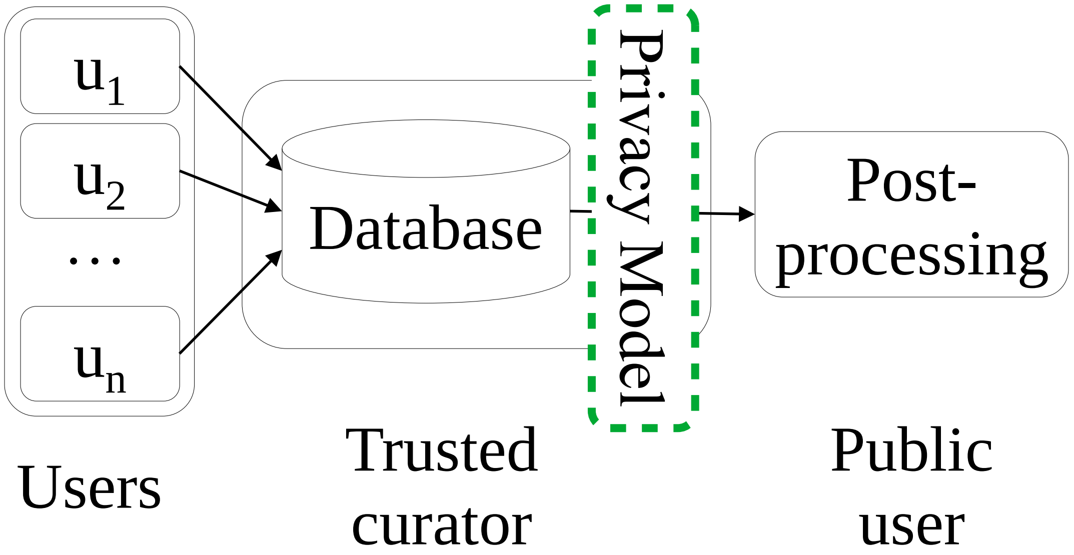
Random Sampling Plus Fake Data (RS+FD): Multidimensional Frequency Estimates With Local Differential Privacy*

Héber H. ARCOLEZI, Jean-François COUCHOT, Bechara AL BOUNA, Xiaokui XIAO

* SIGIR Student Travel Grant

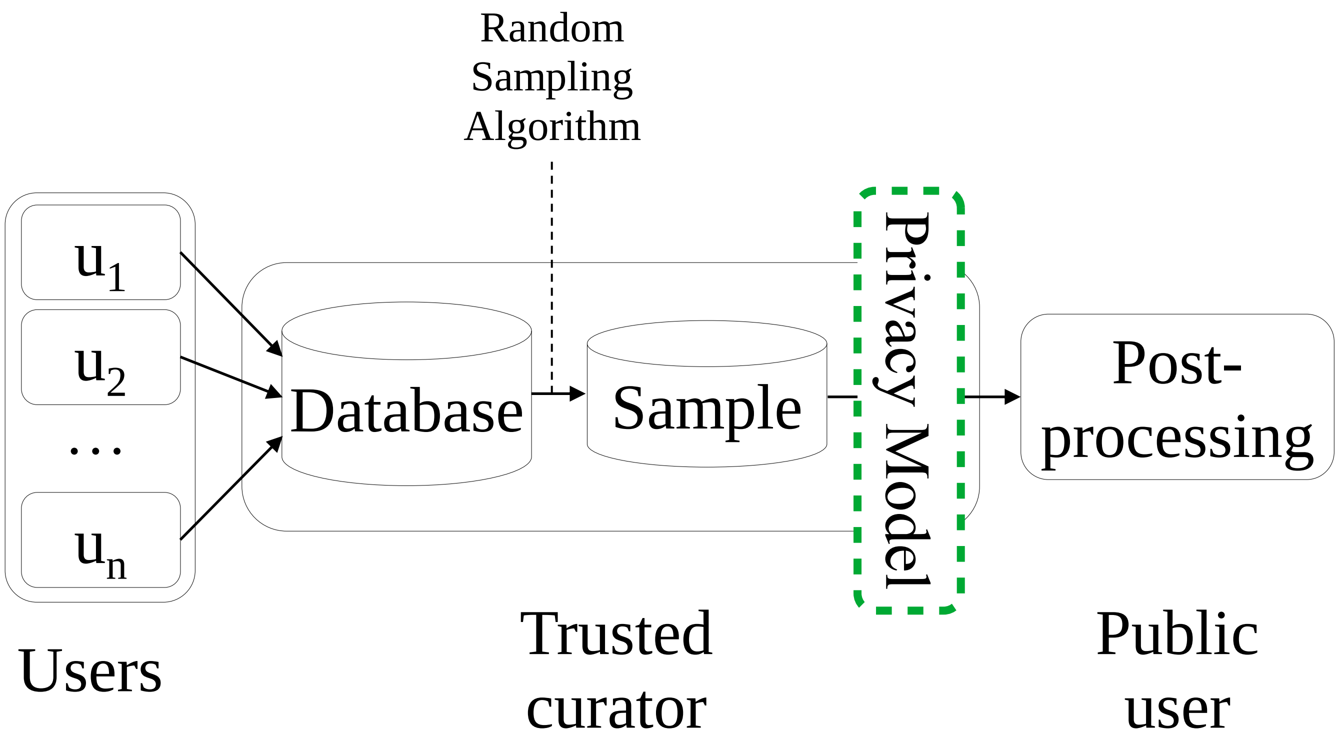
Differential Privacy (DP)[1]



