ETH zürich

Optimization of Privacy-Utility Trade-offs under Informational Self-determination*

Presented

by Thomas Asikis, asikist@ethz.ch

Supervisors:

Prof. Dr. Dirk Helbing

Dr. Evangelos Pournaras

*https://arxiv.org/abs/1710.03186

UN Sustainability Goals

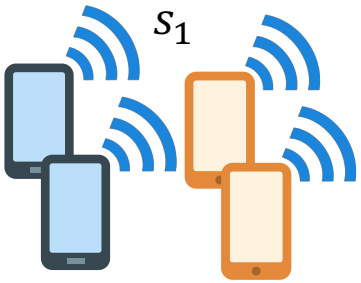
The UN Development Group has issued general guidance on data privacy, data protection and data ethics concerning the use of big data…

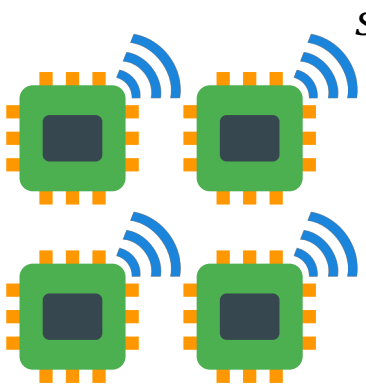
http://www.undp.org/content/undp/en/home/sustainable-development-goals.html
http://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html

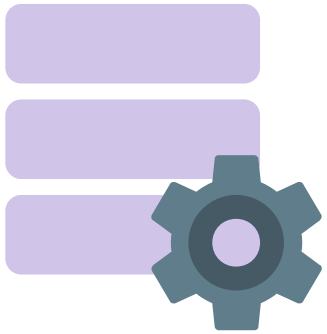
Common Data Flows: No privacy

$s_i \in \mathbb{S}$, actual values

$g(\mathbb{S}) \in \mathbb{R}^1$
aggregation







Data Producers

Data Consumer

Masking Data: Privacy Setting

Masking is an elementwise operation:

$$f(s_i; \theta) = m_i \iff \mathrm{F}(\mathbb{S}; \theta) = \mathbb{M}$$

f :*masking mechanism*

s_i: sensor value

θ: masking parameters

m_i: masked value

Privacy Preserving Masking

$$f(s_i; \theta) = s_i + \varepsilon_i$$

Privacy evaluation metrics (q) are calculated based on ε_i

E.g. Variance of ε_i values

Utility

Preserved benefit of data consumer operation on masked data:

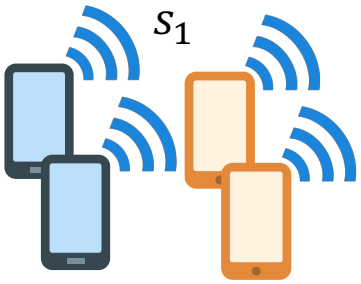
$$g(\mathbb{M}) = g(\mathbb{S}) + \delta, \delta \to 0$$

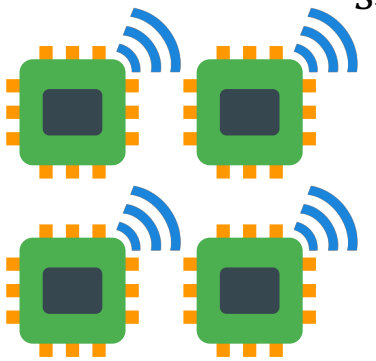
Utility evaluation metrics (u) calculated based on δ.

E.g. accuracy of masked total load in a smart grid

Current Data Flows: Engineering Privacy

$\mathbb{S}$: actual values

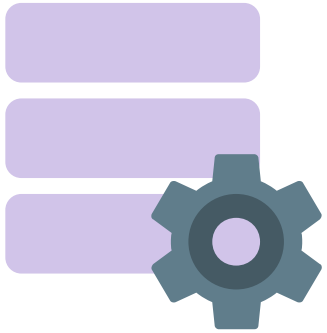




$F(\mathbb{S}; \theta^*)$

$g(\mathbb{M})$

Masked value aggregates



Data Producers

Data Consumer

Selection of Masking Mechanisms

Challenges:

