# Probabilistic Counting

Andrea Marino

University of Pisa

Pisa, March 2015

Part I

# Size Estimation Frameworks

A sketch $S(A)$ is a compressed form of representation for a given set $A$ providing the following operations:

`INIT` $(S(A))$ How a sketch $S(A)$ for $A$ is initialized.

`UPDATE` $(S(A), u)$ How a sketch $S(A)$ for $A$ modifies when an element $u$ is added to $A$.

`UNION` $(S(A), S(B))$ Given two sketches for $A$ and $B$, provide a sketch for $A \cup B$.

`SIZE` $(S(A))$ Estimate the number of distinct elements of $A$.

## Properties

- Given two sketches $S(A)$ and $S(B)$ for any two sets $A$ and $B$, $S(A \cup B)$ can be computed just by looking at $S(A)$ and $S(B)$, i.e.: C=$\texttt{UNION}$ $(S(A), S(B)) \equiv$ ($\texttt{INIT}$ $(C)$; for $u \in A \cup B$, $\texttt{UPDATE}$ $(C, u)$ ; $\texttt{RETURN}$ $C$ )
- If we call $\texttt{UPDATE}$ $(S(A), u)$ and we already did $\texttt{UPDATE}$ $(S(A), u)$ with the same $u$ the sketch does not modify, e.g. the operation $\texttt{SIZE}$ $(S(A))$ return the same value.

# $k$-min approach (Cohen 1994)

A $k$-min sketch includes the item of smallest rank in each of $k$ independent permutations.

- In this case a sketch $S(A)$ is a sequence of exactly $k$ entries, $a_1, \ldots, a_k$, where each entry can be an element of $A$ or $\perp$.
- Let $r_1, r_2, \ldots, r_k$ be ranks $r_i : U \to \{1/n, 2/n, 3/n, \ldots, 1\}$, setting $r(\perp) = \infty$.

`INIT` $(S(A))$: $a_i = \perp$ for any $i$.

`UPDATE` $(S(A), u)$: for every $i$, with $1 \leq i \leq k$, if $r_i(a_i) > r_i(u)$, $a_i$ is replaced with $u$.

`UNION` $(S(A), S(B))$: return $\{c_1, \ldots, c_k\}$ such that $c_i = argmin\{r(a_i), r(b_i)\}$.

`SIZE` $(S(A))$: return $k / \sum_{a_i \in S(A)} r(a_i) - 1$

Choosing $k = \Theta(\epsilon^{-2} \log n)$ the relative error is bounded by $\epsilon$ w.h.p.

- Edith Cohen: Estimating the Size of the Transitive Closure in Linear Time. FOCS 1994: 190-200
- Edith Cohen: Size-Estimation Framework with Applications to Transitive Closure and Reachability. J. Comput. Syst. Sci. 55(3): 441-453 (1997)

# bottom-$k$ approach (Cohen & Kaplan 2007)

Given

- a *ranking* (i.e., a bijective function) $r : U \rightarrow \{1/n, 2/n, \ldots, 1\}$
- and a subset $A$ of $U$

we denote as

- $H_k(A)$ the first $k$ elements of $A$ according to $r$ and
- with $k_{th}(A)$ the rank of the $k$-th element of $A$ according to $r$

A bottom-$k$ sketch includes the $k$ items with smallest rank in a single permutation, that is, $H_k(A)$.
In this case a sketch is simply a subset of the set.

Given a rank $r : U \rightarrow \{1/n, 2/n, 3/n, \ldots, 1\}$:

`INIT` $(S(A))$: $S(A)$ is the empty set.

`UPDATE` $(S(A), u)$: If $|S(A)| < k$, add $u$ to $S(A)$. Otherwise, if $r(\max(S(A))) > r(u)$, replace $\max(S(A))$ with $u$.

