Privacy & Data De-identificaiton:

Status of Research and Application in Taiwan

Dr. Yennun Huang
CITI, Academia Sinica
Oct. 26th, 2017

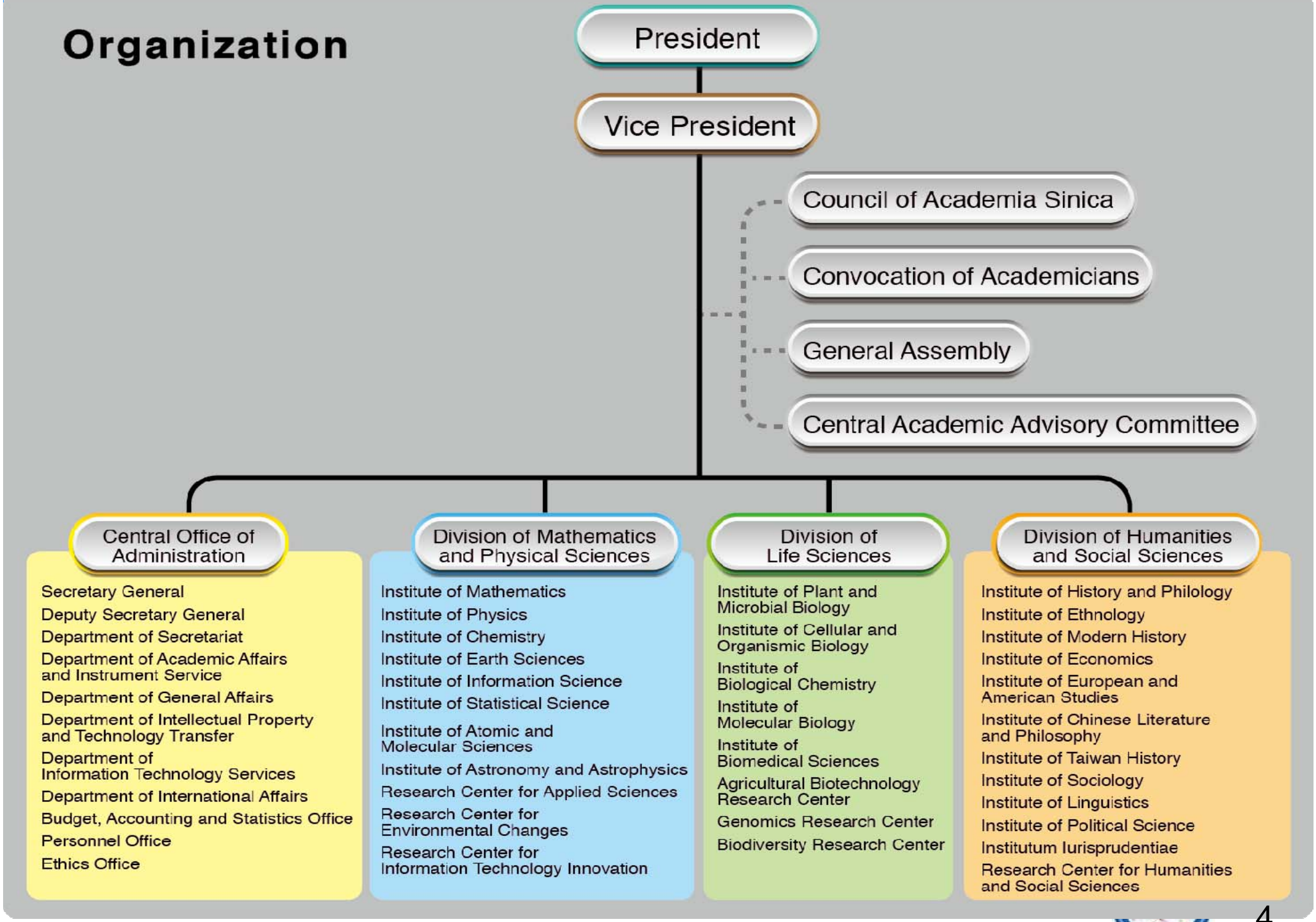
Yennun Huang IEEE Fellow

- Director, Center for Information Technology Innovation (CITI), Academia Sinica
- CEO, Taiwan Information Security Center (TWISC), Academia Sinica
- Chairman, Asia IoT Alliance
- DMTS of AT&T Bell Labs
- Department Head, Director, Executive Director, AT&T – Labs
- VP of Engineering, PreCache Inc.
- Executive VP, Institute for Information Industry, Taiwan
- President, VeeTIME Corporation, Taiwan
- Distinguished Research Fellow, Academia Sinica, Taiwan
- Deputy Executive Secretary, Office of Science and Technology, Executive Yuan, Taiwan