- *Centralized* setting of DP.
- Interpretation: The addition (or removal) of anyone's record has a minimal (ϵ) influence on the outcome.
- Small $\epsilon \rightarrow$ stronger privacy
- $\epsilon \rightarrow$ a.k.a. "privacy budget"
- Robust to post-processing.

[1] Dwork, C., Roth, A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science (3–4), 211–407 (2014).

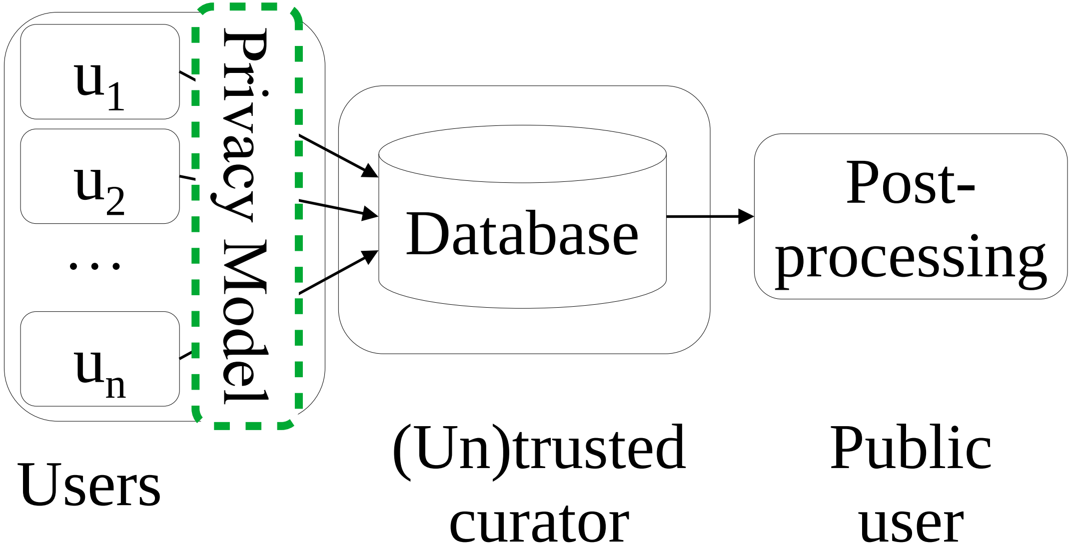
Amplification by Sampling[2]



- Randomly subsample the database w/ sampling rate β.
- Interpretation: an attacker is unable to distinguish which data samples were used in the analysis.
- Amplification: $\epsilon' \geq \epsilon$
- $\epsilon = \ln\left(1 + \beta\left(e^{\epsilon'} - 1\right)\right)$

[2] Ninghui Li, Wahbeh Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. ASIACCS'12 (2012).

Local Differential Privacy (LDP)[3]



- *Local* setting of DP.
- Interpretation: Any two items have close probability (controlled by ϵ) to be mapped to the same perturbed value.
- Several LDP implementation in practice.

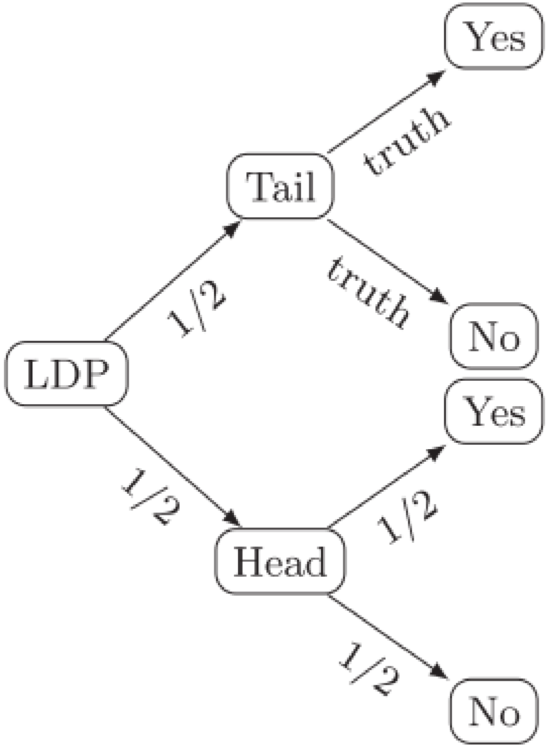
[3] Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A. What can we learn privately? In: 49th Annual FOCS (2008).

