Applying Differential Privacy to Search Queries in a Policy Based Interactive Framework

Palanivel Kodeswaran*
University of Maryland, Baltimore County (UMBC)
1000, Hilltop Circle
Baltimore, MD 21250
palanik1@cs.umbc.edu

Evelyne Viegas
Microsoft Research
One Microsoft Way
Redmond, WA 98052
evelyne v@microsoft.com

ABSTRACT

Web search logs are of growing importance to researchers as they help understanding search behavior and search engine performance. However, search logs typically contain sensitive information about users and therefore considerable caution must be exercised when considering releasing the logs to the research community. Current approaches to releasing search logs focus on either protecting the privacy of users or enhancing the utility of data to researchers. In this work, we address the privacy-utility tradeoff by providing safe access to search logs, instead of releasing them. We propose a policy based safe interactive framework built on semantic policies and differential privacy to allow researchers access to search logs, while maintaining the privacy of the users. Semantic policies are used to infer the higher levels of information that can be mined from a dataset based on the fields accessed by a researcher. The accessed fields are then used to build research profile(s) that guide the amount of privacy to be enforced using differential privacy. We show the additional utility that can be obtained in our framework by two demonstrative experiments that involve access to user level information. Our results indicate that valid research can be conducted in our framework without forgoing the privacy of individuals.

Categories and Subject Descriptors

H.1.m [**Information Systems**]: Miscellaneous

General Terms

Management, Measurement, Documentation, Experimentation, Security.

Keywords

Privacy, Policy, Semantics.

1. INTRODUCTION

Search query logs are of growing importance to academic researchers interested in a variety of fields such as studying user search behavior, search engine performance. However, these logs are currently closed behind the vaults of large corporations as there are currently no well accepted practices for sharing such data with academic researchers. Sharing search logs needs to be a cautious exercise, since logs contain user information and, therefore could potentially undermine the privacy of the individuals in the dataset.

Existing approaches to releasing query logs (after anonymisation and appropriate sanitization) to the research community follow a non-interactive access framework and are subject to data breaches. Once the data is released in the public domain, the data owners have no control over the use of the dataset, and privacy violations have been reported in the media in the past [9]. Other approaches have given higher priority to privacy and restricted the kinds of information released in these datasets. For example, the Microsoft Request for Proposal (RFP) dataset had no user id released, and the data set was protected via a licensing agreement whose terms bound the recipient of the data to not share the data set. However, these approaches restrict the kinds of research that can be performed with these datasets, and researchers have expressed their need for access to user fields such as user id and IP address to broaden the scope of research questions that can be explored by the research community, as discussed in [25].

In this paper, we focus on balancing both the needs of the researchers to pursue scientific research as well as the privacy of individuals in the dataset. Unlike existing approaches, we propose an interactive access framework through which researchers can now query the system for data that they need rather than being provided with a one-size-fits-all anonymised dataset to start with. We argue that the interactive access paradigm benefits both data owners and researchers as follows. In the interactive framework, data owners never release the raw data to researchers and therefore are always in control of their data. Access control can easily be enforced in this framework compared to the non-interactive framework where data owners lose control over the data once it is released. The researchers too are bound to benefit from the interactive framework in that it is flexible as they can now issue queries over all the fields of the dataset rather than having to work with sliced up datasets that cannot support the entire range of research questions.

PAVLAD'09, November 6, 2009, Hong Kong, China.
Copyright 2009 ACM 978-1-60558-804-9/09/11...$10.00.

*Work done during internship at Microsoft Research

Our interactive framework is built using

1. *Semantic Policies* that allow expressing the data owner policy in terms of the higher levels of knowledge that can be inferred from the dataset
2. *Differential Privacy* which enables tuning the amount of privacy guaranteed

The main contributions of this work can be summarized as

1. Proposing an interactive framework to safely access sensitive information
2. The first (as far as the authors are aware) practical demonstration of differential privacy to the study of search logs showing the utility of differential privacy to do scientific research on sensitive data.

2. PRIVACY IN SEARCH QUERY LOGS

Laws abound to protect the privacy of individuals. However, there are no universally accepted definitions of privacy. In this section, we describe our model/interpretation of privacy in search query logs. Search queries can be interpreted to represent an individual's intentions and information needs and could thereby reveal a large amount of personal information about the individual. For example, most users search for their own names (vanity queries), pornography, alarming terms such as murder, specific health conditions as described in [10]. These queries if released could potentially raise suspicions about the user and therefore could be classified as sensitive data from the individual's perspective. If any of the alarming terms are searched in the same session as a vanity search, we may in fact be able to identify the searcher thereby threatening her privacy.

