Differential Privacy in Statistical Databases

Presenters: Jonathan Evans, Victoria Girkins, Sam Sartor

What is Differential Privacy?

- Danger of statistical databases
- One extreme - allow unlimited queries (vulnerable to tracker attacks)
- The other - return random answers to queries (not terribly useful)
- Balance between security and usability
- Solution: Differential Privacy.
  - Add noise to query returns
  - Limit number of queries
  - Result: High probability of privacy

Example - A statistical database with and without DP

- Eve is a user who may make queries to the database.
- Her chosen query is
  `select sum(income) from table`
- She learns that a new entry has been added to the DB and runs the query again. Let's see the effect DP has on her devious plan to learn a user's income.

[Effect of privacy budget]

Implementation

- How do we noisify the data?
- We chose the Laplace distribution - suitable for our purposes but not for categorical data.

Our Approach

- Goal: Epsilon-Differential Privacy

$$\Pr[\mathcal{K}(D) \in S] \leq \exp(\varepsilon) \times \Pr[\mathcal{K}(D') \in S]$$

- *K* is our noise-adding mechanism; in our case, the Laplace distribution.
- Choose Laplace distribution centered at 0.
- Must consider: privacy budget (ε) and sensitivity (Δf)
- Scale parameter = k×Δf/ε
- Guaranteed epsilon-differential privacy

Setup

- Generate large (10,000 rows) CSV dataset
    - Name: from the `names` library
    - Age: uniformly distributed between 18 and 65
    - Income: log-normal distributed, roughly matches US
    - Zip Code: from a random set of ~40
    - Net-worth: from age, income, and some random offset
- Write a program to make useful queries against the dataset
- Protect the privacy of individuals

Implementation

- Reads CSV file
- Implemented functions: sum, count, min, max, mean, variance, sd
  - How to compute statistic (e.g. mean = total / count)
  - How to find Δf (e.g. from most extreme value)
- Filters rows with user-specified where clause (python expression)
- Handles edge cases where 1 or 0 rows are returned
- Adds laplace noise dependent on Δf, and a given ε / query limit
- Displays the privatized result
- DIFFERENTIAL PRIVACY!

Our Results

```
use dp (y/n): n
database csv: bigdata.csv
=================
summarize field: age
summary function: sum
where: zip == 31643
10412
=================
summarize field: age
summary function: sum
where: zip == 31643 and name != "Abel Woods"
10379
```

The total age of people in 31643 is **10412**

0% error

Abel Woods is 10412 - 10379 = **33**

0% error, VIOLATION OF PRIVACY!

```
use dp (y/n): y
epsilon: 5
query limit: 2
database csv: bigdata.csv
=================
summarize field: age
summary function: sum
where: zip == 31643
10409
=================
summarize field: age
summary function: sum
where: zip == 31643 and name != "Abel Woods"
10435
```

The total age of people in 31643 is **10409**

0.2% error

Abel Woods is 10409 - 10445 = **49**

48% error, PRIVACY PROTECTED!

Future work

- Not many implementations outside of academia
- Lots of user data is vulnerable
- Opportunity for software engineers to implement in industry

Conclusion

- What we expected
- What we got
- Limitations of our work
- Limitations of DP
- Lack of industry adoption

References

- Charu Aggarwal. “On k-Anonymity and the Curse of Dimensionality.” In: VLDB 2005 - Proceedings of 31st International Conference on Very Large Data Bases. Vol. 2. Jan. 2005, pp. 901–909.
- Atockar. “Differential Privacy: The Basics”. In: (2014).URL: https://research.neustar.biz/2014/09/08/differential-privacy-the-basics/.
- Cynthia Dwork. “A Firm Foundation for Private Data Analysis”. In: Communications of the ACM (Jan. 2011).URL: https://www.microsoft.com/en-us/research/publication/a-firm-foundation-for-private-data-analysis/.
- Cynthia Dwork. “Differential Privacy”. In: 33rd International Colloquiumon Automata, Languages and Programming, part II (ICALP 2006). Vol. 4052. Springer Verlag, July 2006.ISBN: 3-540-35907-9. URL: https://www.microsoft.com/en-us/research/publication/differential-privacy/.
- Quan Geng and Pramod Viswanath. “The optimal mechanism in differential privacy”. In: 2014 IEEE International Symposium on Information Theory (2014).DOI: 10.1109/isit.2014.6875258. URL: https://arxiv.org/pdf/1212.1186.pdf.
- Michael Hilton. Differential Privacy: A Historical Survey. URL: http://www.cs.uky.edu/~jzhang/CS689/PPDM-differential.pdf.