Randomized Response (RR)[4]

- Motivated by surveying people on sensitive topics.
- Main idea → Providing deniability to users' answer (yes/no → binary).
- Survey people: "Are you a member of the communist party?"
- Each person:
  - Throw a secret coin:
    - If tail throw the coin again (ignoring the outcome) and answer the question honestly.
    - If head, then throw the coin again and answer "Yes" if head, "No" if tail.

[4] Warner, S.L. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 1965.

RR's Unbiased Frequency Estimation



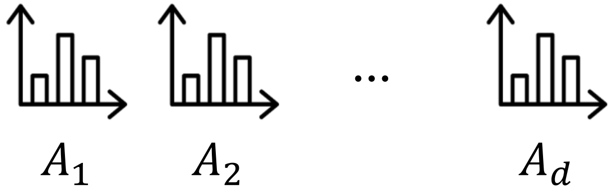
- $O_y \rightarrow$ proportion of *observed yes*
- $O_y \approx \frac{1}{2} t_y + \frac{1}{4} n$
- $t_y \rightarrow$ proportion of *true yes*
- $t_y \approx 2O_y - \frac{1}{2} n$
- Satisfies LDP w/: prob. of 'being honest' prob. of 'lying'
  - $\epsilon = \ln({}^{0.75}/_{0.25}) = \ln(3)$

Problem Statement

- **Key Issue:** Collecting *multidimensional* data under *ϵ-LDP* for the fundamental task of *frequency estimation*.

- **More formally (notation):**

  - d attributes $A = \{A_1, A_2, \dots, A_d\}$;
  - Each attribute A_j has a discrete domain D_j of size $|D_j| = k_j$;
  - Each user u_i for $1 \leq i \leq n$ has a tuple $v^i = (v_1^i, v_2^i, \dots, v_d^i)$;
  - **Analyzer:** estimate a k_j-bins histogram for each attribute $j \in [1, d]$.



Protocols for Single Frequency Estimation

- **Generalized RR (GRR)**[5]: Extends RR to the case of $k_j \geq 2$.

$$\forall_y \in D_j \; Pr\left[\psi_{GRR(\epsilon)}(v) = y\right] = \begin{cases} p = \frac{e^{\epsilon}}{e^{\epsilon}+k_j-1}, if\ y = v \\ q = \frac{1}{e^{\epsilon}+k_j-1}, if\ y \neq v \end{cases} \qquad \epsilon = \ln\left(\frac{p}{q}\right)$$

- **Optimized Unary Encoding (OUE)**[6]: Encode as a bit-vector B and perturb each bit independently into a new bit-vector B'. More specifically:

$$Pr[B'_i = 1] = \begin{cases} p = {}^1\!/_2, if\ B_i = 1 \\ q = \frac{1}{e^{\epsilon}+1}, if\ B_i = 0 \end{cases} \qquad \epsilon = \ln\left(\frac{p(1-q)}{q(1-p)}\right)$$

[5] Kairouz, P., Bonawitz, K. and Ramage, D. Discrete distribution estimation under local privacy. In International Conference on Machine Learning (2016).
[6] Wang, T., Blocki, J., Li, N. and Jha, S. Locally differentially private protocols for frequency estimation. In 26th USENIX Security Symposium (2017).

Protocols for Single Frequency Estimation

- **Unbiased Estimator**: To estimate the frequency $f(v_i)$ that a value v_i occurs for $i \in [1, k_j]$ one calculates

$\hat{f}(v_i) = \frac{N_i - nq}{n(p-q)}$, N_i = number of times the value (or bit) i has been reported.