In this work, we focus on protecting the privacy of individuals in search query logs, which contain both explicit and implicit private information about individuals. In general, data about individuals can be classified as

PII (Personally Identifiable Information): Any piece of information that can uniquely identify an individual such as, Social Security Number, Credit Card number, email address, Full Name if not common. Following the HIPAA [1] safe harbor guidelines, search logs should not reveal PII.
Not PII: Data that is common to a reasonably large subpopulation such as gender, zipcode.
Inadvertent PII: Multiple pieces of Not PII which when joined together and with additional data processing can identify an individual. For example, " Daily Planet reporter working in Metropolis" provides enough implicit information to uniquely identify the individual.

Search queries which are essentially free text can be classified as inadvertent PII and need to be sanitized before releasing them. Current approaches involve removing PII, such as removing phone numbers, social security number, either manually or through automatic techniques. Unfortunately, these approaches are ad hoc and are performed at the expense of enabling scientific research.

3. RELATED WORK

There is a large body of work devoted to the study of query logs [18] looking at user intent [17], query format [21], how users issue search queries [13], creating a taxonomy of web search [3], human interaction with web search engines [20], and [4] for a survey of approaches.

One of the well known approaches for privacy preserving data release of structured records is k-Anonymity [22]. The basic idea is that each record should be indistinguishable from at least k-1 other records in the released data set. Computationally, K-anonymity has shown to be NP-Hard [2]. Furthermore, the effect of k-anonymity on the utility of data is not clearly understood [15]. Also, it is not yet clear how to ensure that the search history of one user is exactly identical to that of k-1 other users. Achieving K-anonymity typically involves generalization and suppression of cells in the dataset. Generalization works well for situations where there is a well defined domain hierarchy such as generalizing male and female to person. However, there is no well defined hierarchy for generalizing query terms across queries and users consistently.

There has also been work in the field of anonymising network logs for the purpose of sharing between network administrators. Slagell et al. [19] propose using multiple levels of anonymisation for different fields of network logs based on what the user/enterprise intends to reveal. These approaches generally exploit the inherent structure and semantics of fields present in network logs. For example, IP addresses have specific formats and interpretations that make them amenable to anonymisation such as releasing only the first 16 bits of an IP address. However, these approaches are not directly applicable to search query logs since search terms are essentially free text with structure and semantics yet to be formalized. In [23] Pang et al. present a high level programming environment for packet trace anonymisation with support for a variety of tasks such as changing HTTP/SMTP headers, retaining the anonymised payload data.

There has recently been active interest in privacy preserving approaches to releasing search query logs. Xiong et al. [15] present an overview of query log analysis applications as well as various granularities of releasing log information and their associated privacy threats. The authors in [14] point out the unsuitability of token based hashing for query log anonymisation. Token based hashing involves tokenizing the search string and securely hashing each token to an identifier. The authors argue that an adversary having access to a reference query log could exploit the statistical properties of the co-occurrence of terms within search queries to invert the hash function and thereby reveal the underlying search strings.

Given that the search queries themselves contain sensitive information that can lead to re-identification, approaches to query release cannot include a user id, even anonymised, to prevent re-identification of individuals from the search queries. However, the lack of a user id across sessions prevents researchers from pursuing questions that involve discovering individuals' query trends (without actually identifying any particular individual) such as the number of individuals searching for a particular product over time. Our approach differs from the above approaches in that we propose an interactive access framework that does not directly depend on anonymisation for privacy, but rather is based on semantic policies and differential privacy.

4. POLICY BASED SAFE INTERACTIVE ACCESS FRAMEWORK

We propose a safe interactive access framework for accessing search query logs based on

1) Semantic Policies
2) Differential Privacy

The amount of privacy enforced by the framework is governed by the "privacy threat" of the research questions asked. We use semantic policies to model higher level abstractions of knowledge that can be inferred by a researcher based on the fields accessed in the search logs. Differential privacy is used as the privacy enforcement mechanism to translate the semantic policies into appropriate privacy guarantees.

4.1 Semantic Policies

Dwork [5] proved that in the general case semantic security is unattainable since the adversary could possess auxiliary information over which the data owner has no control. Consequently, we take a policy based approach in which the semantics of the data is used to control the amount of privacy to be enforced. Different fields in a dataset vary in their privacy revealing property. For example, in query logs the user id field is probably more privacy revealing than the timestamp of when a search is performed. Data owners who are aware of the privacy semantics of the fields and combinations thereof can specify the amount of privacy to be enforced based on the fields accessed by the researchers. In our system, we use semantic policies to specify

1) Higher level abstractions of knowledge such as user level trends, temporal trends
2) The utility of each field in the dataset to the above defined abstractions/trends
3) High level abstractions that can be inferred from the dataset based on the fields accessed