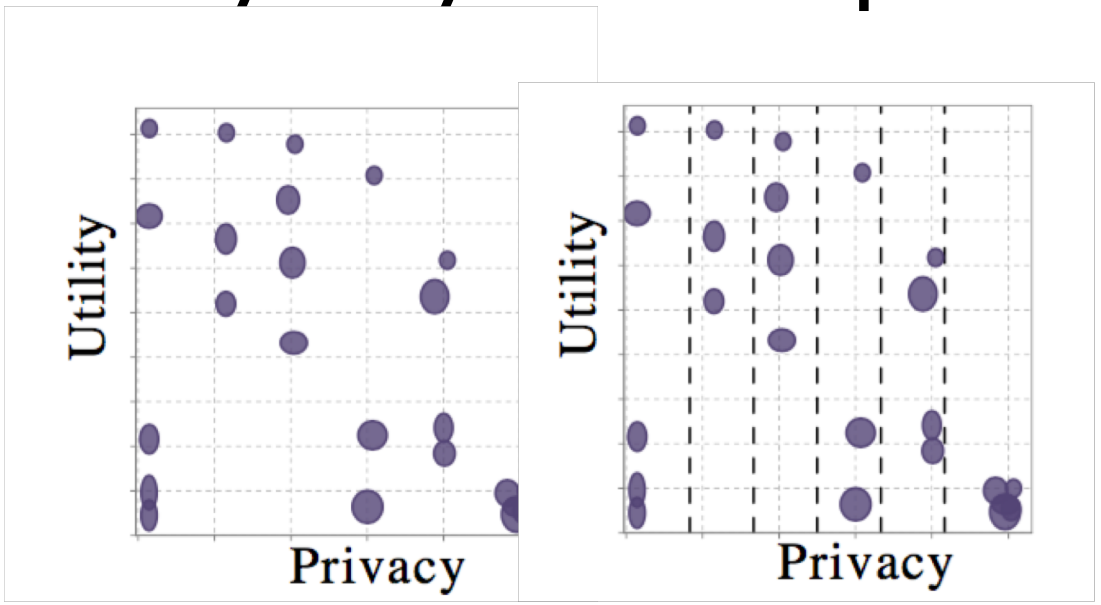
1. Evaluation criteria & comparison
2. Optimization
3. Constraints by user preferences

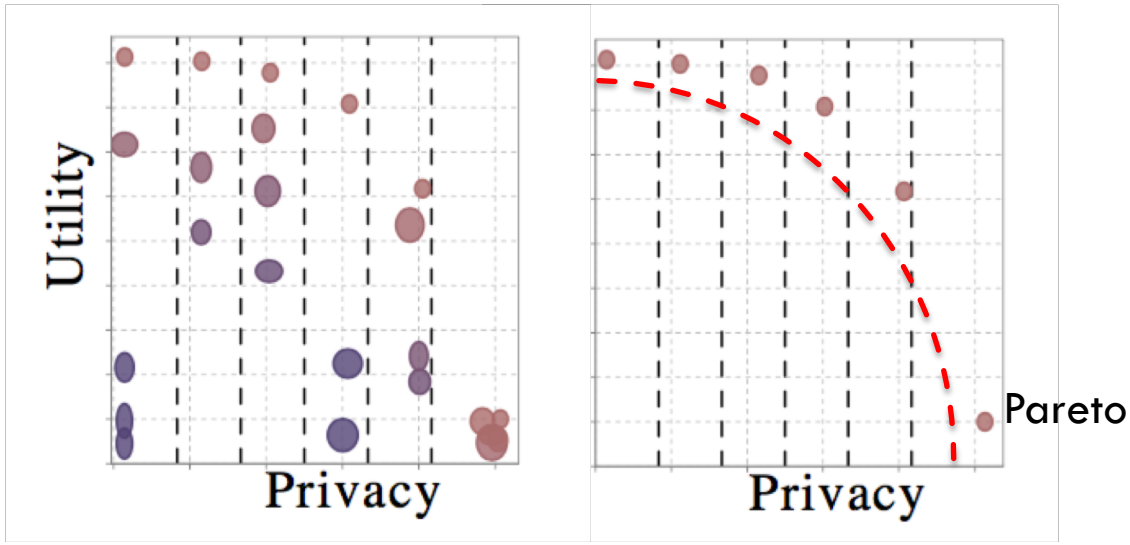
Multiple Masking Mechanisms

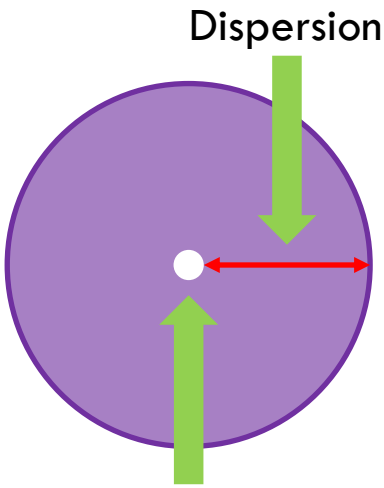
Solutions

1. Privacy q and Utility u, usually q is independent or opposing to u (Aggarwal, 2008)
2. Measure and compare trade-off (Aggarwal, 2008)
3. Optimize trade-off, NP-Hard (Krause, 2008)
4. Self-Determination (Novel)