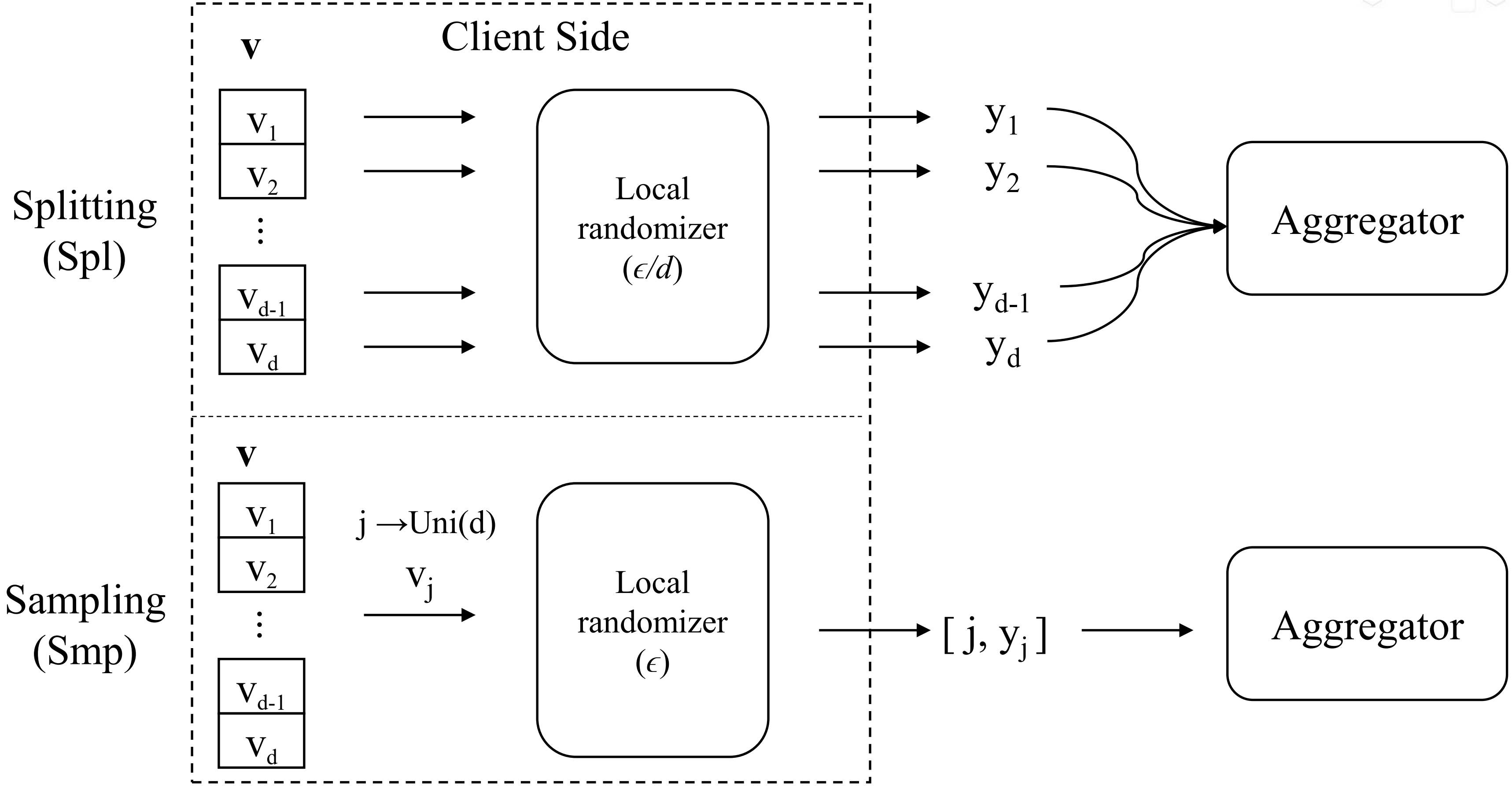
- **Approximate Variances**:

$$Var[\hat{f}_{GRR}(v_i)] = \frac{e^{\epsilon} + k_j - 2}{n(p-q)^2} \qquad Var[\hat{f}_{OUE}(v_i)] = \frac{4e^{\epsilon}}{n(p-q)^2}$$

- **Adaptive LDP protocol**[6]: Given $k_j, p, q, and\ \epsilon$

$$ADP = \begin{cases} GRR & if\ k_j < 3e^{\epsilon} + 2 \\ OUE & otherwise. \end{cases}$$

What is the state-of-the-art?



Why not *Smp*?

