北京大学
PEKING UNIVERSITY

Finding Simplex Items in Data Streams

Zhuochen Fan, Jiarui Guo, Xiaodong Li, Peking University, China
Tong Yang, Peking University & Peng Cheng Laboratory, China
Yikai Zhao, Yuhan Wu, Bin Cui, Peking University, China
Yanwei Xu, Huawei Technologies Co. Ltd., China
Steve Uhlig, Queen Mary University of London, UK
Gong Zhang, Huawei Technologies Co. Ltd., China

Outline

PART 01 Background

PART 02 Problem Statement

PART 03 X-Sketch Design

PART 04 Experimental Results

PART 05 X-Sketch for ML

PART 06 Conclusion

01

PART ONE

Background

01/ Background

Frequency estimation is important

Finding frequent items in data streams have been well studied by the community .

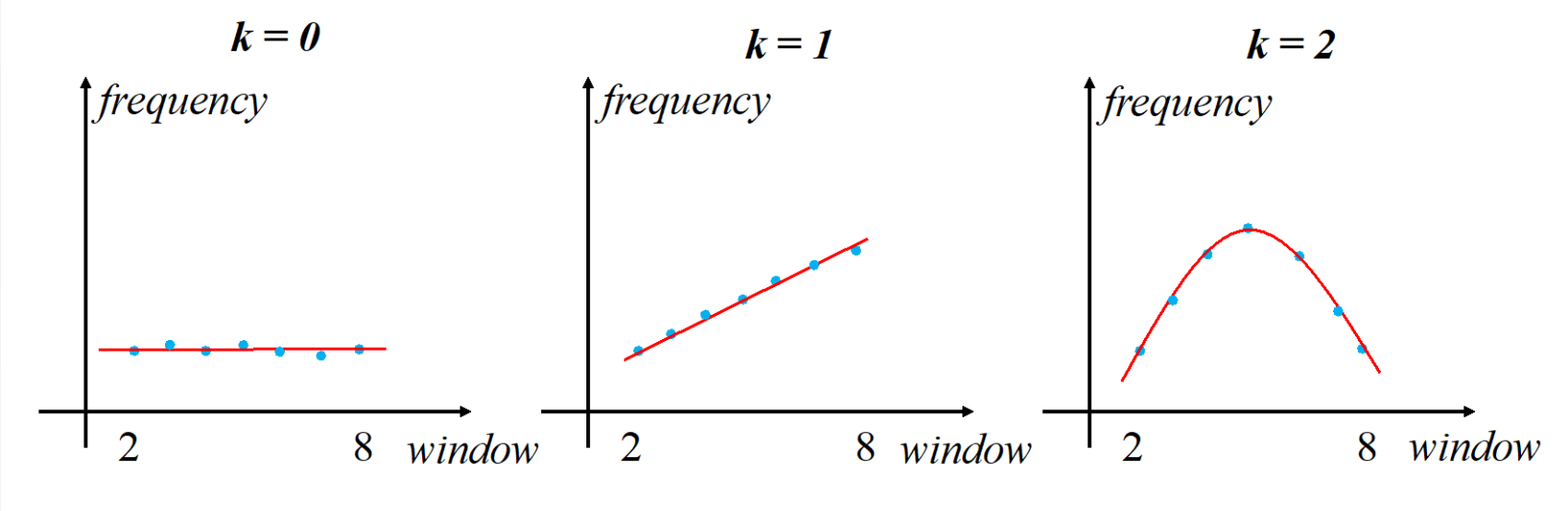
Best solution: Sketch

1) Memory efficient and time efficient
2) Small errors

01/ Background

Patterns in which item frequencies present in a certain number of **consecutive** windows are also worth exploring.

Such patterns in p consecutive windows may be fitted by k-degree polynomials.

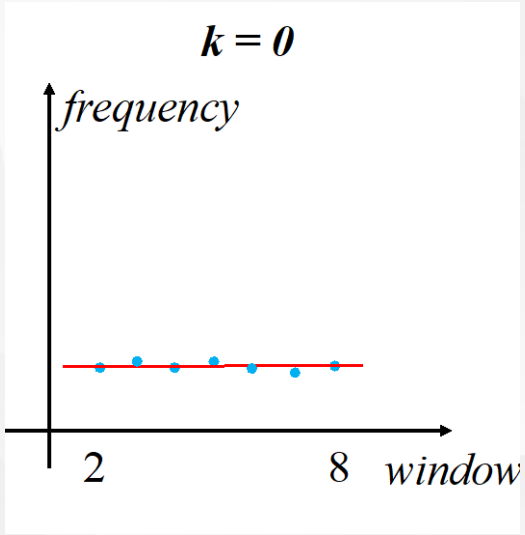


01/ Background

k=0:

1) stable cache lines

2) stable flows in network management

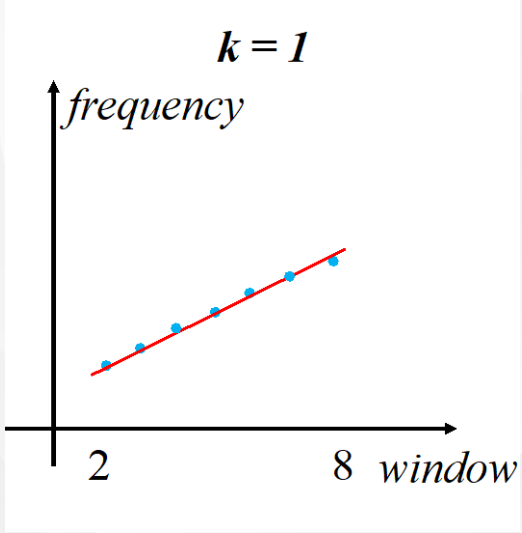


01/ Background

k=1:

1) speed up machine learning models

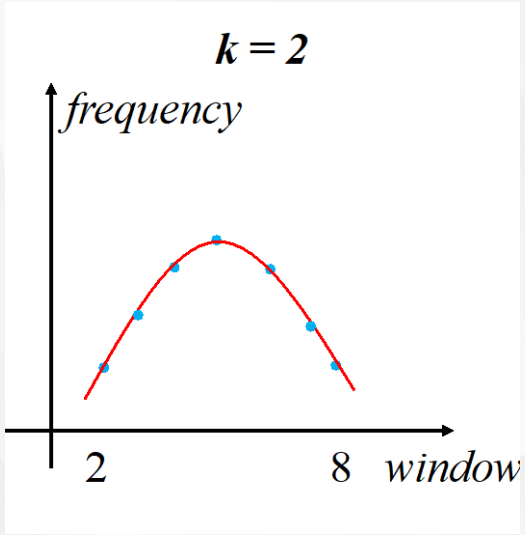
2) detect DDoS attacks



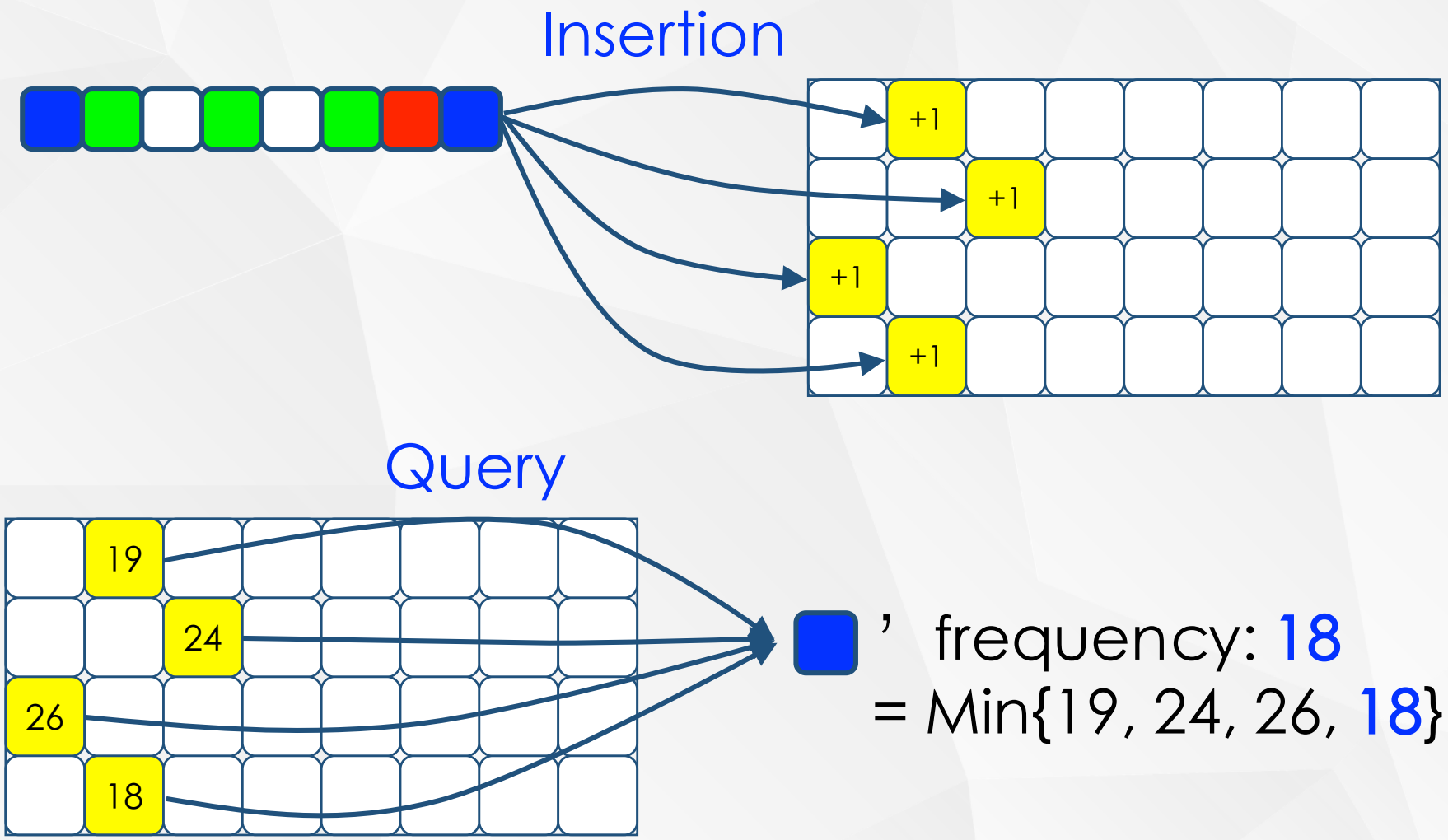
01/ Background

k=2:

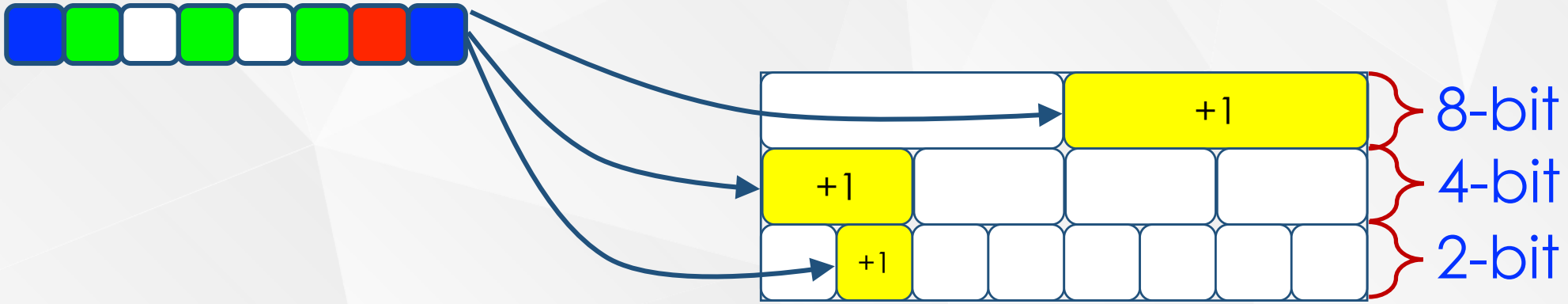
1) monitor traffic in wireless networks



01/ Background- CMSketch



01/ Background- TowerSketch



02

PART TWO

Problem Statement

02 Problem Statement

k-simplex items

An item e is called k-simplex item from window w if and only if its frequencies in p consecutive windows $f_w, f_{w+1}, \cdots, f_{w+p-1}$ satisfy:

(1) $f_{w+i} > 0, \forall i = 0, 1, \cdots, p-1$;

(2) There exists a k^{th}-degree polynomial

$$f(n) = a_0 + a_1 n + a_2 n^2 + \cdots + a_k n^k = \sum_{j=0}^{k} a_j n^j,$$

such that the mean squared error (MSE) ε satisfies

$$\varepsilon = \frac{1}{p} \sum_{i=0}^{p-1} (f(i) - f_{w+i})^2 \leq T;$$

where T is a predefined threshold.

02 Problem Statement

simplex items and persistent items

1) number of windows items appear
2) frequency is important
3) consecutive windows

03

PART THREE

X-Sketch

X-Sketch

01 Baseline Solution

02 An Optimization

03 Data Structure

03 X-Sketch - Baseline Solution

Part 1:

Record frequencies for last p windows using CMSketch

Part 2:

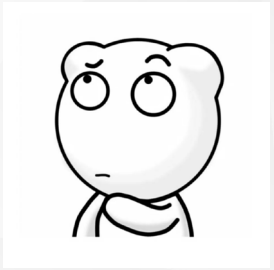
Record lasting time of simplex items using a set and a hash table.

03 X-Sketch - Optimization

The fitting error will not increase if we simply increase k.

Distinguish k-simplex items from (k-1)-simplex items:

Threshold for $|a_k|$



03 X-Sketch - Data Structure

Stage 1:

Record frequencies for last s windows using TowerSketch.

Stage 2:

Record potential simplex items and their lasting time using a hash table.

03 X-Sketch - Data Structure

Stage 1:

1) Short-Term Filtering: quickly filter non-simplex items

2) Potential: measure the probability of becoming simplex items

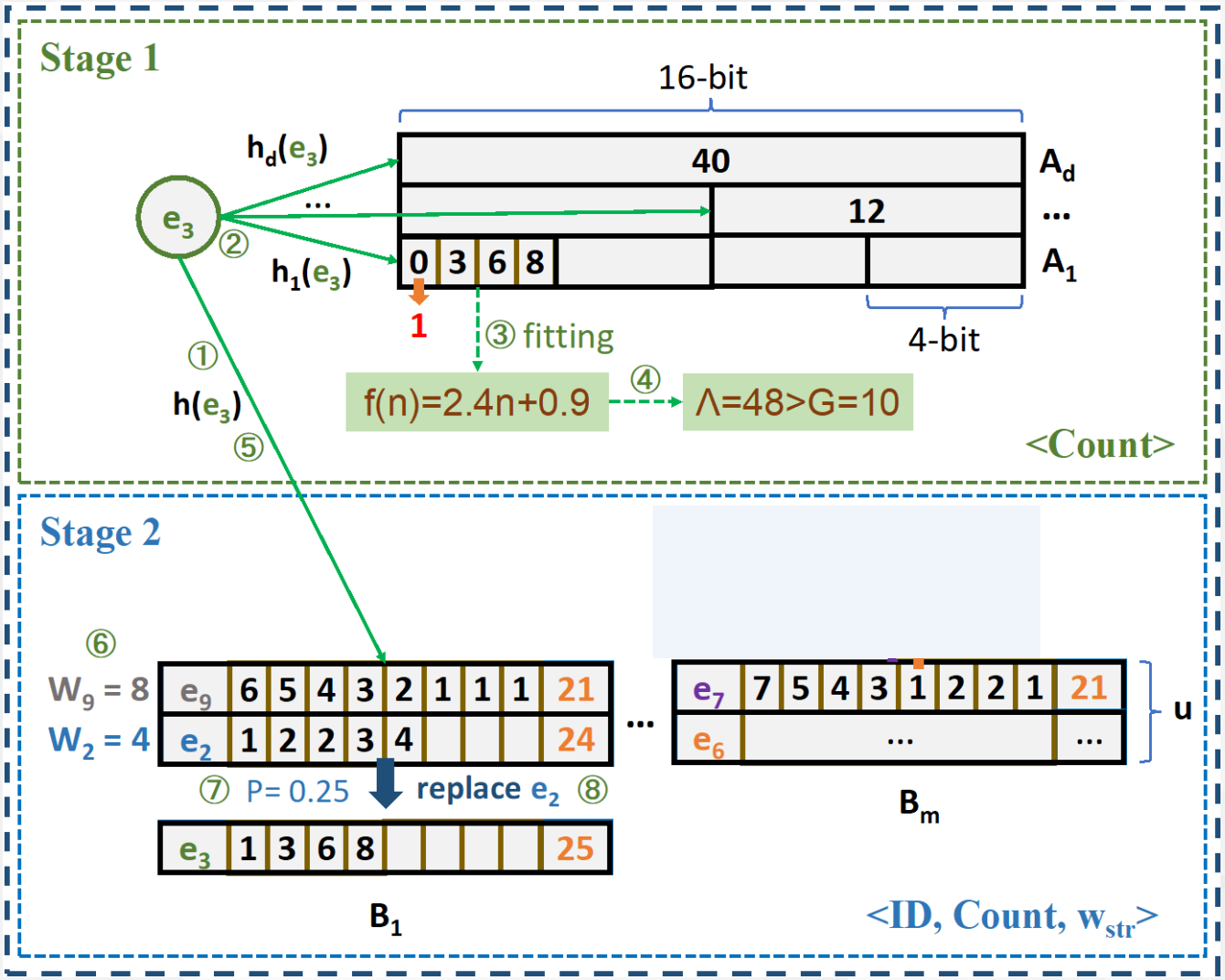
$$\Lambda = \frac{|a_k|}{\varepsilon + \Delta}$$

03 X-Sketch - Data Structure

Stage 2:

Weight Election: a reasonable replacement strategy

03 X-Sketch - Data Structure



04

PART FOUR

Experimental Results

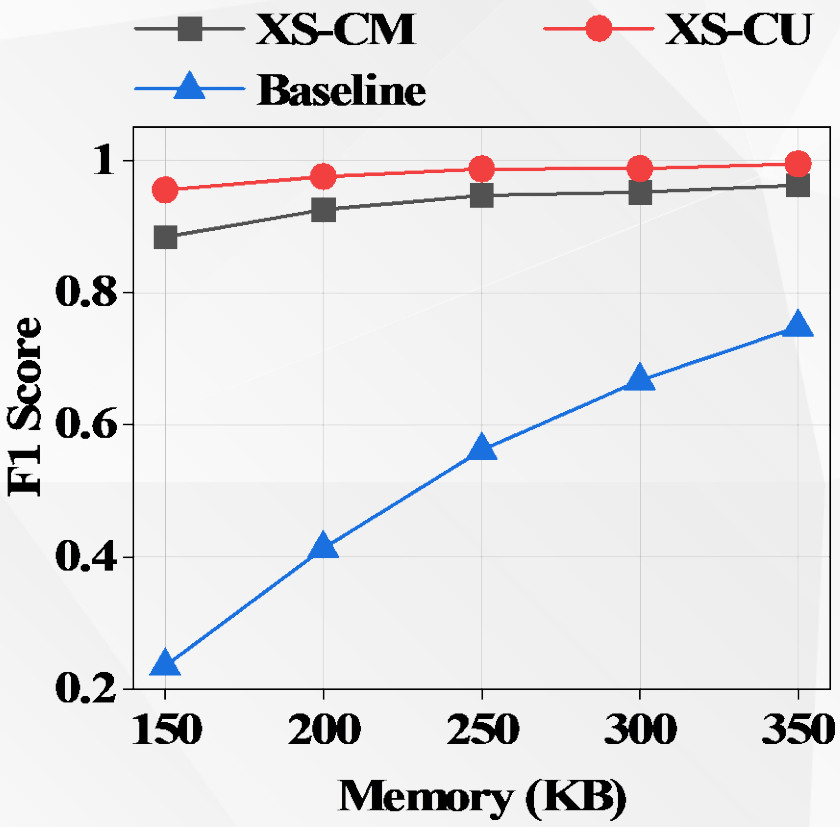
04 Experiments (Setup)

