Delay-allowed Differentially Private Data Stream Release

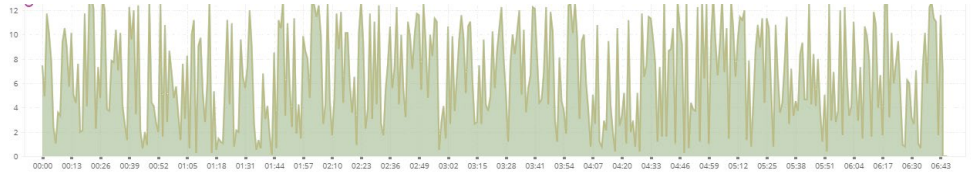
Xiaochen Li , Zhan Qin, Kui Ren, Chen Gong, Shuya Feng, Yuan Hong , Tianhao Wang

Data Stream Release-Traffic Analysis

A **city transportation department** monitors traffic flow to optimize signals and detect congestion.

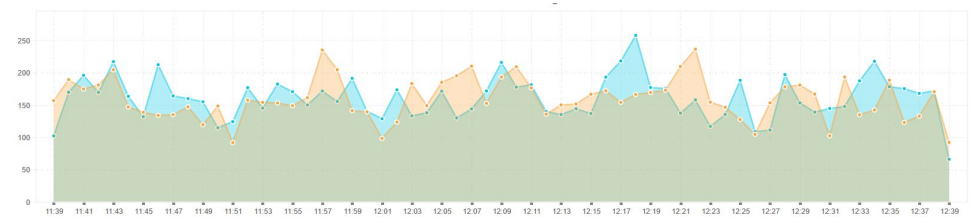
- There are multiple streams continuously collected by sensors.

Vehicle Count



...

Vehicle Speed



Data Stream Release-Traffic Analysis

A **city transportation department** monitors traffic flow to optimize signals and detect congestion.

- The analysts need to query the original stream for each new report.

  How many vehicles are at a certain intersection during the rush hour?
  What is the average speed during rush hour on a certain day?
  When is the highest traffic flow on a certain street?
  …

- If these streams contain sensitive personal information…

  Frequent queries can increase the privacy leakage in the streams.
  Different types of queries require different privacy protection mechanisms.
  Data analysts need to interact frequently with data holders.

Data Stream Release-Traffic Analysis

A **city transportation department** monitors traffic flow to optimize signals and detect congestion.

- The analysts need to query the original stream for each new report.

  How many vehicles are at a certain intersection during the rush hour?
  What is the average speed during rush hour on a certain day?
  When is the highest traffic flow on a certain street?
  …

- If these streams contain sensitive personal information…

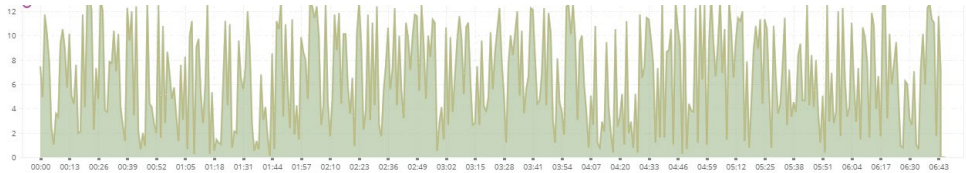
  Frequent queries can increase the privacy leakage in the streams.
  Different types of queries require different privacy protection mechanisms.
  Data analysts need to interact frequently with data holders.

Is there a once-and-for-all solution?

Data Stream Release-Traffic Analysis

- If we continuously release noisy versions of sensitive data streams…

Vehicle Count



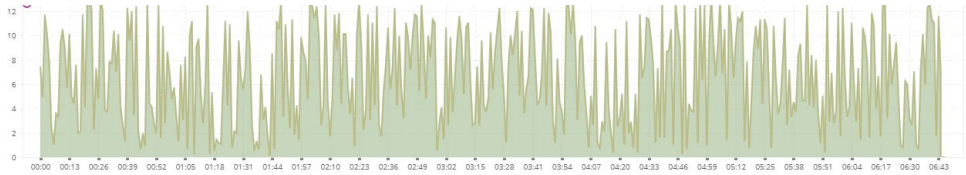
Add noise to each data point

Noisy Vehicle Count

Data Stream Release-Traffic Analysis

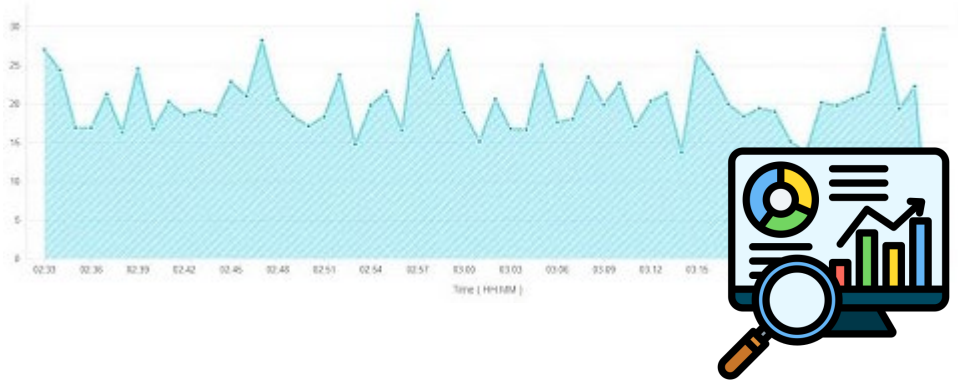
- If we continuously release noisy versions of sensitive data streams…

Vehicle Count



Add noise to each data point

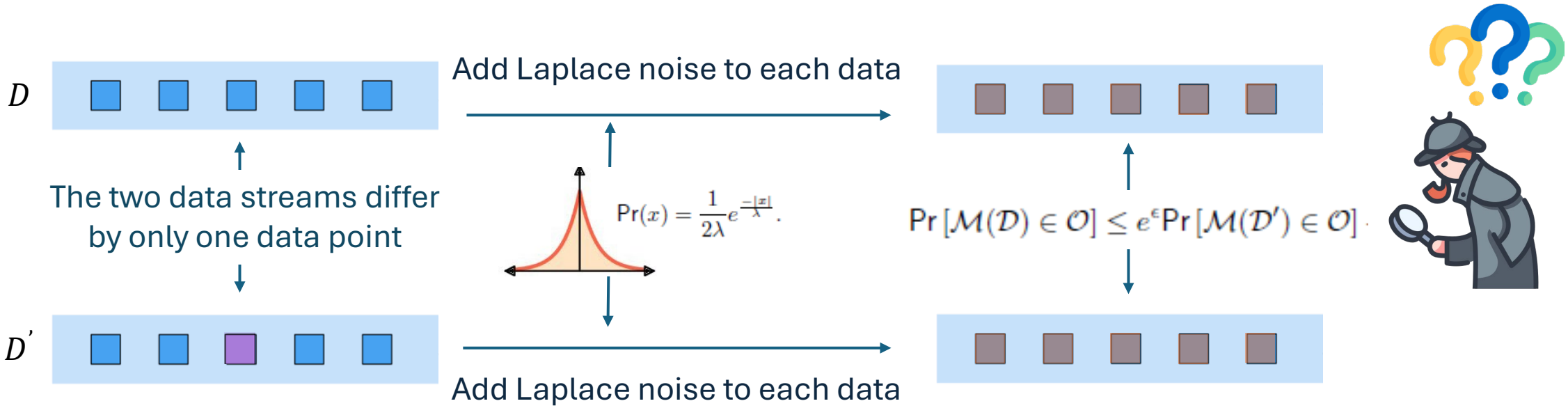
Noisy Vehicle Count



It can be used to conduct all analysis without privacy concerns:
How many vehicles are at a certain intersection during the rush hour?
What is the average speed during rush hour on a certain day?
When is the highest traffic flow on a certain street?
…