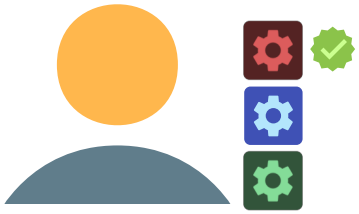
Privacy Utility Trade-off Optimization

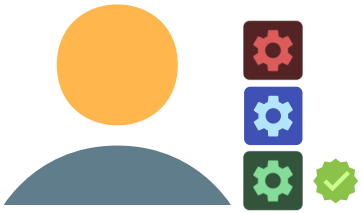


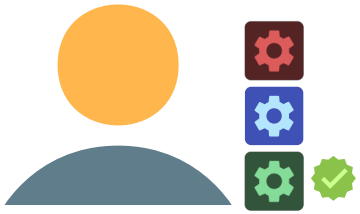


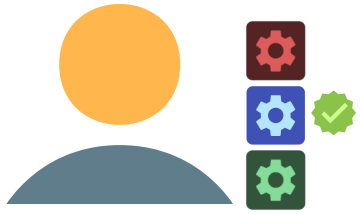


Universal Selection: A Selected Mechanism

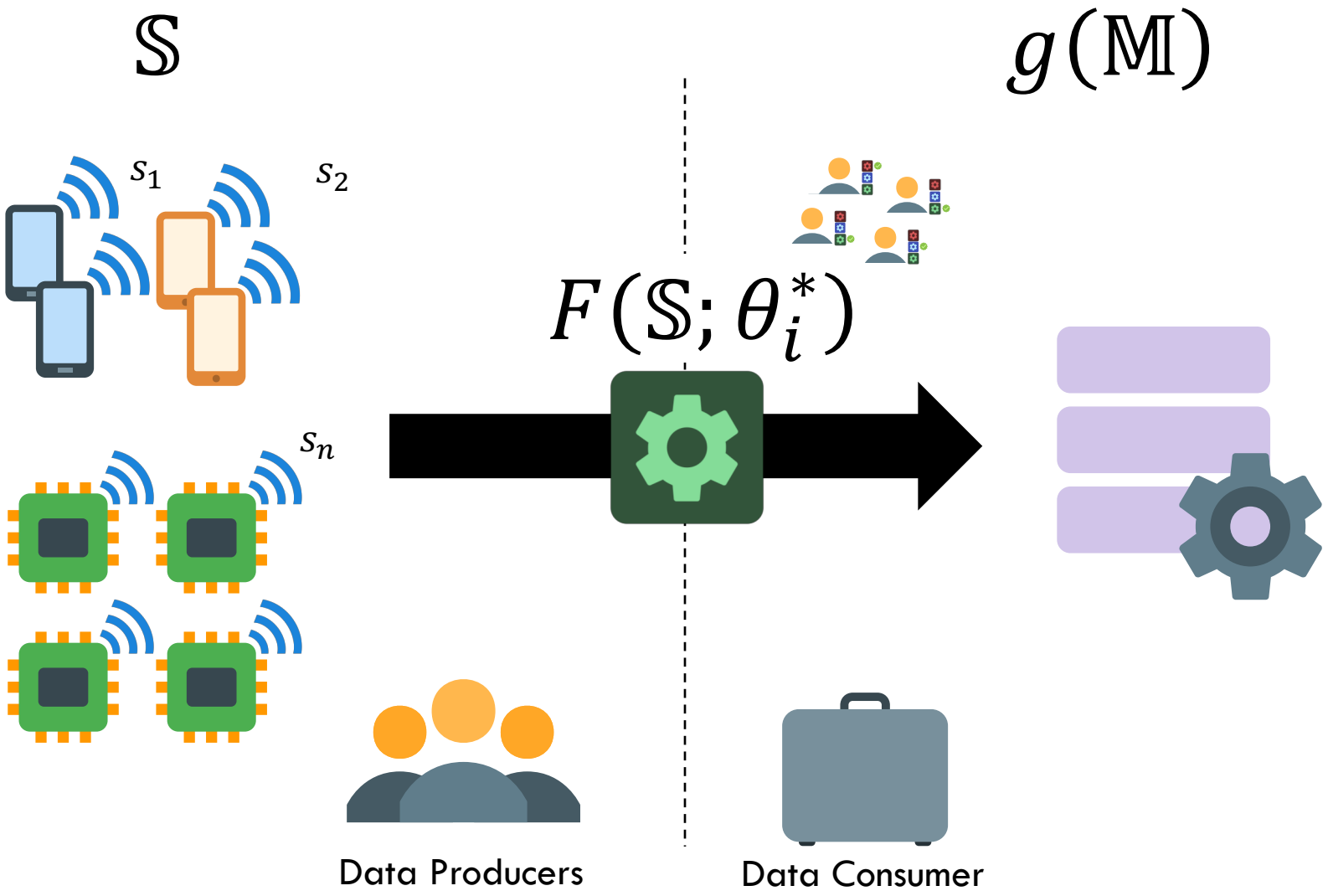








Data Flow: Universal Selection

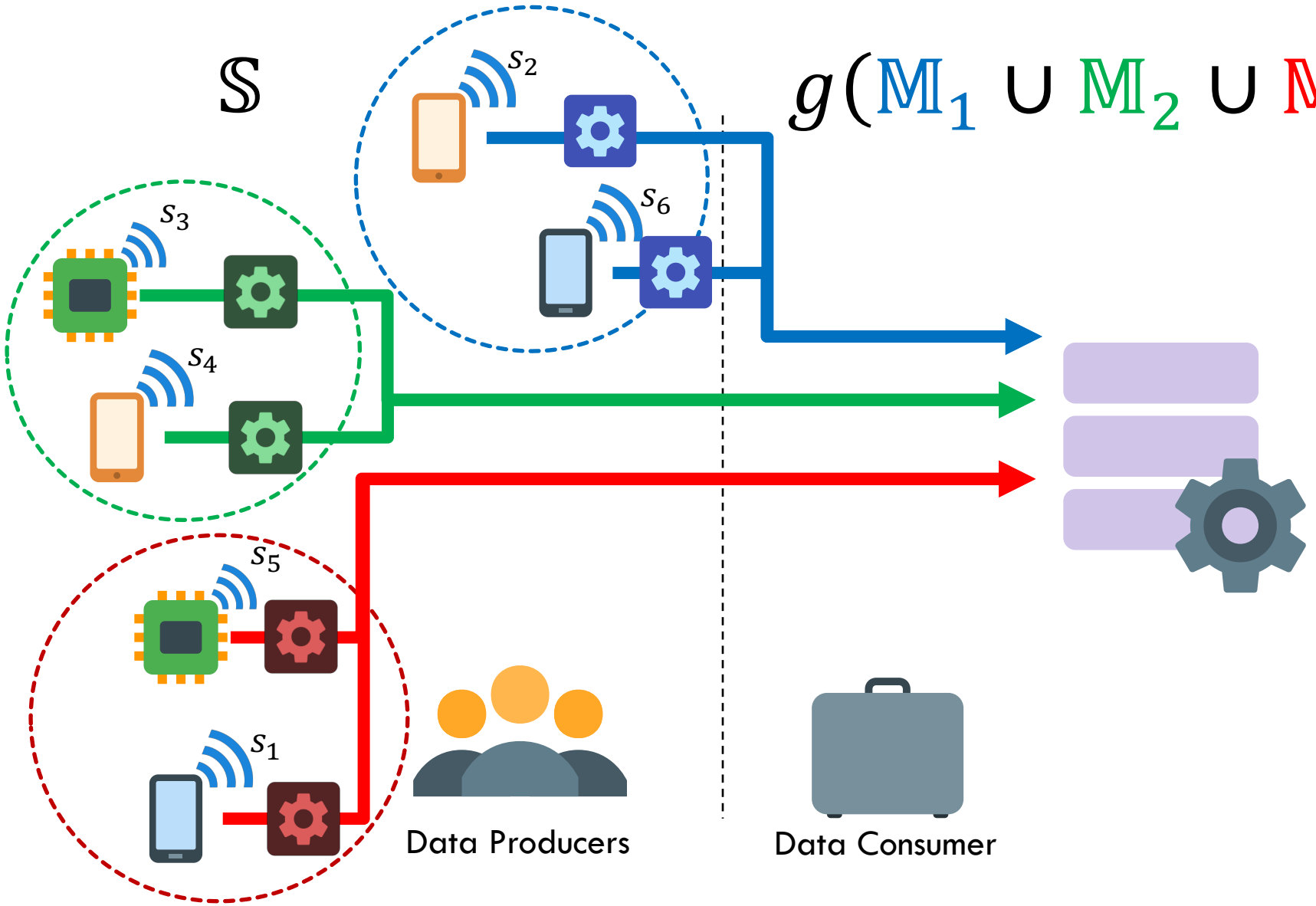


From Universal to Heterogeneous Selections

Self-Determination: Individuals select autonomously

Influence on privacy-utility trade-offs?

