The 34th International Symposium on High-Performance Parallel and Distributed Computing 2025, Notre Dame, IN, USA

Bringing Differential Privacy to HPC: Privacy-Preserving Transformations of HPC Traces

Ana Solórzano

Rohan Basu Roy Benjamin Schwaller

Sara Petra Walton Jim M. Brandt Devesh Tiwari

Bringing Differential Privacy to HPC: Privacy-

Executive Summary

1. Raise awareness about privacy in shared HPC systems
2. Introduce **Differential Privacy (DP)** to the HPC community
3. Present our toolset to apply DP to real HPC monitoring traces

In HPC

- Users run applications in supercomputers → shared environments
- Telemetry data is collected all the time → important for data analysis
- Few public repositories of telemetry data for research → data compliance issues and privacy concerns delay the release of these logs

When data is shared

- Traditional privacy protection methods are applied: encoding, encryption, hashing

What is privacy?

Security: measures taken to protect data from unauthorized access and attacks.
Privacy: what data is collected, how it is used, and who can see it.
Difference between privacy and security[1]:
Security risks arise from loss of confidentiality, integrity and availability of a system
Privacy: risks arise from data manipulation (processing, output, control).

1. CSRC-NIST (Computer Security Resource Center) – (National Institute of Standards and Technology)

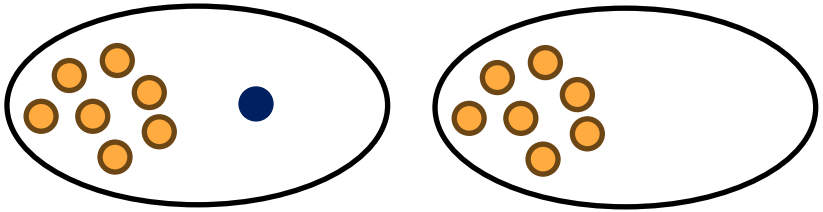
What is sensitive data?

GDPR 4.1 – "Information that can directly or indirectly identify a person can be personal data, location data, an online identifier (IP, MAC...) or to one or more factors specific to the physical, physiological, genetic, economic, cultural or social identity of that natural person;"

Differential Privacy (DP)

Differential Privacy is a mathematical process to ensure data remains private regardless of how much information a malicious analyst has (Dwork et al., 2014).

Databases D and D' are neighbors if | D Δ D'| = 1

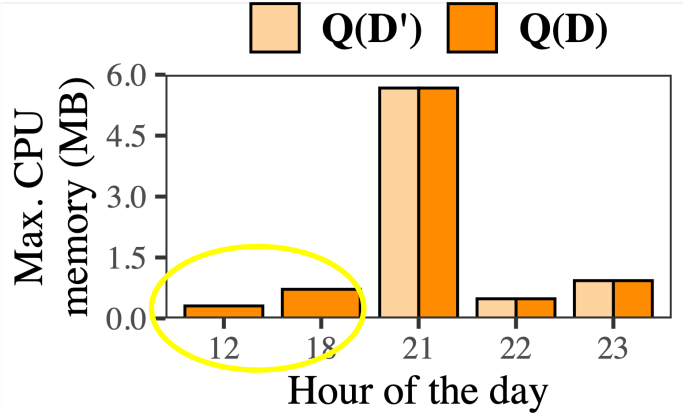


Q is ε-DP if for every neighbor D, D' and every event S:

