PSSketch: Finding Persistent and Sparse Flow with High Accuracy and Efficiency

Jiayao Wang[1], Qilong Shi[2], Xiyan Liang[3], Han Wang[4]
Wenjun Li[4], Ziling Wei[1], Weizhe Zhang[5], Shuhui Chen[1]

1 - National University of Defense Technology
2 - Tsinghua University
3 - Nankai University
4 - Peng Cheng Laboratory
5 - Harbin Institute of Technology

Task

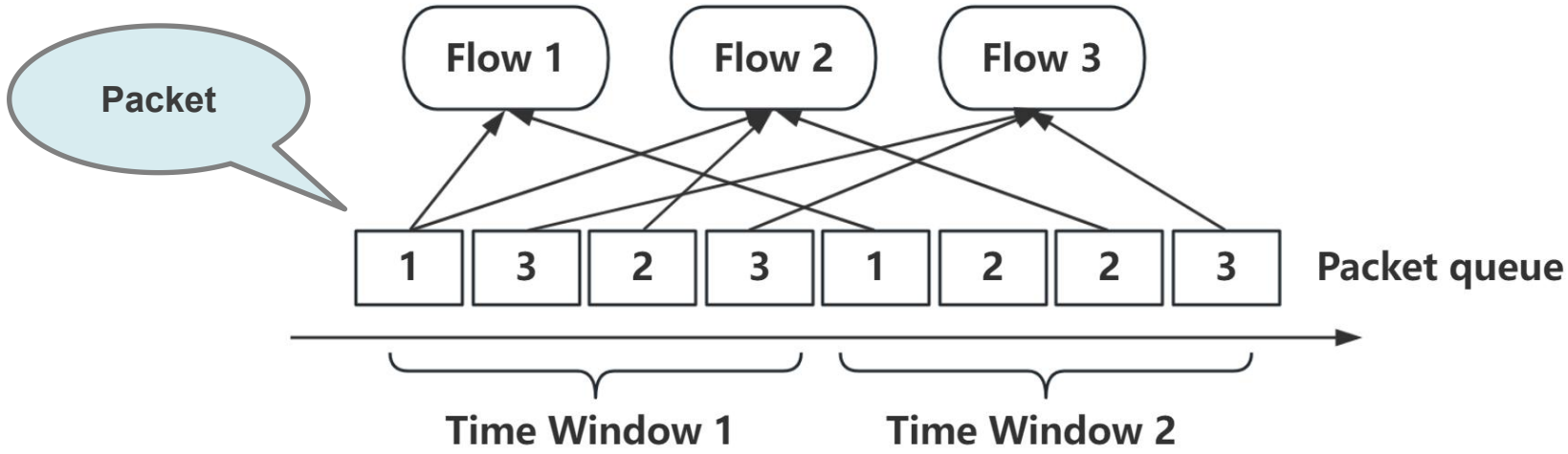
- **Data Flow**

- Cardinality = 3

- Frequency(3) = 3

- Duration(3) = 2

- Density(3) = 3/2 = 1.5



Task

- **Heavy Flows**

★ A flow that contains a large number of packets or data.

★ Majority of flows are tiny flows

★ Large flows carry the majority of data

Flow ID	Packet number	Data size (Bytes)	Duration (Windows)
1	2	11	1
2	3	60	2
3	40	23363	5
4	73	91176	61
5	1	608	1
		...	
N	6	442	2

