Analyzing the Applications of Differential Privacy

Sanya Nema, A. Prasad Sistla

Department of Computer Science

University of Illinois at Chicago, Adlai E. Stevenson High School

Background About Privacy

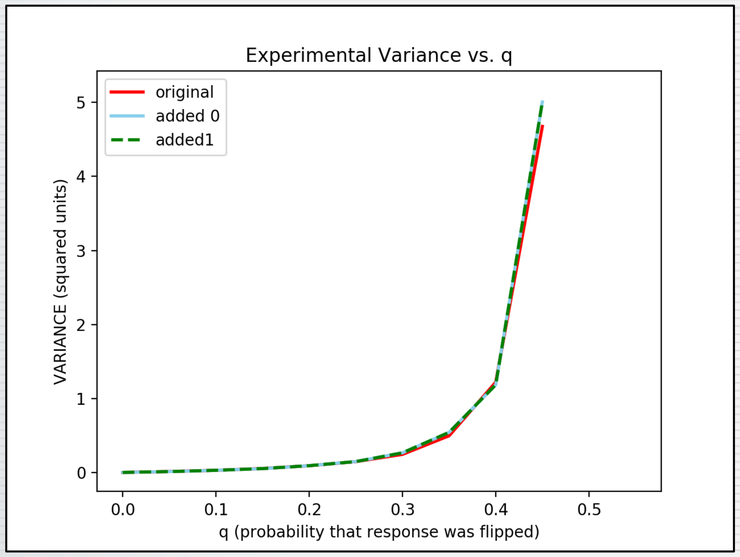
- Has become more important in the day of technology
- Recent data breaches (Facebook, Exactis, Quora, etc.) have raised skepticism regarding true privacy protection
- This has led to discouraging individuals from giving their data to companies in surveys
- Companies want data to gauge their demographics — not having this data hurts them
- Solutions have been proposed to help curb these negative effects, such as anonymization
- This method is vulnerable to linkage attacks and does not work
- Other methods have also been able to be exploited

Background About Differential Privacy

- Developed in 2006
- Privacy is typically measured with epsilon
- This is an algorithm that provides individuals with privacy of their identity & companies with data
- It does this by introducing noise (ex: Laplacian noise or randomized response algorithm) into the data before it sends it back to a data analyst
- This makes it so that the analyst/company can see data trends as a whole, and NOT the participants' identity
- The basic idea is to assure individuals that whether or not they participate, it will not significantly change the overall trends
- Thus, their participation will not be able to be gauged by an outside entity & they should participate

Details of the Algorithm

- Simulate the *randomized response algorithm* with a coin toss algorithm
- Goal of randomized response: to estimate in a population the fraction of people that have property P (e.g. drug use)
- Used 'q', epsilon = (log($\frac{1+q}{1-q}$)
- Provides for random noise distribution in the data
- Use probability 'p' — fraction of individuals that have property 'P' within the dataset
- Value of 'p' is estimated from the responses — P_{est}
- For each individual, answer will be flipped with probability 'q' and stay the same with probability 1-q
- q<=0.5; higher the q, higher the privacy and lower the accuracy
- Highest privacy when q = 0.5
- Used code to simulate this by generating a random number between 0 and 1 and flipping response if <= q



The graph depicts the relationship between variance and probability 'q', showing that as 'q' is increased, variance also increases. This indicates that accuracy is being lost.

```
Yes of OG: 3307.0
pActual: 0.3307
Mean pEst of OG: 0.3306564
vTotal of OG: 218.81781112000073
Variance of OG : 0.1505013980807064
Mean pEst of added 0: 0.330719528047195 23
vTotal of added 0: 218.9062022504879
Variance of added 0: 0.15538978696878303
Mean pEst of added 1: 0.3304241575842415
vTotal of added 1: 218.50888591772724
Variance of added 1: 0.14863808721594296
```