$$\mathbb{P}[Q(D) = S] \leq e^{\epsilon}\mathbb{P}[Q(D') = S]$$

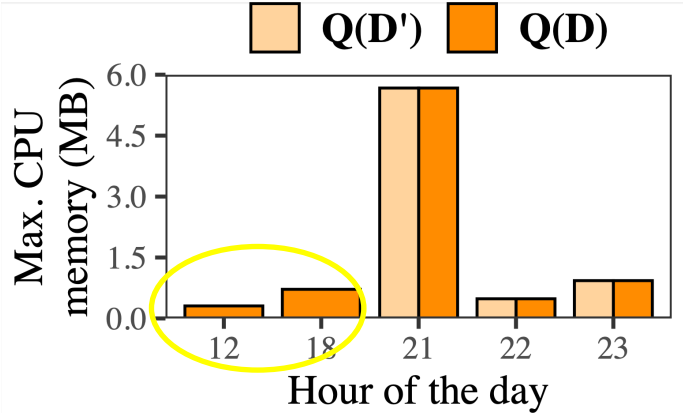
Differential Privacy in HPC

Sharing this information can reveal sensitive data

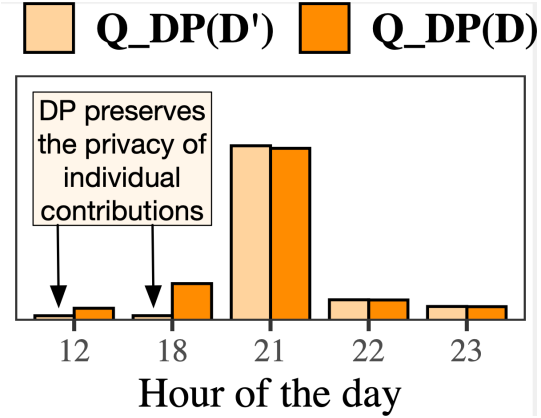


Differential Privacy in HPC

Sharing this information can reveal sensitive data

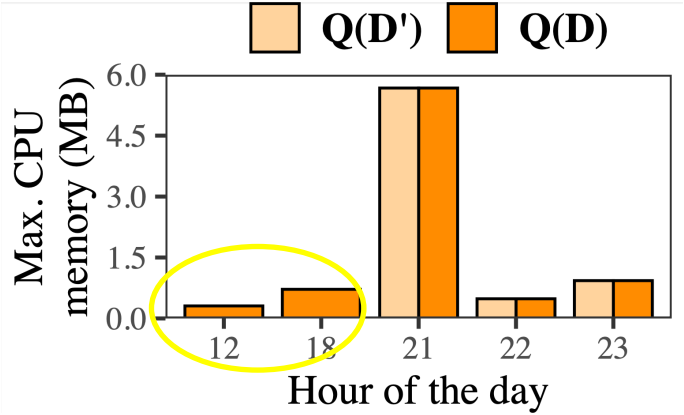


Sharing this information can NOT reveal sensitive data

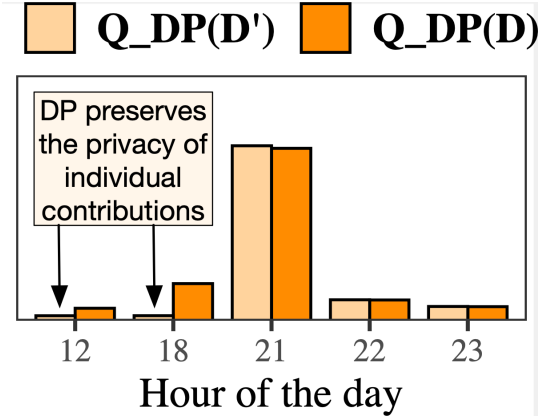


Differential Privacy in HPC

Sharing this information can reveal sensitive data



Sharing this information can NOT reveal sensitive data



Due to the randomness of the DP method a malicious analyst cannot know for sure the contribution of the individual → privacy guarantee due to DP

Differential Privacy (DP) definitions

Privacy Budget (**ε**): controllable parameter defined by the data owner or trusted analyst and shared with the DP dataset.