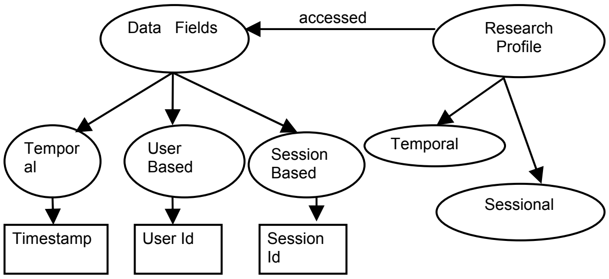


Figure 1. Semantic policies specify higher level knowledge that can be inferred based on the fields accessed

Every researcher who requires access to the data is assigned a research profile based on the fields accessed. The research profile in a way represents the higher level abstractions of knowledge that can be mined by the researcher based on the fields she has access to. For example, a researcher who has access to the "session id", "timestamp" and "queries" fields, could possibly learn session level patterns such as the average number of queries per session, time interval between successive queries in a session.

We enable this nature of semantic reasoning by mapping each field in the dataset as belonging to a higher level knowledge class. For example, in Figure 1, the timestamp field is an instance of a temporal field; session id is an instance of a session based field. Our approach offers extensibility in that the research profiles which are specified at the higher level knowledge abstraction are not directly affected by changes in the lower level fields in the dataset. Thus adding a new field such as IP address involves only mapping the field to the appropriate knowledge class such as Locational, User Based. The research profiles themselves remain the same and are unaffected by the addition or deletion of fields in the underlying dataset. Thus, the same profiles could be used across multiple data sets as long as the research profiles remain consistent with the underlying data set.

We use the research profile thus constructed to set appropriate privacy guarantees as specified by the data owner. For example, the data owner may specify that User Level patterns are more revealing than Session level patterns, and hence be guaranteed higher privacy.

Our semantic policies differ from traditional authorization and privacy policies such as SecPAL [16], P3P [11] in that the intent is to express higher level semantic information rather than focusing on file/data level access control.

4.2 Differential Privacy

In this section we provide an overview of differential privacy [5] and our motivation behind using it as the policy enforcement mechanism in our framework. Differential privacy works along the lines of statistical databases and returns only numerical aggregates and thereby, the underlying raw data is not released in public as such. Unlike privacy preserving mechanisms such as k-anonymisation which attempt to guarantee absolute privacy, differential privacy focuses on reducing the increased privacy risk of an individual participating in a dataset. In other words, differential privacy addresses the change in privacy rather than absolute privacy. This definition of differential privacy enables us to tune the amount of privacy that is guaranteed based on the semantic policies described in the previous section. Our results indicate that in most cases, accurate results about the dataset can be released without threatening the privacy of the individuals in the underlying dataset.

Differential privacy essentially guarantees that the presence or absence of an individual does not greatly alter the resulting output distribution. Mathematically, for any random computation k, ε-differential privacy guarantees

$$\Pr[k(D_1) = X] \leq \Pr[k(D_2) = X] \times e^{\varepsilon}$$

for all datasets D_1 and D_2 differing by at most one element. For example, assuming D_1 to be the database including a particular individual and D_2 to be the database without the individual, differential privacy guarantees then that the outputs of a computation such as average height on both datasets does not differ from each other by more than e^{ε}, where ε is tunable. With small values of ε, the difference between the two outputs becomes negligible making it impossible for an adversary to guess whether the output corresponds to D_1 or D_2. Thus differential privacy guarantees an individual's privacy by masking their presence or absence in the dataset.

Intuitively, differential privacy can be viewed as adding random noise to the real output of a computation before returning the result to the user. The magnitude of the noise added determines the privacy-utility tradeoff. Larger the noise, higher the privacy but lower the utility. Also, the noise added is independent of the underlying dataset and depends only on the query function.

In our system, the semantic policies specified by the data owners are used to set the ε in differential privacy, thereby controlling the amount of privacy guaranteed**.**

5. SYSTEM ARCHITECTURE

Figure 2 presents an overview of our system architecture. In this section, we describe the building blocks of our architecture and how they are tied together in our framework.

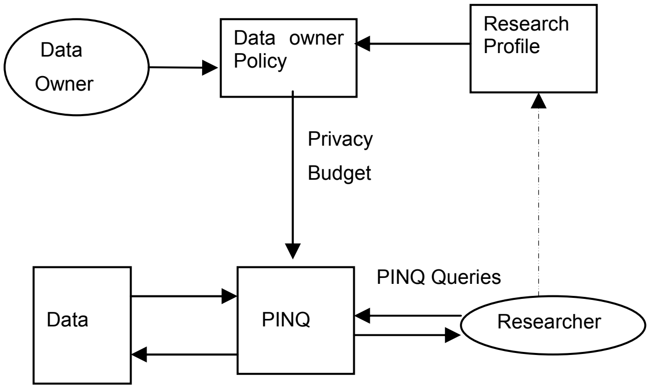


Figure 2. System Architecture

Our system is based on the interactive access paradigm wherein researchers query the system as opposed to the non-interactive paradigm in which data as a whole (after filtering and anonymisation) is released to the research community. The various building blocks in our system are as follows

DATA: The data that needs to be protected. In our case, this refers to the search query logs.