A few heavy flows

A lot of tiny flows

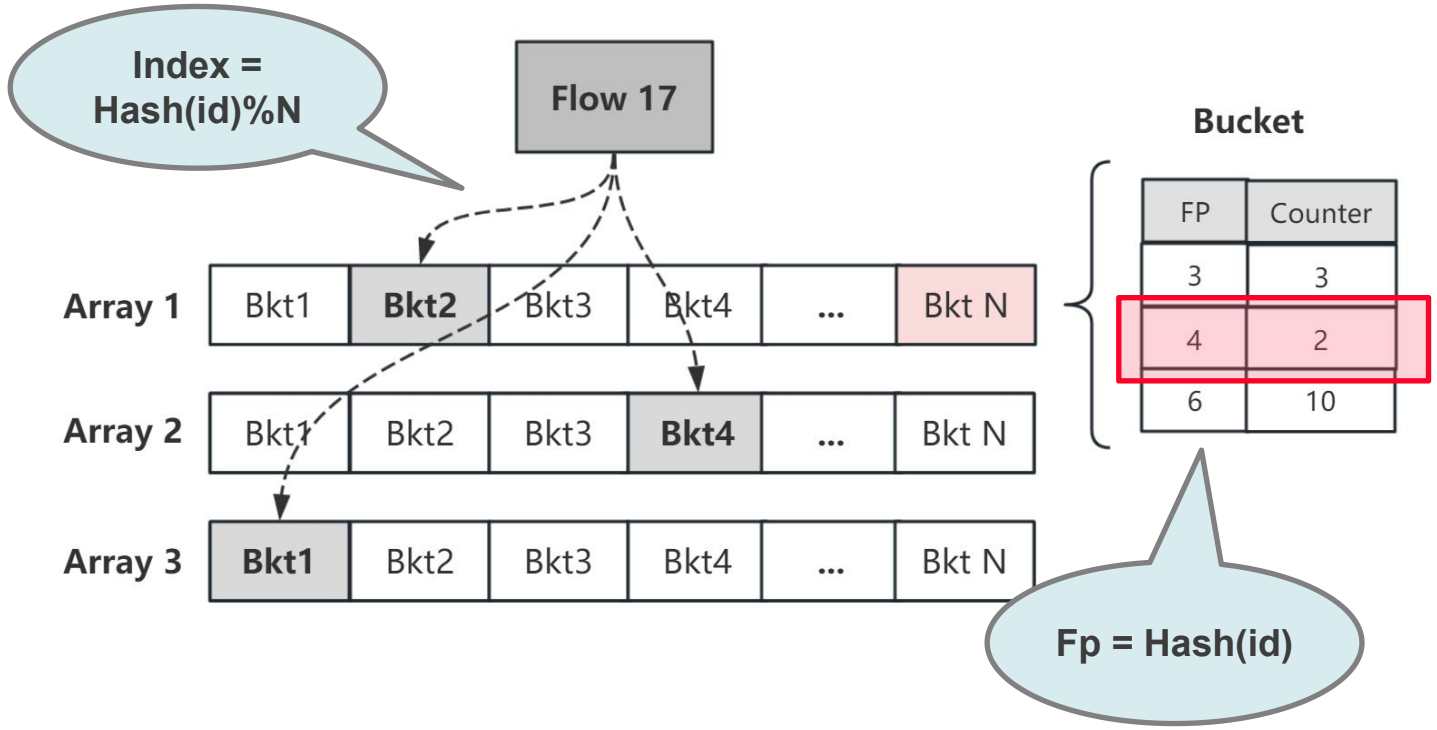
Sketch

- **How Sketch Work**

★ Use Hash Function(s) to find index

★ Store Fingerprint in Bucket

★ Kick out the least valuable item



New Task

- **Persistent Flows**

★ A flow that temporally long in duration.

★ The duration of most of the flows is short

★ Persistent flows can last a very long time and usually correspond to some kind of behavior worth analyzing

Flow ID	Packet number	Data size (Bytes)	Duration (Windows)
1	2	11	1
2	3	60	2
3	40	23363	5
4	73	91176	61
5	1	608	1
		...	
N	6	442	2