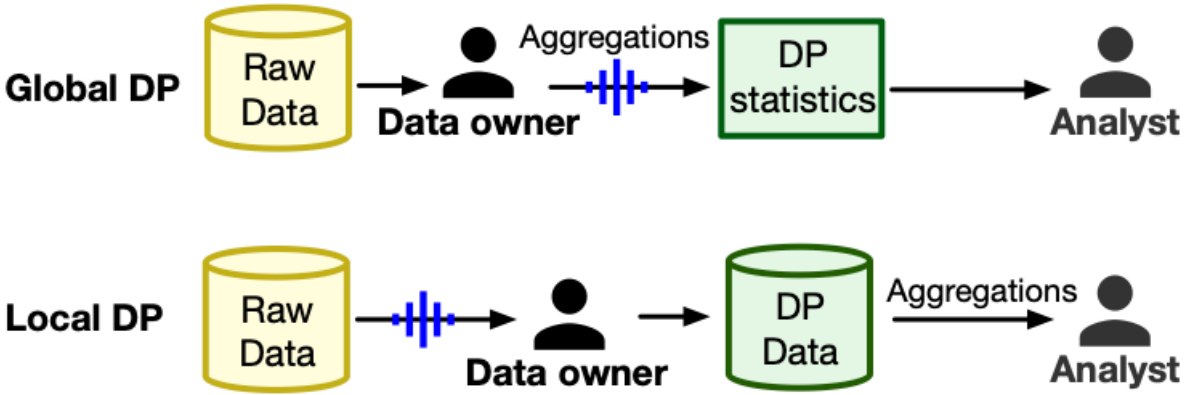
Sensitivity: the maximum change in the output of a query when a single individual's data is added to or removed from the dataset.

$$\Delta f = \max_{\substack{D \text{ and } D' \in \\ \textit{neighboring datasets}}} \left\| f(D) - f(D') \right\|$$

Privacy mechanism: algorithm that adds calibrated noise to data or query results to satisfy the DP condition

$$Q_{privacy\ mechanism} = f(D) + Lap\left(\frac{\Delta f}{\epsilon}\right)$$

Privacy preservation



The process and source code information are public, the data is auditable. The "noisy-measurements" are private (privacy loss budget).

Widely adopted…

United States Census Bureau

Google

Uber

Our contributions

1. Propose the first tool to apply Differential Privacy to HPC system traces
2. Support to applying DP mechanisms to both aggregated and raw data
3. Evaluate the toolset using real-world HPC traces

Who is this useful for?

HPC system administrators, operators, and researchers.

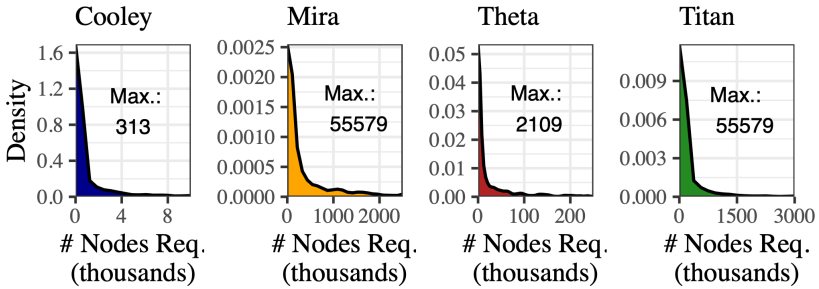
Why is this useful?

Add new layers of privacy for public HPC traces release.

Datasets and Metrics

Datasets from ALCF[1] and OLCF2[2]

System	Years	Size	# Users	# Jobs	% NaN
Cooley	2015 - 2023	88.9 MB	1,890	687,380	4.66
Theta	2017 - 2023	70.3 MB	2,586	466,344	13.06
Mira	2013 - 2020	69.5 MB	2,671	452,747	0
Titan	2015 - 2019	2.8 GB	3,656	12,981,186	0

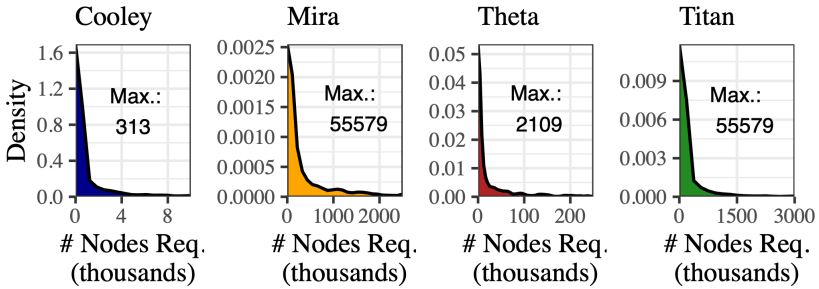


1. https://reports.alcf.anl.gov/data/
2. https://doi.ccs.ornl.gov/ui/doi/334

Datasets and Metrics

Datasets from ALCF[1] and OLCF2[2]