PINQ: The Privacy Integrated Queries (PINQ) framework is an implementation of differential privacy developed at Microsoft Research [24]. PINQ handles all the underlying mathematics of differential privacy. All programs written in PINQ guarantee differential privacy. PINQ supports a number of aggregation operations such as NoisyCount, NoisySum, NoisyAverage. PINQ aggregate operators take $\mathcal{E}$ as an argument which determines the amount of noise that is added, and thereby the accuracy of the returned results. A sample PINQ query looks as follows

```
Count = SearchData.select(x=>x.Split('\t')).
              .Where(x=> x[1].Equals("Britney
Spears"))
              .NoisyCount(0.1);
```

To enforce the differential privacy semantics of various operators such as where, join, group by, the researcher supplied ε is internally amplified by PINQ before applying to the underlying data.

PRIVACY BUDGET: PINQ also allows data providers to specify their data access policy. This is enabled through the privacy budget mechanism. Data providers can specify the maximum allowed accuracy of the returned results by appropriately setting the ε in the privacy budget. When a user issues a query, PINQ consults the privacy budget (after appropriate amplification of ε) to determine whether the query is acceptable. Only if the privacy budget constraints are satisfied i.e. the supplied ε is within the bounds of the privacy budget does PINQ perform the user query.

In our framework, the privacy budget is set by the data owner to reflect the sensitivity of the data. For example, the data owner could specify that user level trends are more revealing that session level trends and therefore are set a lower privacy budget. A lower privacy budget ensures higher privacy. The data owner policy in SPARQL [12] would be expressed as

```
construct{?s ex:hasPrivacyBudget 0.1}
where {
            ?s rdf:type ex:QueryLevel.
          ?s rdf:type ex:UserLevel
}
```

Whenever a researcher issues a query, based on the sensitivity of the query, PINQ decreases the privacy currency from the researcher's budget. When the budget reaches zero, no more queries from the user will be answered. The researcher could either ask a single query with high accuracy or multiple queries with lower accuracy. Thus, we can control the amount of information that can be learned from the dataset.

RESEARCHER: In our framework, we require that the researcher be authenticated prior to data access. In addition to the sensitivity of the data, the privacy budget could also be determined based on the organization of the researcher to reflect institutional agreements.

RESEARCHER PROFILE: Every researcher having access to the data is assigned a researcher profile based on inference from the fields to be accessed by the researcher. For example, we specify that a research profile is of type SessionLevel if it has access to a field of type SessionBased as follows

```
construct {?s rdf:type ex:SessionLevel}
where {?s ex:hasAccessTo ?o .
 ?o rdf:type ex:SessionBased .
}
```

DATA OWNER POLICY: Semantic policies specified by data owner to set the privacy budget.

6. EXPERIMENTS

6.1 Data

We used the following two datasets obtained from Microsoft web search logs for our experiments.

Request for Proposal (RFP) Dataset: Microsoft's RFP [7] dataset consists of about 15 million search queries sampled over one month. The data was filtered to remove PII, and was provided to the winners of the RFP under a limited licensing agreement. Also, more importance was provided to privacy and no user id (even anonymised) is maintained across the logs preventing association of search queries to individuals in the dataset. For each query the published attributes include session ID, timestamp, query string, number of results returned, and results page number. For each result clicked, the released data included the URL, associated query, position on results page and timestamp.

Ad Centre Data: To show how our framework could be used to enhance research without leaking private information we used another data set which, in addition to the search query, included anonymised user id. The dataset consisted of about 5.7 million entries with around 1 million distinct anonymised user ids. It is important to understand that we accessed the data through PINQ and as such researchers do not view any instance of sensitive data.

In all experiments, we used Protégé [6] as the policy editor and the Intellidimension Semantics SDK [9] for specifying semantic queries in SPARQL [12].

6.2 Metrics

One of the goals of our experiments was to show that scientific research could be performed in our framework and that the utility of the data was not compromised. Since utility metrics are not generic and are application specific, we attempt to develop a simple generic metric that can reflect the error characteristics of the data obtained in our framework. The average query error is defined as

$$\text{Average Query Error} = \sum_i \left|(R_i - P_i)\right| \text{ / No. of data points}$$

where, R_i - value of ith data point in real dataset

P_i - value of ith data point returned in our framework.

6.3 Framework Validation

In these experiments, we replicate an experiment from the RFP. The goal is to validate that scientific research can be performed in our framework yielding valid results, and empirically understand the ε-to-noise mapping.