Data Flow: Heterogeneous Selections



Heterogeneous Selections

Analytically and experimentally proven:

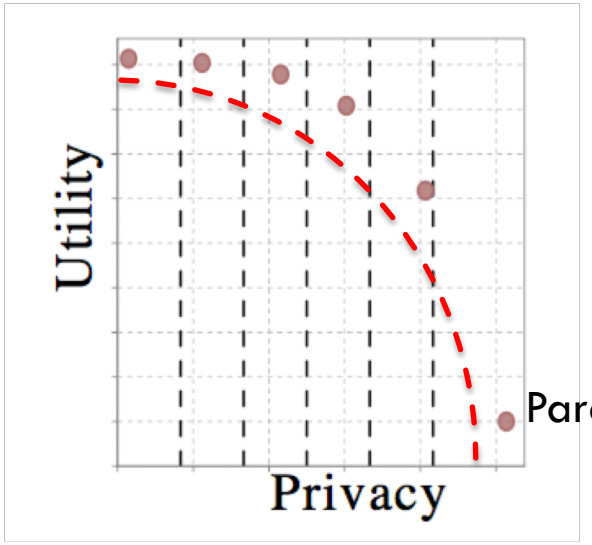
"The aggregation of masked values using different masking mechanisms, approximates the aggregation of all actual values."

$$g\left(\bigcup_i \mathbb{M}_i\right) = g(\mathbb{S}) + \delta, \delta \to 0$$

Experimental Settings

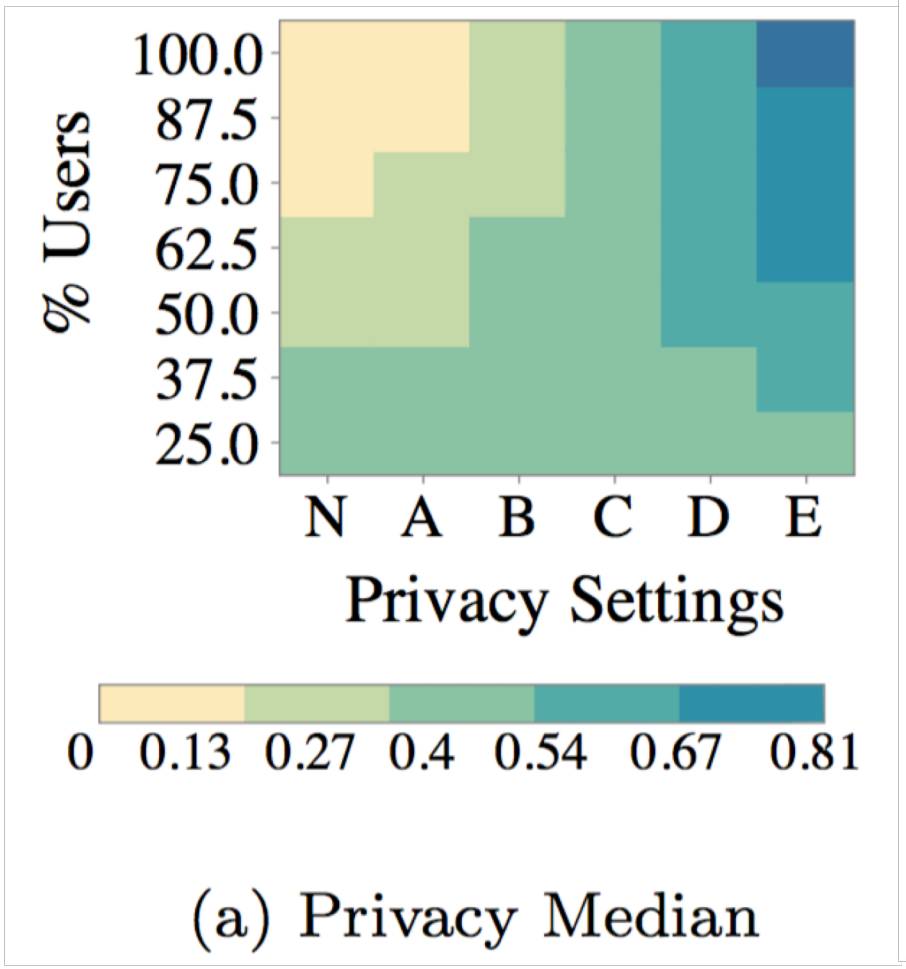
- Real World data: **6,435** users in **536** days
- **165,559,680** sensor values
- Over **20,000** privacy settings
- Multiple Repetitions per experiment for statistical significance
- Over **200,000** experiments
- Over 1 real world month of Euler runtime

