Differential Privacy on Fully Dynamic Streams

Yuan Qiu[1] and Ke Yi[2]

[1]Southeast University

[2]Hong Kong University of Science and Technology

Backgrounds – Fully Dynamic Streams

- A data stream is a sequence of pairs (s_t, x_t)
  - Insertion $(s_t = +) \implies D_t = D_{t-1} \cup \{x_t\}$
  - Deletion $(s_t = -) \implies D_t = D_{t-1} - \{x_t\}$
  - No-op $(s_t = \perp) \implies D_t = D_{t-1}$
- Notation
  - N_t: number of *updates* up to time t
  - n_t: *data size* at time t

$$N_t^+ + N_t^- = \underbrace{N_t}_{\text{Increases over time}} \gg \underbrace{n_t}_{\text{Fluctuates and even hits zero}} = N_t^+ - N_t^-$$

Timestamp	Update	Dataset
0	-	$\emptyset$
1	$+, a$	$\{a\}$
2	$+, b$	$\{a, b\}$
3	$+, c$	$\{a, b, c\}$
4	$\perp$	$\{a, b, c\}$
5	$-, b$	$\{a, c\}$
6	$+, d$	$\{a, c, d\}$
7	$-, c$	$\{a, d\}$
8	$-, a$	$\{d\}$
9	$-, d$	$\emptyset$
…	…	…

Backgrounds – Differential Privacy

- DP: for any neighboring instances $D \sim D'$ and any subset of outputs Y,
$$\Pr[\mathcal{M}(D) \in Y] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D') \in Y] + \delta\,.$$
- For streams, neighboring instances differ by one update.

Timestamp	Update	Dataset
0	-	$\emptyset$
1	$+, a$	$\{a\}$
2	$+, b$	$\{a, b\}$
3	$+, c$	$\{a, b, c\}$
4	$\perp$	$\{a, b, c\}$
5	$-, b$	$\{a, c\}$
...	...	...

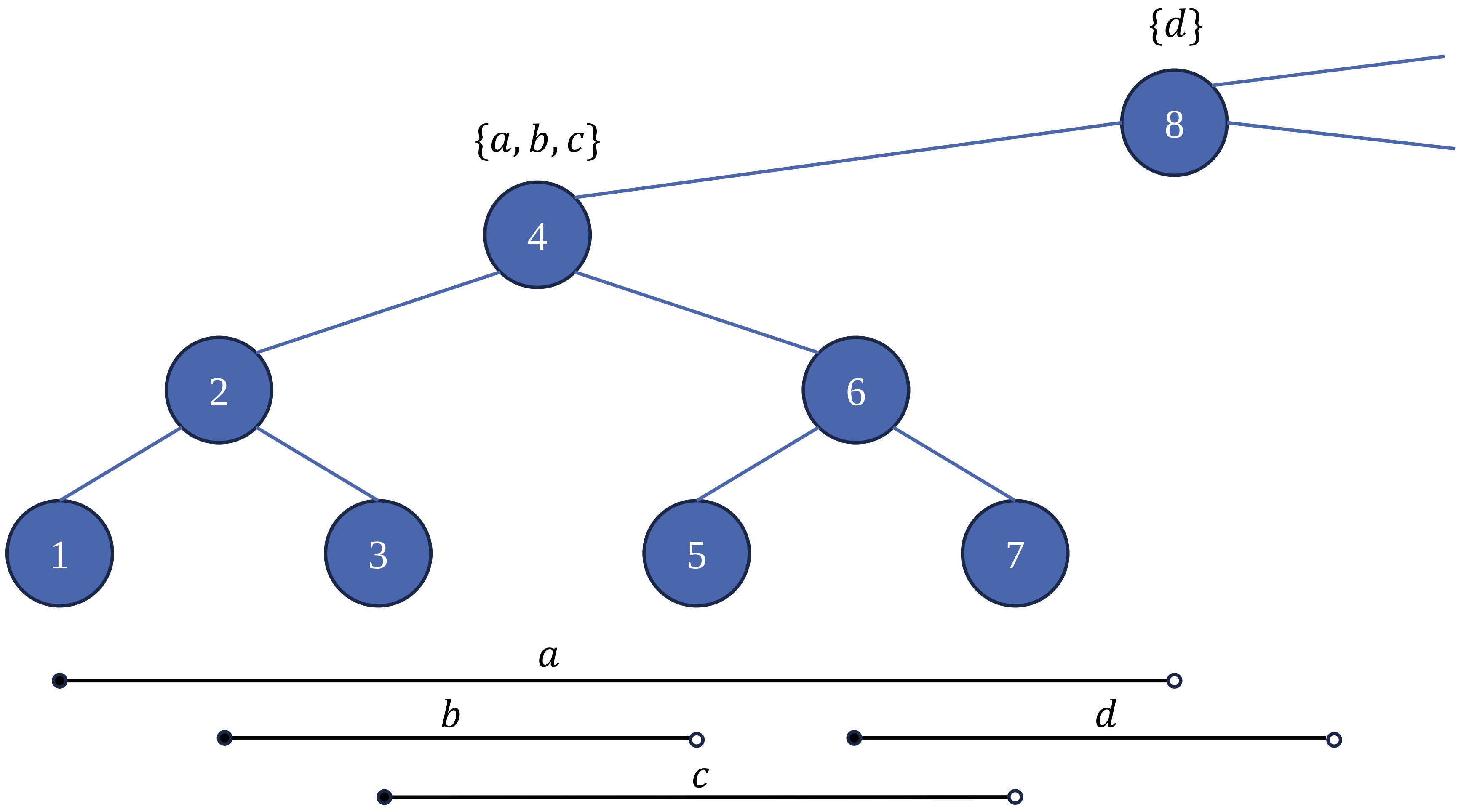
Timestamp	Update	Dataset
0	-	$\emptyset$
1	$+, a$	$\{a\}$
2	$+, b$	$\{a, b\}$
3	$+, c$	$\{a, b, c\}$
4	$+, \boldsymbol{e}$	$\{a, b, c, \boldsymbol{e}\}$
5	$-, b$	$\{a, c, \boldsymbol{e}\}$
...	...	...