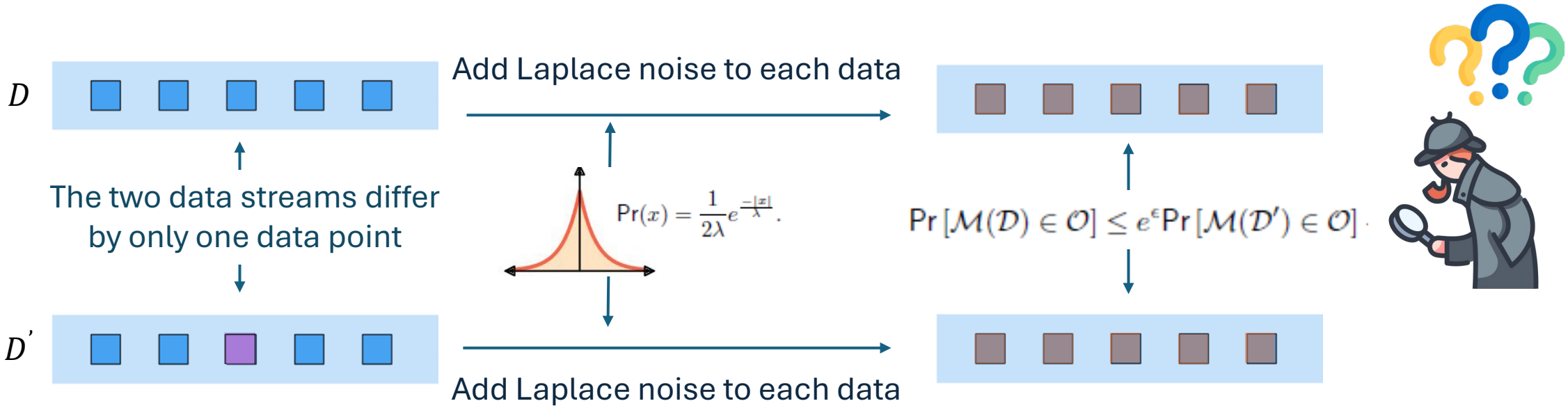
Differential Privacy Guarantee

We provide event-level DP protection for releasing private data stream.



Differential Privacy Guarantee

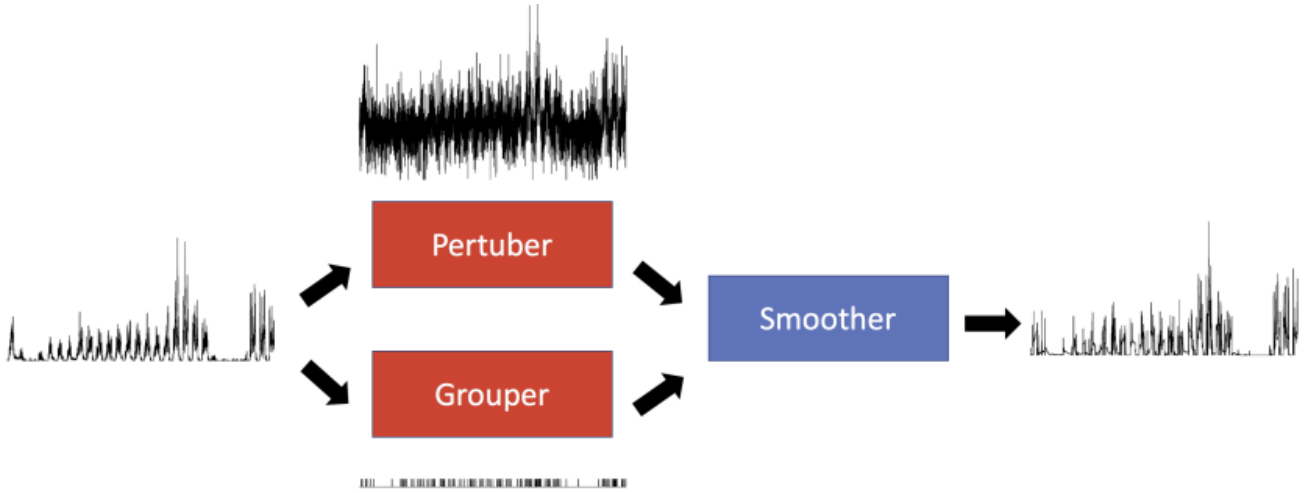
We provide event-level DP protection for releasing private data stream.



Adding noise to each data point introduces excessive noise into the data stream.

Existing Solutions with DP

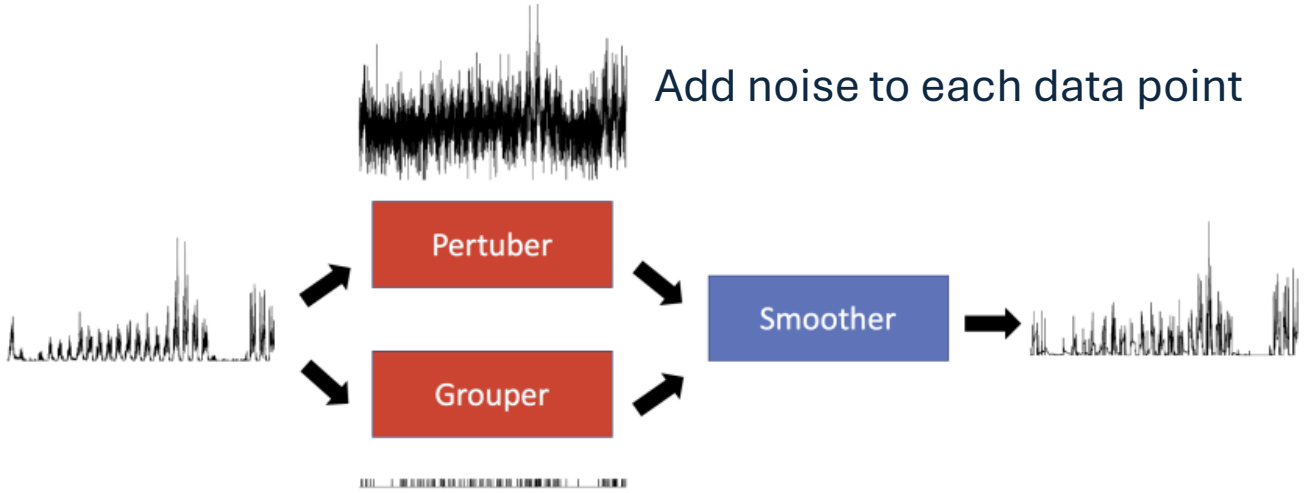
Group-based Post-Processing Framework



[1] Chen, Yan, et al. "Pegasus: Data-adaptive differentially private stream processing." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.

