Intent-aware Query Obfuscation for Privacy Protection in Personalized Web Search

Wasi Uddin Ahmad
University of California, Los Angeles

Kai-Wei Chang
University of California, Los Angeles

Hongning Wang
University of Virginia

Motivation

- Personalization is everywhere

Previous Solutions

- Identifiability aspect of privacy
  - Secured communication, encrypted data storage

- Linkability aspect of privacy
  - Plausible deniable search
    - Submit proxy query instead of the true query
  - Obfuscation-based private web search
    - Submit cover-up queries along with the true query

Motivation

Do users submit isolated queries during web search?

Assumption

- Topics of search queries are sensitive
  - Indicate a user's (private) search intent

- All search query topics are sensitive
  - Leads to stronger privacy protection

Definitions

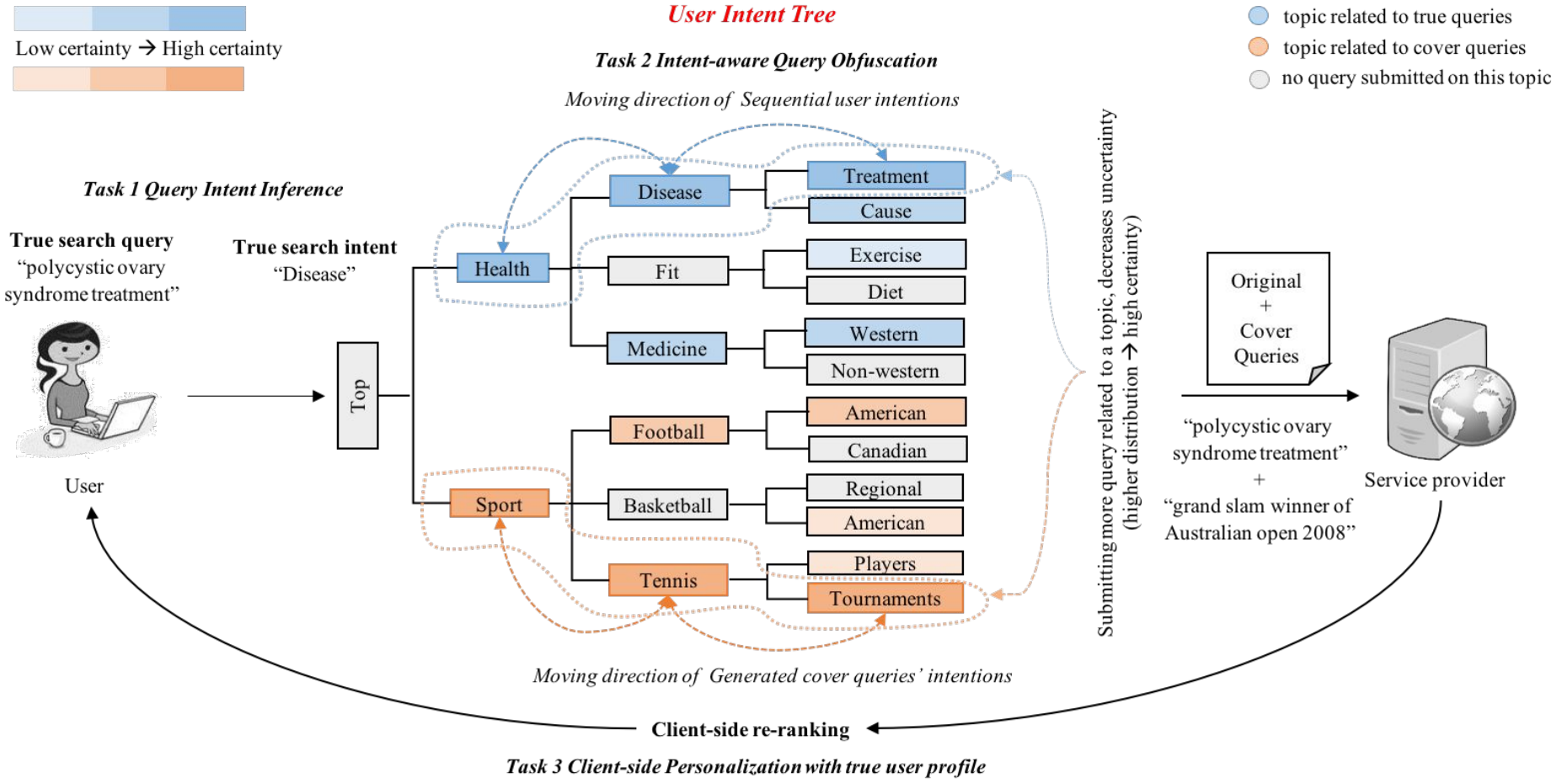
- User profile - a hierarchically organized tree where,
  - Each node represents a topic (a.k.a intent)
  - Each topic contains N-gram language models (LM)
    - LMs are approximated based on submitted queries and clicked documents
- Search task - A sequence of queries submitted in the same search session
  - Assumption: associated topics must form a sub-tree in the original topic tree

