OTTAWA 2023

64TH WORLD STATISTICS CONGRESS

CPS 07
Statistical estimation II

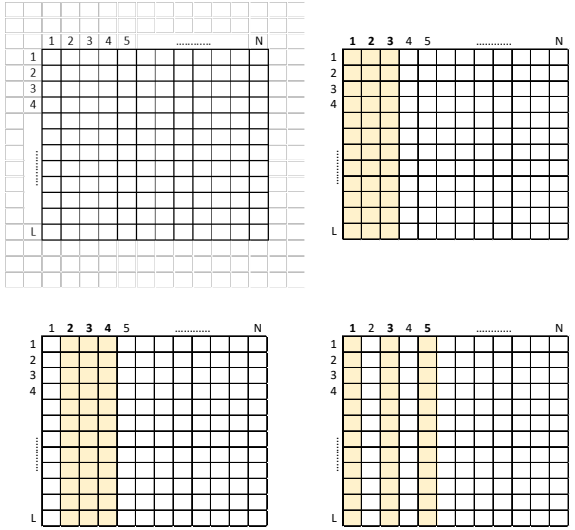
Quantifying the contribution of individual records to the reidentification risk of (pseudo)anonymized datasets

Michel Béra[1], Vasiliki Daskalaki[2], Gilbert Saporta[1], Kimon Spiliopoulos[2], Konstantinos Spinakis[3] and Photis Stavropoulos[2]

[1]CNAM [2]Quantos S.A. [3]EPFL

Monday 17 July, 8:30AM - 9:40AM

Dataset risk according to the QaR method (1)



- $L \times N$ tabular dataset
- Each set of p columns, $1 \leq p \leq N$, is a quasi-identifier of size p

- $N_p = \binom{N}{p} = \frac{N!}{p!(N-p)!}$ quasi-identifiers of size p

Dataset risk according to the QaR method (2)

Reidentification risk associated with quasi-identifier $Q(j), j = 1,2, \dots, N_p$

$$\theta(j) = \frac{H(j)}{L}$$

$H(j)$: number of distinct values assumed by $Q(j)$ in the dataset

Sex	Age group	Nationality
Female	55-64	Greek
Male	25-34	Greek
Female	25-34	Italian
Male	25-34	Greek
Male	35-44	Greek
Female	55-64	Greek
Male	25-34	Greek
Male	25-34	Greek

Female	55-64	Greek	
Male	25-34	Greek	
Female	25-34	Italian	
Male	35-44	Greek	

$$\theta(j) = \frac{4}{8} = 0.5$$

Dataset risk according to the QaR method (3)

- Compute the risks $\theta(j), j = 1,2, \ldots, N_p$, associated with all quasi-identifiers of size p
- Compute the empirical $1 - \pi_u$ quantile, u, of the risks
- Retain the N_u risks that are greater than u; **Extreme risks**
- Fit a Generalised Pareto Distribution (GPD) to a logit transformation of the extreme risks
- Based on the estimated GPD parameters, estimate the **$1 - \alpha$ quantile of the distribution of risks**

AFNOR SPEC Z90-030

$$\hat{T}(p,\alpha) = exp\left(ln\left(\frac{u}{1-u}\right) + \frac{\hat{\beta}}{\hat{\xi}}\left[\left(\frac{N_p \cdot \alpha}{N_u}\right)^{-\hat{\xi}} - 1\right]\right) / \left\{1 + exp\left(ln\left(\frac{u}{1-u}\right) + \frac{\hat{\beta}}{\hat{\xi}}\left[\left(\frac{N_p \cdot \alpha}{N_u}\right)^{-\hat{\xi}} - 1\right]\right)\right\}$$

- This is the **dataset's reidentification risk**

A record's contribution to risk

- Contribution of record $i, i = 1,2, \dots, L$: $\hat{T}(p, \alpha)$- $\hat{T}(p, \alpha: i)$

- An indication of the risk of identifying the respective statistical unit
- Removal of a few "largest contributors" may reduce considerably the reidentification risk of the dataset

- 'Proper' backward elimination of records very expensive computationally
- Number of dataset risk computations involved in S eliminations

$$1 + \sum_{s=1}^{S} (L - s + 1)$$

E.g., removal of 10 out of 12000 records: 119956 risk computations

A record's contribution to risk: proxies

- Contribution to the initial dataset's risk: $DT(i) = \hat{T}(p, \alpha)$- $\hat{T}(p, \alpha : i)$
- Contribution to the initial dataset's risk as estimated by PLS regression of the sign of $DT(i)$ on the record's contents (numerical values)
- A record entropy