Academia Sinica

- The most preeminent academic institution in Taiwan
- Founded in Republic of China in 1928 by the Nationalist government.
- Directly under President Office
- Roughly 10% of the Taiwan National science and technology R&D budget goes to Academia Sinica each year (10 Billion NT dollars).
- Promote international cooperation and scholarly exchanges that will accelerate the overall development of academic research in Academia Sinica and Taiwan.

Organization



Center for information technology innovation(CITI): Background

- The Research Center for Information Technology Innovation (CITI) was formally founded in 2007, and started to operate in Sep. 2008 (with various committees formed)
  - The 31st and the youngest research unit in Academia Sinica

Center for information technology innovation(CITI): Mission

- To promote the innovation and application of information technologies, with emphases on exploring the enabling technology for essential infrastructure and also on integrating inter-disciplinary technologies so as to provide the key ingredients that are invaluable for the upcoming knowledge-based and service-based societies.

CITI - Three Thematic Centers

- Grid & Scientific Computing Thematic Center
  - Distributed Cloud Computing
- Taiwan Information Security Thematic Center
  - **Headquarter of National TWISC centers**
- Intelligent & Ubiquitous Computing Thematic Center
  - FinTech/RegTech
  - IoT platforms and applications

TWISC – Taiwan information security center

- Mission:
  - Boost research and development activities in information security,
  - Promote public awareness,
  - Facilitate international collaboration,
  - Train and educate security experts and
  - Foster partnership among government, academia, and private sector.

Privacy Research in CITI

The Basic Problem for Data Science

How to do de-identification to

- Enable “desirable uses” of the data while protecting the “privacy” of the data subjects?
  - Political policy
  - Academic research
  - Study drug trial
  - Security: searching for terrorists/crimials
  - Market analysis, ….

Data De-identification: Why it is important?

- To facilitate the development of data analysis industry and collaborative analysis of data between business partners
- To unleash the power of open data: leveraging public wisdom. Need privacy protection to make it happen
- To fulfill regulation requirement
  - Europe: GDPR, General Data Protection Regulation
  - Taiwan: Personal Data Protection Act
  - etc.

Data De-Identification is difficult

- Encryption
  - How to analyze and compute on encrypted data?
- Anonymization
  - Re-identification is possible
- Access mediation/control
  - With multiple queries, re-identification is possible
- Adding noise
  - Permutation
  - Differential Privacy

Privacy Breach Events (1)

- IN 2010, NetFlix included 100 million movie ratings, along with the date of the rating, a unique ID number for the subscriber, and the movie info.

De-anonymize Netflix data

- “Sparsity” of data: In Netflix data, not two records are similar more than 50%.
- If the profile can be matched up to 50% similarity to a profile in IMDB , then the adversary knows with good chance the true identity of the profile.
  - Found that if you knew a few movies a Netflix subscriber had rented in a given time period, you could reverse-engineer the data and find out the rest of their viewing history.

A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets (how to break anonymity of the netflix prize dataset),” in Proc. 29th IEEE Symposium on Security and Privacy, 2008.

Privacy Breach Events (2)

- The state of Massachusetts distributed a research dataset containing de-identified insurance reimbursement records of Massachusetts state employees that had been hospitalized. To protect the employees' privacy, their names were stripped from the dataset, but the employees' date of birth, zip code, and sex was preserved to allow for statistical analysis.
- Sweeney was able to re-identify the governor's records by searching for the "de-identified" records that matched the Governor's date of birth, zip code, and sex. She learned this information from the Cambridge voter registration list, which she purchased for $20. Sweeney then generalized her findings, arguing that up to 87% of the U.S. population could be uniquely identified by their 5-digit ZIP code, date of birth, and sex based on the 1990 census.

