Bottom-k and Priority Sampling, Set Similarity and Subset Sums with Minimal Independence

Mikkel Thorup

University of Copenhagen

Min-wise hashing [Broder, '98, Alta Vita]

Jaccard similary of sets A and B is $f(A, B) = |A \cap B|/|A \cup B|$.

With truly random hash function h

$$f(A, B) \quad = \quad \Pr_h [\, \min h(A) = \min h(B) \,]$$

Min-wise hashing [Broder, '98, Alta Vita]

Jaccard similary of sets A and B is $f(A,B) = |A \cap B|/|A \cup B|$.

With truly random hash function h

$$f(A,B) \quad = \quad \Pr_h [\, \min h(A) = \min h(B) \,]$$

For concentration, repeat k times:

k-Mins use k indepedent hash functions $h_1, ..., h_k$.
For set A store signature $M^k(A) = (\min h_1(A), .., \min h_k(A))$.

Min-wise hashing [Broder, '98, Alta Vita]

Jaccard similary of sets A and B is $f(A,B) = |A \cap B|/|A \cup B|$.

With truly random hash function h

$$f(A,B) \quad = \quad \Pr_h [\, \min h(A) = \min h(B) \,]$$

For concentration, repeat k times:

k-Mins use k indepedent hash functions $h_1, ..., h_k$.
For set A store signature $M^k(A) = (\min h_1(A), .., \min h_k(A))$.

To estimate Jaccard similariy $f(A,B)$ of A and B, use

$$|M^k(A) \cap M^k(B)|/k = \sum_{i=1}^{k} [\min h_i(A) = \min h_i(B)]/k$$

Min-wise hashing [Broder, '98, Alta Vita]

Jaccard similary of sets A and B is $f(A,B) = |A \cap B|/|A \cup B|$.

With truly random hash function h

$$f(A,B) \quad = \quad \Pr_h [\, \min h(A) = \min h(B) \,]$$

For concentration, repeat k times:

k-Mins use k indepedent hash functions $h_1, ..., h_k$.
For set A store signature $M^k(A) = (\min h_1(A), .., \min h_k(A))$.

To estimate Jaccard similariy $f(A,B)$ of A and B, use

$$|M^k(A) \cap M^k(B)|/k = \sum_{i=1}^{k} [\min h_i(A) = \min h_i(B)]/k$$

Expected relative error below $1/\sqrt{f(A,B) \cdot k}$.

Bias issues

We do not have space for truly random hash functions.

Bias issues

We do not have space for truly random hash functions.

We say h is ε-minwise independent if for any S, $x \in S$,:

$$\Pr_{h \in \mathcal{H}}[h(x) = \min h(S)] = \frac{1 \pm \varepsilon}{|S|}$$

Bias issues

We do not have space for truly random hash functions.

We say h is ε-minwise independent if for any S, $x \in S$,:

$$\Pr_{h \in \mathcal{H}}[h(x) = \min h(S)] = \frac{1 \pm \varepsilon}{|S|}$$

Such bias ε takes independence $\Theta\left(\lg \frac{1}{\varepsilon}\right)$
[Indyk'99, Patrascu Thorup '10].

Bias issues

We do not have space for truly random hash functions.

We say h is ε-minwise independent if for any S, $x \in S$,:

$$\Pr_{h \in \mathcal{H}}[h(x) = \min h(S)] = \frac{1 \pm \varepsilon}{|S|}$$

Such bias ε takes independence $\Theta\left(\lg \frac{1}{\varepsilon}\right)$
[Indyk'99, Patrascu Thorup '10].