System	Years	Size	# Users	# Jobs	% NaN
Cooley	2015 - 2023	88.9 MB	1,890	687,380	4.66
Theta	2017 - 2023	70.3 MB	2,586	466,344	13.06
Mira	2013 - 2020	69.5 MB	2,671	452,747	0
Titan	2015 - 2019	2.8 GB	3,656	12,981,186	0



Defining privacy and accuracy

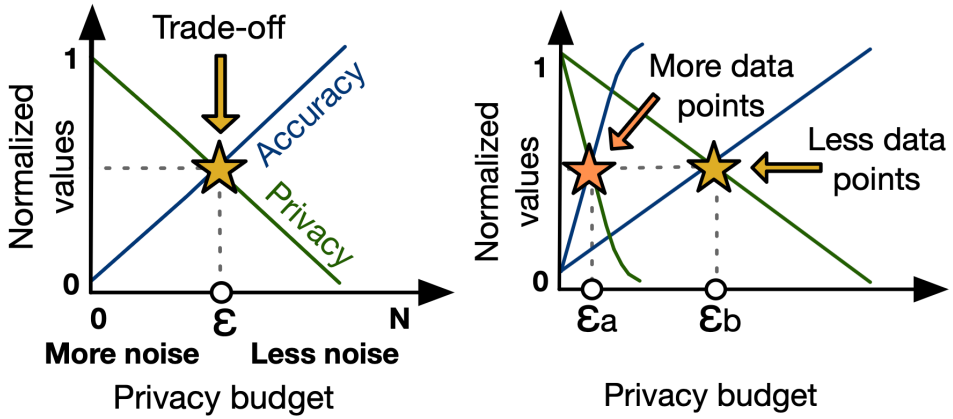
$$Privacy = \sqrt{\sum_{i=1}^{n} (Q(\mathbf{D_i}) - Q_{DP}(\mathbf{D_i}))^2 \quad + (Q(\mathbf{D}) - Q_{DP}(\mathbf{D}))^2}$$