Existing Solutions with DP

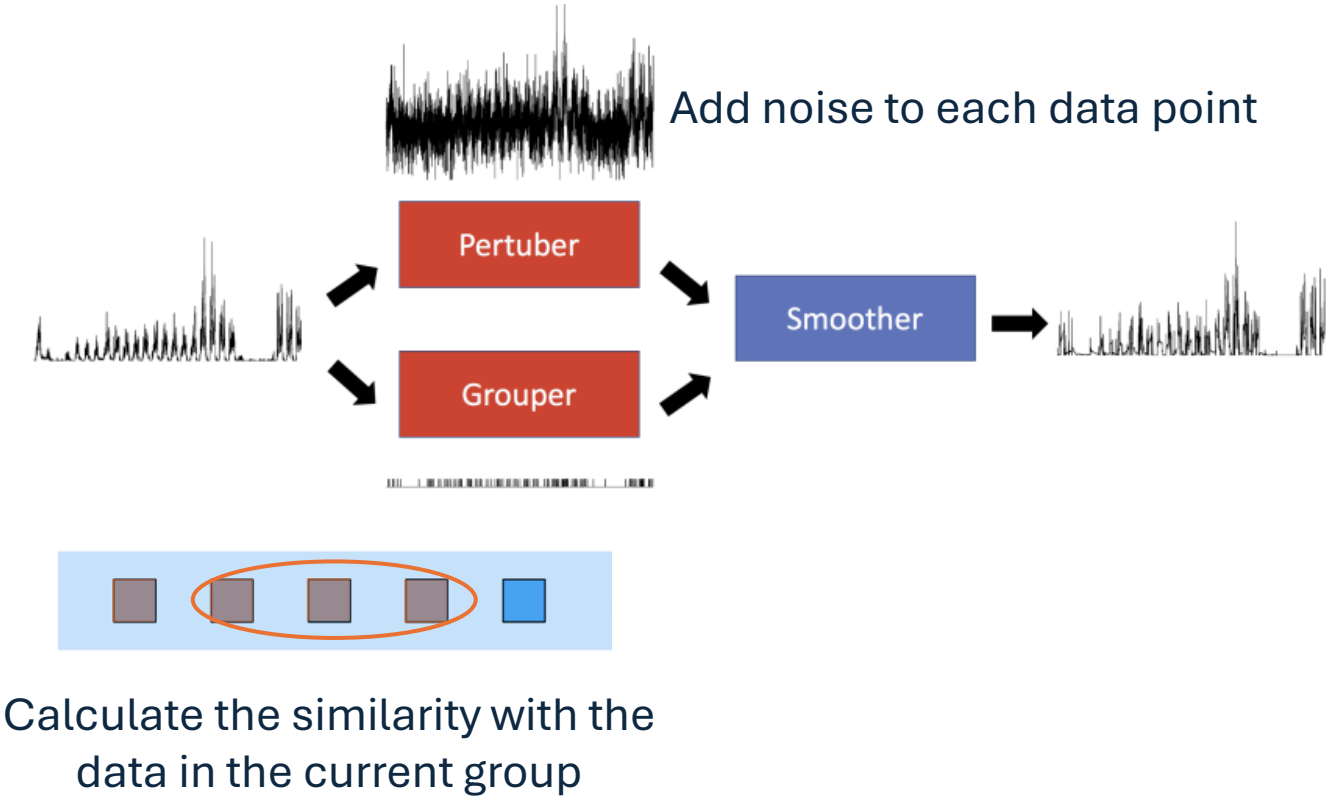
Group-based Post-Processing Framework



[1] Chen, Yan, et al. "Pegasus: Data-adaptive differentially private stream processing." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.

Existing Solutions with DP

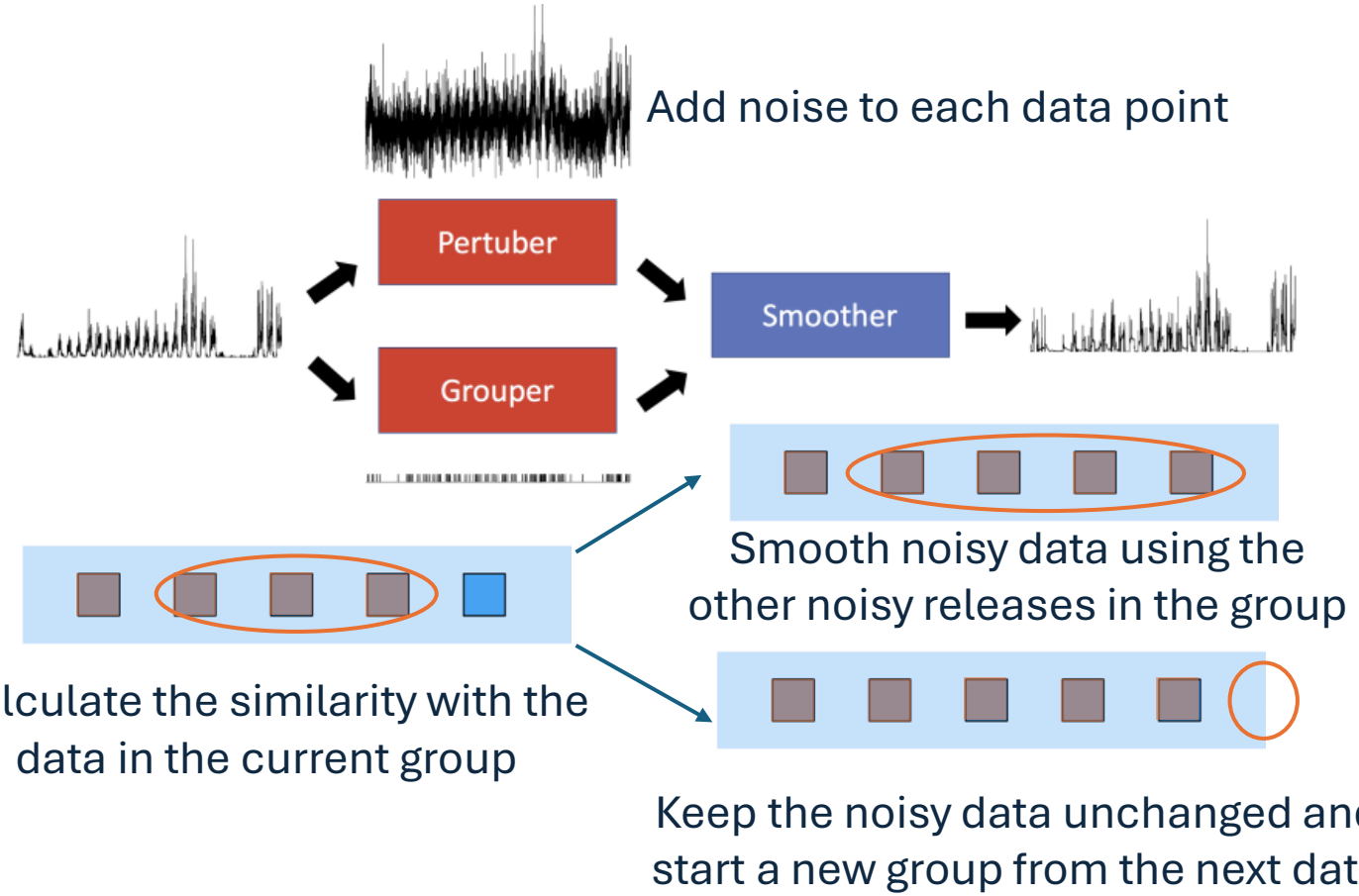
Group-based Post-Processing Framework



[1] Chen, Yan, et al. "Pegasus: Data-adaptive differentially private stream processing." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.

Existing Solutions with DP

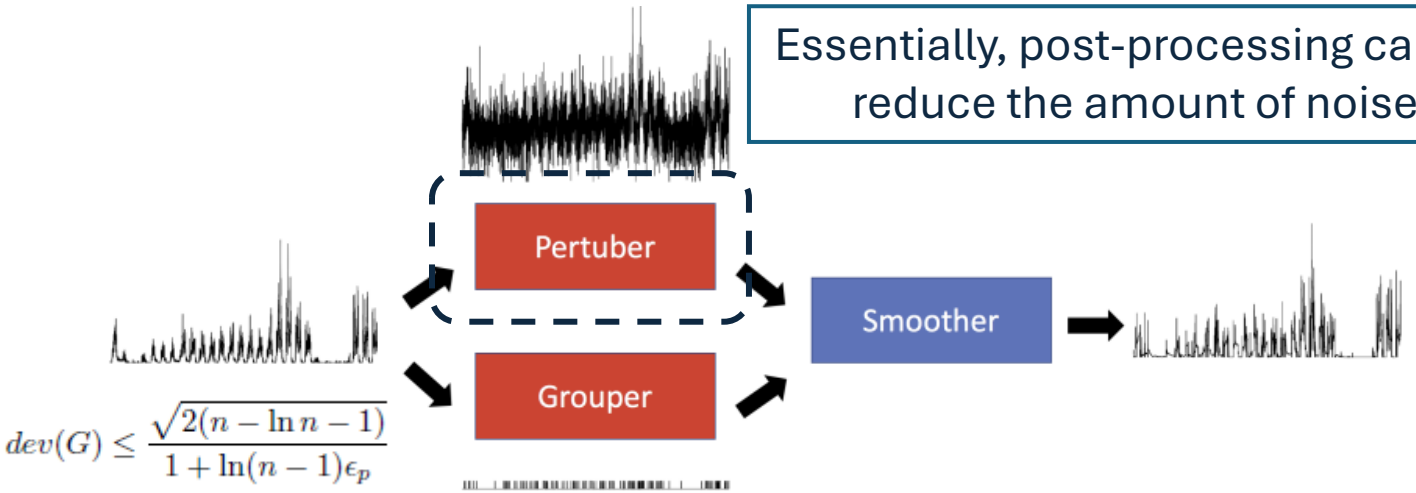
Group-based Post-Processing Framework



[1] Chen, Yan, et al. "Pegasus: Data-adaptive differentially private stream processing." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.