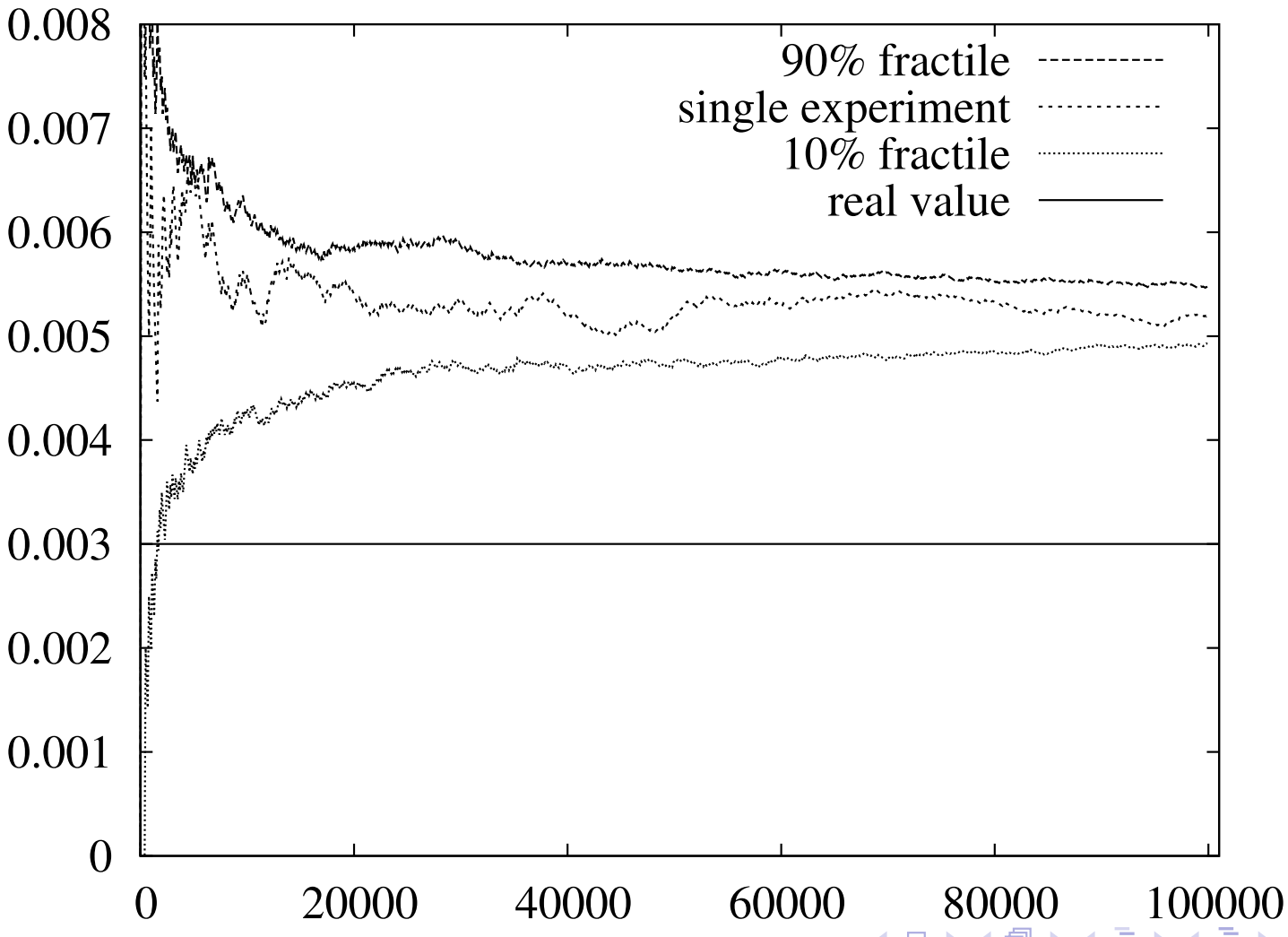
Bias ε not improved by k-mins no matter repetitions k.

Practice

Mitzenmacher and Vadhan [SODA'08]: with enough entropy, 2-independent works as good as random, but

Practice

Mitzenmacher and Vadhan [SODA'08]: with enough entropy, 2-independent works as good as random, but real world full of low-entroy data, e.g., consecutive numbers:



Bottom-k

With single hash function h, the bottom-k signature of set X is

$$S_k(X) = \{k \text{ elements of } X \text{ with smallest hash values}\}$$

Bottom-k

With single hash function h, the bottom-k signature of set X is

$$S_k(X) = \{k \text{ elements of } X \text{ with smallest hash values}\}$$

For subset $Y \subseteq X$, estimate frequency $f(Y, X) = |Y|/|X|$ as

$$|S_k(X) \cap Y|/k$$

Bottom-k

With single hash function h, the bottom-k signature of set X is

$$S_k(X) = \{k \text{ elements of } X \text{ with smallest hash values}\}$$

For subset $Y \subseteq X$, estimate frequency $f(Y, X) = |Y|/|X|$ as

$$|S_k(X) \cap Y|/k$$

For sets A and B, the signature of union $A \cup B$ computed as

$$S_k(A \cup B) = S_k(S_k(A) \cup S_k(B))$$

and Jaccard similarity $f(A, B)$ is estimated as

$$|S_k(A \cup B) \cap S_k(A) \cap S_k(B)|/k$$

Bottom-k

With single hash function h, the bottom-k signature of set X is

$$S_k(X) = \{k \text{ elements of } X \text{ with smallest hash values}\}$$

For subset $Y \subseteq X$, estimate frequency $f(Y, X) = |Y|/|X|$ as

$$|S_k(X) \cap Y|/k$$

For sets A and B, the signature of union $A \cup B$ computed as

$$S_k(A \cup B) = S_k(S_k(A) \cup S_k(B))$$

and Jaccard similarity $f(A, B)$ is estimated as

$$|S_k(A \cup B) \cap S_k(A) \cap S_k(B)|/k$$

Our result:

Theorem If h is 2-independent, even including bias, the expected relative error is $O(1/\sqrt{f \cdot k})$.

Bottom-k

With single hash function h, the bottom-k signature of set X is

$$S_k(X) = \{k \text{ elements of } X \text{ with smallest hash values}\}$$

For subset $Y \subseteq X$, estimate frequency $f(Y, X) = |Y|/|X|$ as

$$|S_k(X) \cap Y|/k$$

For sets A and B, the signature of union $A \cup B$ computed as

$$S_k(A \cup B) = S_k(S_k(A) \cup S_k(B))$$

and Jaccard similarity $f(A, B)$ is estimated as

$$|S_k(A \cup B) \cap S_k(A) \cap S_k(B)|/k$$

Our result:

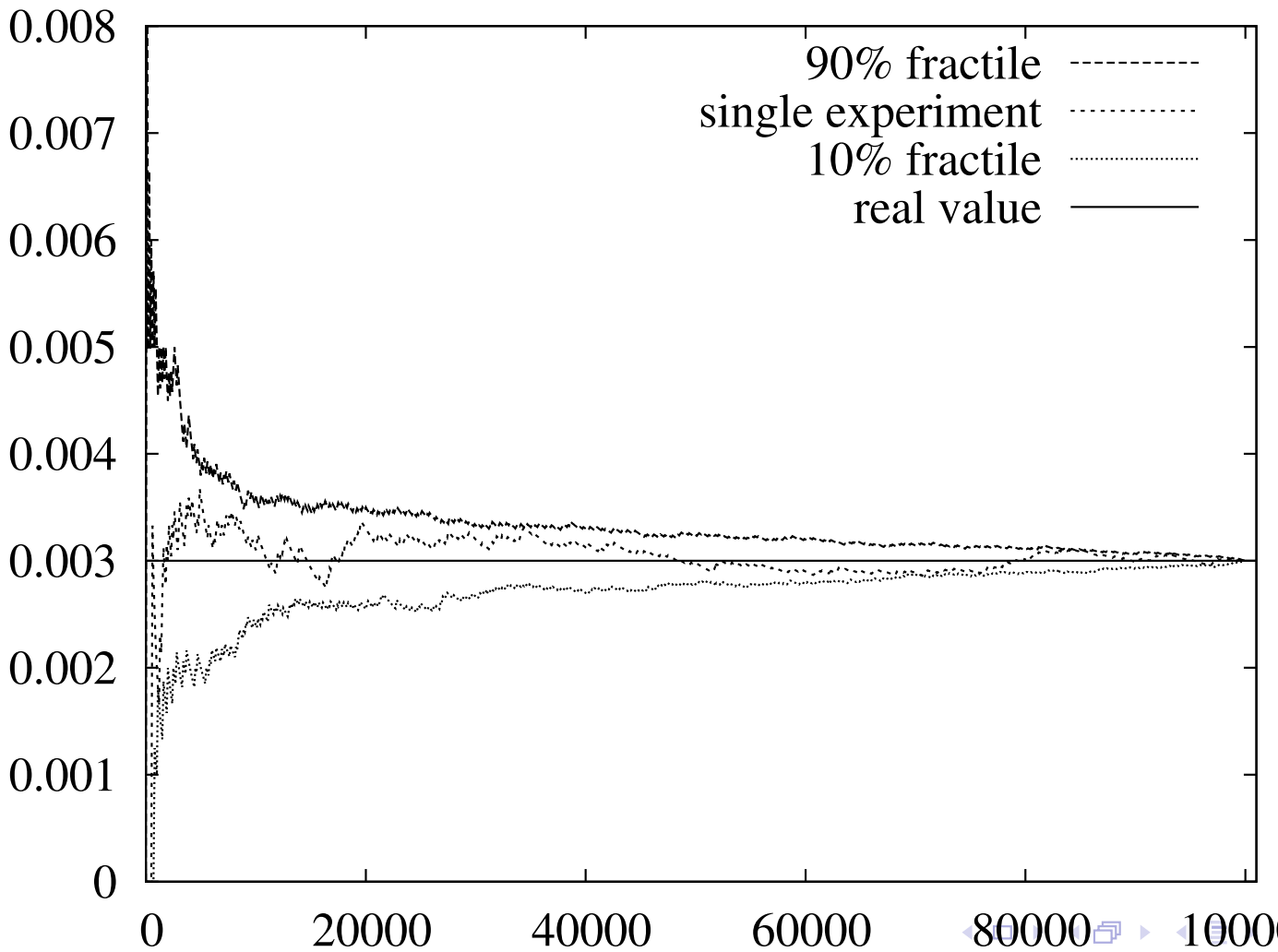
Theorem If h is 2-independent, even including bias, the expected relative error is $O(1/\sqrt{f \cdot k})$.