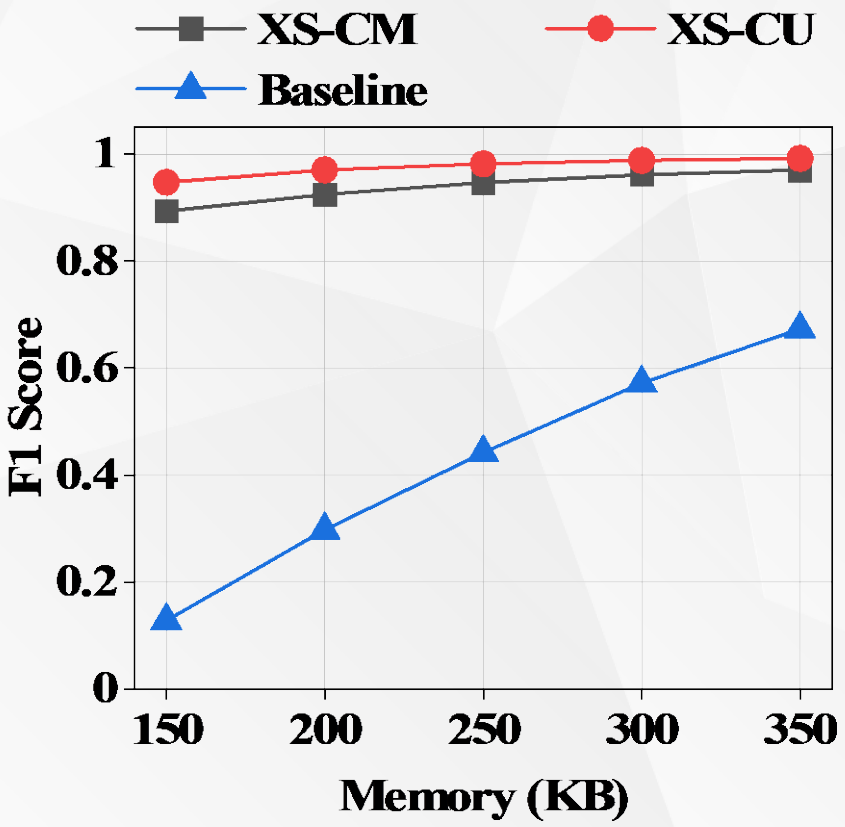
Datasets: CAIDA, MAWI, Data Center, Synthetic.

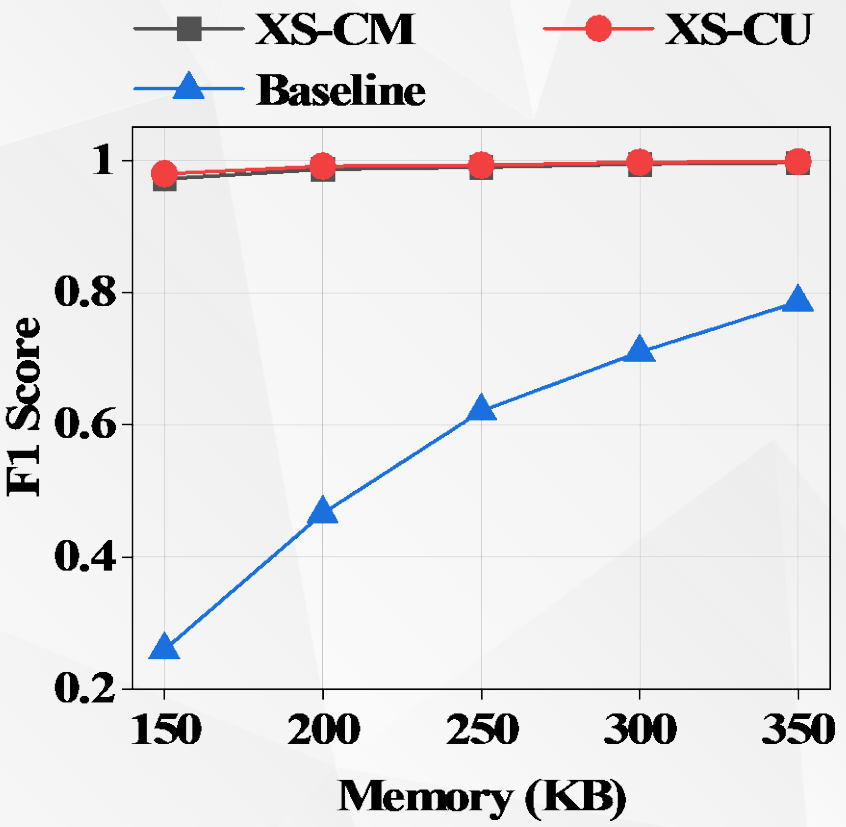
Metrics:

PR, RR, F1 Score, ARE, Throughput

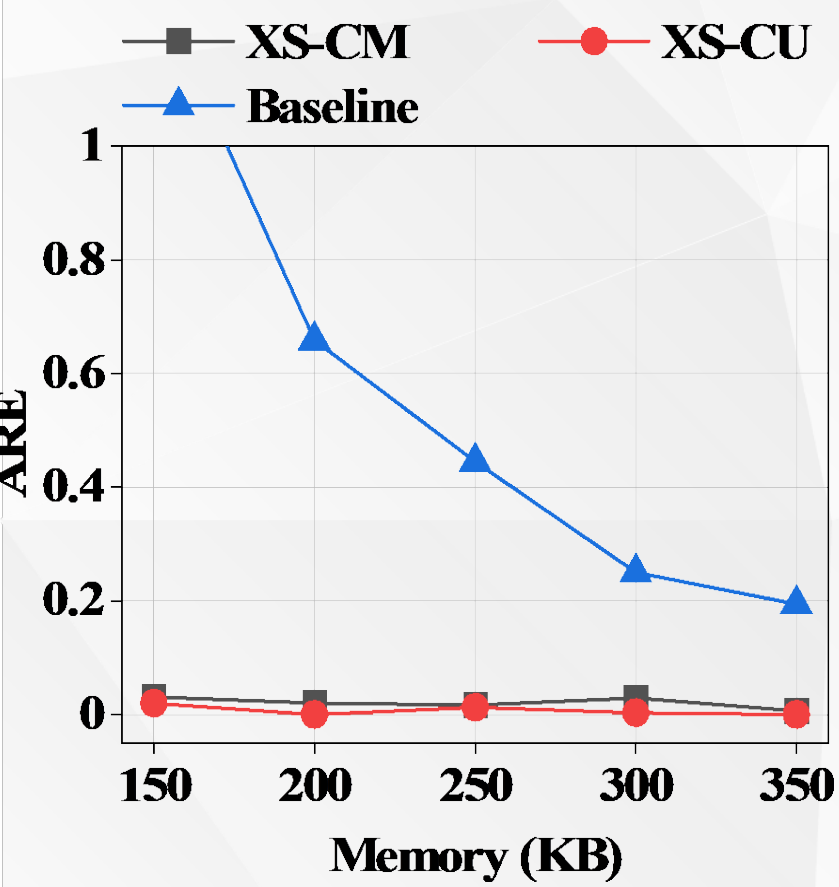
04 Experiments (F1 Score)