*PS Flow

* **P**ersistent and **S**parse

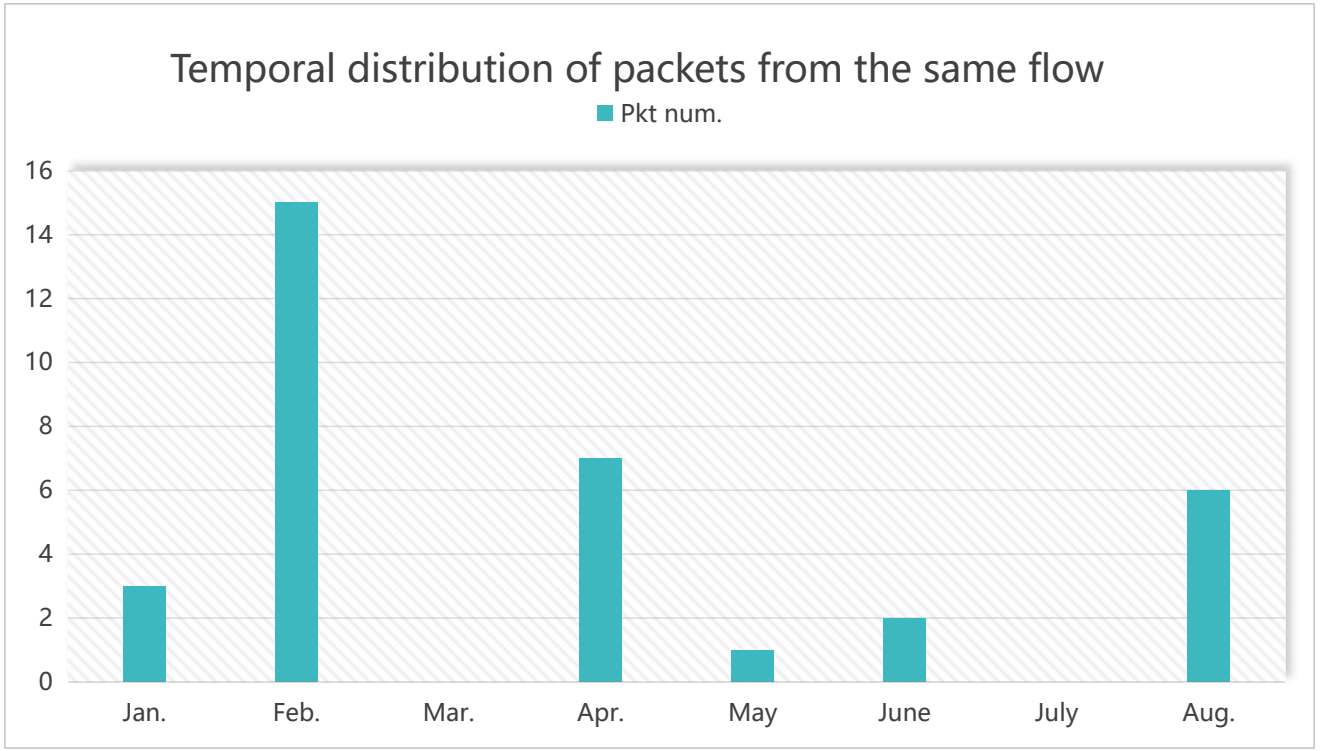
- **Hidden Anomaly Behavior!**

★ Backdoors, APT

★ Mass production account, Proxy

Most normal flows are of short duration. Some flows that carry large amounts of data have higher persistence, as well as density

What about **Persistent and Low-Desity flows?**



Related Work

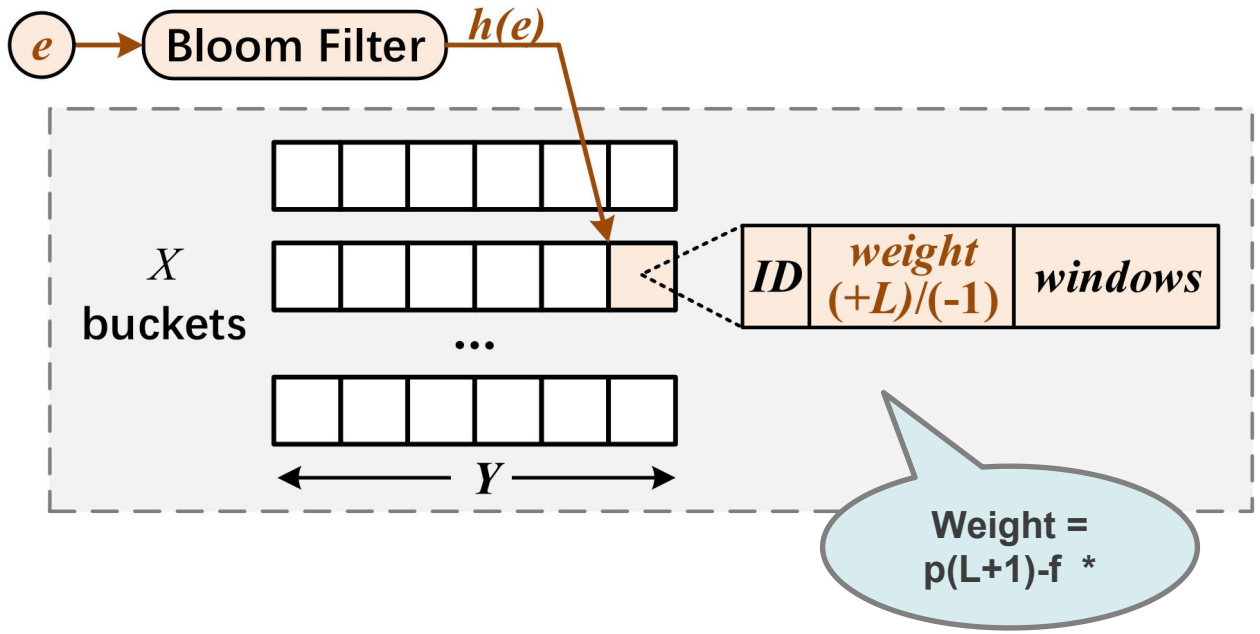
- **PISketch**

★ Use a binary function (Weight) to deal with ‘ Frequency and Persistency.