Using K-anonymity as data releasing mechanism for research purpose

- K-anonymity make sure that no combination of data attributes can be aggregated to identify a particular person. It checks that if there're at least K people with the same combination. If the K is not large enough there're two ways to increase the K in a dataset:
  - Generalization: To aggregate two or more possible values in an attribute into one value.
  - Suppression: To remove the data entries that doesn't satisfy K-anonymity al-toghether.
- Challenges
  - "Generalization" sometimes requires a lot of efforts. E.g. to combine "musician" and "writer" into "artist"? to mark numbers between 5 ~ 10 as 8?
  - Therefore right now suppression is the route chosen. However sometimes too many records are removed and render the data useless.
  - A good utility measurement is still lacking; the few researchers that went this route have reported bad utility on the result data.

Approach: Synthetic data

Sex	Blood	…	Cancer?
F	B	…	Y
F	A	…	N
M	O	…	N
M	O	…	Y
F	A	…	N
M	B	…	Y

Sex	Blood	…	Cancer?
M	B	…	N
F	B	…	Y
M	O	…	Y
M	A	…	N
F	O	…	N

"fake" people

Utility: preserves statistics with *every* set of attributes!

Problem: computation time

Differential Privacy: Basic Concept

Name	Has AIDS
Tom	1
John	0
Eric	1
Ross	0
Steve	1

f(i) : the partial sum of the first i rows

f(5) – f(4) : reveals Steve has AIDS

Generating Synthetic Data

- It is a hard problem – too many dimension of a contingency table

	Age	Gender	Cancer
1	20	M	1
2	30	F	0
3	20	M	1

20,M,0	20,M,1	20,F,0	20,F,1	30,M,0	30,M.1	30,F,0	30,F,1
0	2	0	0	1	0	0	0

Differential Privacys

- Ensure that the removal or addition of any record in the database does not change the outcome of any analysis by much.

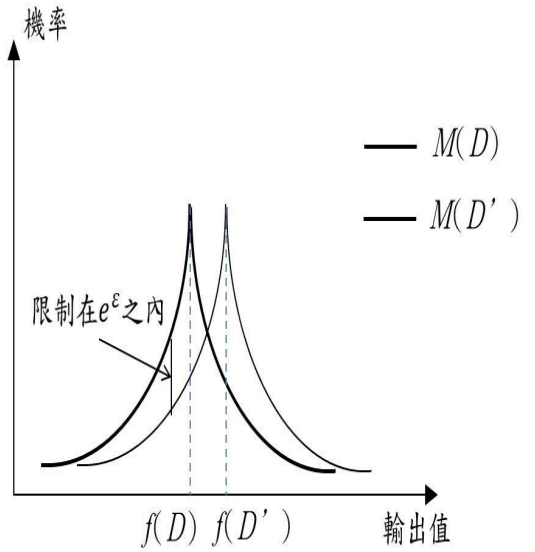
$$\Pr[K_f(D) \in s] \leq e^{\varepsilon} \Pr[K_f(D') \in s]$$