Existing Solutions with DP

Group-based Post-Processing Framework

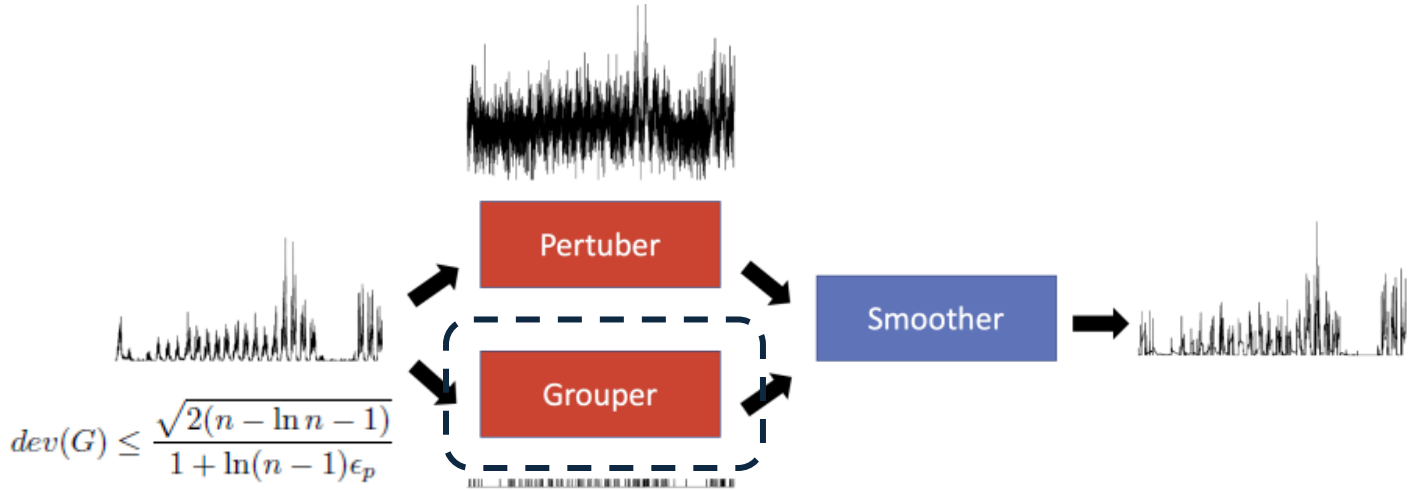


Essentially, post-processing cannot reduce the amount of noise.

[1] Chen, Yan, et al. "Pegasus: Data-adaptive differentially private stream processing." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.

Existing Solutions with DP

Group-based Post-Processing Framework



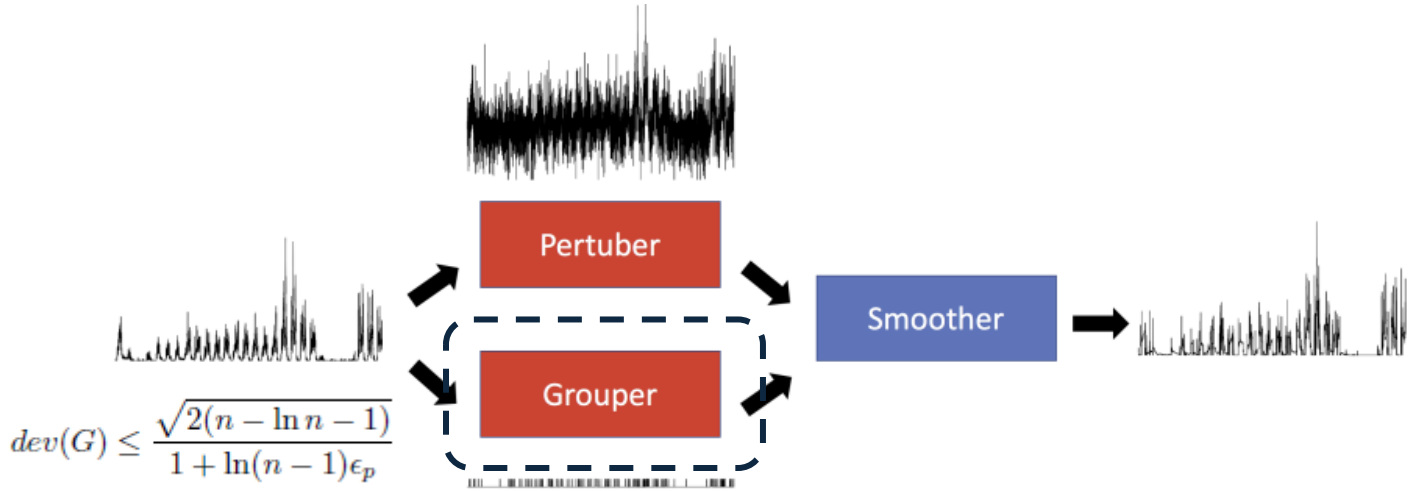
Accuracy Improvement Requires:

Sufficiently large number of data points **Or** High similarity among the data points

[1] Chen, Yan, et al. "Pegasus: Data-adaptive differentially private stream processing." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.

Existing Solutions with DP

Group-based Post-Processing Framework



Accuracy Improvement Requires:

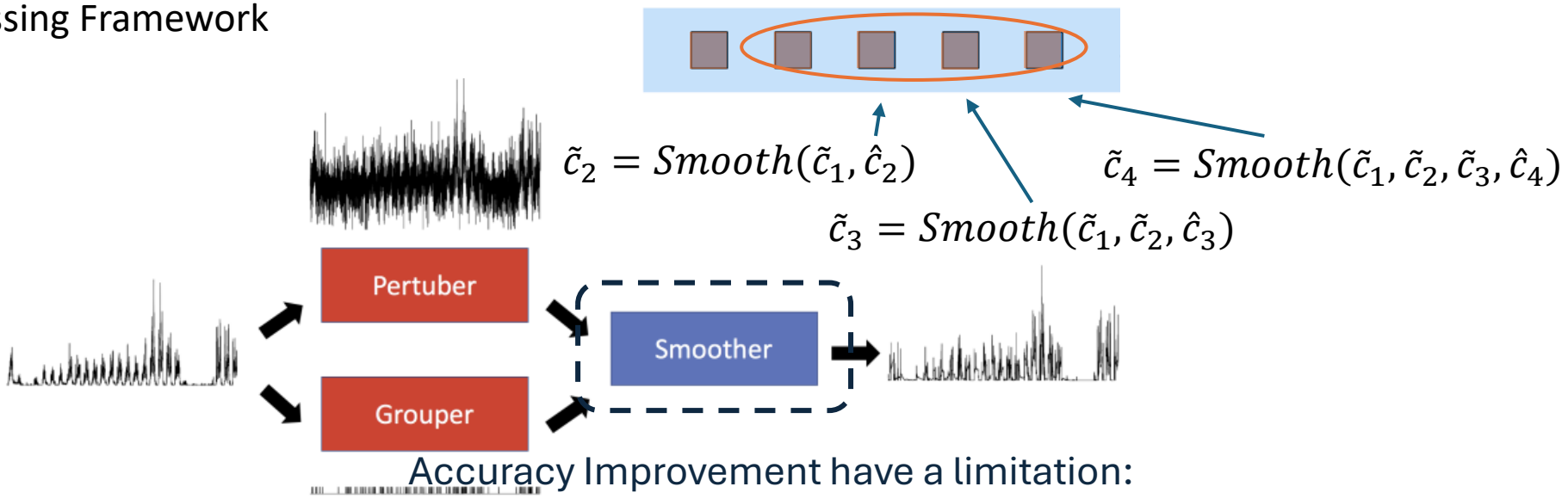
Sufficiently large number of data points **Or** High similarity among the data points

The condition for accuracy improvement through post-processing is difficult to satisfy.