Experimental Evaluation

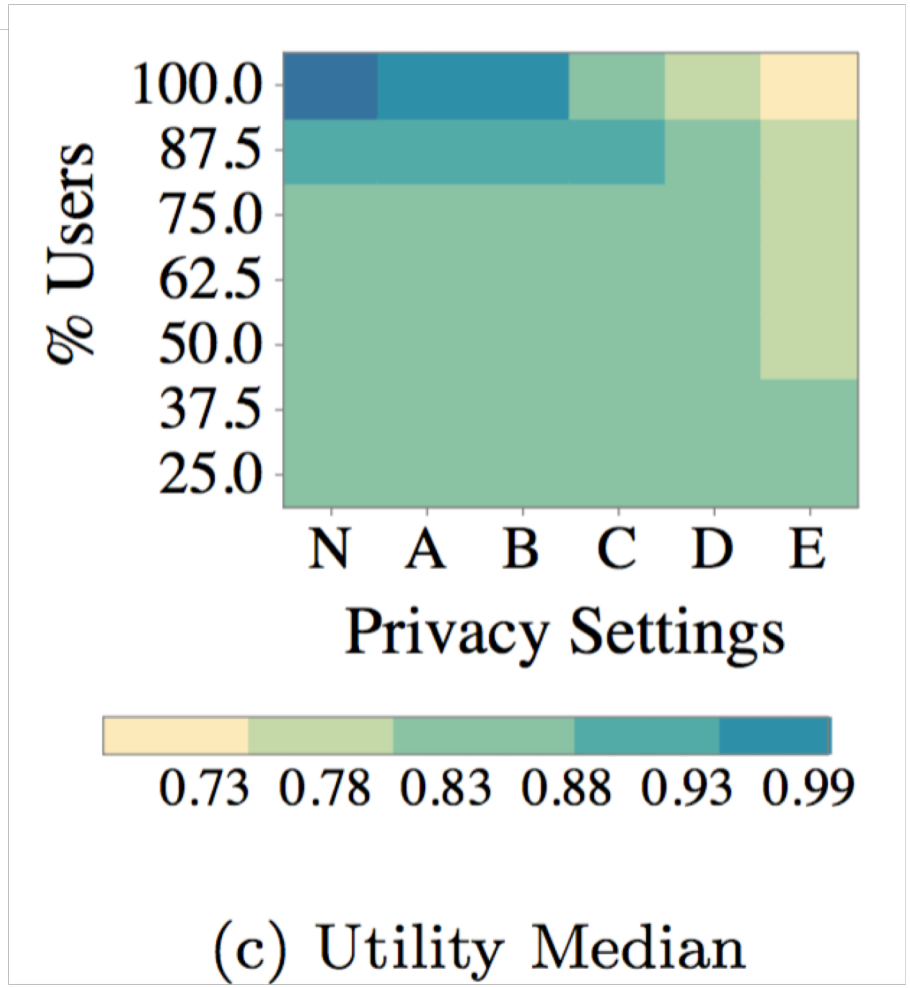


Masking Mechanism	Privacy	Utility
A	0.01	0.99
B	0.20	0.98
C	0.40	0.84
D	0.60	0.76
E	0.80	0.68
N (No Masking)	0.00	1.00