Porat proved this for 8-independent hashing and $k \gg 1$.

Practice

We can use Dietzfelbinger's super fast 2-independent hashing

```
random int64 a,b;
h(int32 x) return (a*x + b) >> 32;
```

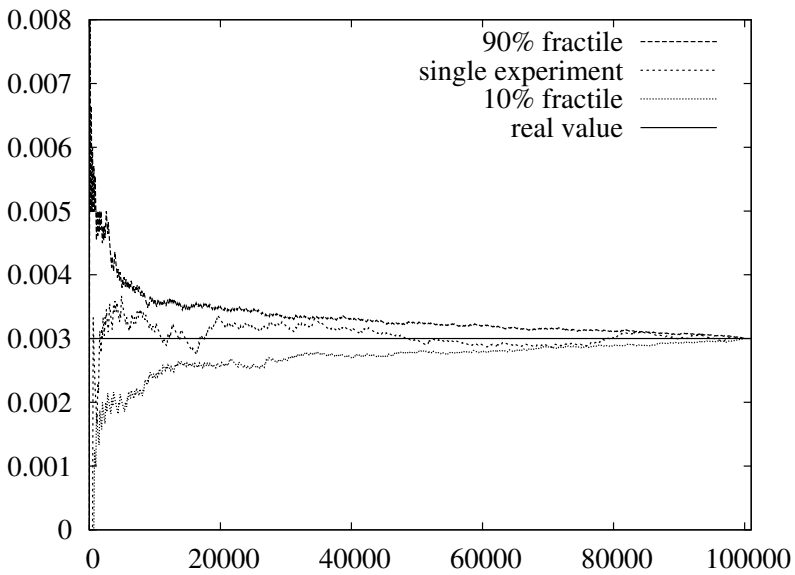


Practice

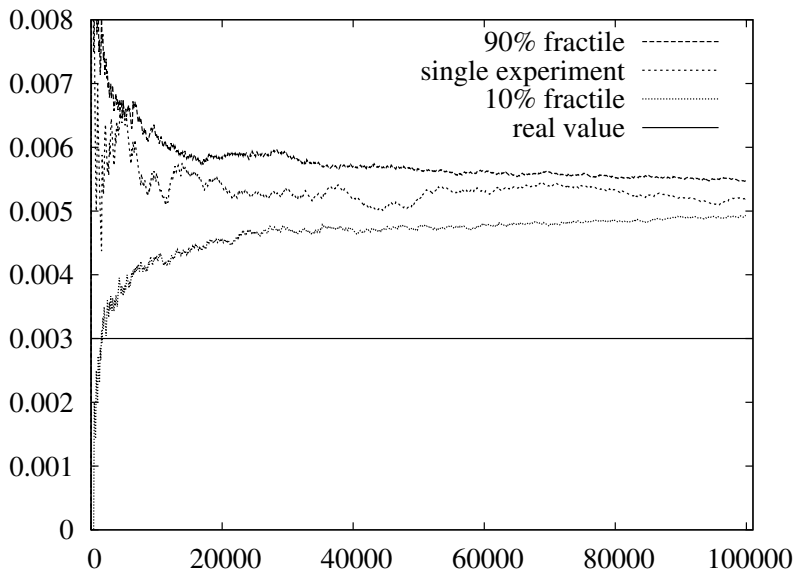
We can use Dietzfelbinger's super fast 2-independent hashing

```
random int64 a,b;
h(int32 x) return (a*x + b) >> 32;
```