$$Accuracy = \|Q(\mathbf{D}) - Q_{DP}(\mathbf{D})\|$$

1. https://reports.alcf.anl.gov/data/
2. https://doi.ccs.ornl.gov/ui/doi/334

Global DP

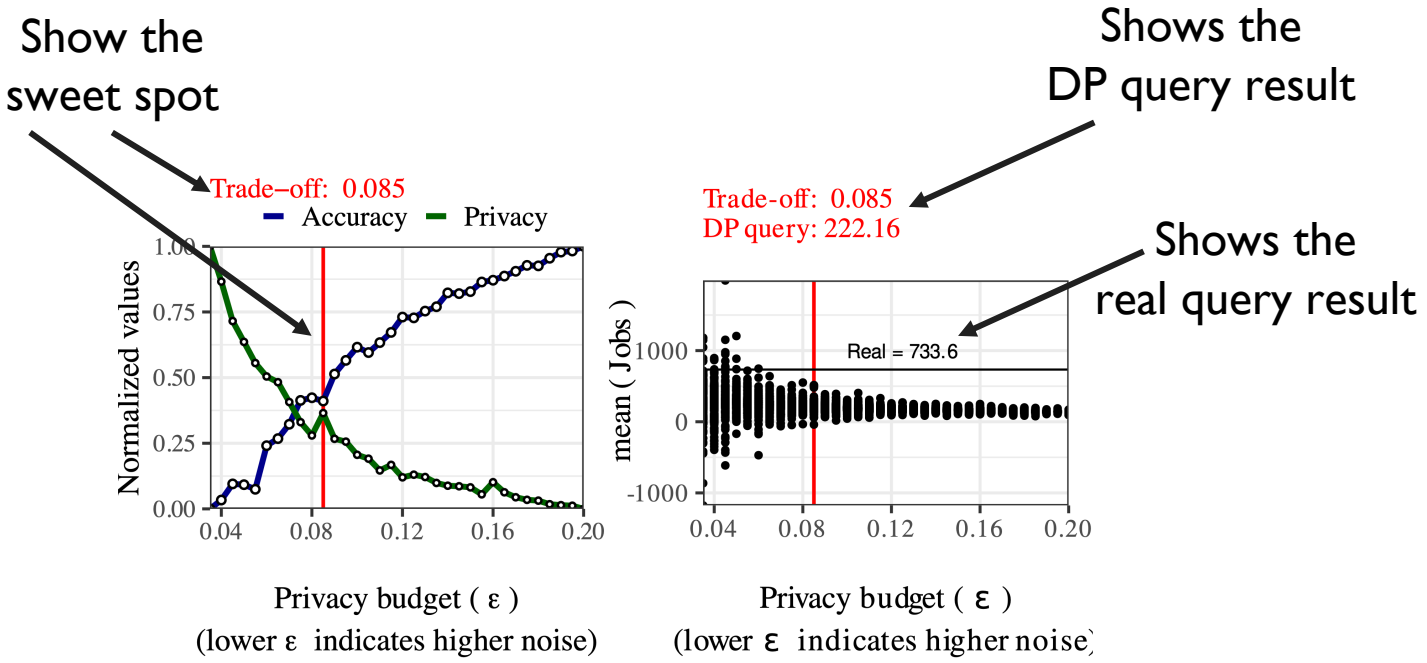
Interpretating *accuracy* and *privacy*



(a) *Privacy and Accuracy*

(b) *Impact of data points*

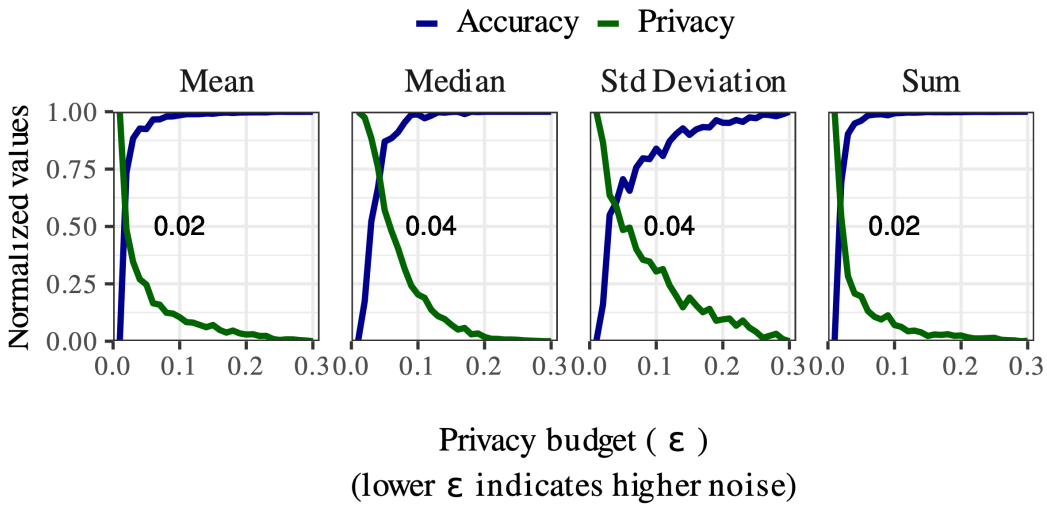
Global DP: toolset visualization output



Command line toolset.