★ Use a simple Sketch to store Weight and treat the task as a Heavy Flow Task

- **Strawman**

★ Use Heavy Flow Sketch + Persistent Flow Sketch



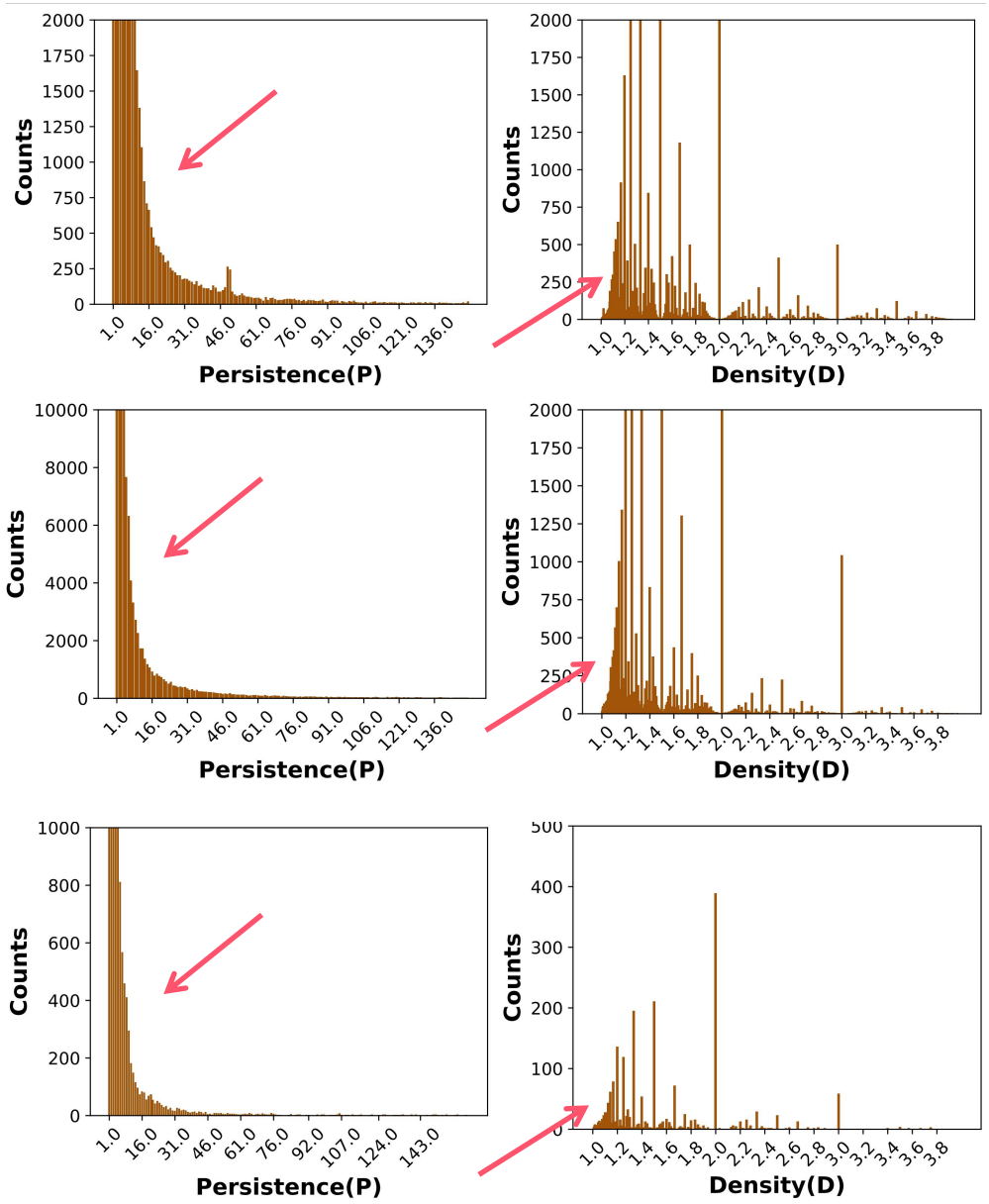
* p - Persistency. Time window count of the flow
f - Frequency. Occurrence count of the flow
L - A preset value

Observation

- **Cliff Feature**

★ The persistence of most flows is very low, and as the persistence increases, the proportion of flows decreases significantly

★ Most flow density(D)* do not approach 1, but are somewhere between 1.2 and 2. The proportion of flows with density close to 1 is extremely small