[1] Chen, Yan, et al. "Pegasus: Data-adaptive differentially private stream processing." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.

Existing Solutions with DP

Group-based Post-Processing Framework



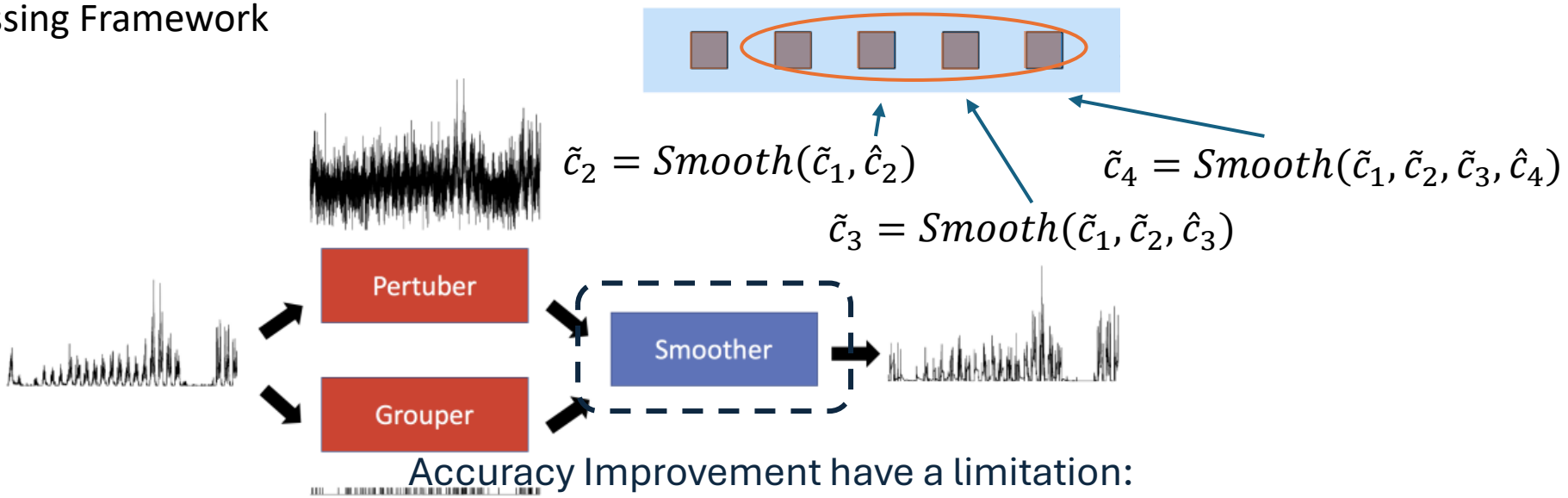
Accuracy Improvement have a limitation:

Data earliest involved in each group cannot benefit from post-processing.

[1] Chen, Yan, et al. "Pegasus: Data-adaptive differentially private stream processing." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.

Existing Solutions with DP

Group-based Post-Processing Framework



Accuracy Improvement have a limitation:

Data earliest involved in each group cannot benefit from post-processing.

Merely addressing the length of the group doesn't provide much help.

[1] Chen, Yan, et al. "Pegasus: Data-adaptive differentially private stream processing." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.

Questions Regarding Real-Time Setting

Before proposing solutions, let's first consider two questions.

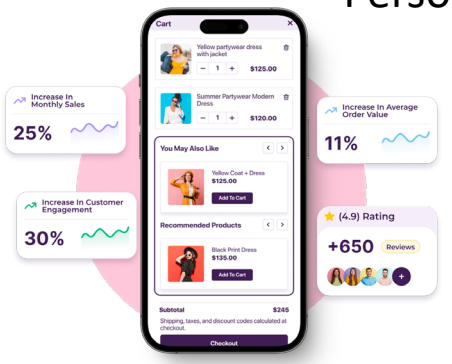
Is real-time release required for all data streams in practice?

Questions Regarding Real-Time Setting

Before proposing solutions, let's first consider two questions.

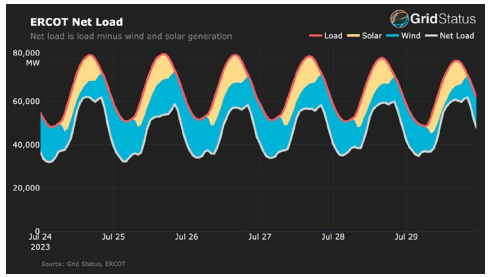
Is real-time release required for all data streams in practice?

Personalized Product Recommendations

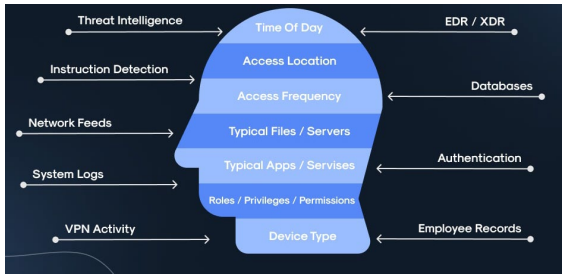


Smart Grid Monitoring

Net Load Forecasting



User Behavior Analysis



Questions Regarding Real-Time Setting

Before proposing solutions, let's first consider two questions.

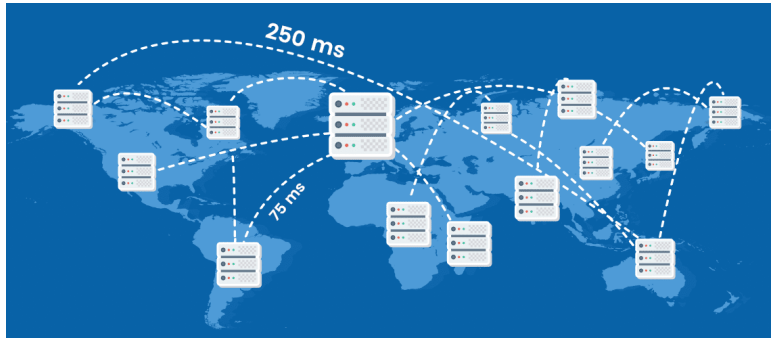
Can approaches theoretically designed for real-time truly achieve real-time?

Questions Regarding Real-Time Setting

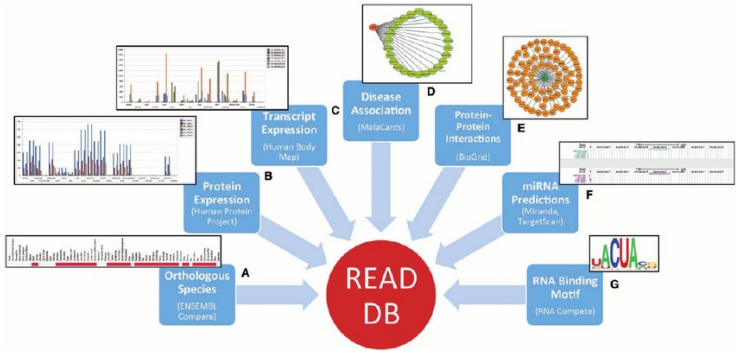
Before proposing solutions, let's first consider two questions.

Can approaches theoretically designed for real-time truly achieve real-time?

