Operations on distributed data sketches

$\mathcal{M} = (S, m)$

- S - fundamental set
- $m: S \to \mathbb{N}_{\geq 1}$
  number of occurrences of an element

1) Data sketch

$1, 2, 3, 1, 2, 3$ $\rightsquigarrow$ $\boxed{\text{SKETCH}}$

stream

number of unique elements in the stream

$n = |S| = 3$

- memory required to have the exact answer is $O(n)$ (store each unique element)

- we look for a solution with memory of size like $O(\log(N))$ or even better $O(\log \log N)$

$(1, w_1), (2, w_2) (3, w_3), (1, w_1) \ldots$ – stream of pairs (val, size)

$\boxed{\text{SKETCH}}$ $\rightsquigarrow$ total size of unique elements $\boxed{|S|_w = w_1 + w_2 + w_3}$

where $S = \{(1, w_1), (2, w_2) \ldots\}$

and $|S|_w = \sum_{(i, w_i)} w_i$

---

2) Operations on data sketches

node A $\rightsquigarrow$ set A $\rightsquigarrow$ sketch A

node B $\rightsquigarrow$ set B $\rightsquigarrow$ sketch B

- what's $|A \cap B|$, $|A \cup B|$, $|A \setminus B|$, $|A \cap B|_w$, $|A \setminus B|_w$ ...
- the idea: construct a solution that allows to estimate/mimic set theory operations based on data sketches and use any number of sketches and any sequence of operations on those sketches.

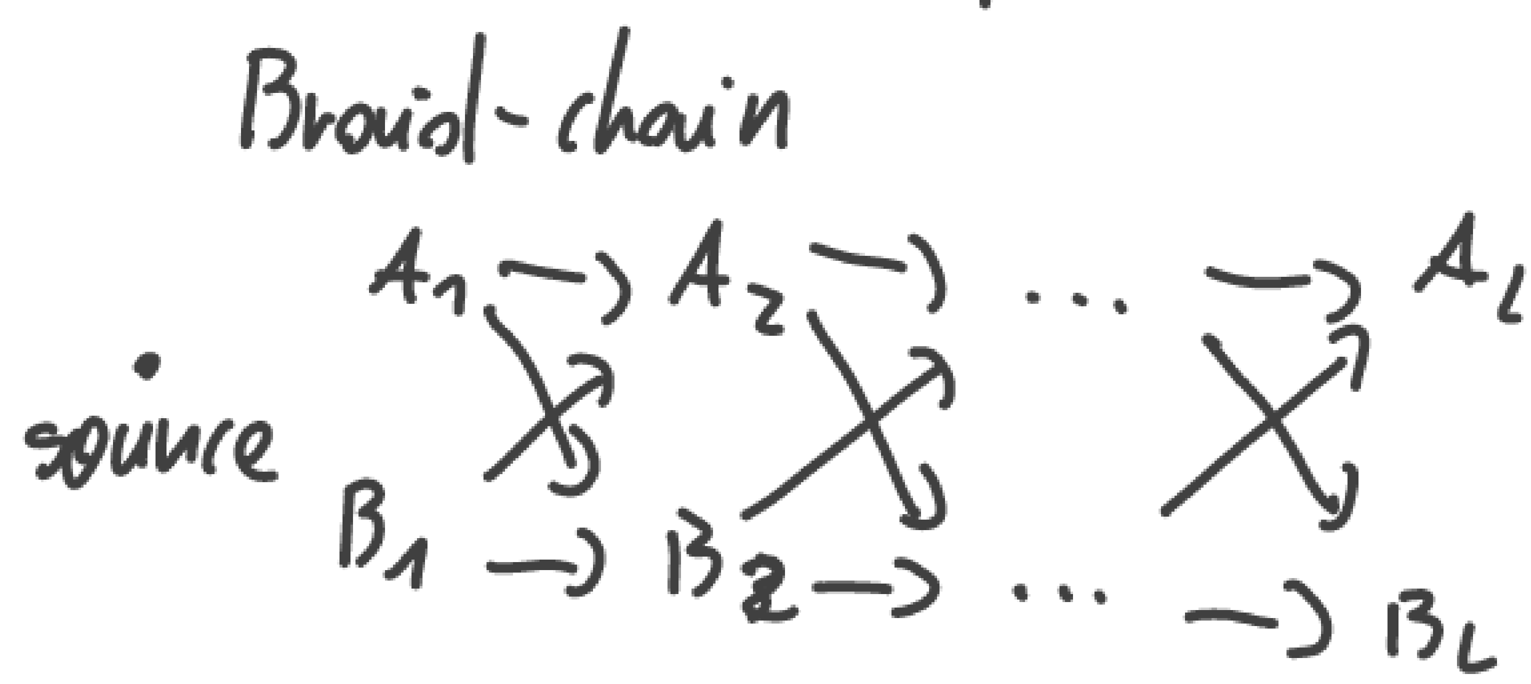
---

3) Applications

(i, w_i)