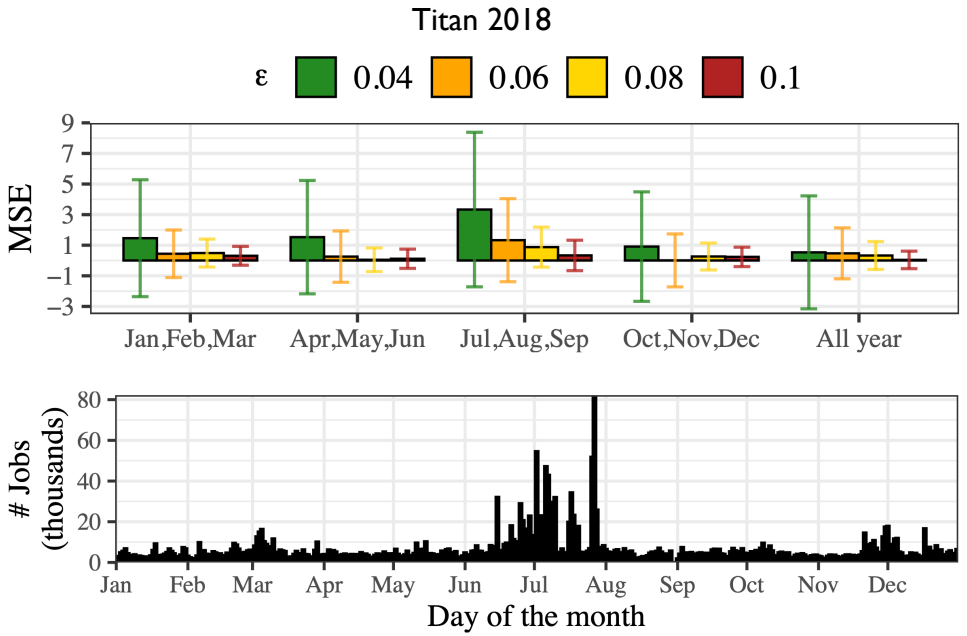
Global DP: using the same ε

- Different DP-queries about the CPU memory usage.
- Data owners can use a general ε to multiple aggregations



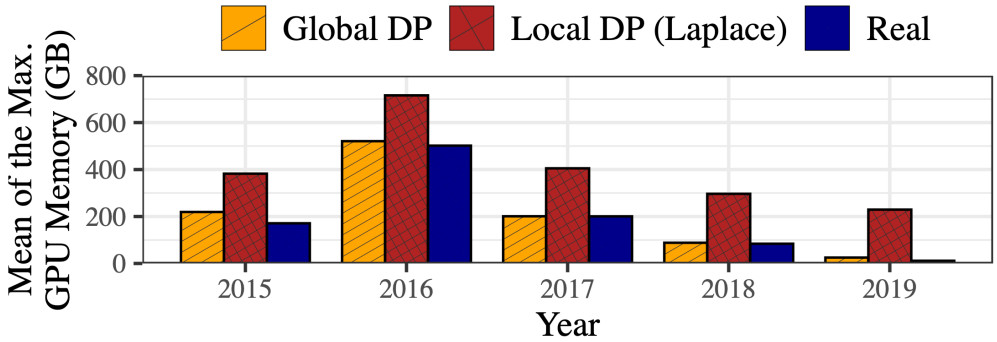
Global DP: in practice

- Major trends are preserved with DP perturbation
- Variations in the original data will impact the amount of noise added to preserve privacy at the same budget.



Local DP

- Does not know future aggregations – more noise needs to be needed
- Ideal to share raw datasets with untrusted parties



Local DP: flipping probability

Only obfuscate the most sensitive values of the dataset.

