UNSW
THE UNIVERSITY OF NEW SOUTH WALES

NICTA

LOKI:
A Privacy-Conscious Platform for Crowdsourced Surveys

Thivya Kandappu, Vijay Sivaraman (UNSW)
Arik Friedman, Roksana Boreli (NICTA)

Crowdsourcing Platforms

- Amazon Mechanical Turk (AMT)
  - >500K users
  - Widely used in psychology and social studies
  - Trusted curator
  - User's unique id, ip-address, city and country are revealed to the surveyor

- Google Consumer Surveys
  - Trusted curator
  - One question at a time

User de-anonymization is easy!

- We launched a series of survey tasks in AMT
- Survey 1: astrology services
  - Star sign, date/month of birth, beliefs in astrology, …
- Survey 2: online match making services
  - Gender, age, marital status, usage of match-making, …
- Survey 3: mobile phone coverage
  - Zip code, phone signal strength and quality, …
- 100 respondents for each survey, 3 hours, $30
- 72 users took all 3 surveys: got their DoB, gender, Zip
  - Can de-anonymize these users with high probability (~76%)

Private Information is Easy to Extract

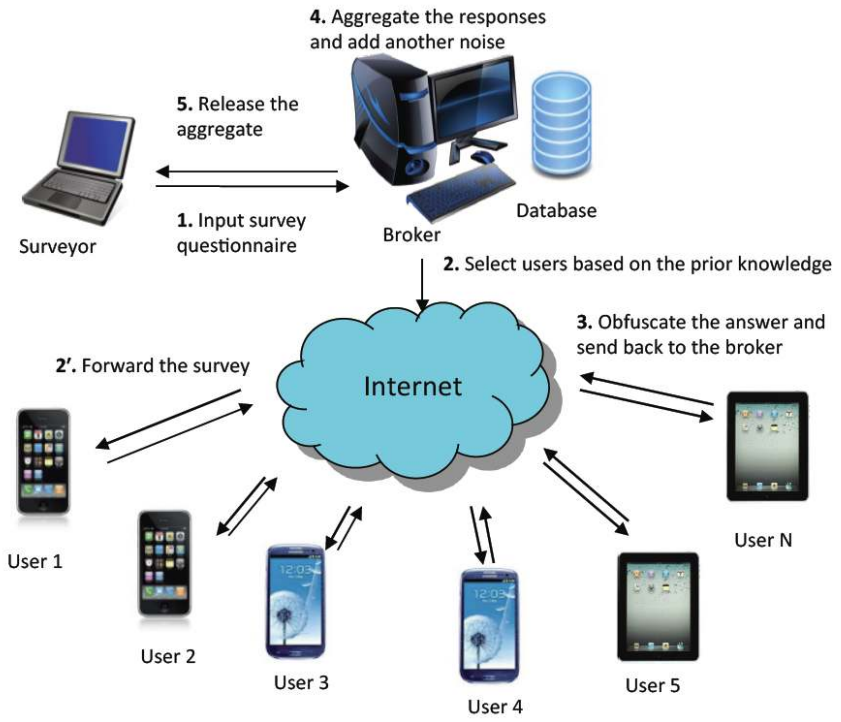
- Survey 4: smoking habits
  - Smoking intensity, coughing frequency, income,…
- 18 of the 72 de-anonymized participants took this survey
- Got highly personal information for these individuals
  - Respiratory health, income, …
- Easy to obtain personal information on these platforms!
- Survey 5: user perception of privacy in such platforms (would you do this if you know you can be de-anonymized?)
  - 73 out of 100 users said they would not have participated

Available Solutions

- Anonymize the user
  - Can still deduce from device-id, IP address
- Trust the surveyor (curator)
  - E.g. trust Google surveys not to sell your data!
  - Or trust lawyers to offer you legal protection
- Obfuscate your answer (hide in the crowd!)
  - Add noise to individual responses
  - Surveyor cannot get accurate individual information but can get accurate "on average" information about the population