6.3.1 Variation of URL Rank

The experiment focuses on understanding the change in the ranking of a URL in the search engine results for a one month time period, using the data from the RFP directly versus accessing the data via PINQ. For our experiments, we consider the URL as http://www.yellowpages.com and “yellow pages” as the corresponding query. Figure 3 shows the number of times in the real dataset (i.e. without PINQ), http://www.yellowpages.com occurs in Position 1 and Position 3 in the returned result set for the query “yellow pages”.

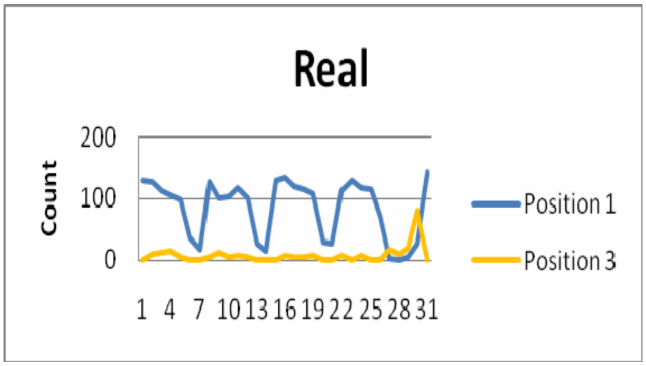


Figure 3. Position vs. day of month using real data

Figure 4. shows the results obtained in our framework with ε equal to one in PINQ’s aggregate operator NoisyCount. From figures 3 and 4, it is clear that the trends in the real data are retained in our framework and scientific research can be performed. We experimented with different values of ε in our framework and obtained the results as shown in Figures 5 and 6. We show these results to illustrate that the underlying distribution is retained in our framework not only among the categories (positions in the result set) but also within a category.

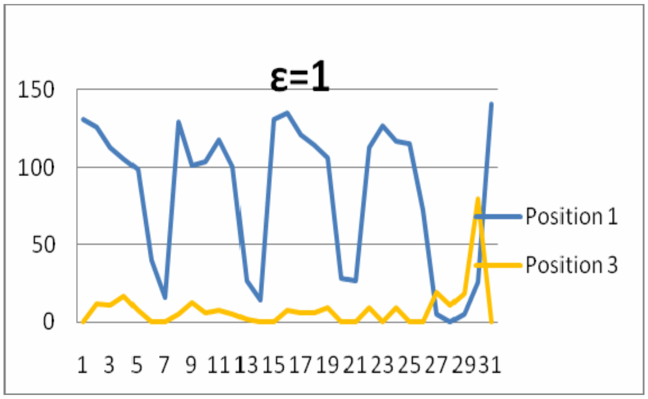
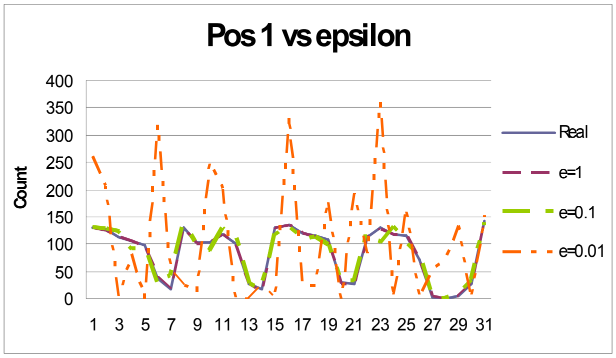


Figure 4. Position vs. day of month in our framework with ε=1



p

Figure 5. No. of times URL was returned in position 1 for a one month period in our framework

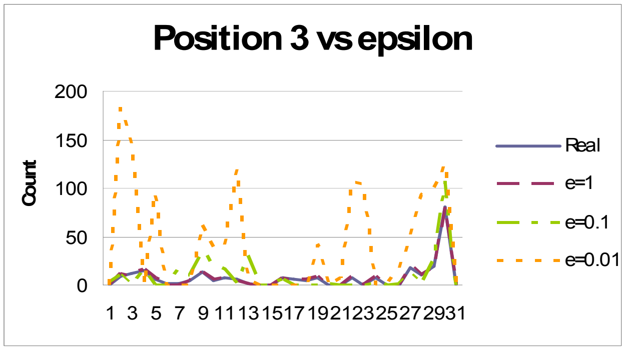


Figure 6. No. of times URL was returned in position 3 for a one month period in our framework

6.3.2 Empirical evaluation of the effect of ε on the accuracy of the returned results

One of the goals of this experiment was to understand empirically the ε-to-noise mapping. Based on the above results, we conclude that for ε > 0.01 the data returned by our framework closely reflects actual data and hence scientific research can be performed with high accuracy. For values of ε >1, the returned results are very close to the actual data, and may be privacy threatening. On the other hand, for ε ≤ 0.01, higher privacy is guaranteed; however the resulting data distribution may not be representative of the actual data although major trends may still be reflected.