$\mathrm{F}(i,j) = \frac{1}{L}\sum_{r=1}^{L} \mathbb{1}\left(x_{r,j_1} \leq x_{i,j_1},\ x_{r,j_2} \leq x_{i,j_2}, \dots,\ x_{r,j_p} \leq x_{i,j_p}\right)$ empirical cdf of $Q(j)$

$$\mathrm{H}(i,j) = \frac{1}{L}\sum_{r=1}^{L} \mathbb{1}\left(\mathrm{F}(r,j) \leq \mathrm{F}(i,j)\right)$$

$E(i) = \sum_{m=1}^{\sqrt{N_p}}(-\, p_m^i \, log(p_m^i))$, where p_m^i the proportion of $\mathrm{H}(i,j)$ that fall in $[(m-1) * 1/\sqrt{N_p}\ ,\ m * 1/\sqrt{N_p}\ ]$

A record's contribution to risk: proxies

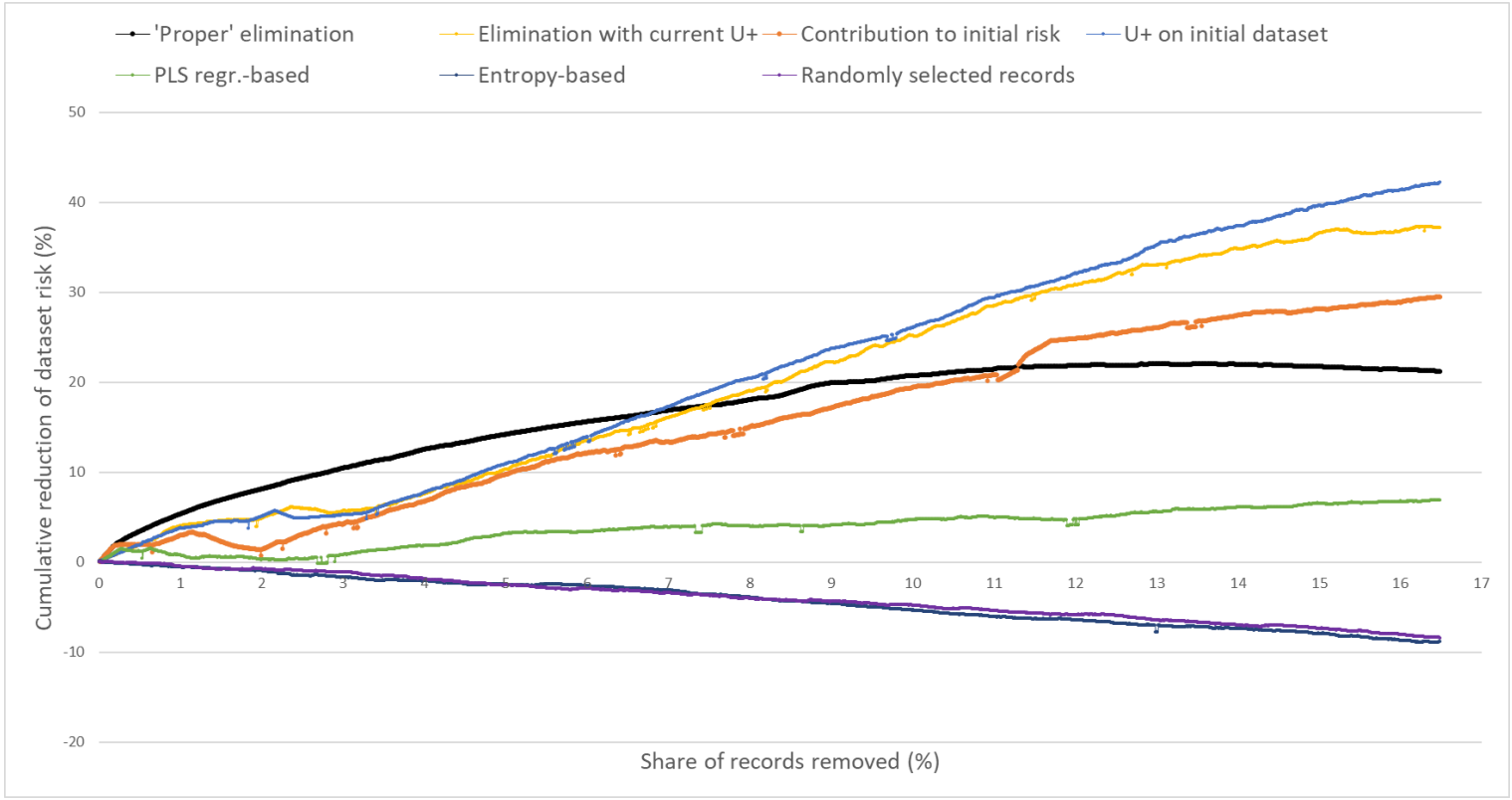
- Uniqueness pattern of a record

$$U(i,j) = \begin{cases} 1 & \text{if record } i \text{ is unique w.r.t. } Q(j) \\ 0 & \text{if record } i \text{ is not unique w.r.t. } Q(j) \end{cases}$$

Proxy: $U^{+}(i) = \sum_j U(i,j)\theta(j)$

Removing a record that is unique w.r.t. $Q(j)$ reduces $\theta(j)$. Records that are unique w.r.t. to a lot of "risky" $Q(j)$ should reduce dataset risk.