System Architecture

- Three entities:
  - Surveyors
  - Users
  - Broker
- Three knobs:
  - Privacy
  - Utility
  - Cost
- Design Choices:
  - Obfuscation technique
  - User privacy levels
  - Privacy loss quantification
  - User privacy depletion
  - Cost



User selection Algorithm

- Minimize estimation Error:
  - Choose a subset of users carefully to minimize the error in the estimation
- User and Group Error History:
  - "value" of a user/group depends on how accurately the user's responses reflect those of the population at large
  - Mean and variance of the user/group error can be estimated
- Balance cost, accuracy and privacy fairness:
  - Monetary constraint
  - Privacy constraint

Optimizing across multiple surveys

- "Fairness parameter" $\alpha \in [0,1]$ – combines monetary and privacy cost of user '*i*' into an overall cost F_i

$$F_i = (1-\alpha)\frac{c_i}{C} + \alpha \cdot \max\left[\frac{\epsilon_i}{R_i^{(\epsilon)}}, \frac{\delta_i}{R_i^{(\delta)}}\right]$$

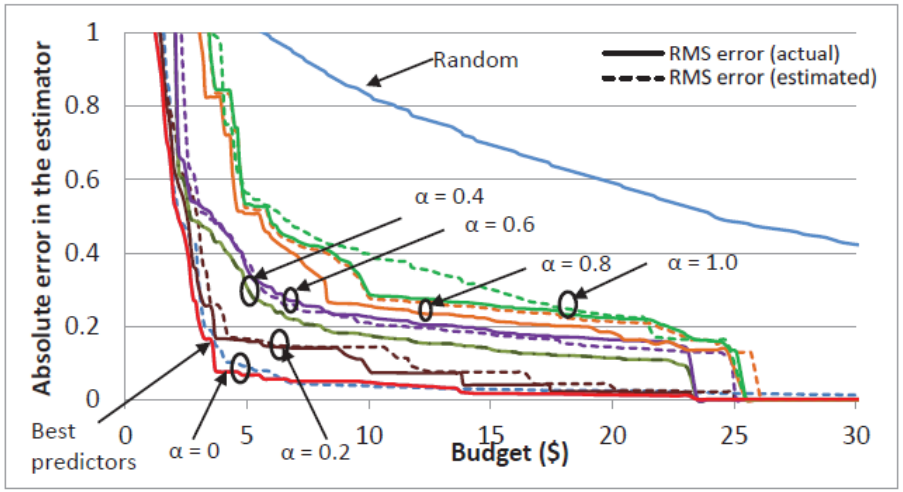
- The improvement in RMS error per unit of cost for the user '*i*':

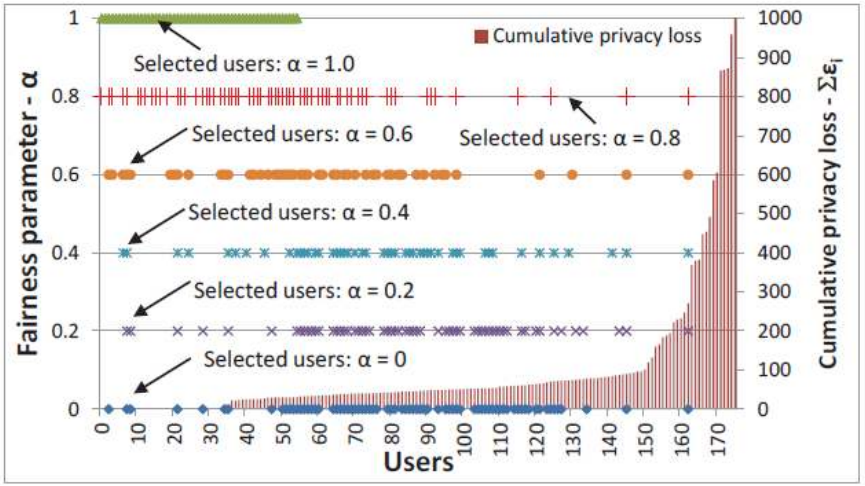
$$\beta_i(S) = \frac{\Delta RMSE(S,i)}{F_i}$$