bottom-k



k-mins



A Simple Union Bound

With $n = |X|$, $S = S_k(X)$, $Y \subseteq X$, estimate $f = |Y|/|X|$ as

$$|Y \cap S|/k$$

Overestimate: For parameters $a < 1$, b, bound probability

$(*)$ $$|Y \cap S| > \frac{1+b}{1-a} f k.$$

A Simple Union Bound

With $n = |X|$, $S = S_k(X)$, $Y \subseteq X$, estimate $f = |Y|/|X|$ as

$$|Y \cap S|/k$$

Overestimate: For parameters $a < 1$, b, bound probability

$$(*) \qquad |Y \cap S| > \tfrac{1+b}{1-a} f k.$$

Define threshold probability (independent of random choices)

$$p = \frac{k}{n(1-a)}.$$

Hash values uniform in $(0, 1)$ so $\Pr[h(x) < p] = p$.

A Simple Union Bound

With $n = |X|$, $S = S_k(X)$, $Y \subseteq X$, estimate $f = |Y|/|X|$ as

$$|Y \cap S|/k$$

Overestimate: For parameters $a < 1$, b, bound probability

$$(*) \qquad |Y \cap S| > \tfrac{1+b}{1-a}\, f\, k.$$

Define threshold probability (independent of random choices)

$$p = \frac{k}{n(1-a)}.$$

Hash values uniform in $(0, 1)$ so $\Pr[h(x) < p] = p$.

Proposition The overestimate $(*)$ implies one of

$$\text{(A)} \qquad |\{x \in X | h(x) < p\}| < k$$
$$\text{(B)} \qquad |\{x \in Y | h(x) < p\}| > (1+b)\, p |Y|.$$

so $\Pr[(*)] \leq \Pr[(\text{A})] + \Pr[(\text{B})]$.

A Simple Union Bound

With $n = |X|$, $S = S_k(X)$, $Y \subseteq X$, estimate $f = |Y|/|X|$ as

$$|Y \cap S|/k$$

Overestimate: For parameters $a < 1$, b, bound probability

$$(*) \qquad |Y \cap S| > \tfrac{1+b}{1-a}\, f\, k.$$

Define threshold probability (independent of random choices)

$$p = \frac{k}{n(1-a)}.$$

Hash values uniform in $(0,1)$ so $\Pr[h(x) < p] = p$.

Proposition The overestimate $(*)$ implies one of

(A) $|\{x \in X | h(x) < p\}| < k$ — $\mathbf{E}[\cdot] = k/(1-a)$

(B) $|\{x \in Y | h(x) < p\}| > (1+b)\, p |Y|$. — $\mathbf{E}[\cdot] = p|Y|$.

so $\Pr[(*)] \leq \Pr[(A)] + \Pr[(B)]$.

