CSC4524: Streaming algorithms

Count-min Sketch

Mini-project 3

- **Compute the number of views for each (domain name, page title) couple**
  - Input: domain name + page title
  - Output: #views
  - Work due to 01-12- 2019 23:59

Solution

Count-min sketch

Count-min sketch

- The Count-min sketch is a probabilistic data structure that serves as a frequency table of events in stream of data. It uses hash functions to map events to frequencies. It uses only sub-linear space at the expense of overcounting some events due to collisions.

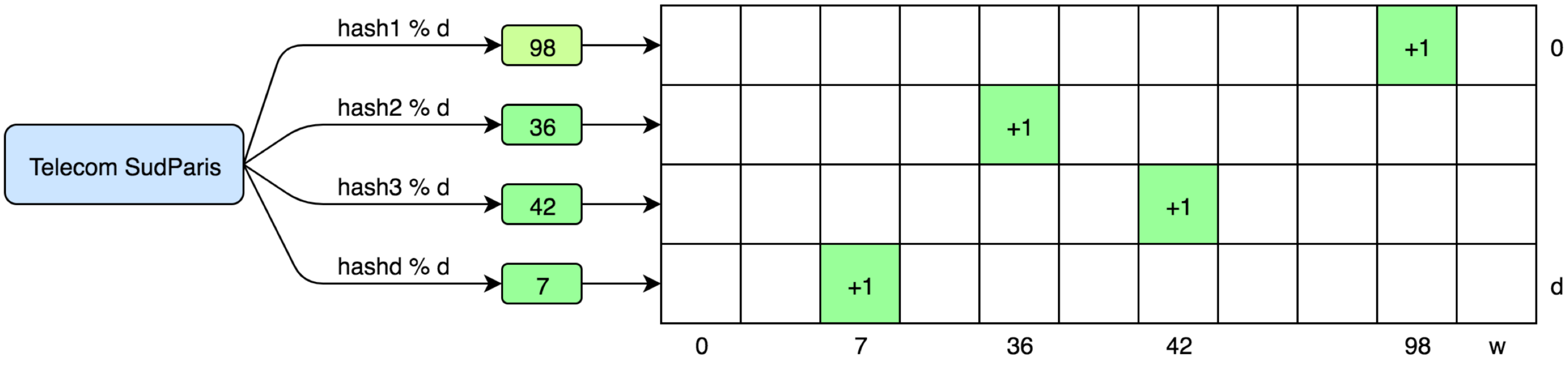
Count-min sketch

- **Int32 table with**
  - w columns
  - d rows

- **d hashing functions**

Add item

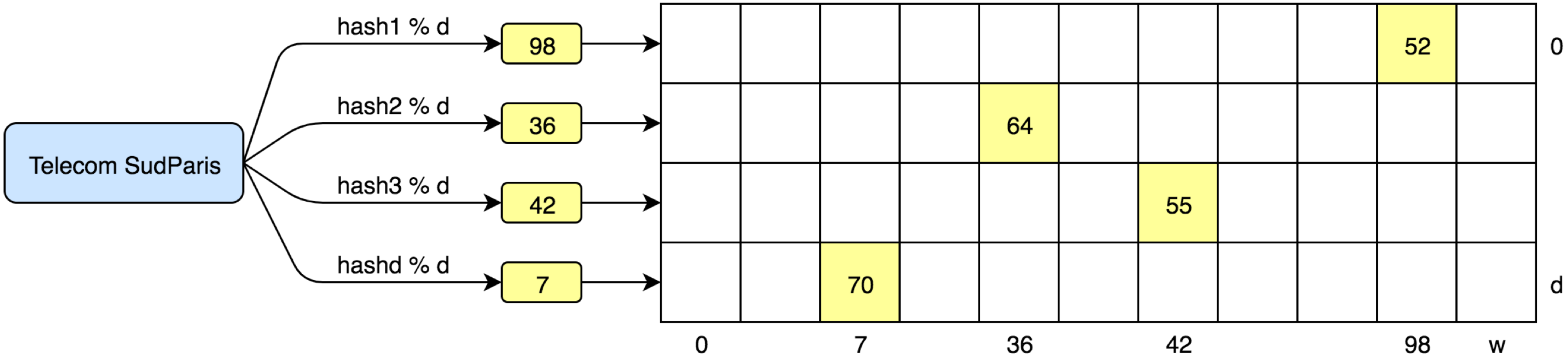
- Let’s add the item “streaming_algo”



w columns

d rows

Retrieve item frequency



min(70 64 55 52) = 52

Complexity

- **Time**

O(#elements)

- **Memory**

O(w.d)

Application

- **Finding heavy hitters**
  - Social networks (Facebook, Twitter …)
  - E-commerce web sites (Amazon …)
  - Entertainment web sites (YouTube …)

Mini-project 4

- **Sub-sample the stream to a fixed size representative sub-set**
  - Input: k the size of the desired sub-set
  - Output: representative sub-set of size k
  - Work due to 09-12-2019 23:59