Backgrounds – DP Linear Queries

- Linear queries: $f(D) = \sum_{x \in D} f(x)$
  - Counting: $f(x) \equiv 1 \quad \Longrightarrow f(D) = |D|$
  - Mean: $f(x) = \frac{x}{n} \quad \Longrightarrow f(D) = \frac{\sum_{x \in D} x}{n} = \mathbb{E}[x]$
  - Variance: $f(x) = \frac{x^2}{n} \quad \Longrightarrow f(D) = \frac{\sum_{x \in D} x^2}{n} = \mathbb{E}[x^2]$
- For *static data*: PMW mechanism has error $\tilde{O}(\sqrt{n})$
- For *insertion-only streams*: Binary Tree mechanism has error $\tilde{O}\left(\sqrt{N_t}\right)$
- Direct extension: Separate the stream into D_t^+ and D_t^-
  - Problem: The error becomes $\tilde{O}\left(\sqrt{N_t^+} + \sqrt{N_t^-}\right) = \tilde{O}\left(\sqrt{N_t}\right)$
  - Our target is $\tilde{O}\left(\sqrt{n_t}\right)$

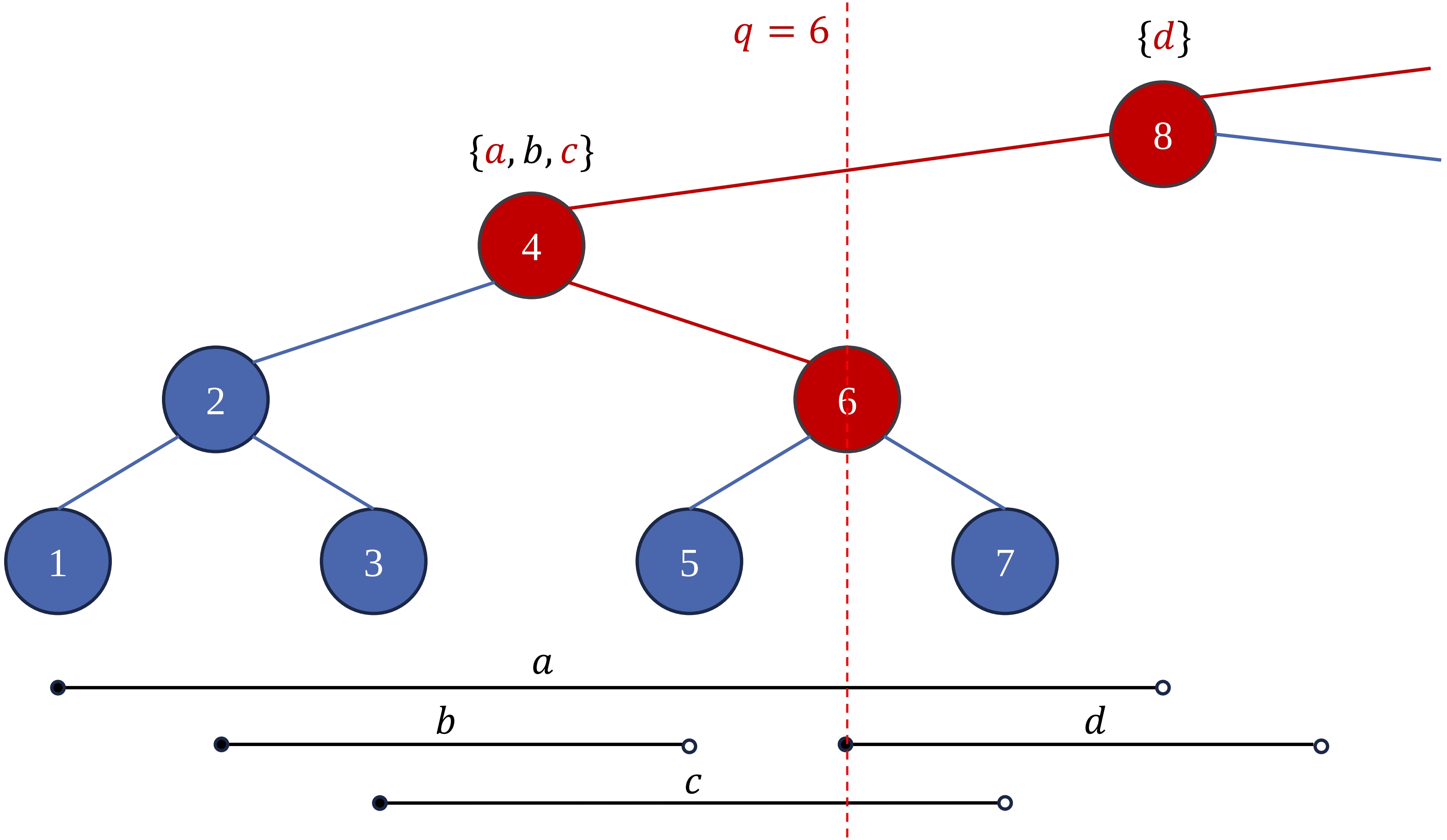
Interval Tree for Stabbing Queries

- Each (insertion, deletion) of an item can be considered an interval.