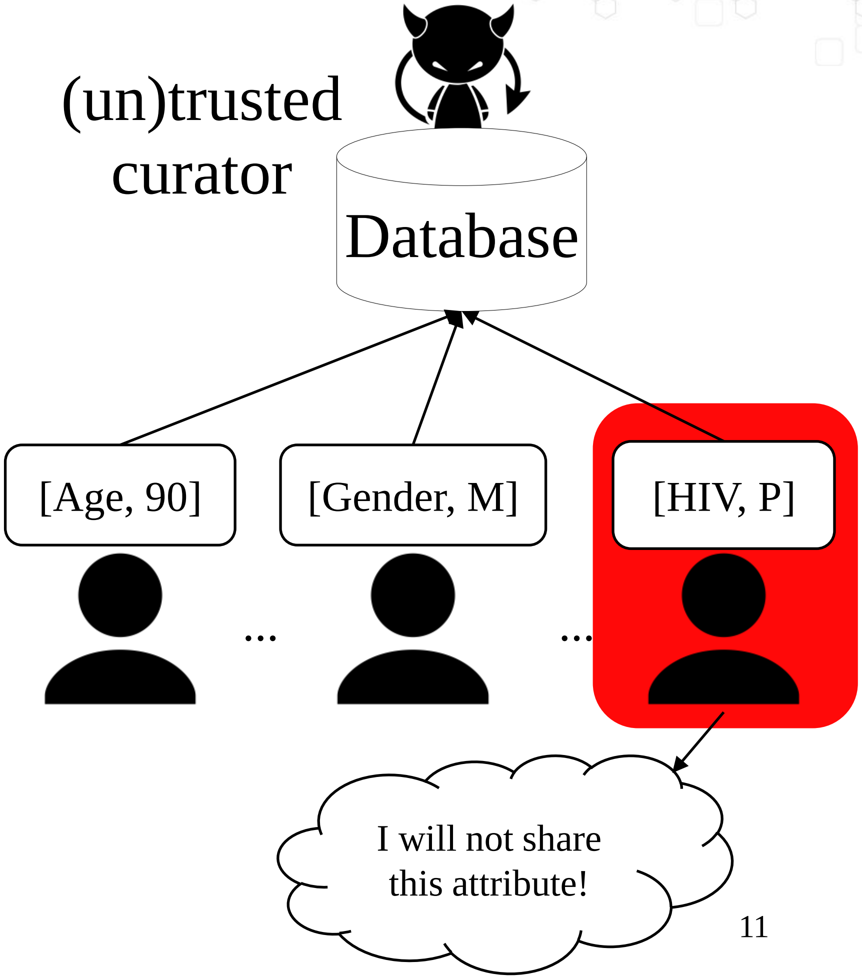
GRR for attributes with small domain
OUE otherwise

Example:

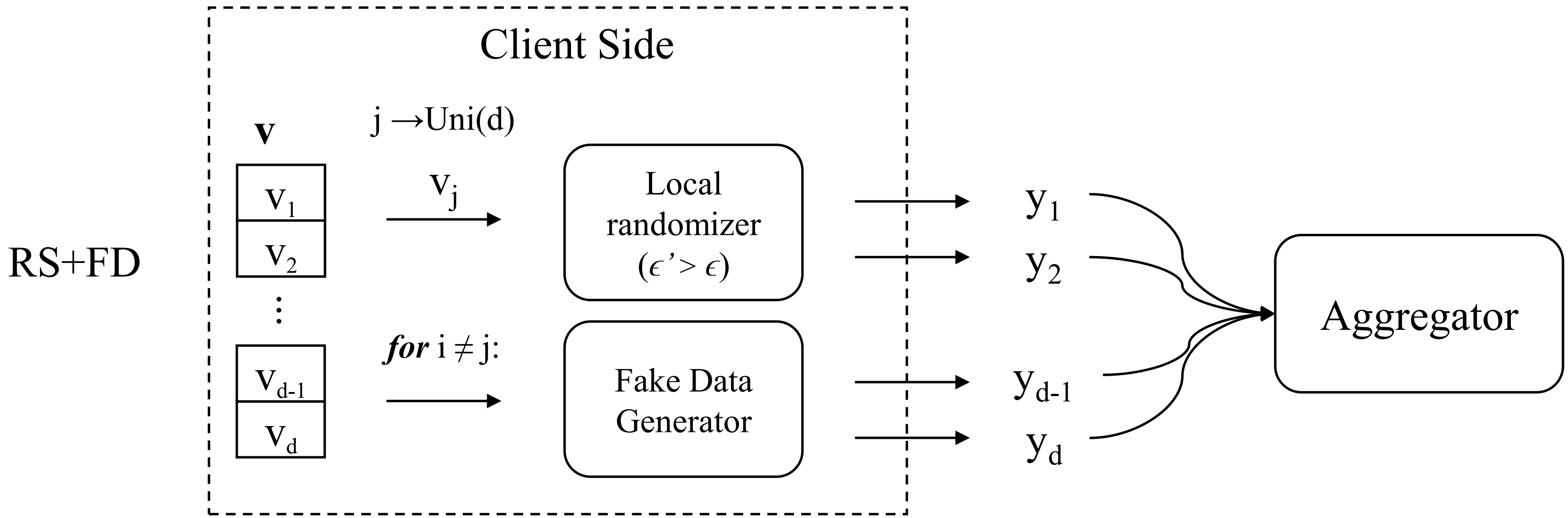
- *Smp*[ADP] → (attribute, ϵ-LDP value)
- Application scenario: health data
- $\epsilon = 2$, $d = 3$ attributes: age ($k_1 = [1, \dots, 100]$), gender ($k_2 = [\mathrm{M}, \mathrm{F}]$), and HIV ($k_3 = [\mathrm{P}, \mathrm{N}]$).