Adding random noise to query

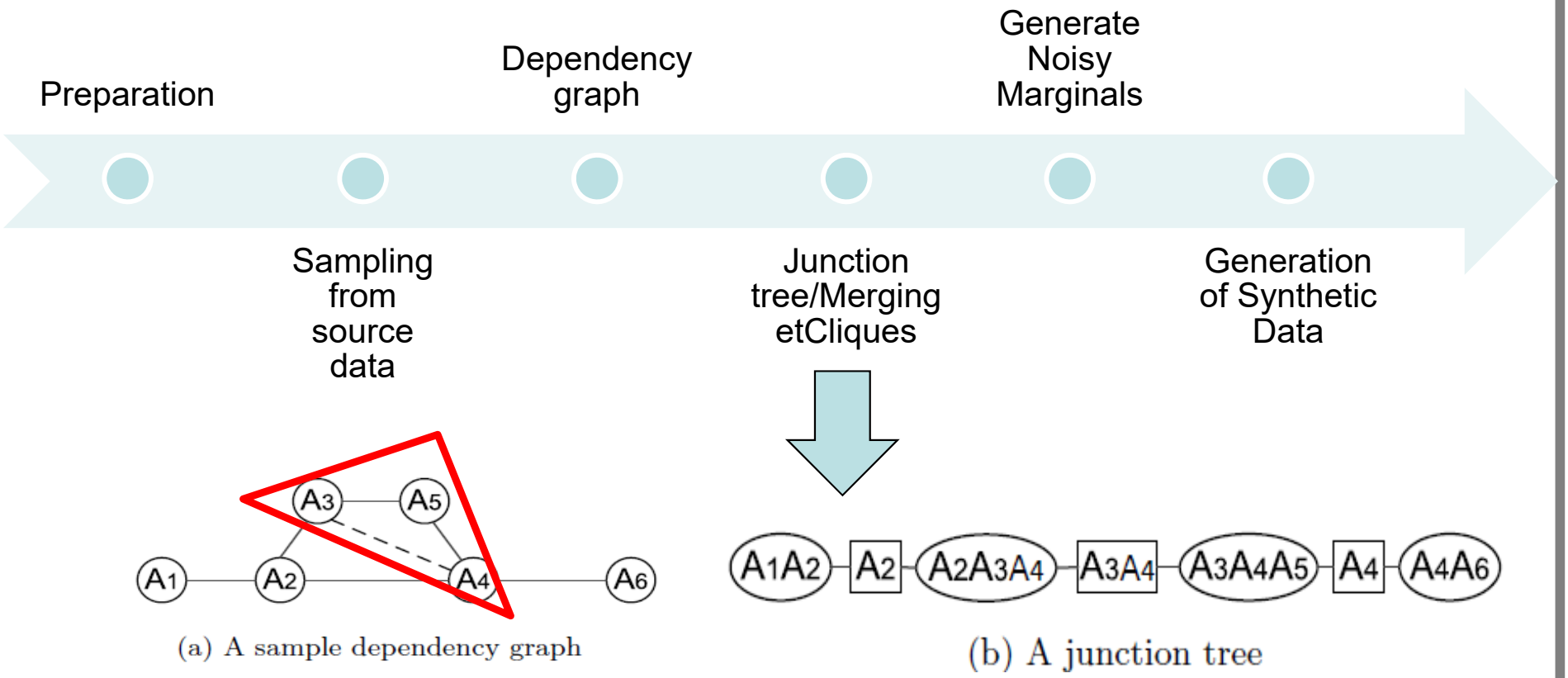
f(X) + Lap (Δf/Ɛ)

Ɛ: differential budget

Δf: sensitivity of query function f



Using "DPTable" to generate synthetic dataset for further analysis



(a) A sample dependency graph

(b) A junction tree

- A series of contingency/marginal tables can be generated from the junction tree. The tables are served as a joint-distribution of data.
- One can draw arbitrary # of data entries from the joint-distribution to get a new and privacy protected synthetic dataset.

Differential privacy as an option to transfer IoT data securely

- IoT devices collect information from all kinds of sensors and send them through a wireless connection to remote servers, so it is always possible that someone try to intercept those information and decode them.
- Differential privacy can also be used to transfer data securely.

Google RAPPOR for IoT Data Privacy

- Low CPU requirement compared to AES and other common encryption / decryption mechanism
- Use bloom filter and differential privacy mechanisms to send values to remote server
- Remote server can aggregate and produce statistical estimates over all collected data, but not a particular entry

Example

- Set f=0.5
- Do a survey, “Do you have AIDS?”
- toss a coin
  - If Head, always say yes
  - If Tail, tell the truth
- The true percentage of people who participate in the survey who has is $2(Y-0.5)$

Raptor Algorithm

1. **Signal.** Hash client's value v onto the Bloom filter B of size k using h hash functions.

2. **Permanent randomized response.** For each client's value v and bit $i, 0 \leq i < k$ in B, create a binary reporting value B'_i which equals to

$$B'_i = \begin{cases} 1, & \text{with probability } \frac{1}{2}f \\ 0, & \text{with probability } \frac{1}{2}f \\ B_i, & \text{with probability } 1 - f \end{cases}$$

   where f is a user-tunable parameter controlling the level of longitudinal privacy guarantee.