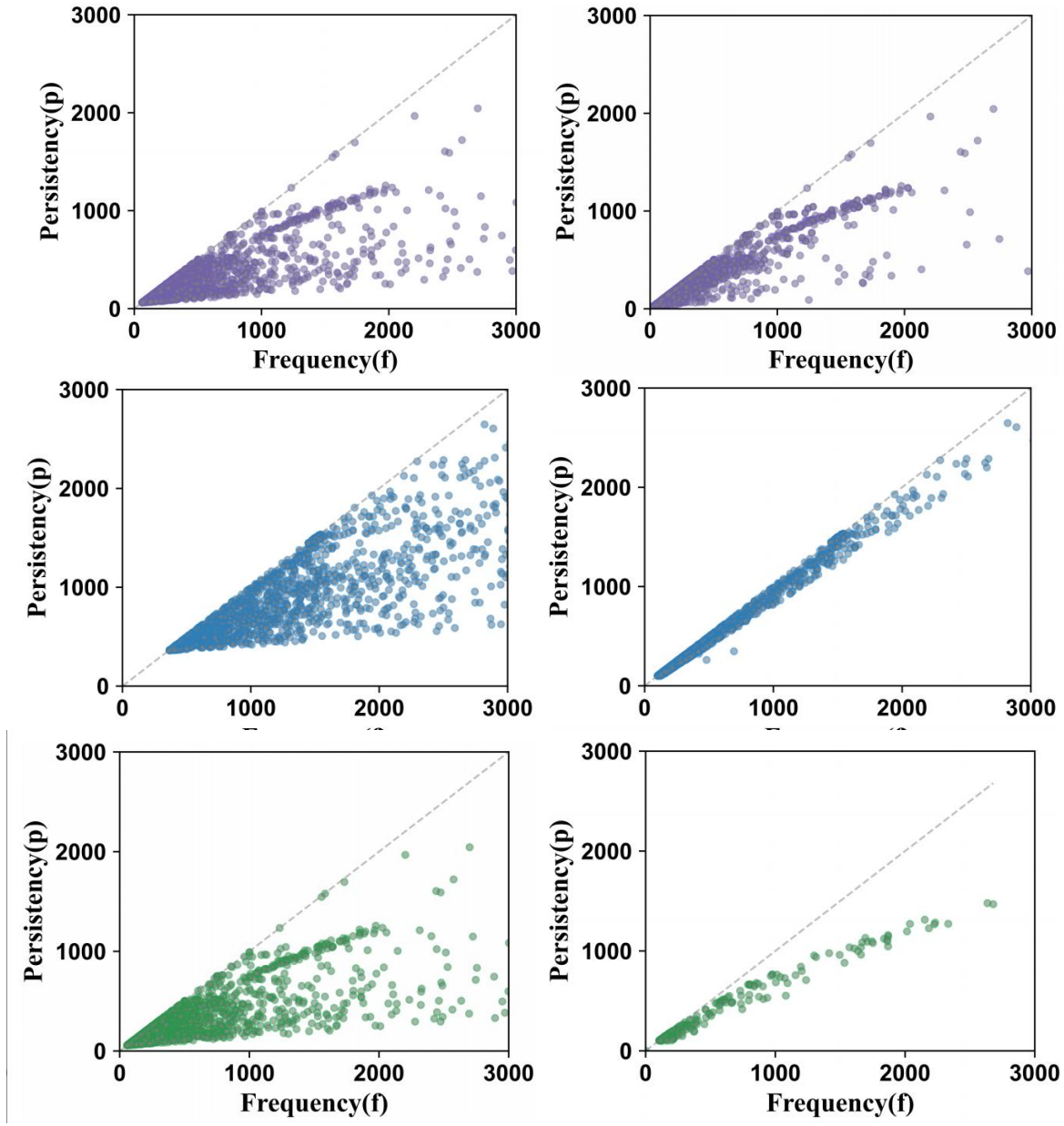
* **D - Density, where D = f/p**
p - Persistency. Time window count of the flow
f - Frequency. Occurrence count of the flow

New Criteria

- **Anomaly Boundaries**

★ At the cliff of two dimensions (persistence and density), regular and irregular flows are distinguished