$p_{grr} = \frac{e^{\epsilon}}{e^{\epsilon}+k_j-1} \approx 0.88$ (probability of 'being honest')

$q_{grr} = \frac{1-p_{grr}}{k_j-1} \approx 0.12$ (probability of 'lying')



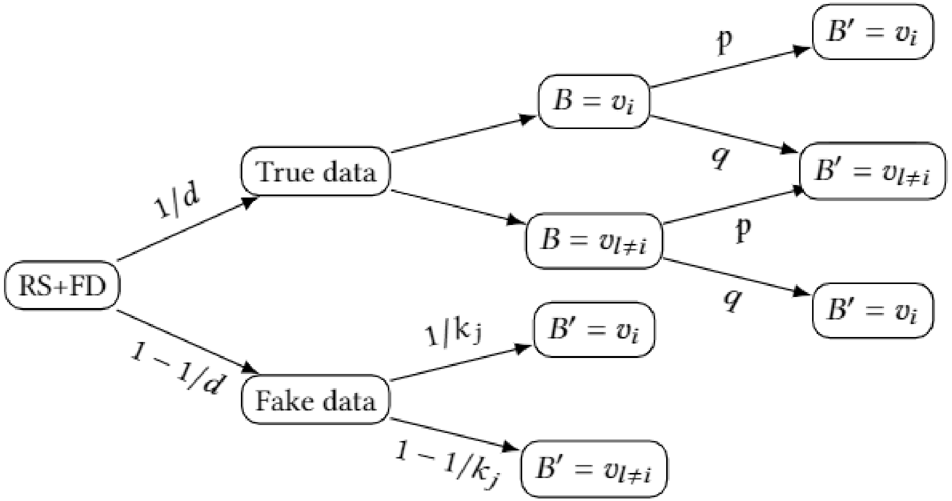
RS+FD: Random Sampling + Fake Data



RS+FD[GRR]

Algorithm 1 RS+FD[GRR]: Client Side

```
Input : tuple v = (v_1, v_2, ..., v_d), domain size of attributes k = [k_1, k_2, ..., k_d], privacy
parameter ε, local randomizer GRR.
Output : privatized tuple y = (y_1, y_2, ..., y_d).
ε' = ln (d · (e^ε − 1) + 1)                    ▷ amplification by sampling [31]
j ← Uniform({1, 2, ..., d})                     ▷ Selection of attribute to privatize
B_j ← v_j
y_j ← GRR(B_j, k_j, ε')                         ▷ privatize data of the sampled attribute
for i ∈ {1, 2, ..., d}/j do                     ▷ non-sampled attributes
    y_i ← Uniform({1, ..., k_i})                ▷ generate fake data
end for
return : y = (y_1, y_2, ..., y_d)               ▷ sampling result is not disclosed
```



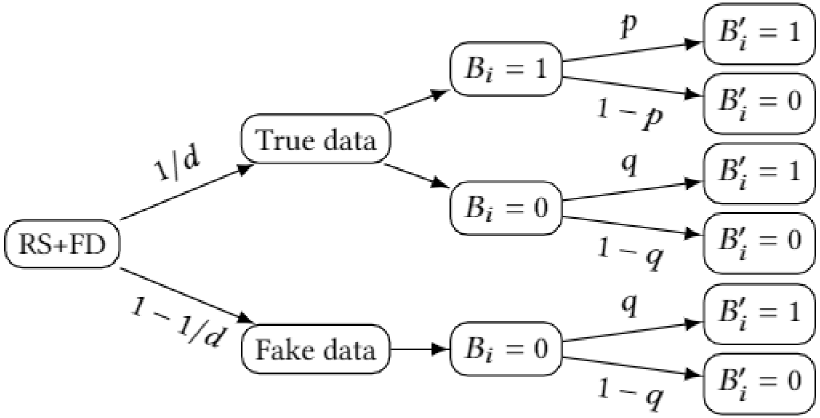
Aggregator → For each attribute, estimate: $\hat{f}(v_i) = \dfrac{N_i d k_j - n\left(d - 1 + q k_j\right)}{n k_j (p - q)}$

RS+FD[OUE-z]

Algorithm 2 RS+FD[OUE-z]: Client Side

Input : tuple $\mathbf{v} = (v_1, v_2, ..., v_d)$, domain size of attributes $\mathbf{k} = [k_1, k_2, ..., k_d]$, privacy parameter ϵ, local randomizer OUE.
Output : privatized tuple $\mathbf{B}' = (B'_1, B'_2, ..., B'_d)$.