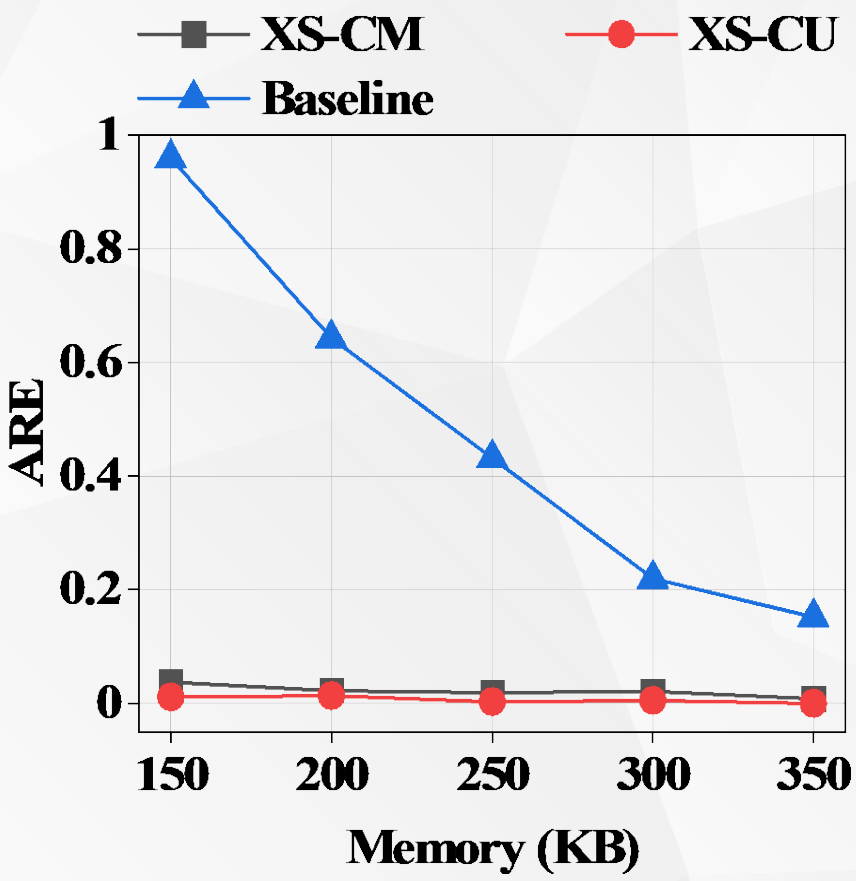


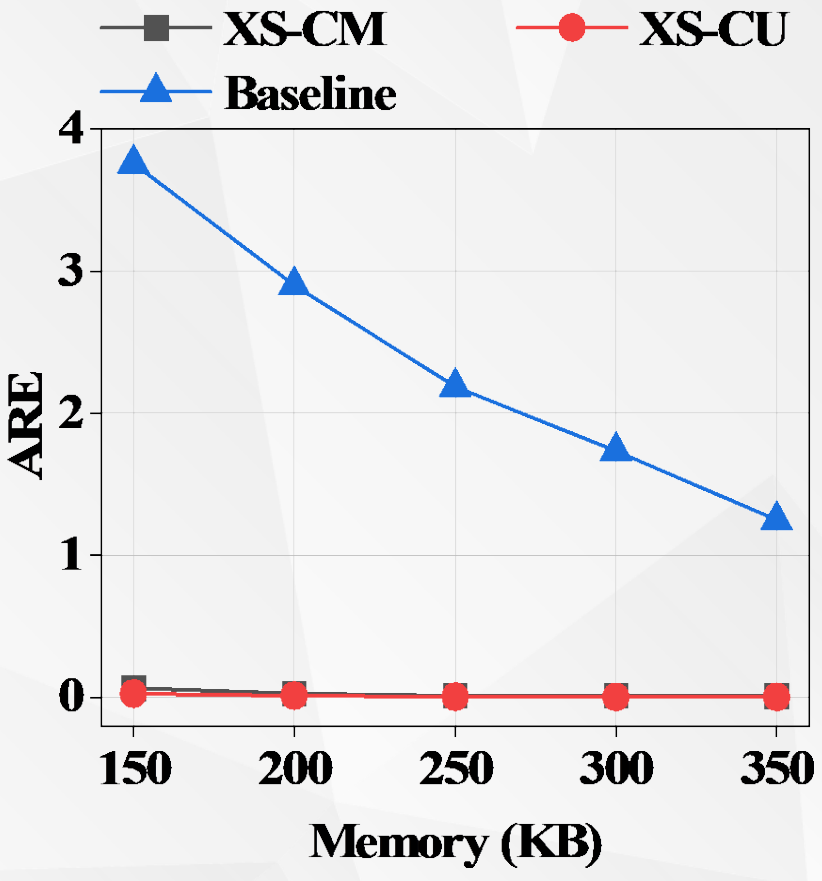




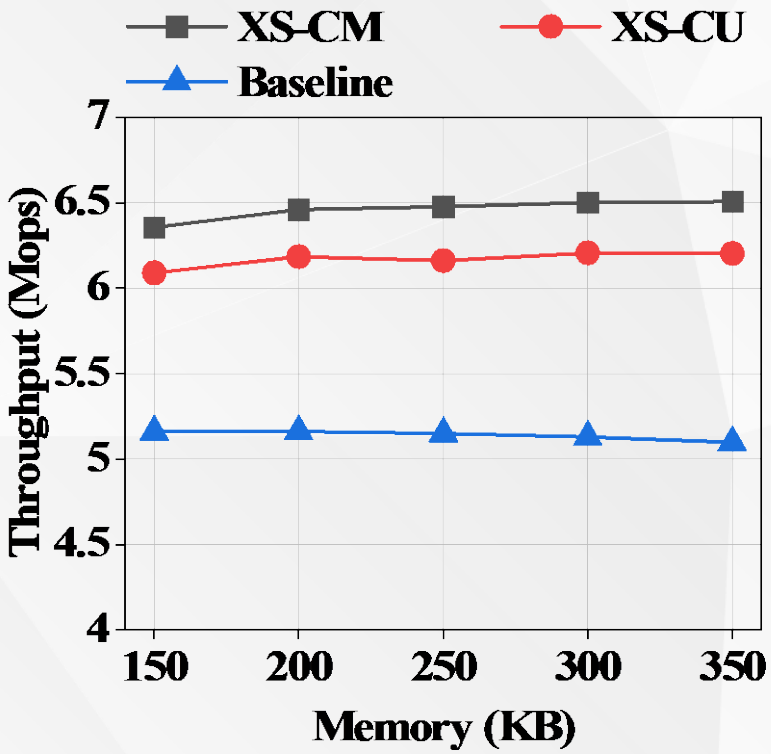
04 Experiments (ARE)