★ We first find persistence anomalies and then analyze density anomalies



PISketch Ours

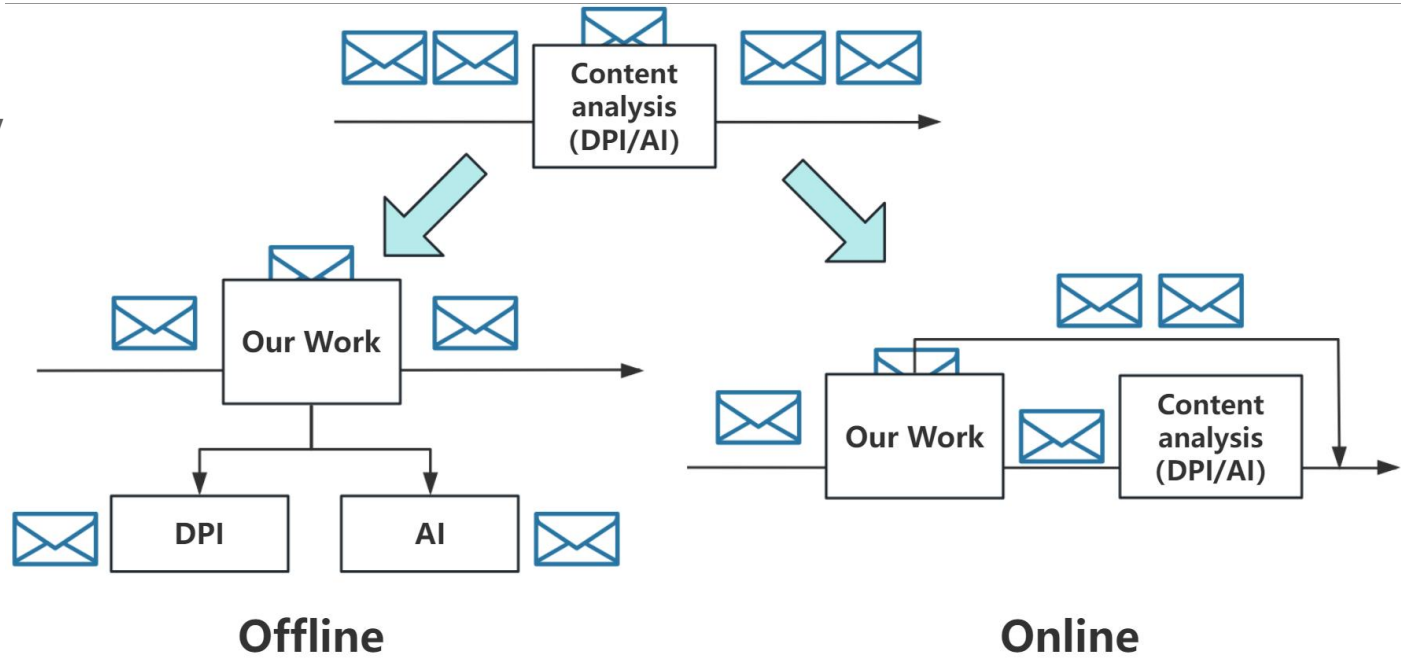
Application Scenario

- **Pre-filter for Content Analysis**

★ Full-flow content analysis is extremely costly

★ Pre-filter the PS flows that look abnormal, usually less than 1% of the whole flow, avoiding a lot of meaningless analysis.

★ Like the thought of Sketch, we may miss some anomalous behavior, but we greatly reduce the workload of content analysis