Results



(a) Privacy Median

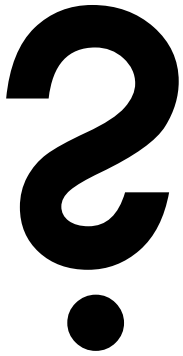


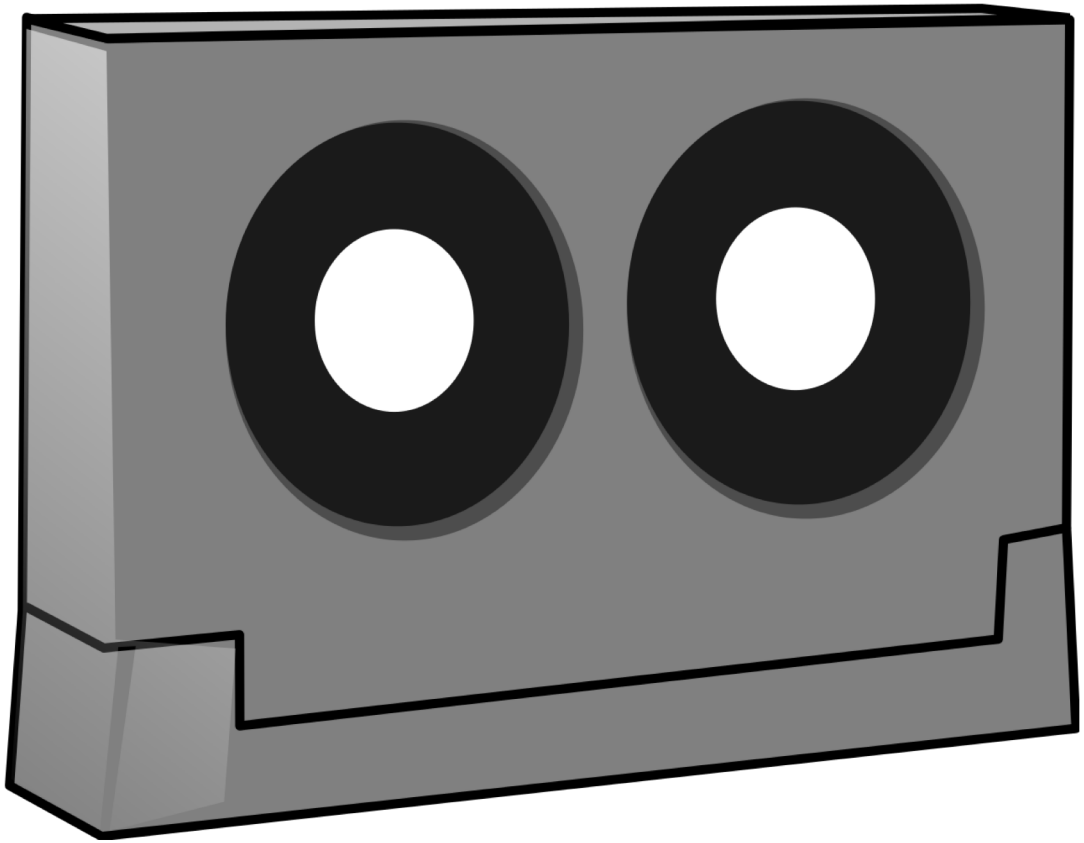
(c) Utility Median

Conclusions

- Privacy and utility can be optimized and personalized
- Measurable trade-offs between privacy and utility
- Analytical, empirical and experimental evidence that heterogeneous privacy setting selection is feasible