- network traffic analysis

  A_j, B_j - denote the set of packets observed by A_j, B_j

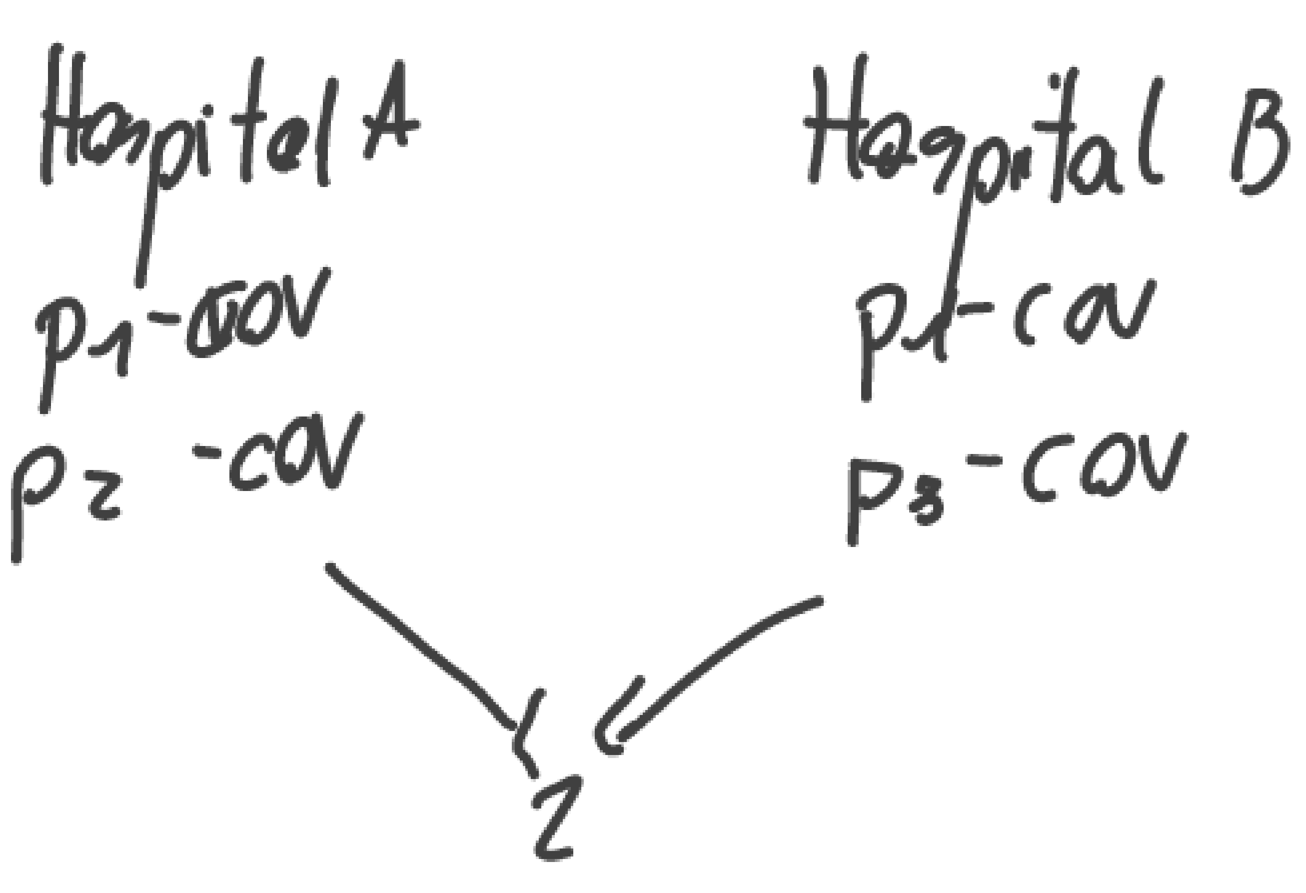


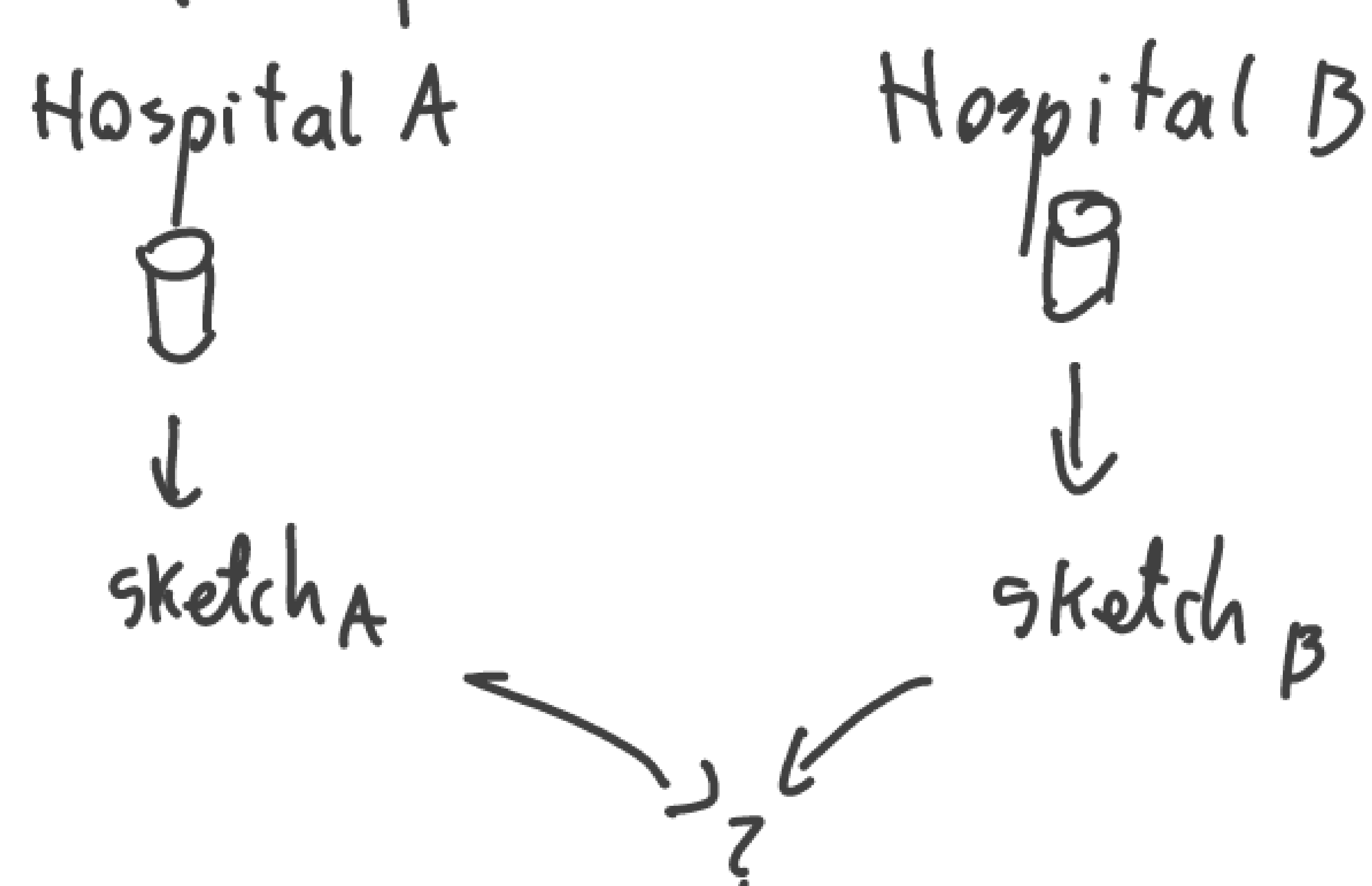
  size of packets that go through a given path:

  $|A_1 \cap A_2 \cap \ldots \cap A_L|_w$

- Query to distributed database

  sketch $A \cup B \rightarrow$ estimate $\pm 10\%$ for up to $4 \cdot 10^8$ element (?)





4) Problem formalization

- Let $\mathcal{M} = (S, m)$ be a multiset, where S is a fundamental set and $m: S \to \mathbb{N}_{\geq 1}$ a multiplicity function
- Problem: We observe element of $\mathcal{M}$ sequentially and we try to estimate $n = |S|$ using memory of size $o(n)$, e.g. $O(\log n)$ or $O(\log\log(n))$
- MinCount, HyperLogLog ⇝ counting problem (number of unique elements)
  1990, 2007, 2016
- ExpSketches ⇝ total weights of unique elements.
  2019, 2021

5) MinCount($k, h, \mathcal{M}$)

// $k \geq 3$

$\mathcal{M} = (S, m)$

$h: S \to \{0,1\}^B \in [0,1]$

$h(s_i) \sim$ Uniform dist

$h(s_1), h(s_2), \dots$ independent

<u>h is hash</u>

$M \leftarrow (1, 1, \dots, 1)$ $\quad |M| = k$

for each $s \in \mathcal{M}$

  if $h(s) \notin M \wedge h(s) < M[k]$

    $M[k] \leftarrow h(s)$

    sort(M) // increasing order

anytime we can estimate $n = |S|$

if $M[k] == 1$ return $|\{i : M[i] \neq 1\}|$

else return $\hat{n} = \frac{k-1}{M[k]}$ ⇝ $E[\hat{n}] = n$

6) Memory consumption

- we have a fixed size of array M, k doesn't depend on n.
- what is the length B of hash value to avoid collisions?