Proof: $\neg(A) \wedge \neg(B) \implies \neg(*)$

Suppose (A) is false, that is,

$$|\{x \in X | h(x) < p\}| \geq k.$$

Proof: $\neg(A) \wedge \neg(B) \implies \neg(*)$

Suppose (A) is false, that is,

$$|\{x \in X | h(x) < p\}| \geq k.$$

Then, since S contains k smallest,

$$\forall x \in S : h(x) < p.$$

so

$$Y \cap S \subseteq \{x \in Y | h(x) < p\}$$

Proof: $\neg(A) \wedge \neg(B) \implies \neg(*)$

Suppose (A) is false, that is,

$$|\{x \in X | h(x) < p\}| \geq k.$$

Then, since S contains k smallest,

$$\forall x \in S : h(x) < p.$$

so

$$Y \cap S \subseteq \{x \in Y | h(x) < p\}$$

Suppose (B) is also false, that is,

$$|\{x \in Y | h(x) < p\}| < (1 + b)\, p\, |Y|$$

Proof: $\neg(A) \wedge \neg(B) \Longrightarrow \neg(*)$

Suppose (A) is false, that is,

$$|\{x \in X | h(x) < p\}| \geq k.$$

Then, since S contains k smallest,

$$\forall x \in S : h(x) < p.$$

so

$$Y \cap S \subseteq \{x \in Y | h(x) < p\}$$

Suppose (B) is also false, that is,

$$|\{x \in Y | h(x) < p\}| < (1 + b)\, p\, |Y|$$

With $p = k/(n(1 - a))$ and $|Y| = f\, n$, we get

$$|\{x \in Y | h(x) < p\}| < \frac{1 + b}{1 - a} f\, k$$

so the overestimate $(*)$ did not happen.

2-independence

Proposition With $p = \frac{k}{n(1-a)}$, the overestimate

$$(*) \qquad |Y \cap S| > \frac{1+b}{1-a} f k.$$

implies one of

$$(A) \qquad |\{x \in X | h(x) < p\}| < k \qquad \text{— } \mathbf{E}[\cdot] = k/(1-a)$$

$$(B) \qquad |\{x \in Y | h(x) < p\}| > (1+b)\, p |Y| \qquad \text{— } \mathbf{E}[\cdot] = p|Y|.$$

so $\Pr[(*)] \leq \Pr[(A)] + \Pr[(B)]$.

2-independence

Proposition With $p = \frac{k}{n(1-a)}$, the overestimate

$$(*) \qquad |Y \cap S| > \frac{1+b}{1-a} f k.$$

implies one of

$$(A) \qquad |\{x \in X | h(x) < p\}| < k \qquad \text{— } \mathbf{E}[\cdot] = k/(1-a)$$
$$(B) \qquad |\{x \in Y | h(x) < p\}| > (1+b)\, p |Y| \qquad \text{— } \mathbf{E}[\cdot] = p|Y|.$$

so $\Pr[(*)] \leq \Pr[(A)] + \Pr[(B)]$.

Application If h is 2-independent, by Chebyshev,

$$\begin{aligned} \Pr[(A)] &\leq 1/(a^2 k) \\ \Pr[(B)] &\leq 1/(b^2 f k) \end{aligned}$$

2-independence

Proposition With $p = \frac{k}{n(1-a)}$, the overestimate

$$(*) \qquad |Y \cap S| > \frac{1+b}{1-a} f k.$$

implies one of

$$\text{(A)} \qquad |\{x \in X | h(x) < p\}| < k \qquad \text{— } \mathbf{E}[\cdot] = k/(1-a)$$
$$\text{(B)} \qquad |\{x \in Y | h(x) < p\}| > (1+b)\, p |Y| \quad \text{— } \mathbf{E}[\cdot] = p|Y|.$$

so $\Pr[(*)] \leq \Pr[(\text{A})] + \Pr[(\text{B})]$.

Application If h is 2-independent, by Chebyshev,

$$\begin{aligned} \Pr[(A)] &\leq 1/(a^2 k) \\ \Pr[(B)] &\leq 1/(b^2 f k) \end{aligned}$$

For any $\varepsilon < 1/3$, appropriate choice of a and b yields

$$\Pr\left[|Y \cap S| > (1+\varepsilon) f k\right] \leq \frac{4}{\varepsilon^2 f k}.$$

2-independence

Proposition With $p = \frac{k}{n(1-a)}$, the overestimate

$$(*) \qquad |Y \cap S| > \frac{1+b}{1-a} f k.$$

implies one of

$$(A) \qquad |\{x \in X | h(x) < p\}| < k \qquad \text{— } \mathbf{E}[\cdot] = k/(1-a)$$
$$(B) \qquad |\{x \in Y | h(x) < p\}| > (1+b)\, p |Y| \quad \text{— } \mathbf{E}[\cdot] = p|Y|.$$

so $\Pr[(*)] \leq \Pr[(A)] + \Pr[(B)]$.