Main Idea

Intent-aware Query-obfuscation for Private-protection (IQP)

- Obfuscate search tasks to achieve task-level privacy
- Map a search task to a subtree of the intent tree
  - Intent tree: a predefined tree of topics
- Maintain the difference in prior and posterior belief of a search engine for true and cover search tasks

IQP Framework



Step 1: Query Intent Inference

- Query intent (a.k.a topic) is approximated using hierarchical language model
  - Hierarchical Dirichlet prior smoothing is performed
- Search intent is predicted by the maximum a posterior inference
- The prior of a topic is proportional to the #nodes in the subtree rooted at the topic node

Step 2: Intent-aware Cover Query and Click Generation

1. Select cover query topics
   a. Specificity of the true query intent
   b. Transition between previous and current query intent
2. Generate cover query
   a. Rejection sampling is utilized
   b. Conditioned on entropy difference between true and cover queries
3. Trained positional click model is employed to generate cover clicks

Step 3: Client-side Personalization

- Client-side reranking using an uncontaminated user profile
- Borda's method for rank aggregation
- Personalization score is computed based on client-side user profile
  - An estimated language model is utilized

$$UPScore(d) = \sum_{w \in q \cap d} \log \frac{(1 - \lambda) P_{ml}(w|d) + \lambda P(w|C)}{\lambda P(w|C)} tf(w)$$

Where, q, d and C denotes query, documents and client-side user profile respectively.

Example

- Session is sampled from AOL search log, 2006

Session queries	pregnancy symptoms in the first month abortion pills florida abortion clinics
Query topics	Home/Family/Pregnancy Society/Issues/Abortion Society/Issues/Regional
Cover queries	Low priced compact flash GPS receivers global hotel investment trends commercial real estate development
Cover query topics	Shopping/Consumer_Electronics/Accessories Business/Real_Estate/Property_Management Business/Real_Estate/Development

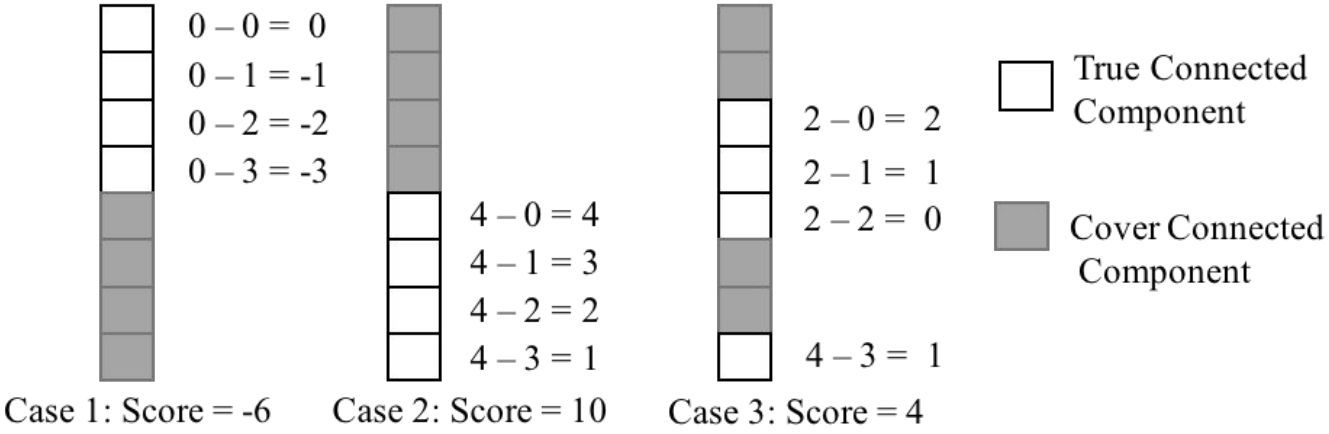
Measuring Task-level Privacy

- Prior works focused on query-level privacy evaluation metrics
  - KL-divergence, normalized mutual information etc.