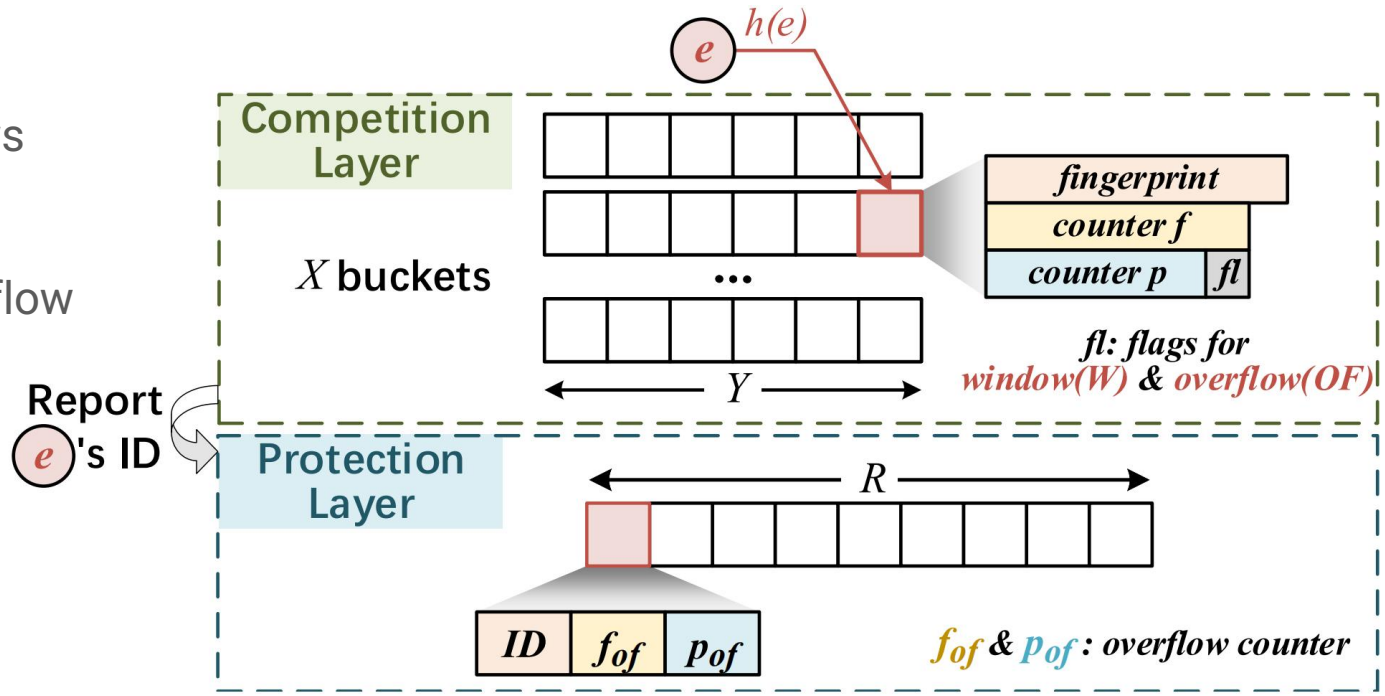


Data Structure

- **PSSketch**

★ The **competition layer** find out persistent flows from the full flow

★ The **protection layer** screens the low-density flow from the persistent flow

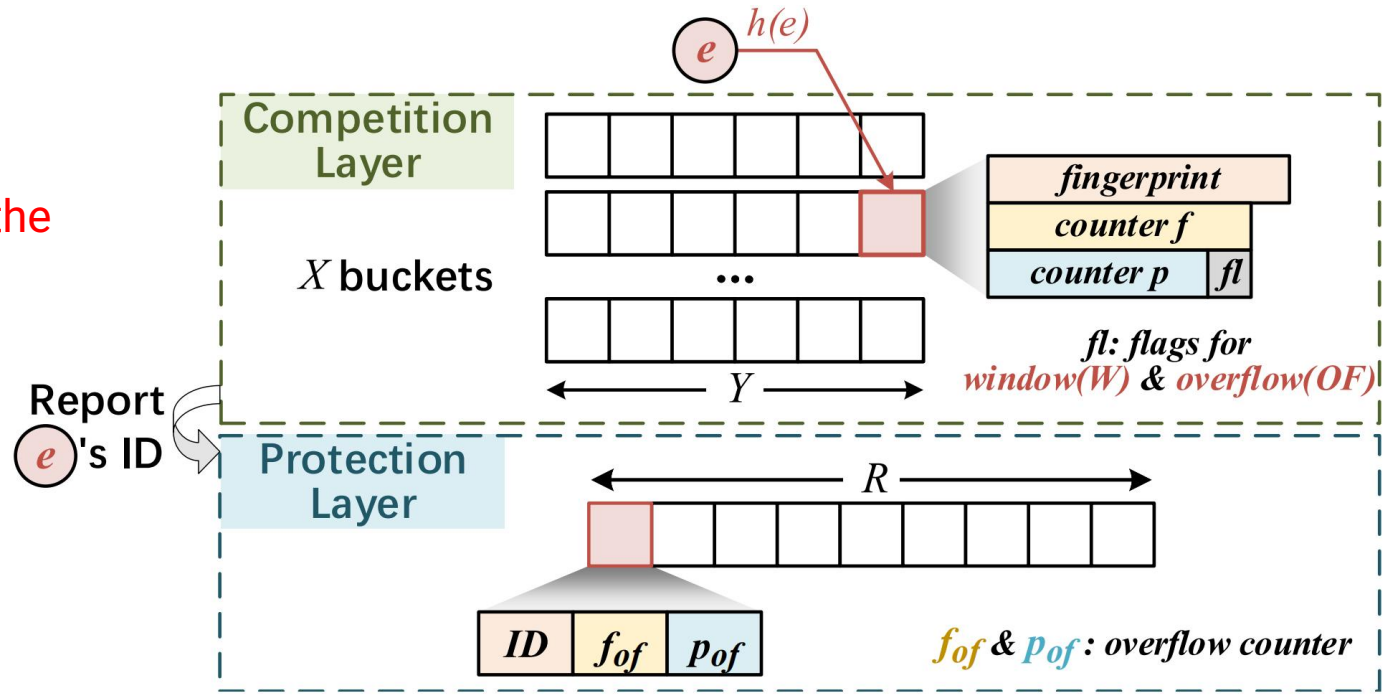


Optimizations

- **PSSketch**

★ The contention layer uses small counters and allows overflow. The protection layer records the number of contention layer counter overflows

★ Typically, data flows only from the contention layer to the protection layer, unless a flow needs to be removed from both layers



Example:

CL = 3 PL = 2 Counter_width = 3bit

cnt = $3+2*2^3=19$

Experiments

- **Accuracy**

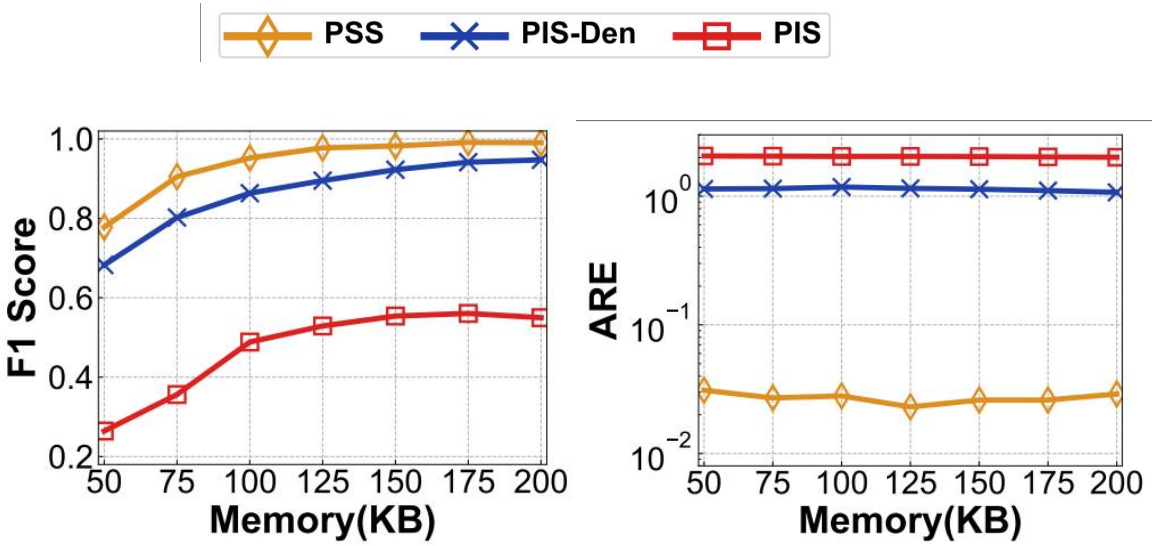
★ Significantly better than existing work