Evaluation: Netflix Dataset

- Netflix dataset
  - 0.5 million users
  - 17,000 movie titles
- History information
  - 1436 movies released in 2004
  - Users rated >50 movies
- Measures
  - User error
  - Group error
  - Accumulated privacy loss

Results: Trade-offs b/w cost, accuracy & privacy

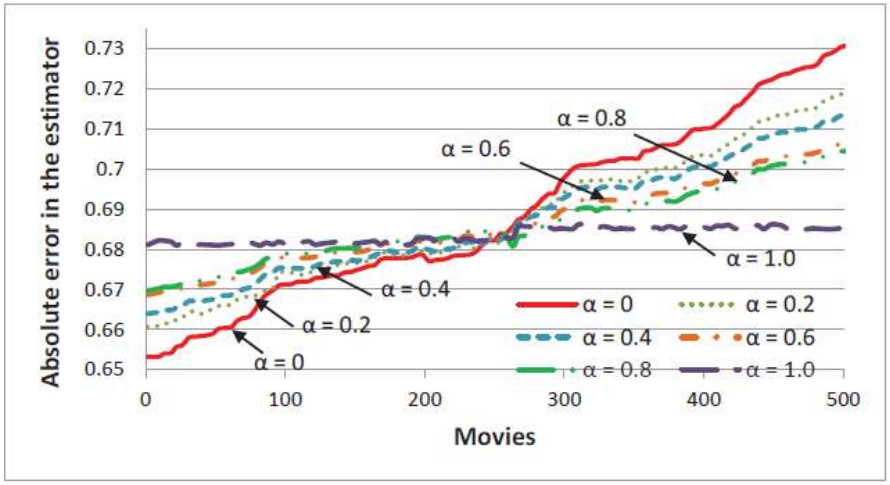


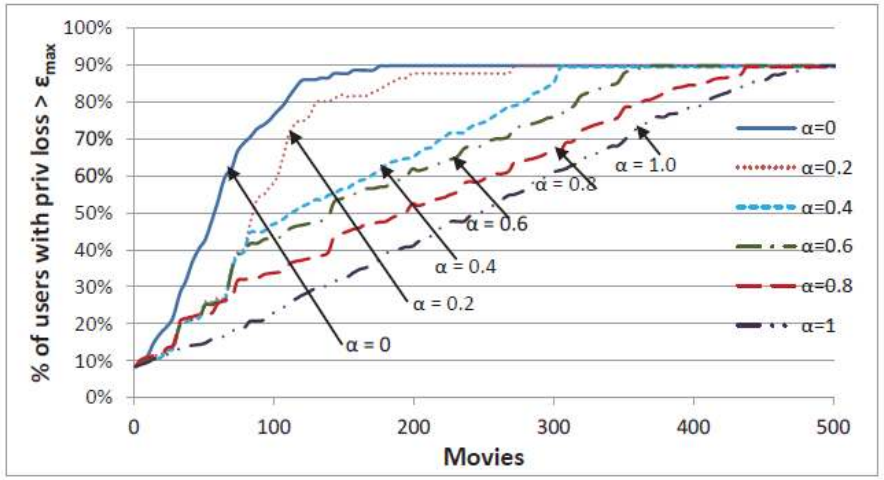


- "best predictors" gives near-perfect estimation – surveying only 37% of the population
- $\alpha = 0$ is identical to best predictors
- As α increases, the error increases

- Loss in accuracy is compensated for in privacy fairness
- $\alpha = 1$ is biased towards the users who have low privacy loss
- As α decreases, selection gives less regard to prior privacy depletion

Results: Long-term performance





- when α is low, the error in initially low but increases with successive movies
- high α results in fairer depletion in privacy and prolongs the lifetime of a user

Loki: Prototype Implementation

- User: https://itunes.apple.com/tr/app/loki/id767077965?mt=8
- Surveyor: http://loki.eng.unsw.edu.au/
- Evaluate the system with 130 volunteer students