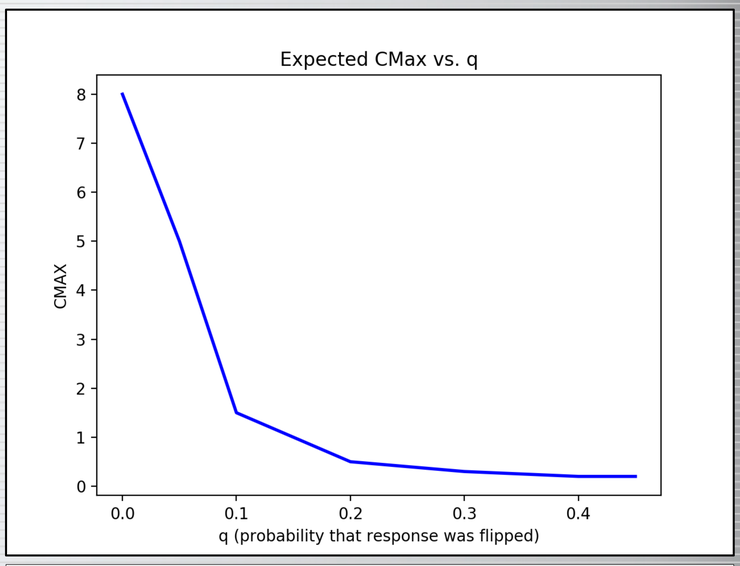
The outcome of the written code

$$P_{est} = \frac{\frac{\text{\# of yes answers}}{n} - (1-q)}{2q-1}$$

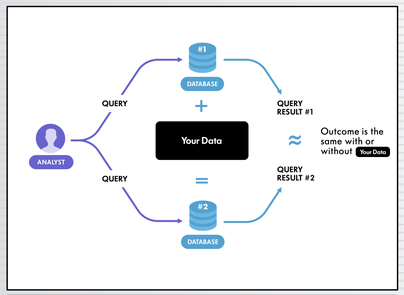
An equation used to calculate P_{est}

$$\text{Variance} = \frac{1}{n\left(16(q-0.5)^2\right) - 0.25}$$

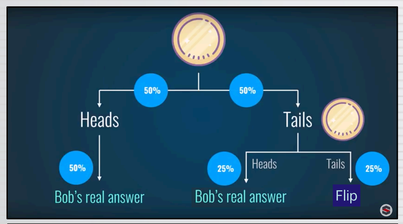
An equation used to calculate variance



This graph shows the relationship between CMax and probability 'q', showing that as 'q' increases, CMax decreases. This indicates that as 'q' increases, privacy increases.



Shows that result of an analyst's query will be the same with or without your data in the database



Shows the randomized response algorithm and with what probabilities the 'real' answer will be reported

Apple

- Uses it for info on emoji use, new words, HealthKit and more
- Data scientists' analyzed their reported data and concluded an epsilon of 14
- This is too high of a value to be considered truly private
- Apple does not publicly release the code behind their differential privacy algorithms
- The opacity of this makes data scientists question the honesty behind their method
- More transparency → More plausible

Google

- Uses it for info on web searches, Google Maps traffic data
- Their reported epsilon value is 2
- System is called RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response)
- Much better than Apple's differential privacy according to scientists
- It has an open source library which is much more transparent
- Accessible to most individuals

U.S. Census Bureau

- Uses it to gather statistics about Americans — salary, drug use, age, etc.
- Unsure of what the epsilon is
- Main reason behind using this is so that individuals can give an accurate picture of US
- Census does not want to discourage participation
- Previously have been criticized for lack of privacy
- One of the most influential usages

References

Differential privacy in the real world: The 2018 end-to-end census test. (2019, February 17). Retrieved July 31, 2019, from https://www2.census.gov/programs-surveys/decennial/2020/resources/presentations-publications/2019-02-17-abowd-differential-privacy.pdf?

Greenberg, A. (2017, September 15). How one of Apple's key privacy safeguards falls short. Retrieved July 31, 2019, from https://www.wired.com/story/apple-differential-privacy-shortcomings/

Orr, A. (2017, September 15). Google's differential privacy may be better than Apple's. Retrieved July 31, 2019, from https://www.macobserver.com/analysis/google-apple-differential-privacy/