Questions

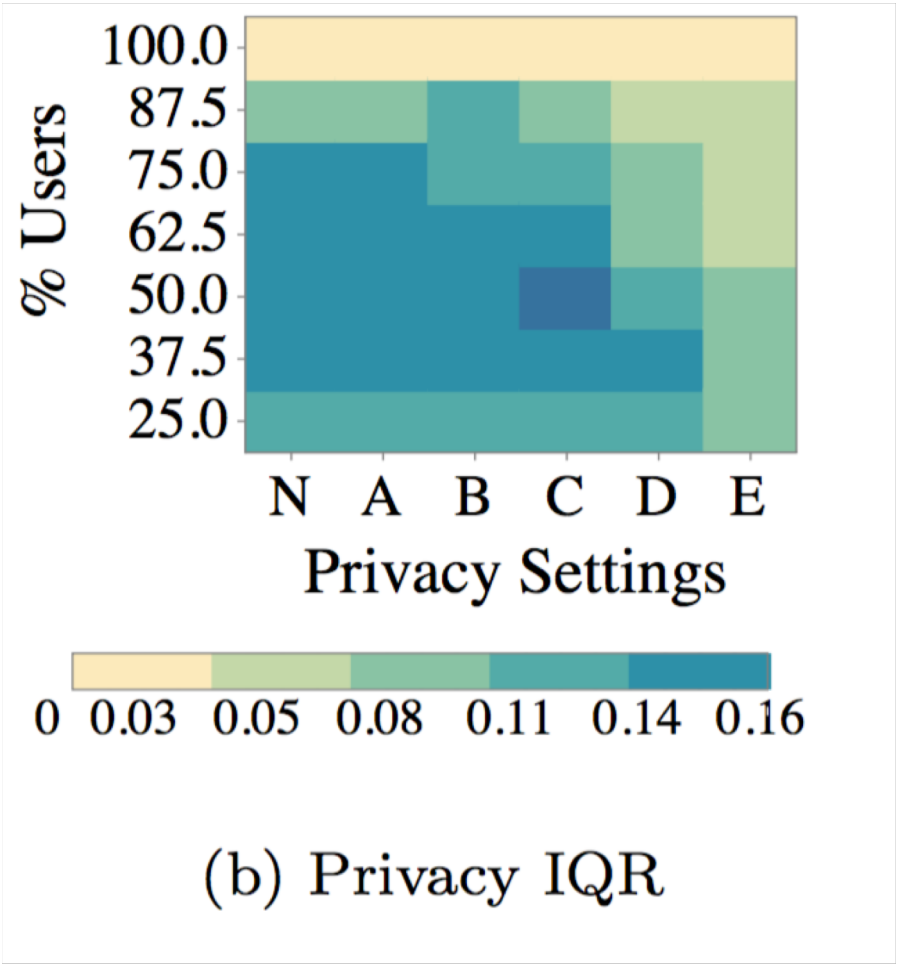




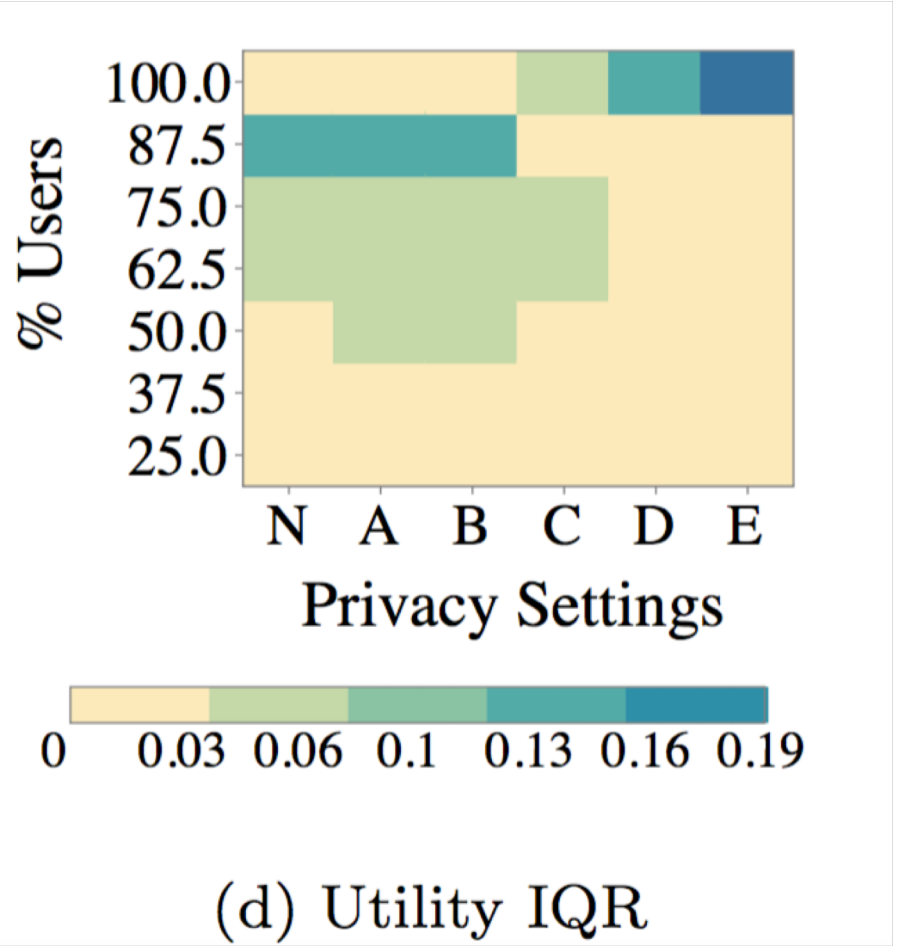
Backup Slides

Just in case

Privacy-Utility Dispersion



(b) Privacy IQR



(d) Utility IQR

How noise is generated

For sum-related aggregations:

noise ε_i sampled from symmetric distribution around zero,

For example in diffential privacy, the noise is generated via a Laplace distribution.

$$\varepsilon = L(\theta), \quad \theta = \{\mu = 0, \beta\}.$$

Privacy Metric

$$q = \alpha_1 \frac{\mu(\mathscr{E}_{\eta,k})}{\max\left(\mu(\mathscr{E}_{\eta,k})\right)} + \alpha_2 \frac{\sigma(\mathscr{E}_{\eta,k})}{\max\left(\sigma(\mathscr{E}_{\eta,k})\right)} + \alpha_3 \frac{H\left(\mathscr{E}_{\eta,k}\right)}{\max\left(H\left(\mathscr{E}_{\eta,k}\right)\right)}$$

Local Error

$$\varepsilon_{n,t} = \left| \frac{f_\eta\left(s_{n,t}, \theta_{\eta,k}\right) - s_{n,t}}{s_{n,t}} \right|$$

Global Error

$$\epsilon_t = \left| \frac{g(M_t) - S_t}{g(S_t)} \right|$$

Utility Metric

$$u = 1 - \left(\gamma_1 \frac{\mu(\epsilon_{\eta,k})}{\max\left(\mu(\epsilon_{\eta,k})\right)} + \gamma_2 \frac{\sigma(\epsilon_{\eta,k})}{\max\left(\sigma(\epsilon_{\eta,k})\right)} + \gamma_3 \frac{H\left(\epsilon_{\eta,k}\right)}{H\left(\max\left(\epsilon_{\eta,k}\right)\right)} \right)$$

Example of Harvard References

According to scientists, the Sun is pretty big.*
The Moon, however, is not so big.*

- [Miller 2005] ^ E. Miller, The Sun, (New York: Academic Press, 2005), 23-5.
- [Smith 1978] ^ R. Smith, "Size of the Moon", Scientific American, 46 (April 1978): 44-6.

*

References and Sources

Selected Related Literature

- Krause, A. and Horvitz, E. (2008). A utility-theoretic approach to privacy and personalization. In Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2, AAAI'08, pages 1181–1188. AAAI Press.
- Aggarwal, C. C. and Yu, P. S., editors (2008). Privacy- Preserving Data Mining, volume 34 of Advances in Database Systems. Springer US, Boston, MA.
- Dwork, C. (2006). Differential privacy. Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, pages 1–12.
- Li, C., Li, D. Y., Miklau, G., and Suciu, D. (2014). A Theory of Pricing Private Data. ACM Transactions on Database Systems, 39(4):1–28.