`UNION` $(S(A), S(B))$: the first $k$ elements of the set $S(A) \cup S(B)$, that is $H_k(S(A) \cup S(B))$

`SIZE` $(S(A))$: return $(k-1)/k_{th}(S(A))$

- Edith Cohen, Haim Kaplan: Summarizing data using bottom-k sketches. PODC 2007: 225-234
- Edith Cohen, Haim Kaplan: Bottom-k sketches: better and more efficient estimation of aggregates. SIGMETRICS 2007: 353-354
- Edith Cohen, Haim Kaplan: Tighter estimation using bottom k sketches. PVLDB 1(1): 213-224 (2008)

# LogLog counters (Flajolet & Martin 1985)

- In this case a sketch $S(A)$ is a sequence of exactly $k$ entries, $a_1, \ldots, a_k$, where each entry is a sequence of $m$ bits.
- Given $k$ partition functions $p_i : U \to \{1, 2, \ldots, m\}$, for each $j$, with $1 \leq j \leq m$ the bit $a_{i,j} = 1$ if there exists an element in $A$ that is mapped to $j$ according to $p_i$, 0 otherwise.
- Each partition function is such that randomly $1/2$ of the elements are mapped to 1, $1/4$ of the elements are mapped to 2, ..., $1/2^i$ of the elements are mapped to $i$.

# LogLog counters

`INIT` $(S(A))$: $a_{i,j} = 0$ for any $i, j$.

`UPDATE` $(S(A), u)$: for any $i$, let $p_i(u) = j$, set $a_{i,j}$ to 1.

`UNION` $(S(A), S(B))$: return $\{c_1, \ldots, c_k\}$ where $c_i$ is the OR between $a_i$ and $b_i$.

`SIZE` $(S(A))$: let $b$ the average position of the least zero bits in $a_1, \ldots, a_k$, return $2^b/.77351$.

# ANF (Palmer et al. 2002)

Palmer, Gibbons, and Faloutsos (KDD 2002) applied this paradigm in the distance distribution context.

// Set $\mathcal{M}(x,0) = \{x\}$
FOR each node $x$ DO
  $M(x,0) =$ concatenation of $k$ bitmasks
    each with 1 bit set ($P(\text{bit } i) = .5^{i+1}$)
FOR each distance $h$ starting with 1 DO
  FOR each node $x$ DO $M(x,h) = M(x,h-1)$
  // Update $\mathcal{M}(x,h)$ by adding one step
  FOR each edge $(x,y)$ DO
    $M(x,h) = (M(x,h)$ BITWISE-OR $M(y,h-1))$
  // Compute the estimates for this $h$
  FOR each node $x$ DO
    Individual estimate $I\hat{N}(x,h) = (2^b)/.77351$
    where $b$ is the average position of the least zero bits
    in the $k$ bitmasks
  The estimate is: $\hat{N}(h) = \sum_{\text{all x}} I\hat{N}(x,h)$

**Figure 2: Introduction to the basic ANF algorithm**

| $x$ | $M(x,0)$ | $M(x,1)$ | $I\hat{N}(x,1)$ | $M(x,2)$ | $I\hat{N}(x,3)$ |
|---|---|---|---|---|---|
| 0 | 100 100 001 | 110 110 101 | 4.1 | 110 111 101 | 5.2 |
| 1 | 010 100 100 | 110 101 101 | 3.25 | 110 111 101 | 5.2 |
| 2 | 100 001 100 | 110 101 100 | 3.25 | 110 111 101 | 5.2 |
| 3 | 100 100 100 | 100 111 100 | 4.1 | 110 111 101 | 5.2 |
| 4 | 100 010 100 | 100 110 101 | 3.25 | 110 111 101 | 5.2 |

**Figure 3: Simple example of *basic ANF***

# HyperLogLog (Flajolet et al. 2007) and HyperANF (Boldi et al. 2011)

Let $\mathscr{D}$ be a fixed domain and $h : \mathscr{D} \to 2^\infty$ be a hash function mapping each element of $\mathscr{D}$ into an infinite binary sequence. The function is fixed with the only assumption that "bits of hashed values are assumed to be independent and to have each probability $\frac{1}{2}$ of occurring" [FFGM07].