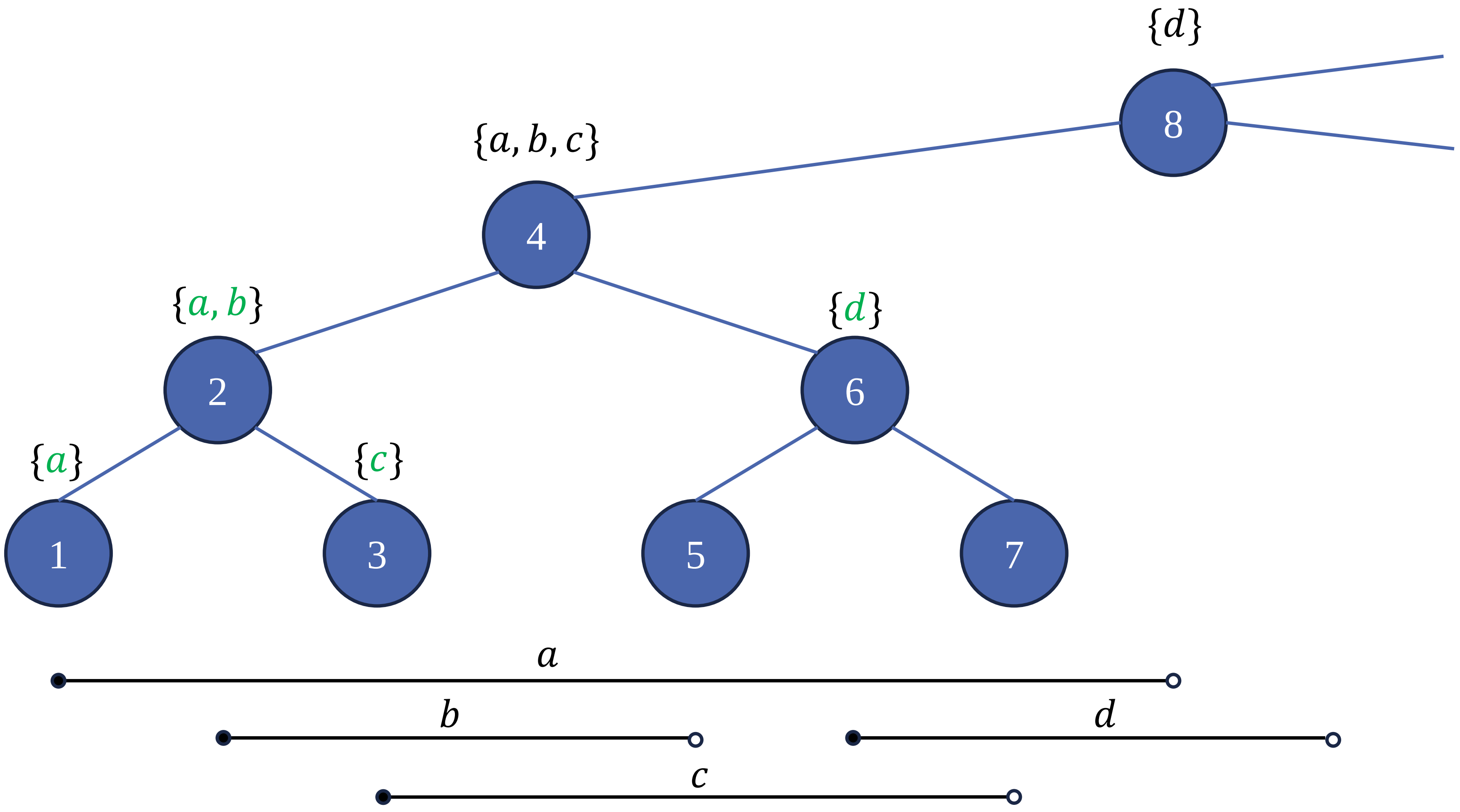
Interval Tree for Stabbing Queries

- Querying: visit nodes on the root-to-node path.