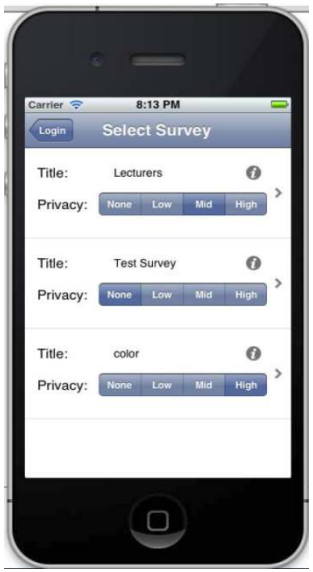


Fig 1. List of surveys and privacy levels available to the users

Fig 2. Questions and rating entered by the user

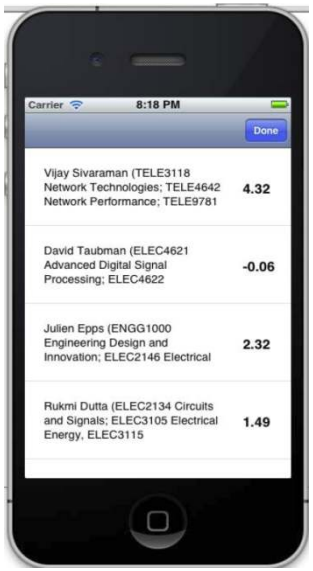


Fig 3. Uploaded user responses after noise addition

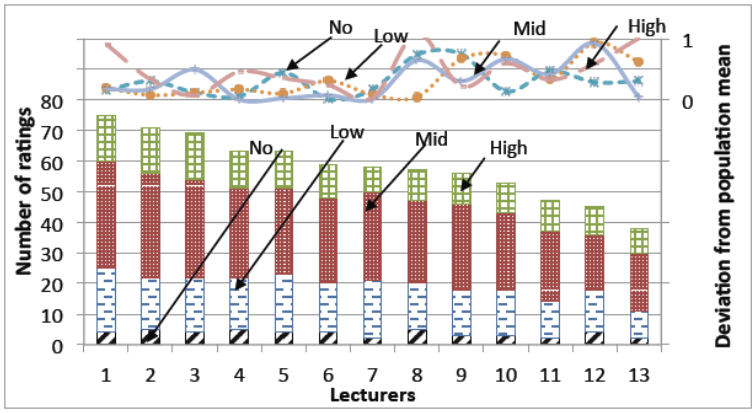


Fig 4. Deviation in mean across the bins for various lecturers

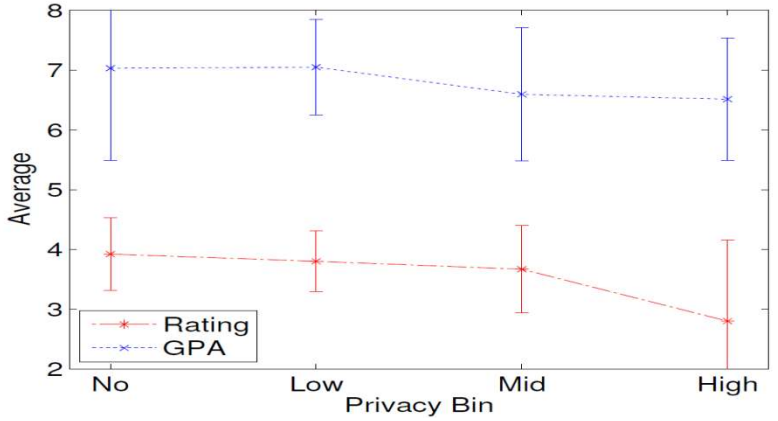


Fig 5. Correlation of privacy choice with average ratings and student GPA

Conclusions and Future Directions

- private information is easy to extract
- giving the means of control to the users
- End-goal: make user data obfuscated so that the users can hide in the crowd while giving meaningful aggregations to the surveyors
- Future directions:
  - user-perception study of LOKI app
  - extension to more question types, e.g., yes/no, multiple choice