- Proposed two new metrics to evaluate task-level privacy protection
  - Transition index (tIndex)
  - Confusion index (cIndex)

Confusion Index (cIndex)

- Measures search engine's belief of a user's search task
  - Search tasks are represented as a sub-tree
- Follows the entropy l-diversity principle
  - Quantifies the difference in prior and posterior distributions of the subtrees associated with true and cover tasks



Transition Index (tIndex)

- Measures task plausibility based on queries' concentration on the intent tree
- A predefined matrix representing transition of intents against the intent tree structure
  - States: {UP1, UP2, DOWN1, DOWN2, SA, MB, Others}
  - Estimated based on a reference search log
- Counts how many cover tasks are ranked ahead of true tasks
  - Score based on intent transition likelihood

Experiments

Data Sources

- Open Directory Project
  - 7,600 topic nodes up to level four
  - 82,020 web documents belonging to the nodes
- AOL search log, 2006
  - 1000 most active users
  - 318,023 testing queries
  - 0.96M web documents indexed
  - Clicked documents are considered as relevant

Experimental Setup

- Apache Lucene-based search engine
  - Ranking algorithm - Okapi BM25
- Server personalizes search result
  - Using language model estimated based on user profiles
  - Borda's method for rank aggregation
- Server returns the top 100 relevant documents
- Sessions are segmented based on 30-minutes inactive time threshold

Evaluation Metrics

- Mean Average Precision (MAP@100)
  - To evaluate ranking quality
- Kullback-Leibler (KL) Divergence
  - Computed between the true and noisy user profiles
  - Measures the effectiveness of privacy protection
- Normalized Mutual Information (NMI)
  - Computed between true and cover query pairs
  - Measures information disclosure by the cover queries

Baseline Details

- Plausible Deniable Search (PDS)
  - Latent semantic indexing to generate cover queries

Baseline Details

- Plausible Deniable Search (PDS)
  - Latent semantic indexing to generate cover queries
- Knowledge-based Scheme (KBS)
  - Cover queries from lexical ontology (WordNet, ODP tree)

Baseline Details

- Plausible Deniable Search (PDS)
  - Latent semantic indexing to generate cover queries
- Knowledge-based Scheme (KBS)
  - Cover queries from lexical ontology (WordNet, ODP tree)
- Topic-based Privacy Protection (TPP)
  - Sample cover query terms using LDA-based topic models

Baseline Details

- Plausible Deniable Search (PDS)
  - Latent semantic indexing to generate cover queries
- Knowledge-based Scheme (KBS)
  - Cover queries from lexical ontology (WordNet, ODP tree)
- Topic-based Privacy Protection (TPP)
  - Sample cover query terms using LDA-based topic models
- Embellishing Search Queries (ESQ)
  - Embellish user query by adding decoy terms

Baseline Details

- Plausible Deniable Search (PDS)
  - Latent semantic indexing to generate cover queries
- Knowledge-based Scheme (KBS)
  - Cover queries from lexical ontology (WordNet, ODP tree)
- Topic-based Privacy Protection (TPP)
  - Sample cover query terms using LDA-based topic models
- Embellishing Search Queries (ESQ)
  - Embellish user query by adding decoy terms
- Anonymizing User Profiles (AUP)
  - Hide individual user identity inside groups' identities