For a given $x \in 2^\infty$, let $h_t(x)$ denote the sequence made by the leftmost $t$ bits of $h(x)$, and $h^t(x)$ be the sequence of remaining bits of $x$; $h_t$ is identified with its corresponding integer value in the range $\{0, 1, \ldots, 2^t - 1\}$. Moreover, given a binary sequence $w$, we let $\rho^+(w)$ be the number of leading zeroes in $w$ plus one[3] (e.g., $\rho^+(00101) = 3$). Unless otherwise specified, all logarithms are

**Algorithm 1** The Hyperloglog counter as described in [FFGM07]: it allows one to count (approximately) the number of distinct elements in a stream. $\alpha_m$ is a constant whose value depends on $m$ and is provided in [FFGM07]. Some technical details have been simplified.

```
h : 𝒟 → 2^∞, a hash function from the domain of items
M[−] the counter, an array of m = 2^b registers
   (indexed from 0) and set to −∞

function add(M: counter, x: item)
begin
   i ← h_b(x);
   M[i] ← max{M[i], ρ⁺(h^b(x))}
end; // function add

function size(M: counter)
begin
   Z ← (Σ_{j=0}^{m−1} 2^{−M[j]})^{−1};
   return E = α_m m² Z
end; // function size

foreach item x seen in the stream begin
   add(M,x)
end;
print size(M)
```

- Flajolet, P.; Nigel Martin, G. (1985). Probabilistic counting algorithms for data base applications. Journal of Computer and System Sciences 31 (2): 182.
- Philippe Flajolet, Eric Fusy, Olivier Gandouet, Frederic Meunier (2007). Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm. Discrete Mathematics and Theoretical Computer Science: 127146
- Christopher R. Palmer, Phillip B. Gibbons, Christos Faloutsos: ANF: a fast and scalable tool for data mining in massive graphs. KDD 2002: 81-90
- Kane, D. M.; Nelson, J.; Woodruff, D. P. (2010). An optimal algorithm for the distinct elements problem. Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems of data - PODS '10. p. 41.
- Paolo Boldi, Marco Rosa, Sebastiano Vigna: HyperANF: approximating the neighbourhood function of very large graphs on a budget. WWW 2011: 625-634
- Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, Sebastiano Vigna: Four degrees of separation. WebSci 2012: 33-42

Part II

# More about Sketches

- *Min-Hash sketches* are a family of statistical technique to represent approximately in little space a subset of items of a universe $U$.
- Cohen (2014) systematizes the description of Min-Hash sketches dividing them in three classes, $k$-min, bottom-$k$, $k$-partition, where the parameter $k$ determines the sketch size.
- Some of these techniques were devised to count the distinct elements of a stream, exploiting the property that by merging the sketches of two streams we can obtain an estimate of their concatenation.
- Some others, like $k$-min (Broder 1998) were focused on the estimation of the Jaccard index.

# Jaccard coefficient

Given two sets $A$ and $B$, their Jaccard index (also called Jaccard similarity coefficient) is defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

- It is a very commonly used measure of similarity, ranging from 0 (for disjoint sets) to 1 (for identical sets).
- The previous techniques can be used or adapted to estimate Jaccard similarity.
- It has a number of applications, for example in information retrieval to estimate document similarity (Broder 1997).

A. Broder. On the Resemblance and Containment of Documents. SEQUENCES '97 Proceedings of the Compression and Complexity of Sequences 1997.

# Notations and Operations

Given:

- a *ranking* (i.e., a bijective function) $r : U \rightarrow \{1, 2, \ldots, n\}$
- and a subset $A$ of $U$,

we denote

- with $H_k(A)$ the first $k$ elements of $A$ according to $r$ (all the elements of $A$, if $|A| < k$), and
- with $k_{th}(A)$ the ranking of the $k$-th element of $A$ ($k_{th}(A) = n$, if $|A| < k$), and
- with $\max(A)$ the element having maximum rank in $A$.

For each method we will revise and specify:

`INIT` $(S(A))$ How a sketch $S(A)$ for $A$ is initialized.