ε-local DP: for every pair of values D(d,d') and for all outputs of A, we have

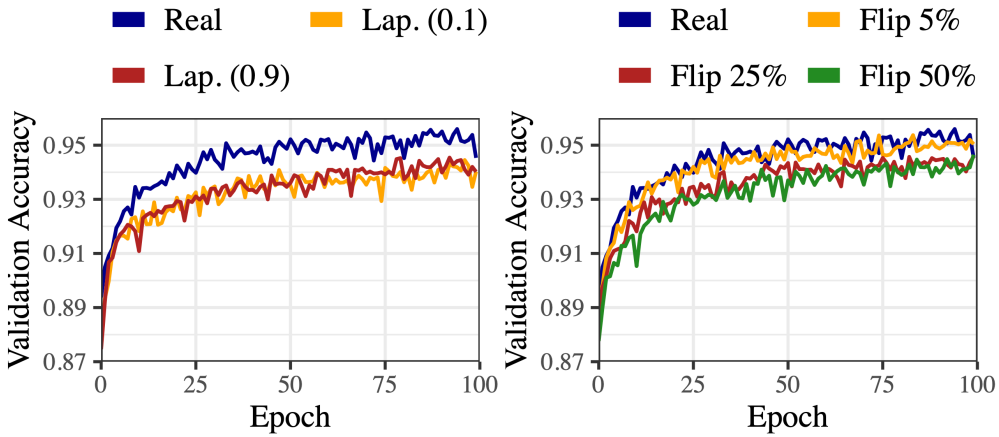
$$\mathbb{P}[A(d_i, P_i) = y] \leq e^{\epsilon}\mathbb{P}[A(d_j, P_j) = y]$$

Profile-based DP: assume a graph G = (P, E) with P profiles of data distributions with edges E. A random process A achieves (G, ε)-DP if for every E connecting the profiles Pi and Pj, with random variables di ~ Pi and dj ~ Pj, and all inputs y ∈ Y:

$$\mathbb{P}[A(d) = y] \leq e^{\epsilon}\mathbb{P}[A(d') = y]$$

Local DP: in practice

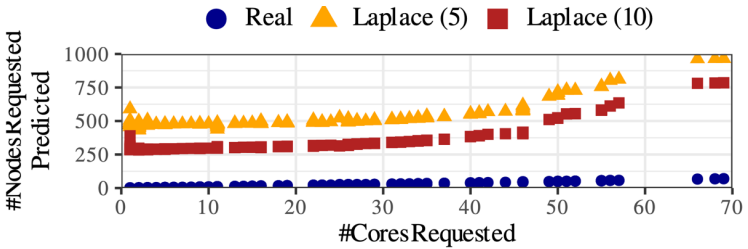
- Classification model (job failure prediction)
- Compared to Laplace mechanism
- Our method adds noise only when necessary

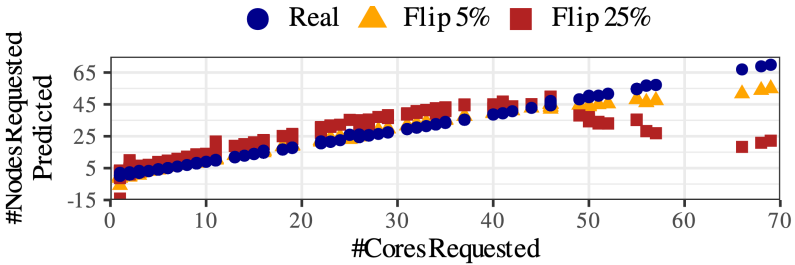


(a) *Laplace mechanism* **(b)** *Flipping mechanism*

Local DP: in practice

- Regression model (nodes requested vs. cores requested)
- Laplace mechanism has bigger impact on the correlation and dimension of results.





Conclusion

1. Differential Privacy mechanisms provide a new layer of individual privacy protection while preserving data utility

2. DP empowers data owners with granular control, allowing them to define a specific privacy budget for noise injection.

3. Our toolset, integrates metrics and visualizations to support the application of DP to both statistical analyses and original datasets

Future work: deployment and integration of ε-DP results to real-time HPC systems monitoring dashboards.

Thank you!

Ana Solórzano

Rohan Basu Roy Benjamin Schwaller

Sara Petra Walton Jim M. Brandt Devesh Tiwari

Contact me:
solorzano.an@northeastern.edu

This work has been partially funded by Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under Contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.