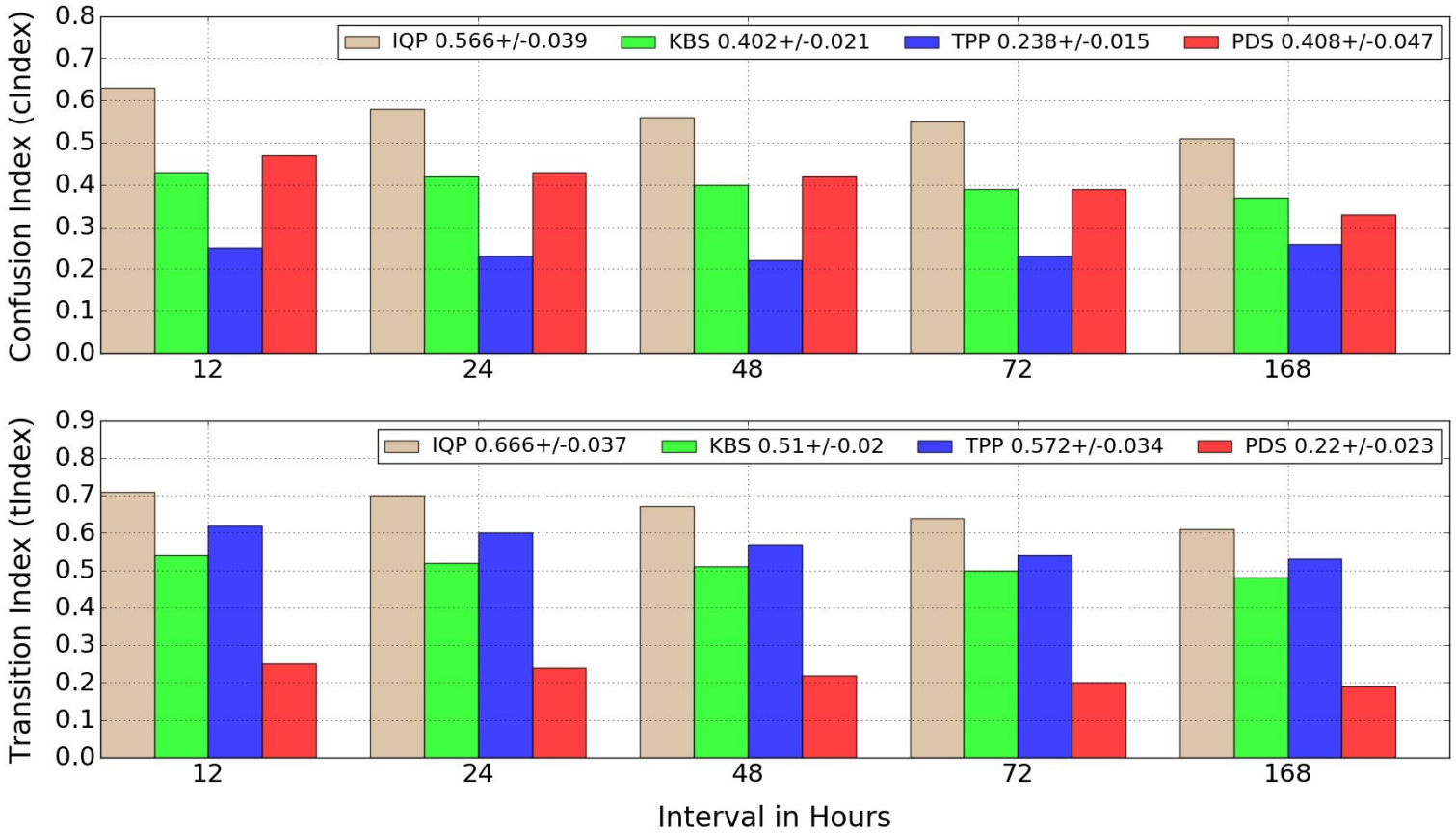
Comparison with Baselines

Model	MAP@100	MAP@100 [client-side personalization]	KL Divergence	NMI
No cover queries				
AUP	0.1088	0.1171	0.9636	
ESQ	0.1161	0.1090	0.0912	
Number of cover queries = 2				
IQP	**0.1387**	**0.1486**	0.6866	**0.2156**
TPP	0.1158	0.1174	**0.7558**	0.3922
PDS	0.1307	0.1391	0.4467	0.4308
KBS	0.1255	0.1474	0.7001	0.2914

* Detailed results can be found in the paper.