`UPDATE` $(S(A), u)$ How a sketch $S(A)$ for $A$ modifies when an element $u$ is added to $A$.

`ESTIMATE` $(S(A), S(B))$ given two sketches, $S(A)$ and $S(B)$, how the Jaccard Index is estimated.

# Threshold Sampling

Given

- a rank $r : U \to \{1, \ldots, n\}$ and
- a threshold $t \in [1, n]$,

the sketch is the set of all the elements $a \in A$ satisfying $r(a) < t$.

- The size of the sketch cannot predicted in advance; for this reason, the other sketches are usually preferred in the applications.

`INIT` $(S(A))$: $S(A)$ is the empty set.

`UPDATE` $(S(A), u)$: If $r(u) < t$ add $u$ to $S(A)$.

`ESTIMATE` $(S(A), S(B))$: The estimate is $J(S(A), S(B))$.

# bottom-$k$ approach (Cohen & Kaplan 2007)

- A bottom-$k$ sketch includes the $k$ items with smallest rank in a single permutation, that is, $H_k(A)$.
- Its application in the context of Jaccard estimation similarity using hash has been studied by Thorup (STOC 2013).
- In this case a sketch is simply a subset of the set.

## bottom-$k$ approach

Given a rank $r : U \to \{1, 2, \ldots, n\}$ (a bijection):

INIT ($S(A)$): $S(A)$ is the empty set.

UPDATE ($S(A), u$): If $|S(A)| < k$, add $u$ to $S(A)$. Otherwise, if $r(\max(S(A))) > r(u)$, replace $\max(S(A))$ with $u$.

ESTIMATE ($S(A), S(B)$): The estimate is

$$\frac{|H_k(A \cup B) \cap H_k(A) \cap H_k(B)|}{k} = \frac{|H_k(S(A) \cup S(B)) \cap S(A) \cap S(B)|}{k}.$$

# A similar process with balls and bins

- Balls and Bins with $R$ red, $W$ white, and $B$ bicolor. Let $N$ be $R + W + B$.
- Sample $k$ balls without replacement.
- Let $X$ be the number of balls bicolor sampled
- Approximate the quantity $\frac{B}{R+W+B}$ as $\frac{X}{k}$.

**Theorem**

$$\mathbb{E}\left[\frac{X}{k}\right] = \frac{B}{R + W + B} = \frac{B}{N}$$

## Hypergeometric Distribution

- Distribution that describes the probability of $h$ successes in $n$ draws, without replacement, from a finite population of size $N$ containing exactly $H$ successes.
- In contrast, the binomial distribution describes the probability of $k$ successes in $n$ draws with replacement.

$$Pr[X = h] = \frac{\binom{H}{h}\binom{N-H}{n-h}}{\binom{N}{n}}$$

The mean is $n \cdot H/N$

- Edith Cohen, Haim Kaplan: Summarizing data using bottom-k sketches. PODC 2007: 225-234
- Edith Cohen, Haim Kaplan: Bottom-k sketches: better and more efficient estimation of aggregates. SIGMETRICS 2007: 353-354
- Edith Cohen, Haim Kaplan: Tighter estimation using bottom k sketches. PVLDB 1(1): 213-224 (2008)
- Edith Cohen, Haim Kaplan: Leveraging discarded samples for tighter estimation of multiple-set aggregates. SIGMETRICS/Performance 2009: 251-262
- Mikkel Thorup: Bottom-k and priority sampling, set similarity and subset sums with minimal independence. STOC 2013: 371-380

# $k$-min approach (Broder 1997)

- A $k$-min sketch includes the item of smallest rank in each of $k$ independent permutations.
- In this case a sketch $S(A)$ is a sequence of exactly $k$ entries, $a_1, \ldots, a_k$, where each entry can be an element of $A$ or $\perp$.

Introduced for cardinality estimation in (Cohen 1994). Applied for Jaccard similarity estimation by (Broder 1997) and studied in (Broder et al. 1998).