6.4 Utility Experiments

Now that we had established that research yielding valid results could be performed, we wanted to show the increased utility attainable in our framework. Since utility metrics are not generic and are highly application specific, we resort to qualitatively demonstrating the increase in utility. Our qualitative approach essentially involves demonstrating that sensitive fields like user id, IP address can be accessed in our framework and research can be performed on the data containing them. The increased utility stems from this very property that data once considered sensitive and which could never be released to the public, can now be accessed by researchers and valid research performed, all the while maintaining the individual's privacy. In the following experiments, we used the Ad Centre Data which contains anonymised user ids. We demonstrate the increased utility by performing analysis using this sensitive field while still guaranteeing privacy.

6.4.1 Trendsetters

In this experiment we focus on the distribution of the number of trendsetters for search queries. We define a trendsetter as the first user to issue a query searched later by other users as well, e.g. first one to search for "Xbox 360". This experiment requires anonymised user ids and would not have been feasible in the RFP dataset. Also, a user may have been the trendsetter for multiple queries such as being the first person to search for "Michael Jackson" and "Farrah Fawcett". We plot the distribution of the number of queries for which a user is a trendsetter as shown in Figures 7 and 8.

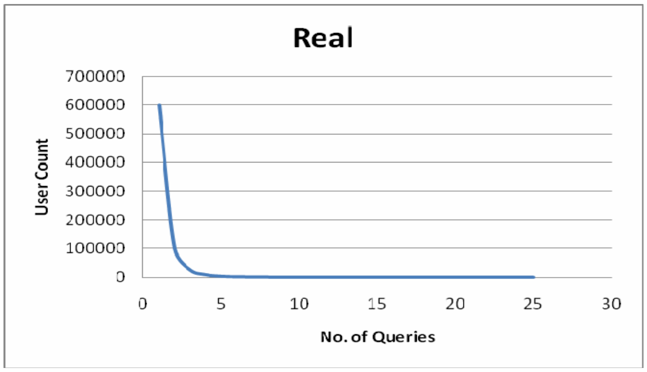


Figure 7. No. of queries for which a user is a trendsetter in real data

Our results show that very few users are trendsetters for more than one or two queries. Figure 8 shows the results obtained in our framework for the same query using different values of ε. Clearly, there is no appreciable difference between the results obtained in our framework and the real data. Table 1 provides a snapshot of the data distribution in the interval of one to five queries.

As we can see from above, the results obtained in our framework closely mirror the real data, validating our approach to perform scientific research while still guaranteeing privacy.

6.4.2 Query Trend

In this experiment, we focus on understanding the distribution of the number of distinct users searching over the period of a day. This experiment also illustrates the tradeoff of using RFP like datasets which contain only query strings and not the associated user ids**.** In the absence of a user id, we are forced to use query count as an approximation of the number of users searching over the period of a day as shown in Figure 9. On the other hand, Figure 10 shows the actual number of users obtained using the anonymised user id field. As is evident from the above graphs, the former experiment although preserving the distribution trends, scales up the actual values by around four to five times thus invalidating any research that would have depended on actual values.

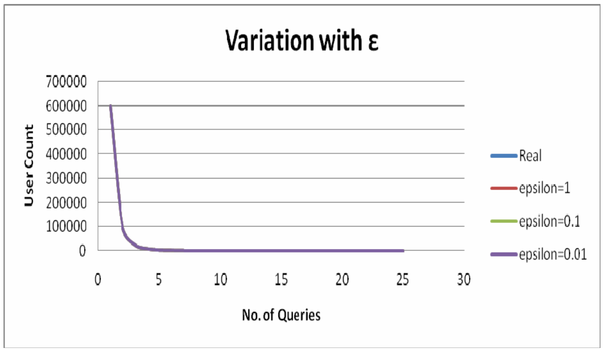


Figure 8. No. of queries for which a user is a trendsetter in our framework

Table 1. Snapshot of the Trendsetter Data

No. of Queries	Real	ε =1	**ε =0.1**	**ε =0.01**
1	598771	598771	598754	598873
2	100084	100085	100088	100131
3	25731	25729.5	25797.4	25960.7
4	9026	9025.58	8996.96	9043.77
5	3701	3699.9	3699.42	3595.3

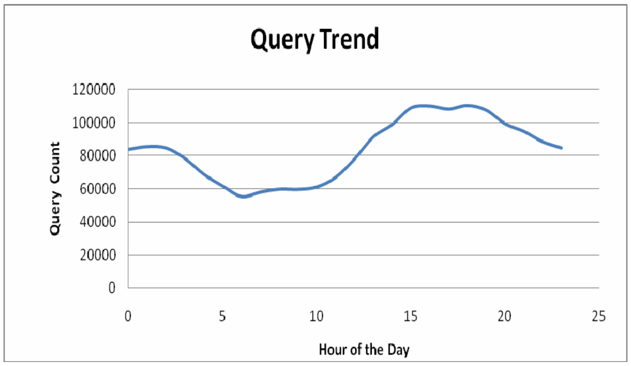


Figure 9. Distribution of the number of queries over the period of a day

Figures 11 and 12 show the results obtained in our framework for the same query. The obtained results follow the real data, and the average error exponentially decreases with increasing ε.