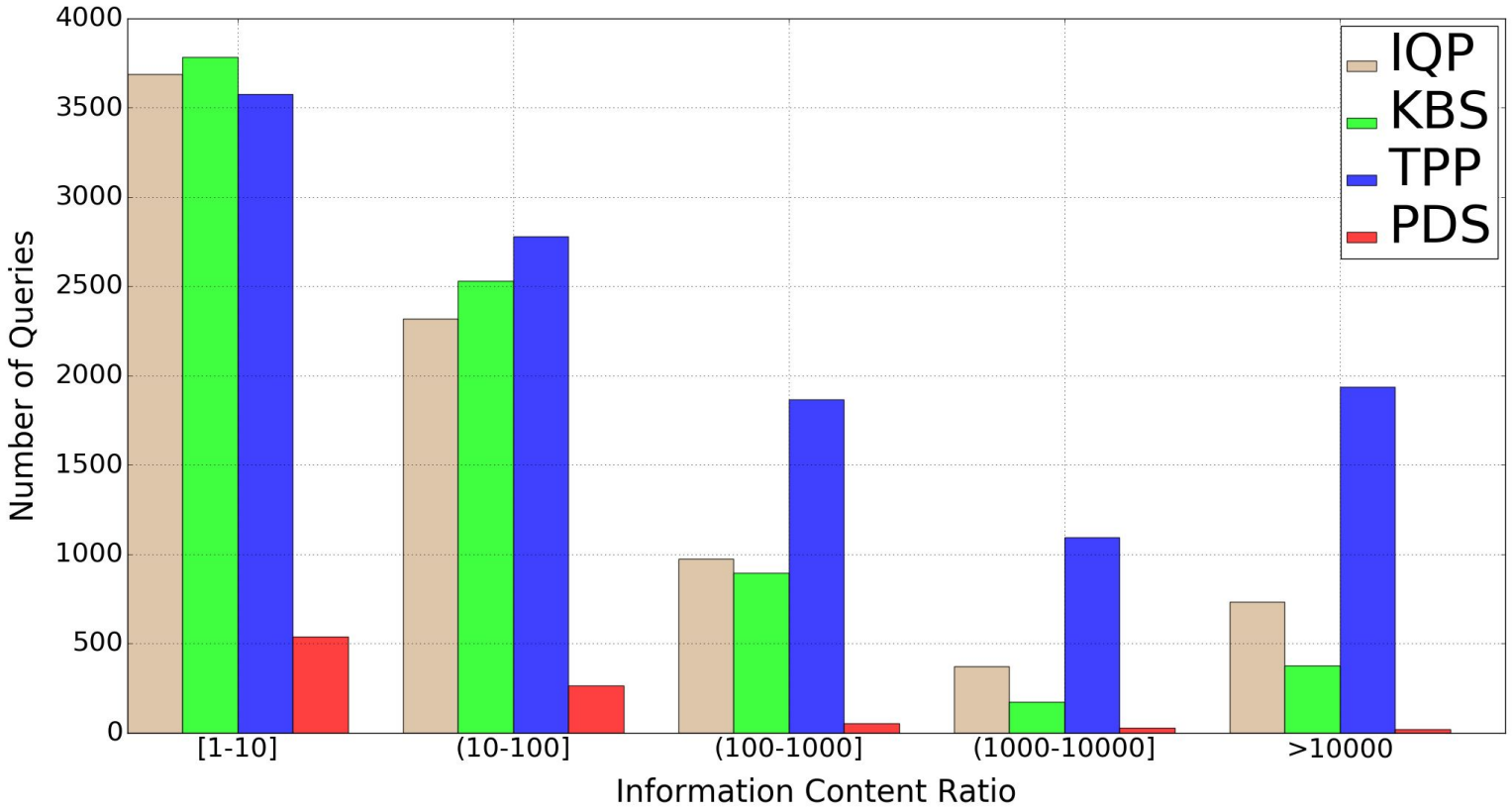
Measuring Task-Level Privacy Protection

- Compares in-session true task and cover task



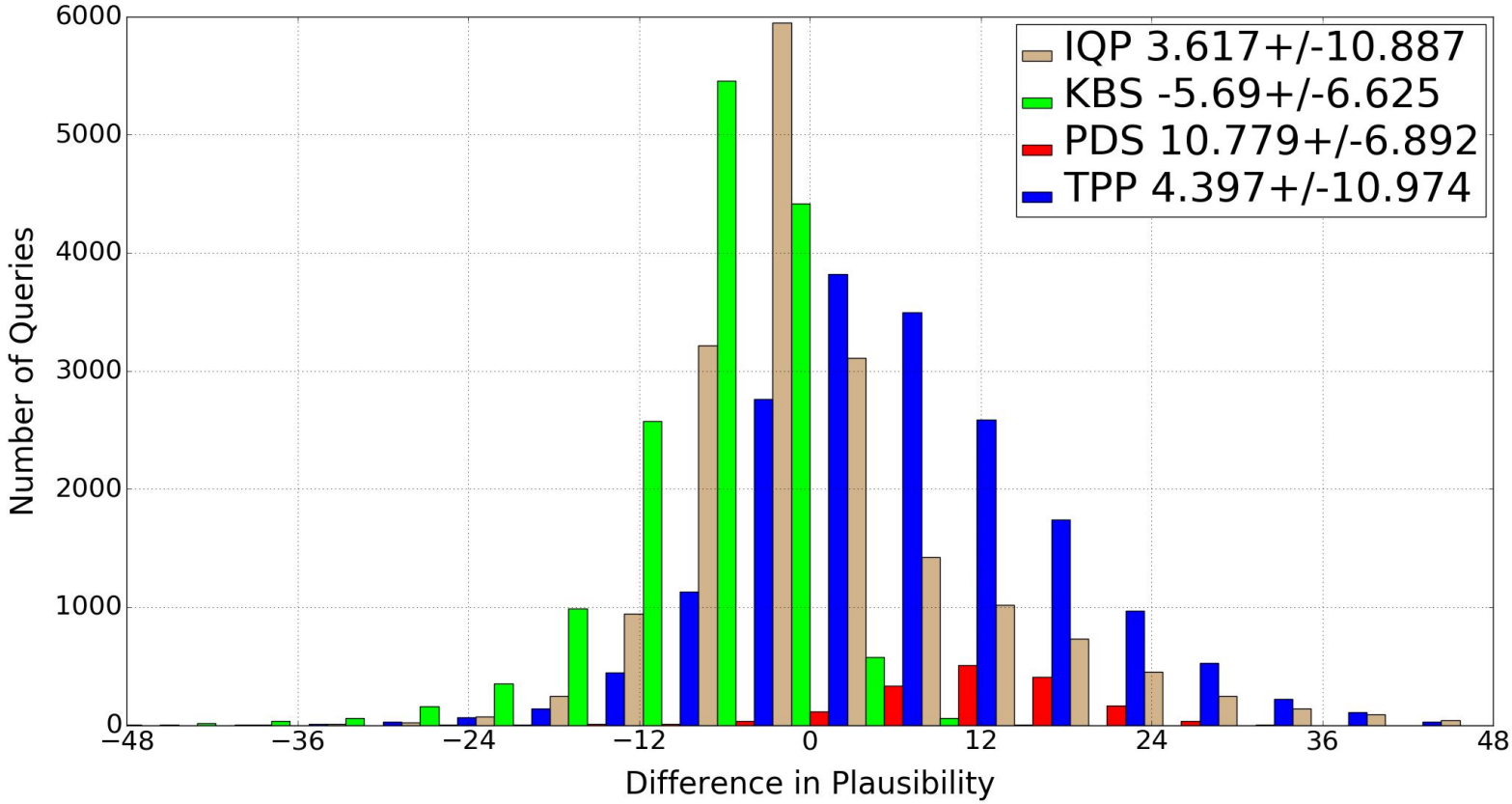
Statistical Query Plausibility

- Measures the ratio of search result hits for a query pair
  - Microsoft Bing API to get the hit count



Statistical Query Plausibility

- Compare true and cover queries at web-scale
  - Microsoft Web Language Model API



Conclusion and Future Works

- Intent-aware query obfuscation solution
  - Handles sequentially developed intents in search tasks
- Proposed two new metrics measuring task-level privacy disclosure
- Future Works
  - Adaptively adjust the number of cover queries
    - Relaxing the assumption that all queries are equally sensitive
  - Perform user studies
    - Understanding real user's satisfaction of privacy protection solutions

Thank You