1: $\epsilon' = \ln(d \cdot (e^{\epsilon} - 1) + 1)$ ▷ amplification by sampling [31]
2: $j \leftarrow Uniform(\{1, 2, ..., d\})$ ▷ Selection of attribute to privatize
3: $B_j = Encode(v_j) = [0, 0, ..., 1, 0, ...0]$ ▷ one-hot-encoding
4: $B'_j \leftarrow OUE(B_j, \epsilon')$ ▷ privatize real data with OUE
5: **for** $i \in \{1, 2, ..., d\}/j$ **do** ▷ non-sampled attributes
6: $\quad B_i \leftarrow [0, 0, ..., 0]$ ▷ initialize zero-vectors
7: $\quad B'_i \leftarrow OUE(B_i, \epsilon')$ ▷ randomize zero-vector with OUE
8: **end for**
return : $\mathbf{B}' = (B'_1, B'_2, ..., B'_d)$ ▷ sampling result is not disclosed



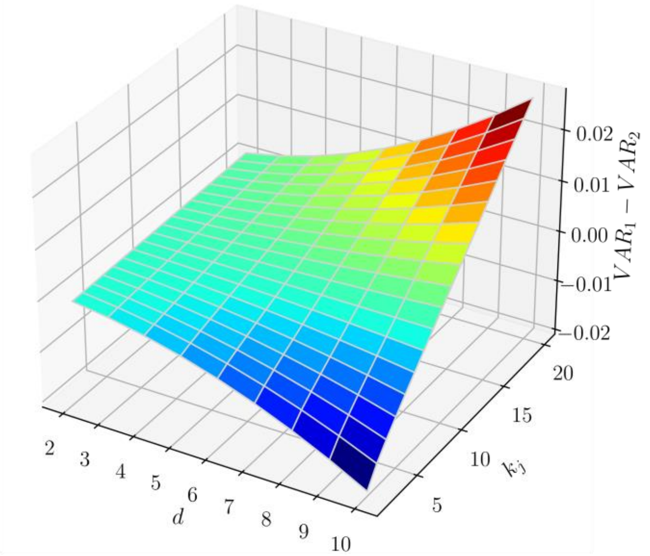
Aggregator → For each attribute, estimate: $\hat{f}(v_i) = \dfrac{d(N_i - nq)}{n\,(p - q)}$

RS+FD[ADP]

- Let $\mathrm{VAR}_1 = \mathrm{VAR}_{\text{RS+FD[GRR]}}$ and $\mathrm{VAR}_2 = \mathrm{VAR}_{\text{RS+FD[OUE-z]}}$
- For each attribute, given d, k_j, and ϵ', select RS+FD[GRR] **if**:

$$\mathrm{VAR}_1 \leq \mathrm{VAR}_2, \text{ i.e., } \textbf{if } \mathrm{VAR}_1 - \mathrm{VAR}_2 \leq 0$$

- Let $\mathrm{n} = 10000$, $\mathrm{d} \in [2, 10]$, $k_j \in [2, 20]$, and $\epsilon' = \ln(3)$



Experiments

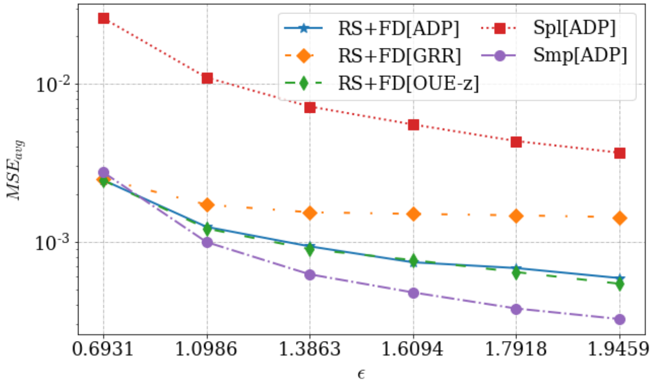
- Datasets:
  - Nursery[7]: $\mathrm{n} = 12960$, $\mathrm{d} = 9$, $\mathbf{k} = [3,5,4,4,3,2,3,3,5]$
  - Adults[7]: $\mathrm{n} = 45422$, $\mathrm{d} = 9$, $\mathbf{k} = [7,16,7,14,6,5,2,41,2]$
  - MS-FIMU[8]: $\mathrm{n} = 88935$, $\mathrm{d} = 6$, $\mathbf{k} = [3,3,8,12,37,11]$
  - Census-Income[7]: $\mathrm{n} = 299285$, $\mathrm{d} = 33$, $\mathbf{k} = [9,52,47,17, \dots, 3,3,2]$
- Evaluation: ϵ = [ln(2), ln(3), ..., ln(7)].
- Metric:

$$\mathrm{MSE}_{avg} = \frac{1}{d} \sum_{j \in [1,d]} \frac{1}{|D_j|} \sum_{v_i \in D_j} \left(f(v_i) - \hat{f}(v_i)\right)^2$$