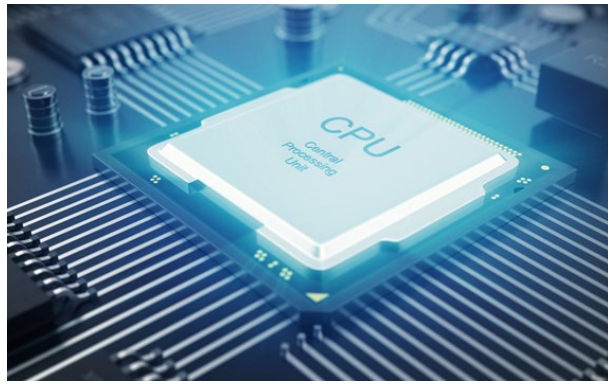
Network Latency



Data Integration



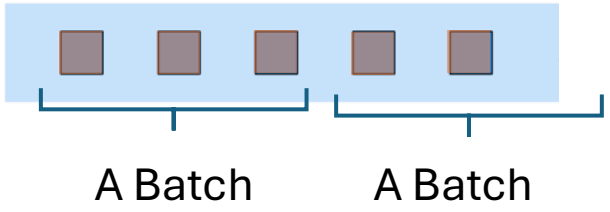
Processing Time



Solutions

Therefore, we consider a stream release setting that allows for a small delay.

- Batch setting



What is the difference from processing batched data ❓

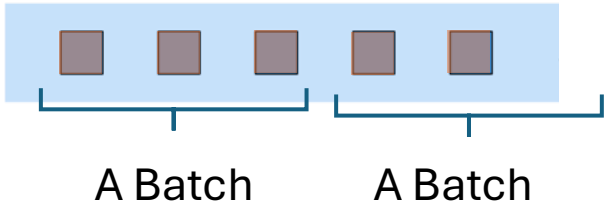
The data within a batch maintains a temporal relationship.

Solutions

Therefore, we consider a stream release setting that allows for a small delay.

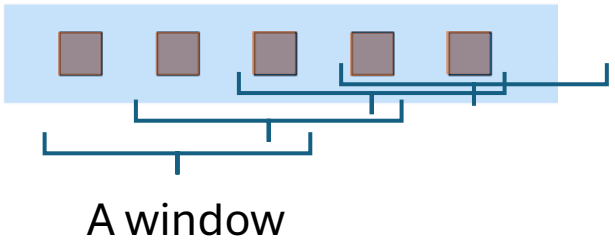
- Batch setting

What is the difference from processing batched data ❓



The data within a batch maintains a temporal relationship.

- Sliding window setting

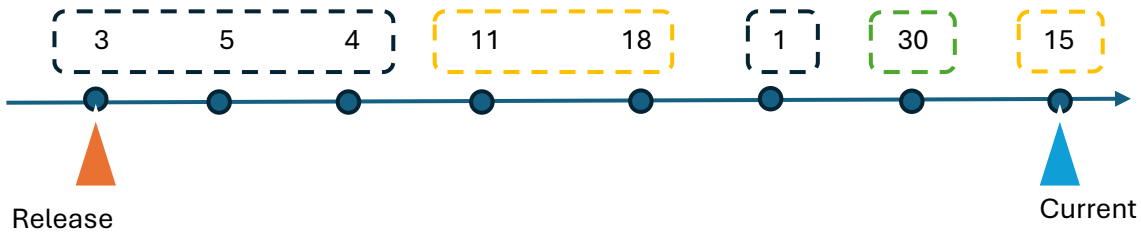


When processing each data, you can look ahead at the next w data points.

Solutions

Batch Setting

- Discontinuous Grouping

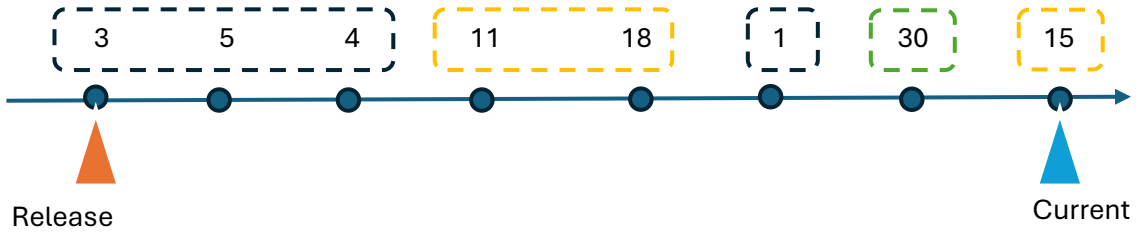


Keeping multiple groups open simultaneously, rather than just one, can **increase the group length**.

Solutions

Batch Setting

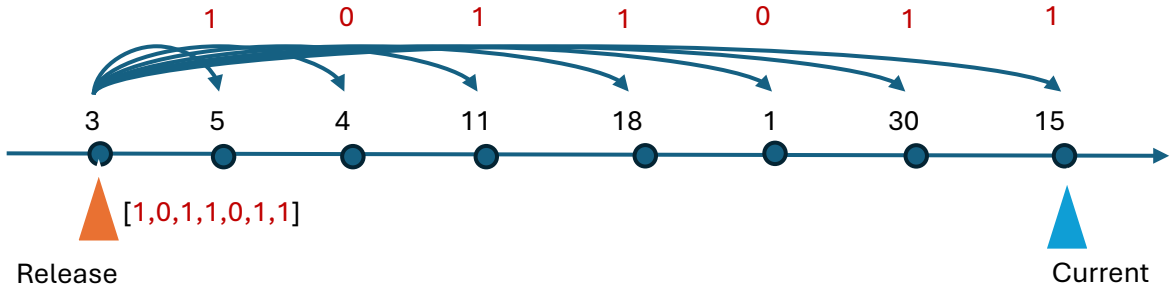
- Discontinuous Grouping



Keeping multiple groups open simultaneously, rather than just one, can **increase the group length**.

Sliding Window Setting

- CompOrder



For each noisy data, **adjust it based on the order relationship** recorded by its previous w data points.

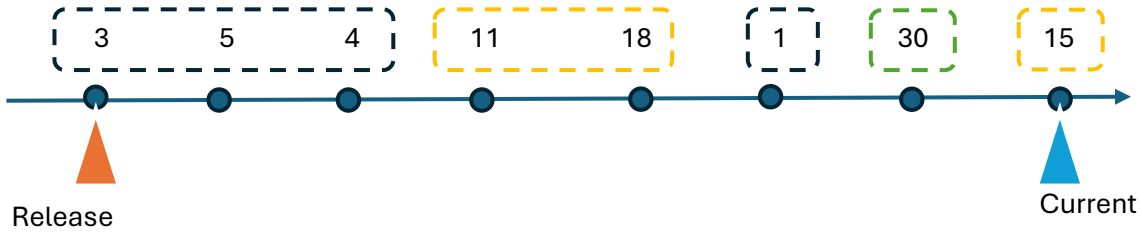
Solutions

Batch Setting

High noise sensitivity

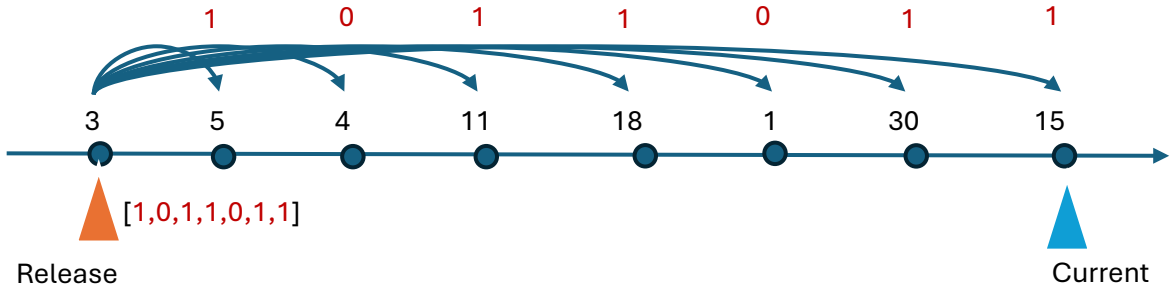
Sliding Window Setting

- Discontinuous Grouping



Keeping multiple groups open simultaneously, rather than just one, can **increase the group length**.

- CompOrder



For each noisy data, **adjust it based on the order relationship** recorded by its previous w data points.

