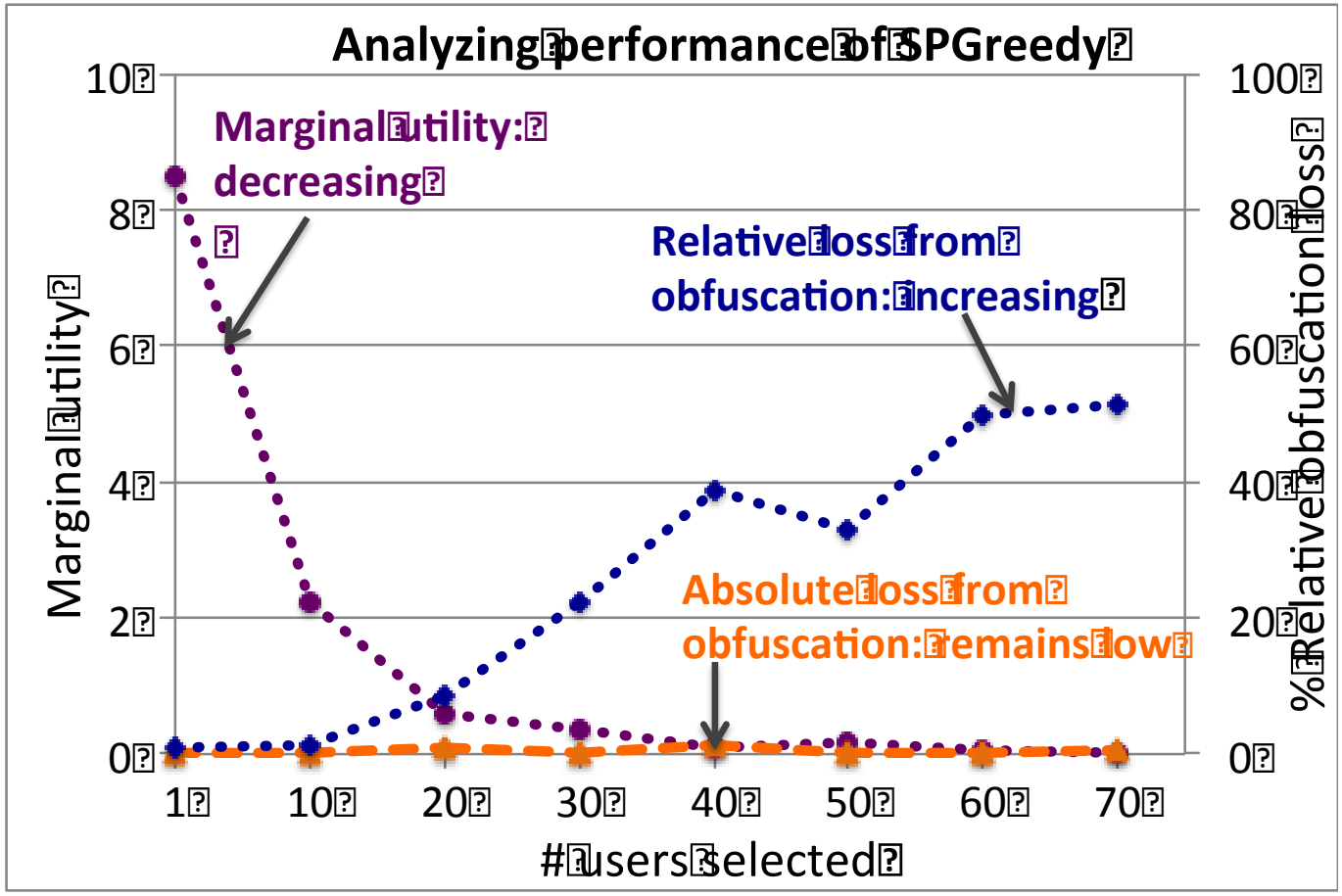
- Both RANDGREEDY and SPGREEDY are competitive w.r.t. GREEDY
- Naïve baseline RANDOM perform poorly

Results: Varying Privacy Risk



- Performance of both RANDGREEDY and SPGREEDY degrades smoothly with decreasing privacy risk (i.e. tighter sampling constraint)

Results: Analyzing Performance



- Absolute obfuscation loss remains low
  - Relative error of obfuscation increase
  - However, marginal utilities decrease because of submodularity

Summary

- Introduced *stochastic privacy:* Probabilistic approach to data sharing
- Tractable end-to-end system for implementing a version of stochastic privacy in online services
- Procedures RANDGREEDY and SPGREEDY for sampling users under constraints on privacy risk
  - Theoretical guarantees on acquired utility
  - Results of independent interest for other applications
- Evaluation of proposed procedures on a case study of personalization in web search