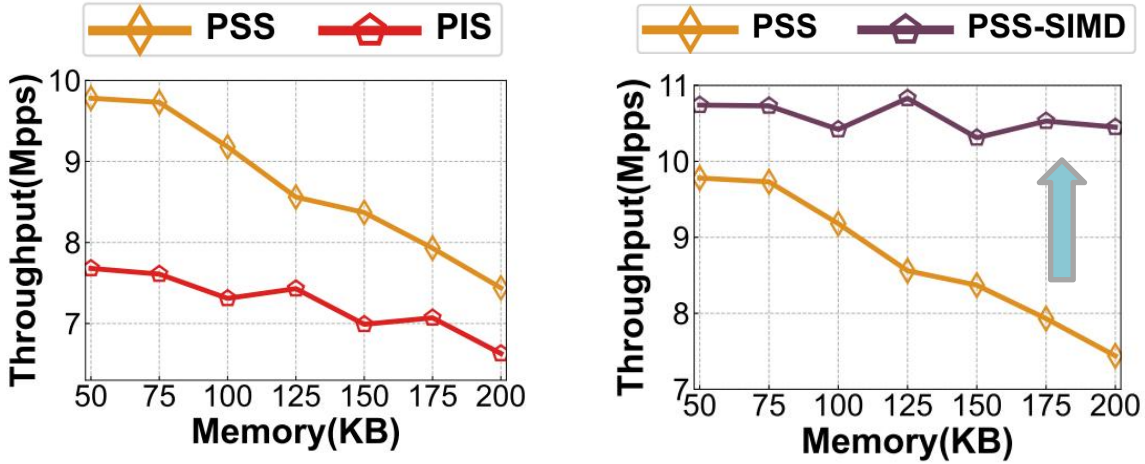
★ Our data structure also has an advantage when using our ceriteria for existing work

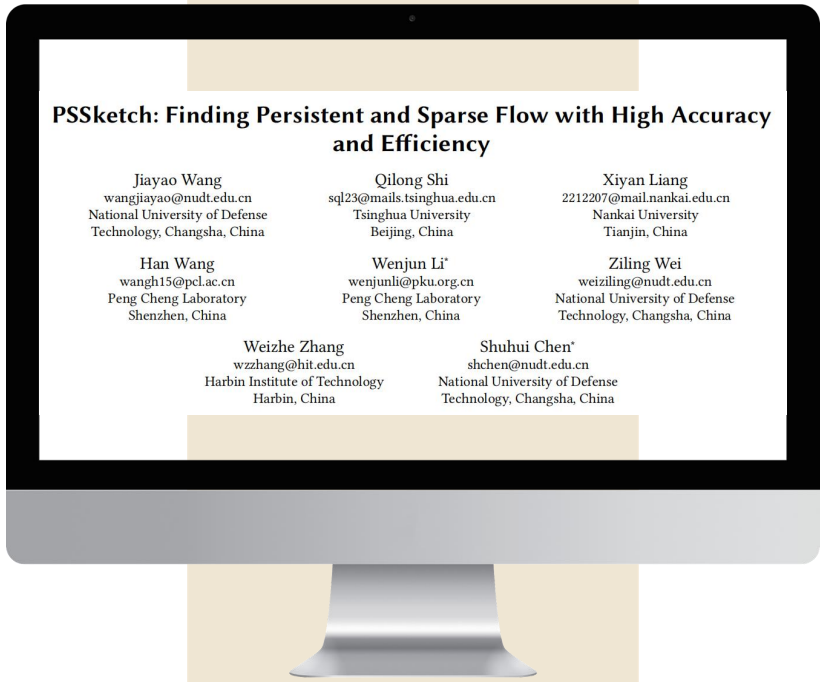
- **Throughput**

★ It outperforms existing work most of the time, declining with memory increase

★ SIMD eliminates this problem.







Summary of PSSketch

★ A strong pre-filter tool for content analysiks.

★ New Criteira concluded from real-world data for describing PS flows

★ Novel two-layer Sketch for reporting PS flows with Storage and algorithm optimization

★ Work With high precision and high throughput

THANK YOU!