Differential Privacy for Supercomputer Sensor Data

Spencer Ortega

University of Southern California

Mentors: Dr. Nathan DeBardeleben (HPC-DES), Dr. Claire Bowen (CCS-6)

LA-UR-19-27097

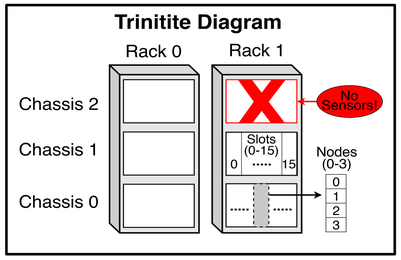
Introduction

- **Problem**: Data-driven research is in need of a secure way to protect sensitive data to release for public use
- Previously used data protection methods contain vulnerabilities
  - E.g. cross-referencing of external data sources to re-identify data
- Differential Privacy (DP) aims to solve this problem
  - Add random “noise” to blur results of statistical queries

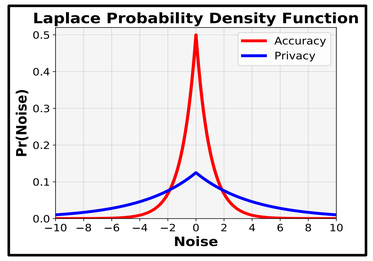
Differential Privacy

- Two-World Privacy
  - Goal: make the results of a query on two datasets, with the presence and absence of **any** record, indistinguishable from one another.
- Privacy-Loss Budget ε
  - Quantifies and bounds how much sensitive information can be leaked
  - Trade-off between privacy and accuracy
- Global Sensitivity (GS) of a Query
  - Max difference of query results on any possible ‘Two-World’ datasets

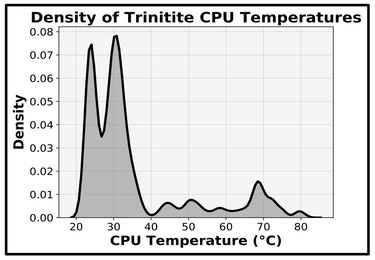
Experimental Setup



- Apply DP to supercomputer sensor data
- To the best of our knowledge, has never been implemented for this domain of data
- Used sensor data from the Trinitite system



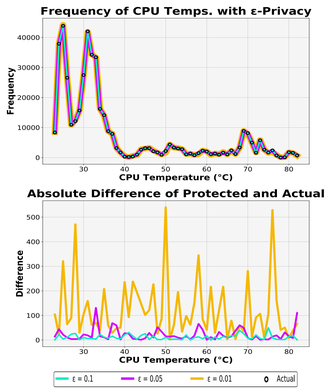
- Sample DP noise from Laplace PDF
- Laplace width affects noise sampling
  - Wide: more random (privacy)
  - Narrow: more deterministic (accuracy)
- (GS / ε) is used to scale the width



- Focused on CPU temperatures
- Easy enough to explain, compared to other more complicated sensors
- Sensitive enough to possibly infer jobs running on the system

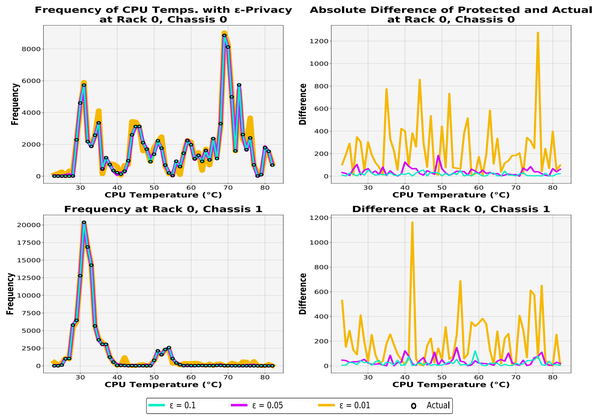
Results

Histogram Query



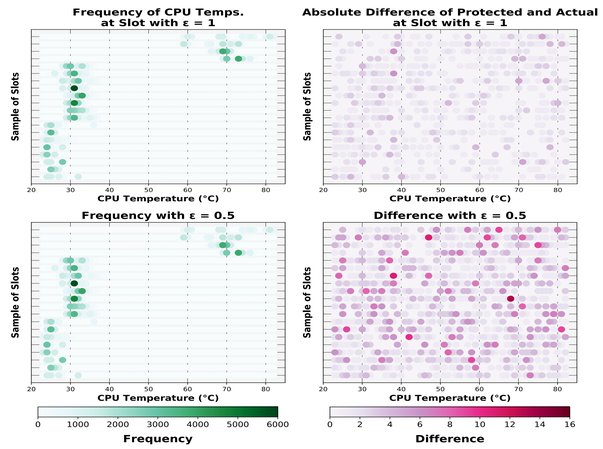
- Illustrates the simplest case of protecting CPU temperature data
- We observe that the smaller ε gets, the more the protected results deviate
- Not a very robust query if we cant analyze aggregate information by group (location)

Histograms by Rack, Chassis



- One approach to protecting aggregate information is by exhaustively asking a query for each group
- ε must be divided for each query, making results for more granular groups very inaccurate
- We see an example of this drop in accuracy when grouping histograms by rack, chassis (divide ε by 5)

Cross Tabulation by Slot and Temp.



- A smarter approach: Cross Tabulation
- Only ask one query about all groups (divide ε by 1)
- Our cross tabulation results with ε set to 1 and .5 insist that we should use .5 instead because it has similar accuracy to 1, but leaks less sensitive information

Future Work

- Apply DP to other types of supercomputer sensors
- Test other DP algorithms on this data
  - Exponential Mechanism
- Explore ε values for appropriate protection of these different types of sensors and mechanisms
- Synthesize datasets by sampling protected histograms
  - Synthetic data can be queried without privacy-loss budget
  - Compare accuracy of analysis to unprotected data
- Potential work with vendors to satisfy NDA’s for data sharing

Acknowledgements

- Research was supported by LANL Laboratory Directed research and Development (LDRD) program at Los Alamos National Laboratory
- Special thanks to my mentors and other USRC staff that helped me throughout my internship