Comparison of the proxies

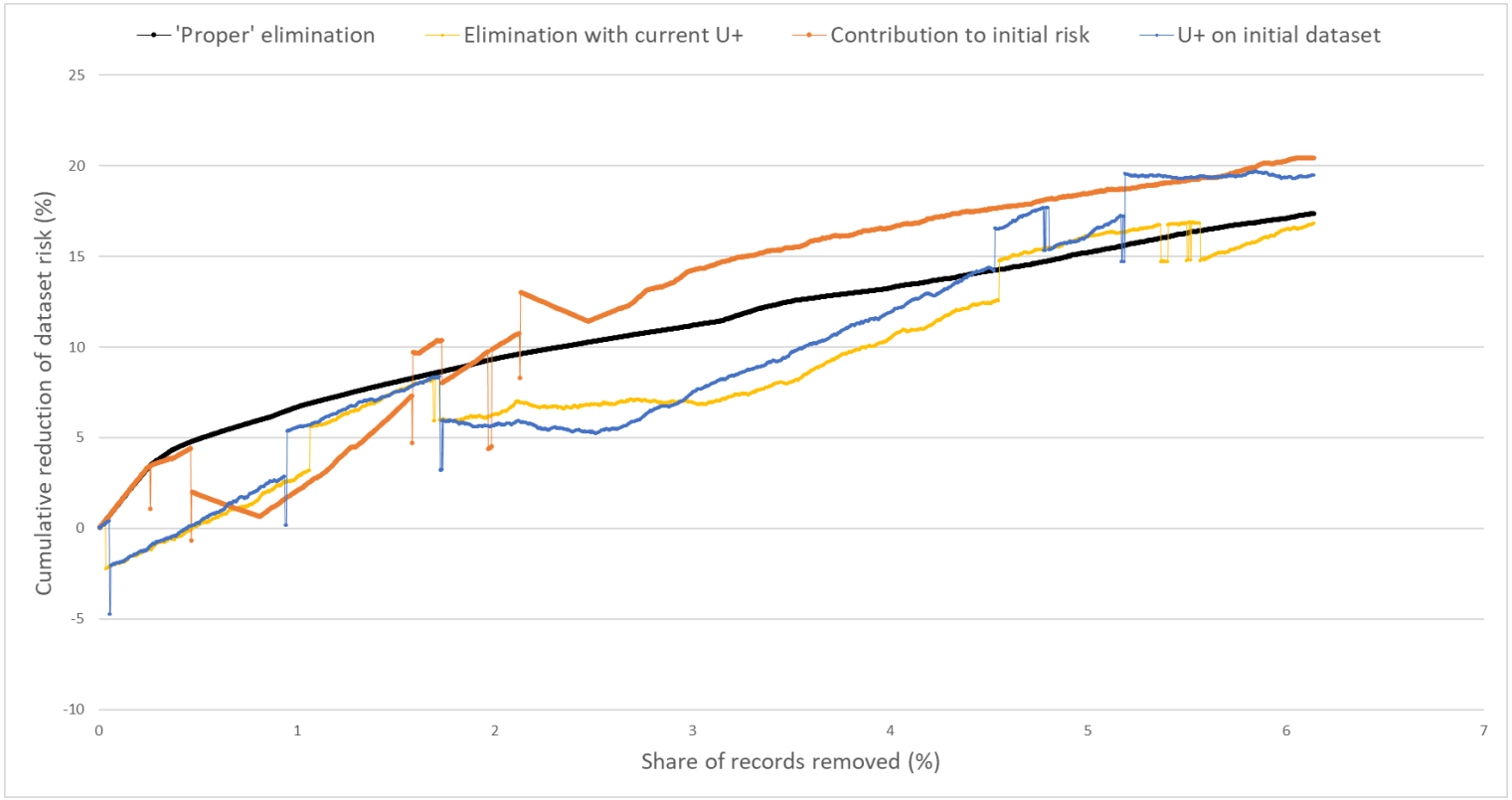


ESS Round 10

12129 records
18 variables

$p = 3$
$\alpha = 0.01$
$\pi_u = 0.05$

Comparison of the proxies



ADULT

32561 records
14 variables

$p = 3$
$\alpha = 0.01$
$\pi_u = 0.05$

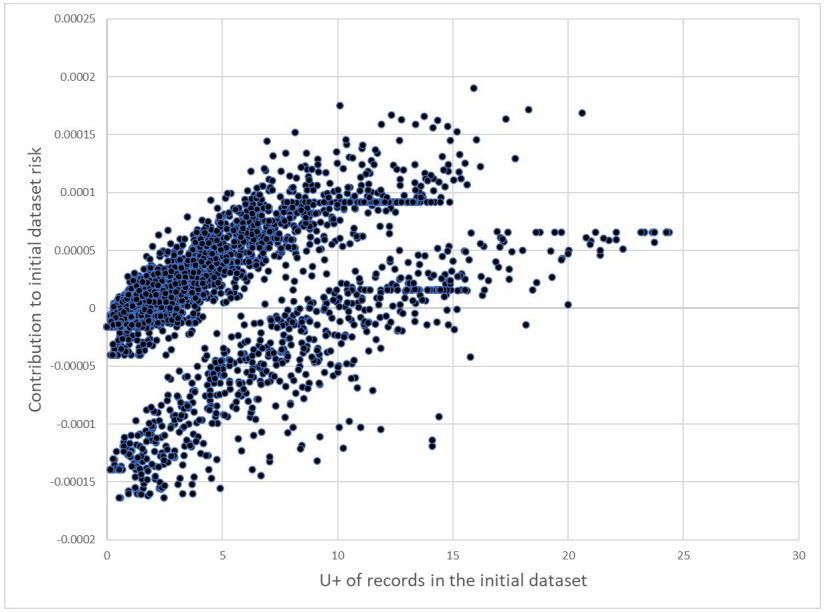
Computing-time gains with the proxies

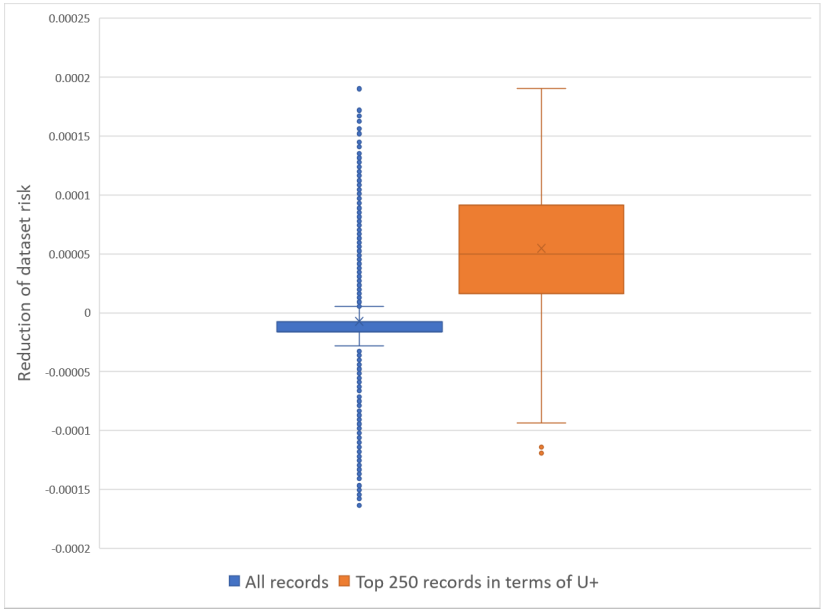
ESS Round 10

No proxy ('proper')	• **12.64 hours** for removal of 1000 records