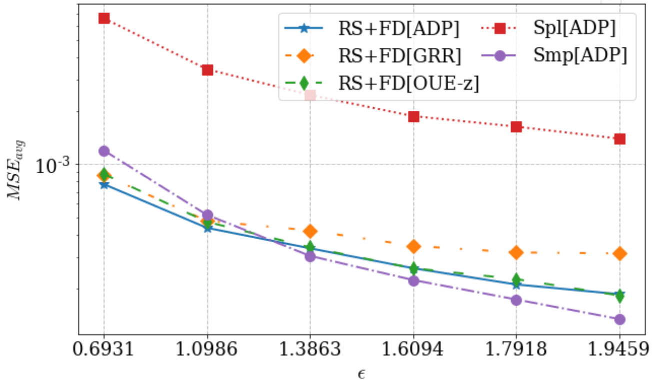
[7] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/index.php

[8] Arcolezi, H.H., Couchot, J.F., Baala, O., et al. Mobility modeling through mobile data: generating an optimized and open dataset respecting privacy. In 16th IWCMC (2020).

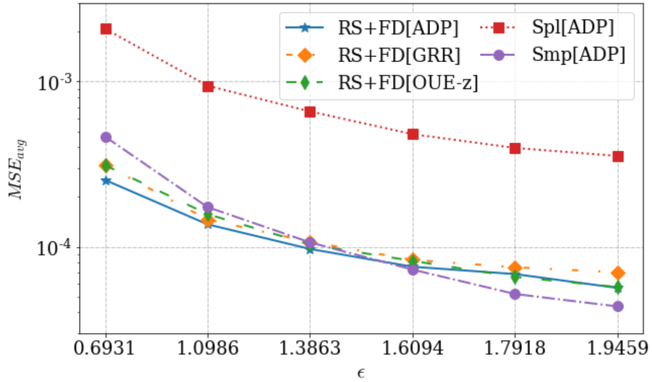
Results



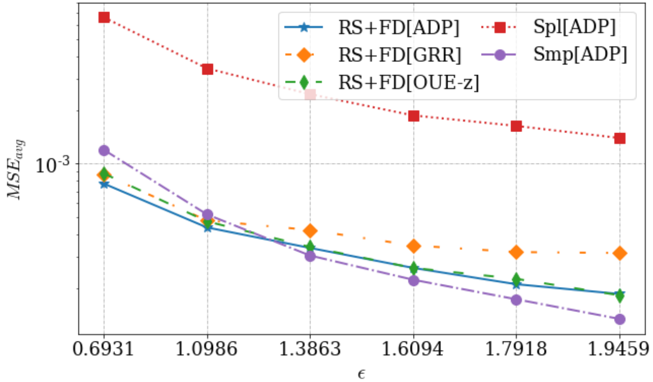
Nursery[7]: n = 12960, d = 9, **k** = [3,5,4,4,3,2,3,3,5]



Adults[7]: n=45422, d=9, **k**=[7,16,7,14,6,5,2,41,2]



MS-FIMU[8]: n = 88935, d = 6, **k** = [3,3,8,12,37,11]



Census-Income[7]: n = 299285, d = 33, **k** = [9,52,47, … , 3,2]

Conclusions

- We propose a generic framework RS+FD for multidimensional frequency estimates under LDP with theoretical proofs.
- RS+FD achieves nearly the same or better utility than *Smp* with higher privacy protection (uncertainty).
- Limitations:
  - Sampling error + noise from fake reports;
  - More computation and communication cost than *Smp*.
- Perspectives:
  - Cast other LDP protocols into RS+FD;
  - Attack: is it possible to state which attribute value is “fake”?

References

1. Dwork, C., Roth, A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science (3–4), 211–407 (2014).
2. Ninghui Li, Wahbeh Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security - ASIACCS'12 (2012).
3. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A. What can we learn privately? In: 49th Annual IEEE Symposium on Foundations of Computer Science (2008).
4. Warner, S.L. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 1965.
5. Kairouz, P., Bonawitz, K. and Ramage, D. Discrete distribution estimation under local privacy. In International Conference on Machine Learning (2016).
6. Wang, T., Blocki, J., Li, N. and Jha, S. Locally differentially private protocols for frequency estimation. In 26th USENIX Security Symposium (2017).
7. Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/index.php
8. Arcolezi, H.H., Couchot, J.F., Baala, O., et al. Mobility modeling through mobile data: generating an optimized and open dataset respecting privacy. In 2020 International Wireless Communications and Mobile Computing (2020).

Thank you very much for your attention!!!

Questions?

Codes → https://github.com/hharcolezi/ldp-protocols-mobility-cdrs

Contact → heber.hwang_arcolezi@univ-fcomte.fr