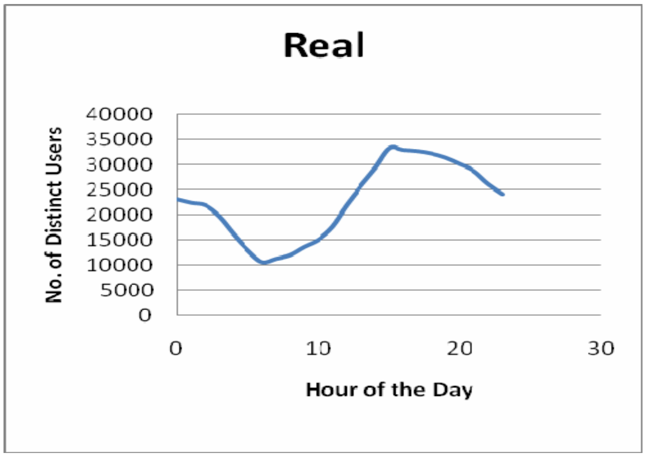


Figure 10. Number of users searching over the period of a day

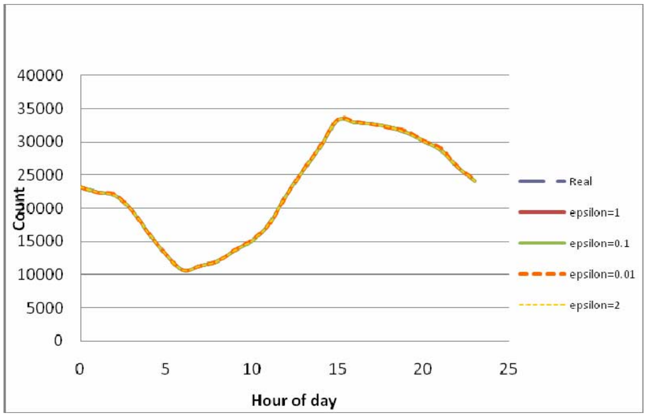


Figure 11. Number of users searching over the period of a day against ε

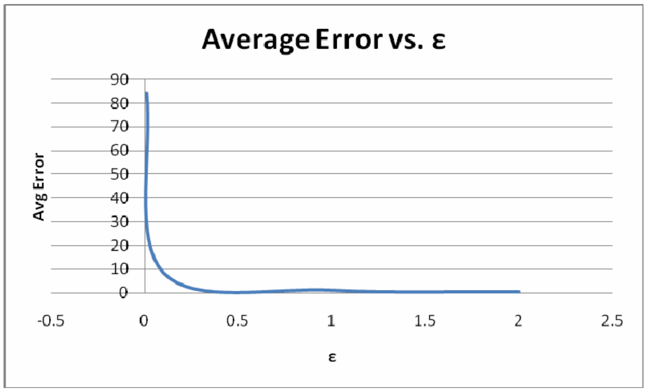


Figure 12. Variation of Average error with ε

7. FUTURE WORK AND CONCLUSION

Based on our experiments, we find that our framework is effective for queries concerning a single entity such as the "number of people searching for starbucks over a period of month". However, the validity of the results when comparing across entities depends on the underlying distributions. If the difference between the data points is large, the result returned would be valid for comparative tasks. However, if the data points are too close to each other, the returned result may not be sufficient to make these decisions. We argue that these use cases which depend on accurate evaluation of the underlying dataset can be handled outside our framework through other licensing agreements. We also argue that query logs are by themselves incomplete data since not all queries of all users are stored by a single search engine. Given the incomplete nature of data in search logs, we argue that the noise added by our framework does not heavily distort the data and hence does not hinder typical research questions. Furthermore, we argue that recent work in synthetic data generation via differential privacy could be used for exploratory data analysis applications, where researchers do not yet have a clear use case for the data. We also note that in addition to the privacy guarantees, our framework is actually flexible to support multiple kinds of queries as opposed to providing a data cube with canned queries.

In future work, we would like to support a Natural Language Processing and ontological interface for allowing researchers to infer knowledge from the dataset. We envision the system being capable of accepting a user defined ontology such as a medical ontology and answering queries such as the number of people searching for medicines, which in current systems would require user access to the raw dataset.

We would also note that though our system has been built for query logs, the framework itself is generic and extensible and therefore can be used for secure information sharing among multiple entities. Given the proliferation of social networks, users would definitely want to be able to control the information they share with others and how that information is used. Our work can be seen as providing the necessary infrastructure to enable users to controllably share information in the public domain without threatening their privacy.