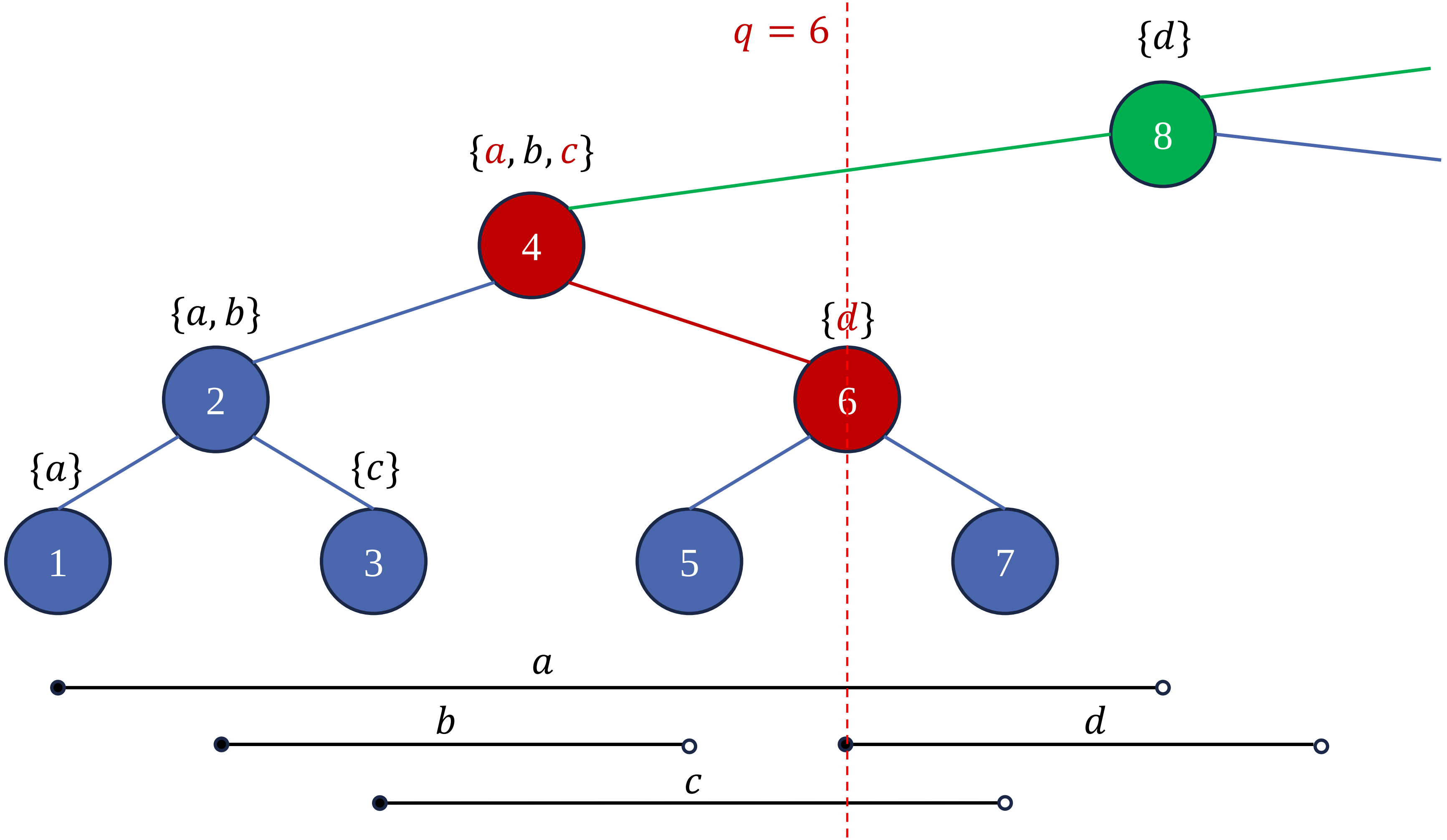
Contribution 1: Online Interval Tree

- Duplicate each node $O(\log t)$ times.



Contribution 1: Online Interval Tree

- Querying: only consider “left” nodes.



Contribution 2: Deletion-only Mechanism

- For each node, it is a deletion-only problem
  - Node i starts with n_i elements, but we can have $n_i \gg n_t$
- Our solution: Track the number of deletions
  - In each round, privately compute: $|D|$, $|D_t^-|$, $F(D)$, $F(D_t^-)$
  - If there are few deletions, then $n_i = O(n_t)$, we simply use $F(D) - F(D_t^-)$
  - If approximately more than $\frac{n_i}{2}$ items are deleted, we allocate new budgets
  - There can only be $O(\log N_t)$ restarts

4	$\{a, b, c\}$	$\Longrightarrow$	$\{a, c\}$	$\Longrightarrow$	$\{a\}$	$\Longrightarrow$	$\emptyset$
	init	…	$(-, b)$	…	$(-, c)$	…	$(-, a)$
	$t = 4$	…	5	…	7	…	8

Conclusion

- There is an (ε, δ)-DP mechanism for arbitrary linear queries on fully dynamic streams with error $\tilde{O}(\sqrt{n_t})$.
- $\tilde{O}$ hides $\frac{1}{\varepsilon}$ and polylog factors in $\frac{1}{\delta}$, $\frac{1}{\beta}$, t, and N_t
- This improves over separating D_t^+ and D_t^-, whose error was $\tilde{O}(\sqrt{N_t})$