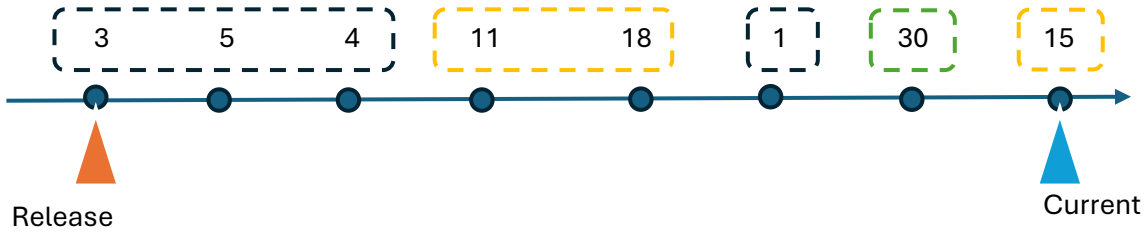
Solutions

Batch Setting

High noise sensitivity

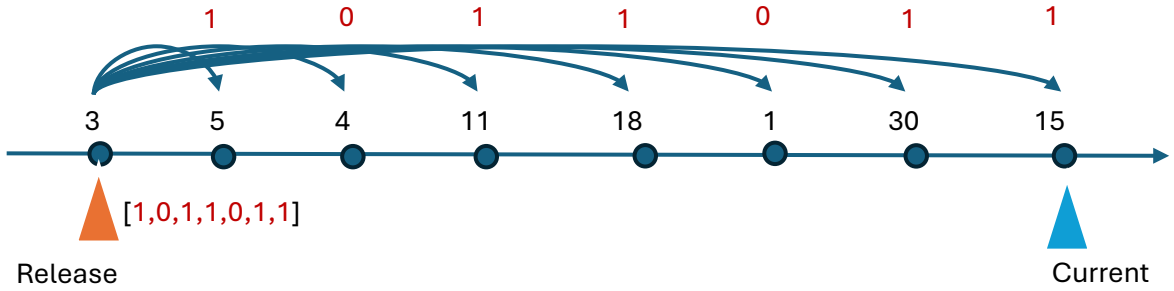
Sliding Window Setting

- Discontinuous Grouping



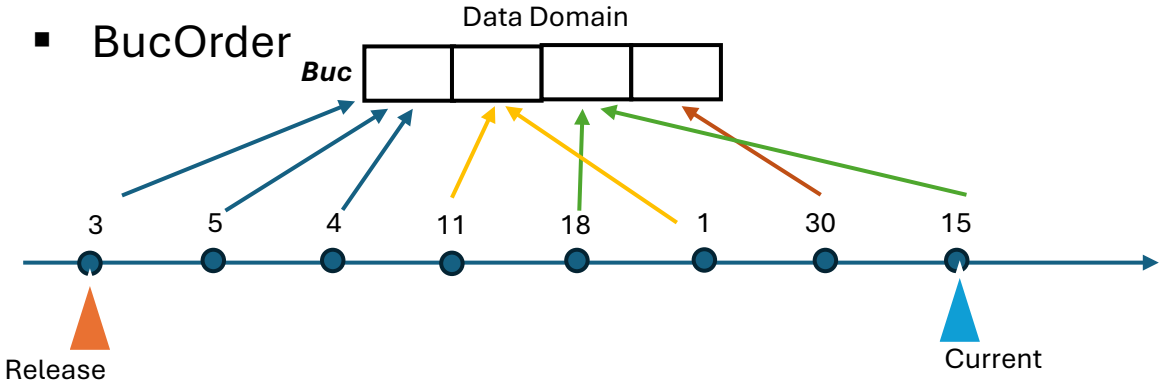
Keeping multiple groups open simultaneously, rather than just one, can **increase the group length**.

- CompOrder



For each noisy data, **adjust it based on the order relationship** recorded by its previous w data points.

- BucOrder



Reduces the impact of a single data point on the grouping/ordering results and saves the privacy budget.

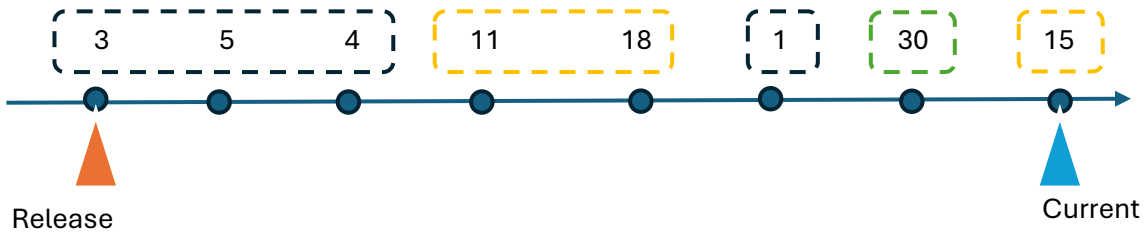
Solutions

Batch Setting

High noise sensitivity

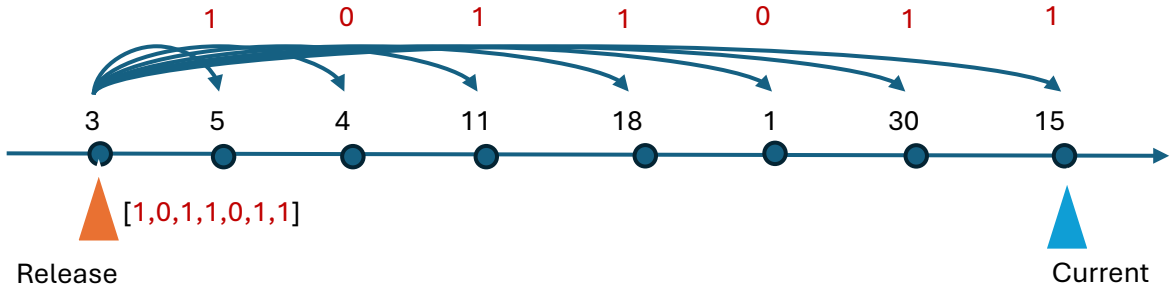
Sliding Window Setting

- Discontinuous Grouping



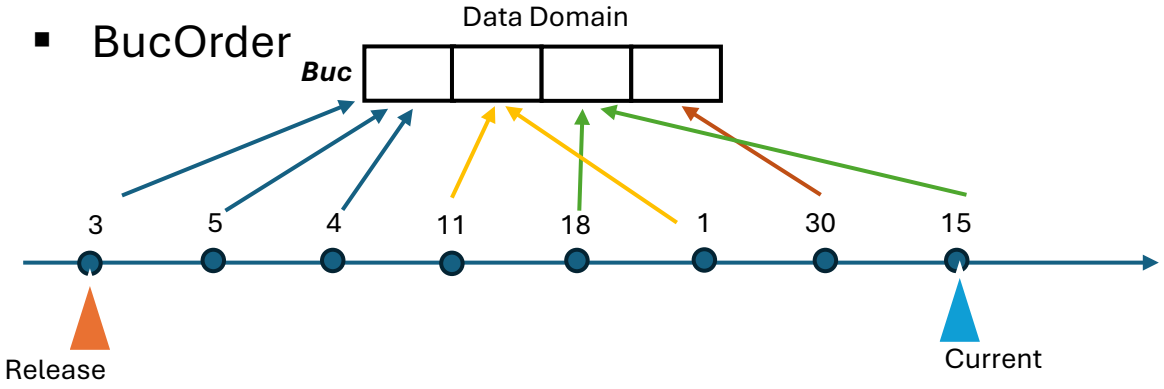
Keeping multiple groups open simultaneously, rather than just one, can **increase the group length**.

- CompOrder



For each noisy data, **adjust it based on the order relationship** recorded by its previous w data points.

- BucOrder



Reduces the impact of a single data point on the grouping/ordering results and saves the privacy budget.

- Add Noise
  - Add noise to data: $x_i + \text{Lap}(\frac{\Delta}{\varepsilon_p}) \rightarrow \hat{x}_i$
  - Add noise to summation: $sum_i + \text{Lap}(\frac{\Delta}{\varepsilon_p}) \rightarrow \widehat{sum}_i$

All the data within the group/bucket benefit from **the same post-processing advantages.**

Solutions

Integrate all the things into a framework.