Current U+ based	• **3.58 hours** for removal of 1000 records
Initial risk contribution	• **10 minutes** to compute all records' contributions • Negligible time to remove 1000 records
Initial U+ based	• **Under 1 minute** to compute the initial U^+ of all records • Negligible time to remove 1000 records
Entropy-based	• **Approx. 1.6 hours** to compute all records' entropies • Negligible time to remove 1000 records
PLS regression-based	• **10 minutes** to compute all records' contributions • Negligible time to fit a PLS model • **User-dependent time** for model checking and adjustments • Negligible time to remove 1000 records

Observations about U+

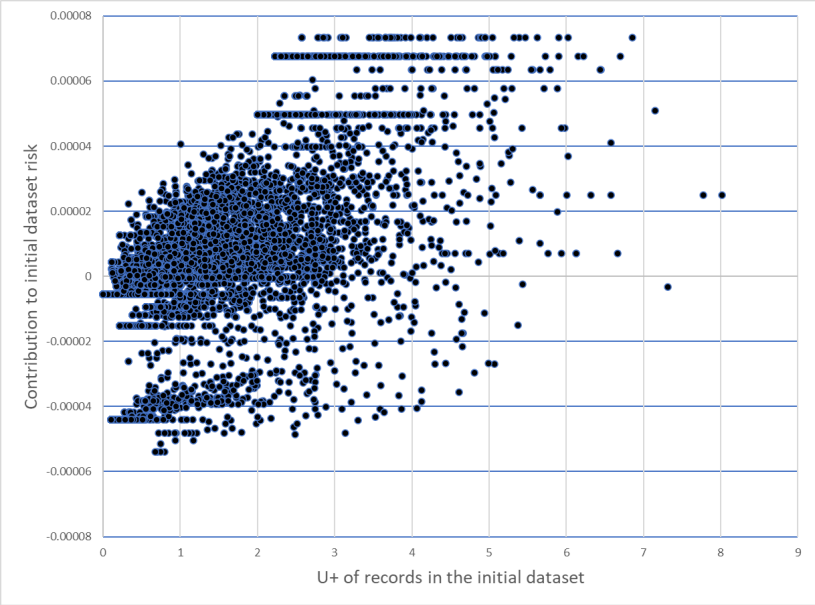
ESS Round 10

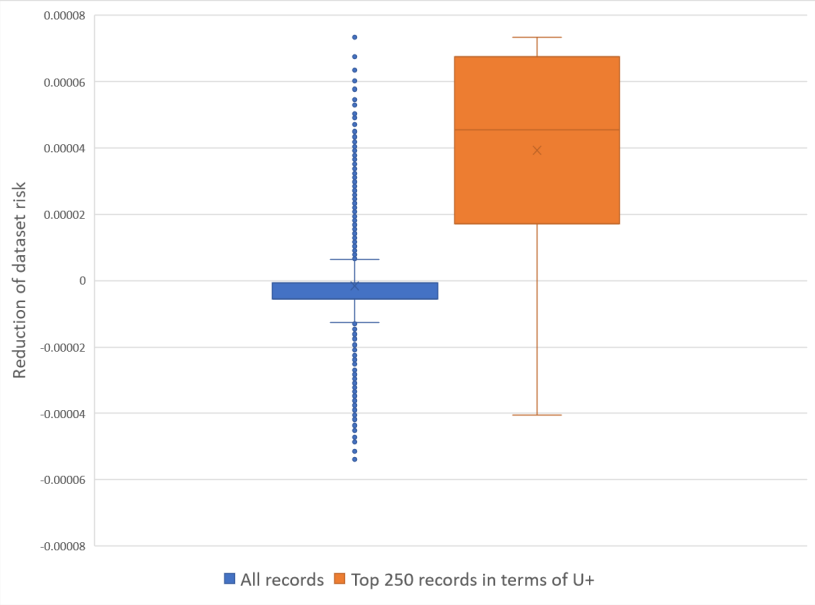




Observations about U+

ADULT





OTTAWA 2023

64TH WORLD STATISTICS CONGRESS

THANK YOU