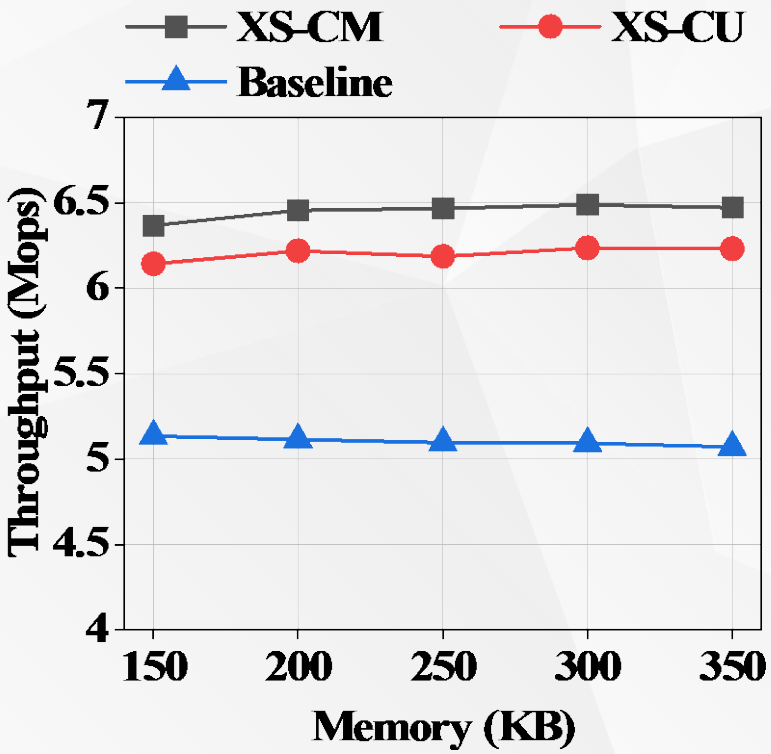


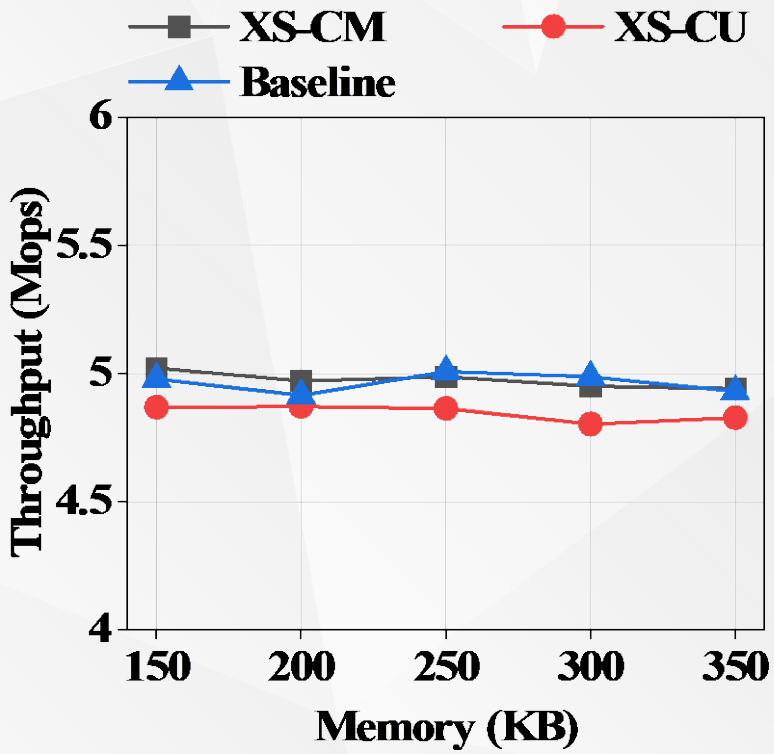




04 Experiments (Throughput)







05

PART Five

X-Sketch for ML

05 X-Sketch for ML

ML for predicting frequency:

1) massive training datasets
2) loops of training epochs
3) not all items are predictable

Solution: X-Sketch + ML

ML

X-Sketch+ML

05 X-Sketch for ML

TABLE III: Experiments on the Transactional Dataset.

	Model	Accuracy (%)	Running Time (s)
$k = 0$	X-Sketch (C++ / py)	98.49	0.014 / 0.225
	Linear Regression	98.38	2.14
	Time Series	98.47	71
$k = 1$	X-Sketch (C++ / py)	90.71	0.013 / 0.222
	Linear Regression	91.67	2.11
	Time Series	93.67	37.6
$k = 2$	X-Sketch (C++ / py)	85.31	0.015 / 0.228
	Linear Regression	85.67	2.13
	Time Series	95.36	71.8

06

PART SIX

Conclusion

06 Conclusion

1. k-simplex items
2. X-Sketch

   Key technique: Short-Term Filtering & Weight Election
3. Theoretical & experimental results
4. Accelerating machine learning

THANKS

Source code: https://github.com/SimpleX-Sketch/X-Sketch

Jiarui Guo
Peking University, China
Email: ntguojiarui@pku.edu.cn
Homepage: https://ntguojiarui.github.io/