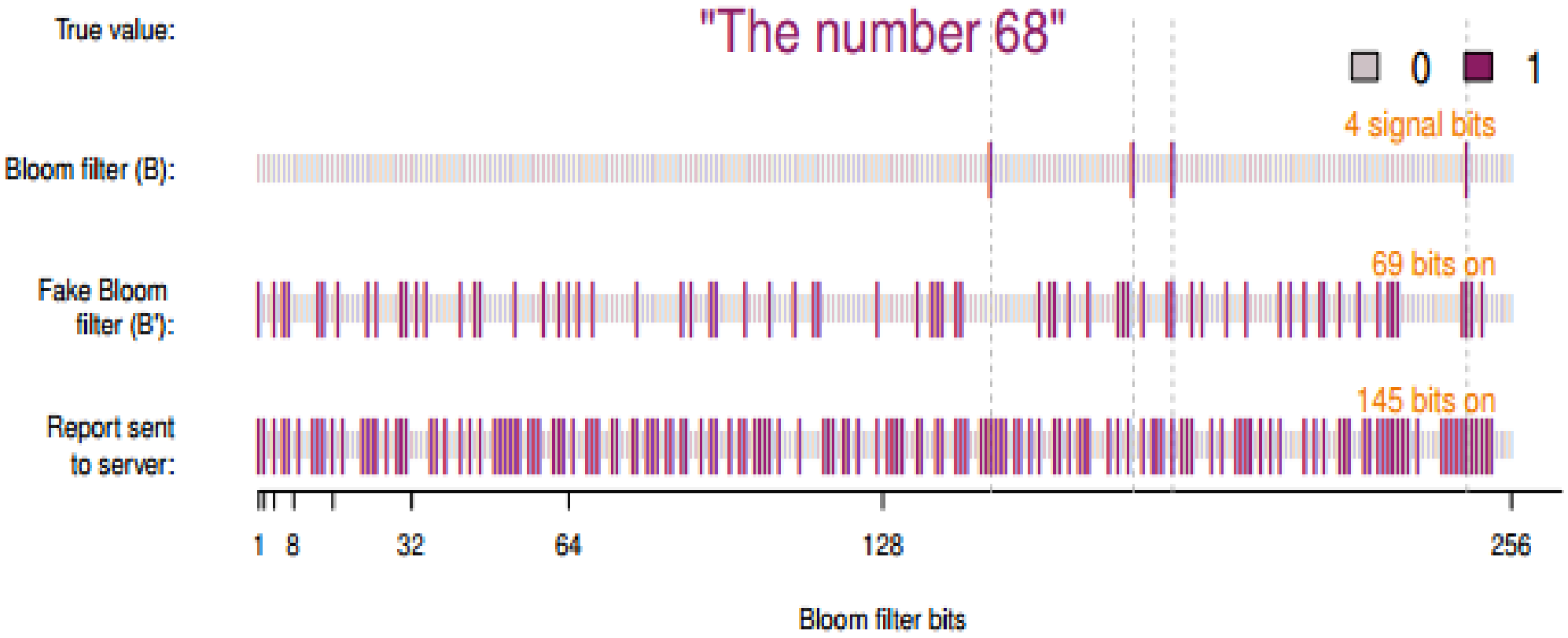
   Subsequently, this B' is memoized and reused as the basis for all future reports on this distinct value v.

3. **Instantaneous randomized response.** Allocate a bit array S of size k and initialize to 0. Set each bit i in S with probabilities

$$P(S_i = 1) = \begin{cases} q, & \text{if } B'_i = 1. \\ p, & \text{if } B'_i = 0. \end{cases}$$

4. **Report.** Send the generated report S to the server.

Steps of Google RAPPOR



f=0.5, q=0.75, p=0.5

On-Going Research Work

- Apply Raptor-based differential privacy algorithm to various IoT applications
  - Choosing f,p,q to maximize privacy for each application
- Implement it in FPGA and improve both client and server side performance
- Modify RAPPOR to allow the sending of multiple attribute labels and attribute values to send more complete data entry.

On-Going Research Work (2)

- Use deep learning to come up with a model for specific string set prediction, to provide even better accuracy than what LASSO regression provided.
- To test the client-server combination on light-weight IoT devices and test its applicability on such device.
- To combine the dimension reduction techniques used in DPTable with RAPPOR, and use more advanced techniques such as EM and modified LASSO regression on the analysis end to preserve relationships between attributes.

Concluding Remarks

- K-anonymity is the most well-known mechanism, but still faced difficulty when used in the real world.
- Differential Privacy provides a provable privacy guarantee at the theoretical level, but is not easy to be applied to real problems.
- Rappor is already being used in Chrome Browser and the likes, and seems quite suitable for privacy preserved analysis. But it can only support simple analysis now.
- Concrete privacy and utility metrics are still the key to further privacy solution development.

Thank you!

- For questions and inquiry for collaboration, please contact:
  - Yennun Huang
  - yennunhuang@gmail.com