8. ACKNOWLEDGEMENTS

We would like to thank the reviewers for their valuable comments. Also, we would like to thank Frank McSherry for his insightful discussions and helping us with the PINQ framework. We are grateful to Denny Lee for his valuable discussions and sharing his experiences on using privacy preserving data analysis techniques on medical data.

9. REFERENCES

[1] HIPAA, http://www.dhhs.gov/ocr/hipaa

[2] Adam Meyerson and Ryan Williams. On the Complexity of Optimal k-Anonymity. PODS, 2004.

[3] Broder, A. 2002. A taxonomy of web search. SIGIR Forum 36, 2 (Sep. 2002), 3-10. DOI= http://doi.acm.org/10.1145/792550.792552Cooper, A.

[4] "A Survey of Query Log Privacy-Enhancing Techniques from a Policy Perspective," ACM Transactions on the Web (TWEB), vol. 2, issue 4, Oct. 2008

[5] C. Dwork. "Differential privacy". In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, ICALP (2), volume

4052 of Lecture Notes in Computer Science, pages 1–12. Springer, 2006.

[6] Protégé, http://protege.stanford.edu/

[7] Microsoft Beyond Search RFP, http://research.microsoft.com/ur/us/ fundingopps/RFPs/BeyondSearch_RFP.aspx

[8] Differential Privacy, http://www.cs.cmu.edu/~CompThink/ mindswaps/oct07/difpriv.ppt

[9] Intellidimension, http://www.intellidimension.com/

[10] A Face is exposed for AOL Searcher No. 4417749, http://www.nytimes.com/2006/08/09/technology/09aol.html?ex=1312776000

[11] The Platform for Privacy Preferences (P3P), http://www.w3.org/P3P/

[12] SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/

[13] Jansen, B. J., Spink, A., Bateman, J., and Saracevic, T. 1998. Real life information retrieval: a study of user queries on the Web. SIGIR Forum 32, 1 (Apr. 1998), 5-17. DOI= http://doi.acm.org/10.1145/281250.281253

[14] Kumar, R., Novak, J., Pang, B., and Tomkins, A. 2007. On anonym zing query logs via token-based hashing. In Proceedings of the 16th international Conference on World Wide Web (Banff, Alberta, Canada, May 08 - 12, 2007). WWW '07. ACM, New York, NY, 629-638. DOI= http://doi.acm.org/10.1145/1242572.1242657

[15] LiXiong, Eugene Agichtein. ”Towards Privacy preserving Query Log Publishing”. In proceedings of WWW 2007.

[16] Moritz Y. Becker, Cedric Fournet and Andrew D. Gordon. Design and Semantics of a Decentralized Authorization Language. In 20th IEEE Computer Security Foundations Symposium (CSF), 3--15, 2007.

[17] Rose, D. E. and Levinson, D. 2004. Understanding user goals in web search. In Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM, New York, NY, 13-19. DOI= http://doi.acm.org/10.1145/988672.988675

[18] Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. 1999. Analysis of a very large web search engine query log. SIGIR Forum 33, 1 (Sep. 1999), 6-12. DOI= http://doi.acm.org/10.1145/331403.331405

[19] Slagell, A.; Yurcik, W., "Sharing computer network logs for security and privacy: a motivation for new methodologies of anonymisation," Security and Privacy for Emerging Areas in Communication Networks, 2005. Workshop of the 1st International Conference on , vol., no., pp. 80-89, 5-9 Sept. 2005

[20] Spink, A. 2002. A user-centered approach to evaluating human interaction with web search engines: an exploratory study. Inf. Process. Manage. 38, 3 (May. 2002), 401-426. DOI= http://dx.doi.org/10.1016/S0306-4573(01)00036-X

[21] Spink, A. and Ozmultu, H. C. 2002. Characteristics of question format web queries: an exploratory study. Inf. Process. Manage. 38, 4 (Jul. 2002), 453-471. DOI= http://dx.doi.org/10.1016/S0306-4573(01)00042-5

[22] Sweeney L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002: 10 (5), pp. 557-570.

[23] Pang, R. and Paxson, V. 2003. A high-level programming environment for packet trace anonymisation and transformation. In Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols For Computer Communications (Karlsruhe, Germany, August 25 - 29, 2003). SIGCOMM '03. ACM, New York, NY, 339-351. DOI= http://doi.acm.org/10.1145/863955.863994

[24] McSherry, F. Privacy Integrated Queries, in Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD), Association for Computing Machinery, Inc., June 2009.

[25] Viegas, E. 2007. Enabling Innovation in Internet Research, In Proceedings of the National Science Foundation Workshop on Data Confidentiality, Sep 5-7, 2007, http://dcws.stat.cmu.edu/viegas.htm