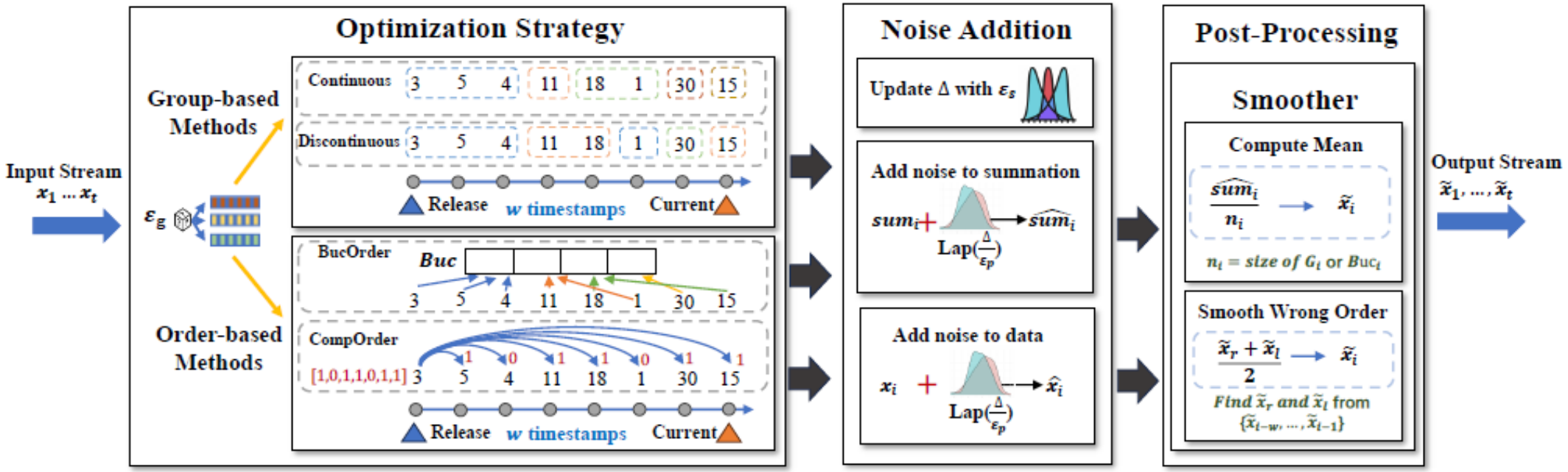
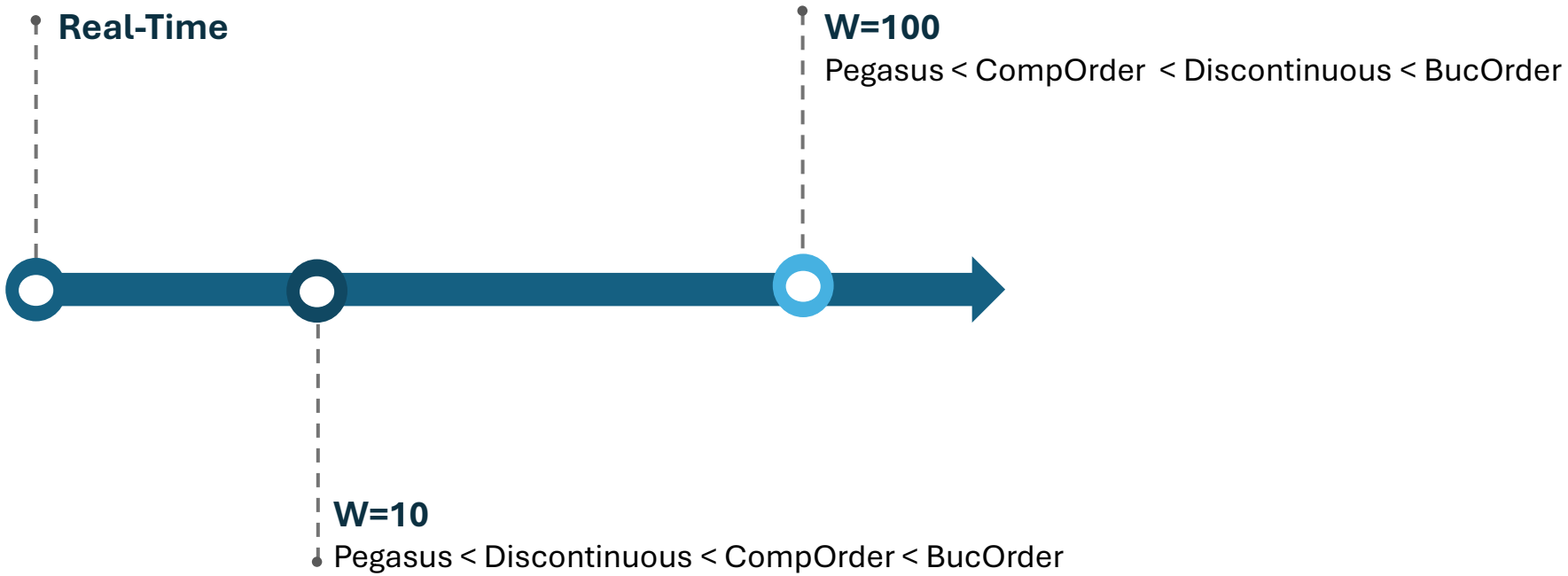


Fig. 1. Overview of the privacy-preserving and delay-allowed data releasing framework. This framework identifies the optimal strategy for the input stream, then introduces noise to comply with differential privacy (DP) standards, and finally post-processes the noised data stream before its release.

Experimental Results

- Comparisons of accuracy between the proposed methods and PeGaSus.



Experimental Results

- Comparisons of accuracy between the proposed methods and PeGaSus.

For the Outpatient dataset containing 1366 timestamps, BucOrder can achieve more than **32-fold** accuracy improvement compared to PeGaSuS when allowing a delay of **10 timestamps**.

[1] Chen, Yan, et al. "Pegasus: Data-adaptive differentially private stream processing." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.

Experimental Results

- Comparisons of accuracy with Methods from other privacy setting.

Methods	Unemployment			Outpatient		
	0.1	0.5	1.0	0.1	0.5	1.0
Naïve	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×
DPI	0.08×	0.38×	0.76×	0.05×	0.25×	0.48×
Adapub	1.26×	1.24×	1.26×	1.23×	1.25×	1.23×
Discontin	0.41×	0.41×	0.41×	0.41×	0.41×	0.41×
CompOrder	0.19×	0.28×	0.40×	0.20×	0.39×	0.56×
BucOrder	0.03×	0.17×	0.35×	0.03×	0.15×	0.29×

[1] Feng, Shuya, et al. "Dpi: Ensuring strict differential privacy for infinite data streaming." *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024.
[2] Wang, Teng, et al. "Adaptive differentially private data stream publishing in spatio-temporal monitoring of IoT." 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC). IEEE, 2019.

Experimental Results

- Comparisons of accuracy with Methods from other privacy setting.

Methods	Unemployment			Outpatient		
	0.1	0.5	1.0	0.1	0.5	1.0
Naïve	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×
DPI	0.08×	0.38×	0.76×	0.05×	0.25×	0.48×
Adapub	1.26×	1.24×	1.26×	1.23×	1.25×	1.23×
Discontin	0.41×	0.41×	0.41×	0.41×	0.41×	0.41×
CompOrder	0.19×	0.28×	0.40×	0.20×	0.39×	0.56×
BucOrder	0.03×	0.17×	0.35×	0.03×	0.15×	0.29×

The proposed methods **show accuracy advantages** over state-of-the-art solutions from **other privacy settings** with a delay of 10 timestamps.

[1] Feng, Shuya, et al. "Dpi: Ensuring strict differential privacy for infinite data streaming." *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024.
[2] Wang, Teng, et al. "Adaptive differentially private data stream publishing in spatio-temporal monitoring of IoT." 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC). IEEE, 2019.

This paper developed a framework for data stream releases that allows for a delay.

Hope it inspires!

Questions are welcome 😊!

xiaochenli@virginia.edu

Paper & Artifact