Birthday paradox $\sqrt{2^B} = n$, we have small probability of collision

↳ $B = 2\log_2 n$

- k hash values of length $B = 2\log_2 n \to$ memory $O(\log n)$

7) Precision of estimate

Definition:

Let $X_1, \dots X_n$ be a sequence of any random variables. We sort their realisations in an ascending order:

$$X_{1:n} \leq \dots \leq X_{n:n}$$

Variable $X_{i:n}$ we call the i-th order statistic.

For example:

$$X_{1:n} = \min(X_1 \dots X_n)$$
$$X_{n:n} = \max(X_1 \dots X_n)$$

We assume that $h(x_i) = V_i \sim U(0,1)$

$V_1, V_2, \dots V_n$ - are independent

Theorem

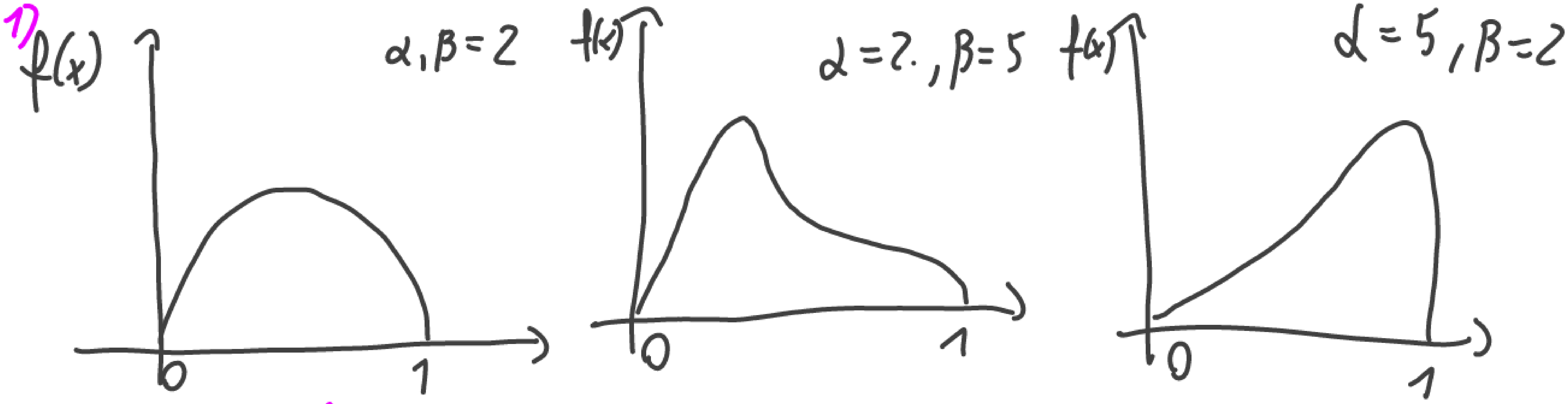
Let us consider $V_1 \dots V_n$, $V_i \sim U(0,1)$, then variable $V_{k:n}$ has beta distribution $Beta(k, n+1-k)$.

$$V_{k:n} \sim M[k] \to \frac{k-1}{M[k]}$$

Proof: exercise 13 and 14

1) $X \sim Beta(\alpha, \beta)$ has density $f(x, \alpha, \beta) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$, $x \in (0,1)$, $\alpha, \beta > 0$

1)



2) $B(\alpha,\beta)$ – beta function

- $B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt$, $Re(\alpha), Re(\beta) > 0$
- $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, $\Gamma(s) = \int_0^\infty e^{-t}\, t^{s-1}\,dt$, $Re(s) > 0$
- $n \in \mathbb{N}$: $\Gamma(n+1) = n!$

3) Ex 13: $X_1 \dots X_n \sim f(x)$, if $F(x)$ have the same density $f(x)$ (dist)

then $X_{k:n} \sim f_k(x) = \dfrac{F^{k-1}(x)\,[1-F(x)]^{n-k}\cdot f(x)}{B(k, n+1-k)}$ *

4) Ex 14: Use * and $f(x)$, $F(x)$ for uni dist. to show that

$U_{k:n} \sim Beta(k, n+1-k)$