- Edith Cohen: Estimating the Size of the Transitive Closure in Linear Time. FOCS 1994: 190-200
- Edith Cohen: Size-Estimation Framework with Applications to Transitive Closure and Reachability. J. Comput. Syst. Sci. 55(3): 441-453 (1997)
- A. Broder. On the Resemblance and Containment of Documents. Proceeding SEQUENCES '97 Proceedings of the Compression and Complexity of Sequences 1997.
- Andrei Z. Broder, Moses Charikar, Alan M. Frieze, Michael Mitzenmacher: Min-Wise Independent Permutations (Extended Abstract). STOC 1998: 327-336

# $k$-min approach

Let $r_1, r_2, \ldots, r_k$ be ranks $r_i : U \to \{1/n, 2/n, 3/n, \ldots, 1\}$, setting $r(\perp) = \infty$.

`INIT` $(S(A))$: $a_i = \perp$ for any $i$.

`UPDATE` $(S(A), u)$: for every $i$, with $1 \leq i \leq k$, if $r_i(a_i) > r_i(u)$, $a_i$ is replaced with $u$.

`ESTIMATE` $(S(A), S(B))$: the estimate is $|\{j \mid a_j = b_j\}|/k$.

$k$-min can be seen as $k$ independent bottom-1 sketches defined with respect to $k$ independent permutations.

- $k$-min corresponds to sampling with replacement.
- bottom-$k$ corresponds to sampling without replacement.

$\Longrightarrow$ bottom-$k$ has less variance than $k$-min.

# $k$-partition (Ping Li et al. 2012)

A $k$-partition sketch first maps items uniformly at random to $k$ buckets and then it includes the item with smallest rank in each bucket.

- Similar to the Flajolet-Martin paradigm, reused in the context of Jaccard estimation similarity in (Li et al. 2012).
- A sketch $S(A)$ is a sequence of exactly $k$ entries, $a_1, \ldots, a_k$, where each entry can be an element of $A$ or $\perp$.

# $k$-partition

Given a partition function $p : U \to \{1, 2, \ldots, k\}$ and a ranking $r : U \to \{1, 2, \ldots, n\}$, setting $r(\perp) = \infty$:

INIT $(S(A))$: $a_i = \perp$ for any $i$.

UPDATE $(S(A), u)$: if $r(u) < r(a_{p(u)})$, then $a_{p(u)}$ is replaced by $u$ in $S(A)$.

ESTIMATE $(S(A), S(B))$: ratio between $|\{j \mid a_j = b_j \neq \perp\}|$ and $k - |\{j \mid a_j = b_j = \perp\}|$.

Ping Li and Arnd Christian Konig. b-bit minwise hashing. In WWW, pages 671-680, 2010.

Ping Li, Art B. Owen, and Cun-Hui Zhang. One permutation hashing. In NIPS, pages 3122-3130, 2012.

- Michael Mitzenmacher, Rasmus Pagh, and Ninh Pham. Efficient estimation for high similarities using odd sketches. In WWW, pages 109-118, 2014.
- Edith Cohen. All-distances sketches, revisited: Hip estimators for massive graphs analysis. PODC, 2014.

# Part III

# Applications for Local Triangle Counting

Given an edge $u, v$ the number of triangles involving this edge is $N(u) \cap N(v)$.
The number of triangles involving a node $u$ is

$$\sum_{v \in N(u)} |N(u) \cap N(v)|$$

Since

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

then

$$|N(u) \cap N(v)| = \frac{(d(u) + d(v)) \cdot J(N(u), N(v))}{J(N(u), N(v)) + 1}$$

Becchetti et al. (TKDD 2010) estimate $J(N(u), N(v))$ by using $k$-min and use the formula above to approximate $|N(u) \cap N(v)|$.

Luca Becchetti, Paolo Boldi, Carlos Castillo, and Aristides Gionis. Efficient algorithms for large-scale local triangle counting. TKDD, 4(3), 2010.

# Current line of research

**Approximating Cardinality of Set Intersection**

Rasmus Pagh, Morten Stoockel, and David P. Woodruff. Is Min-Wise Hashing Optimal for Summarizing Set Intersection?. PODS 2014.