Application If h is 2-independent, by Chebyshev,

$$\begin{aligned} \Pr[(A)] &\leq 1/(a^2 k) \\ \Pr[(B)] &\leq 1/(b^2 f k) \end{aligned}$$

For any $\varepsilon < 1/3$, appropriate choice of a and b yields

$$\Pr\left[|Y \cap S| > (1+\varepsilon) f k\right] \leq \frac{4}{\varepsilon^2 f k}.$$

With more calcluations

$$E[||Y \cap S| - f k|] = O(\sqrt{f \cdot k}).$$

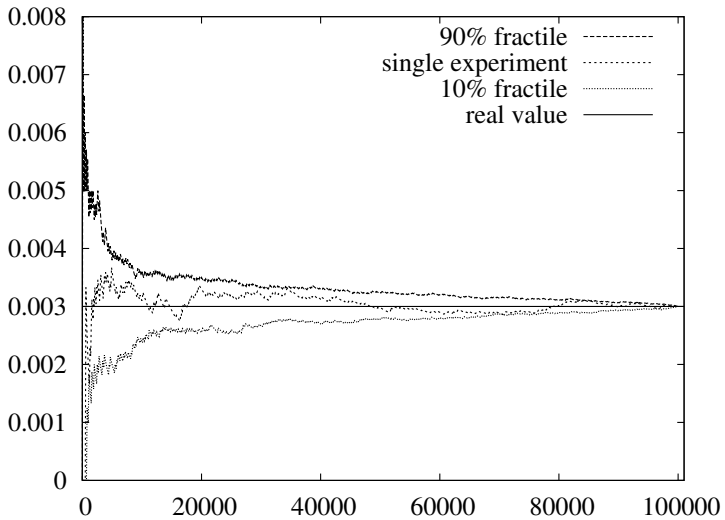
Conclusion

- Both k-mins and bottom-k generalize the idea of storing the smallest hash value.

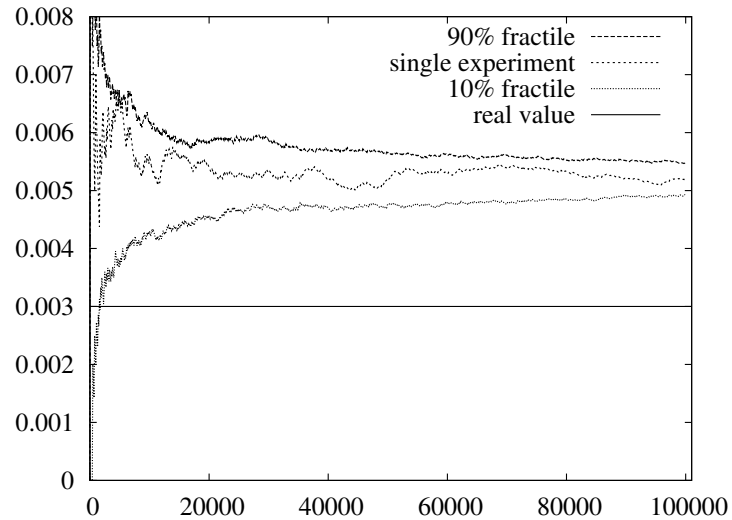
Conclusion

- Both k-mins and bottom-k generalize the idea of storing the smallest hash value.
- With limited independence
  k-mins has major problems with bias whereas bottom-k works perfectly even with 2-indepedence.

bottom-k



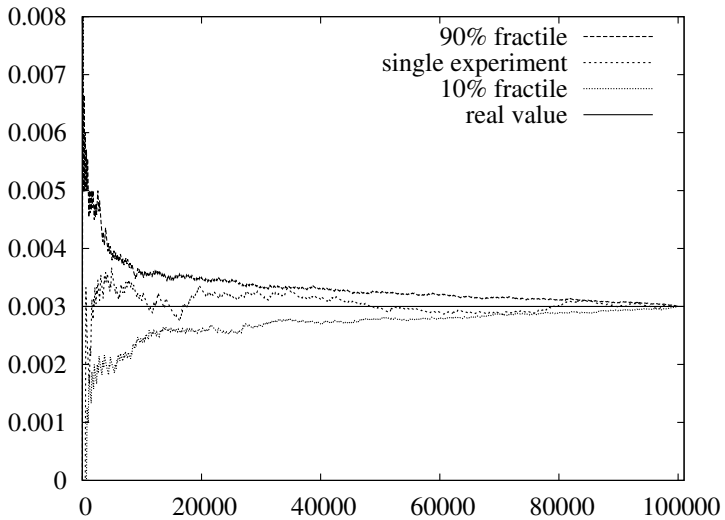
k-mins



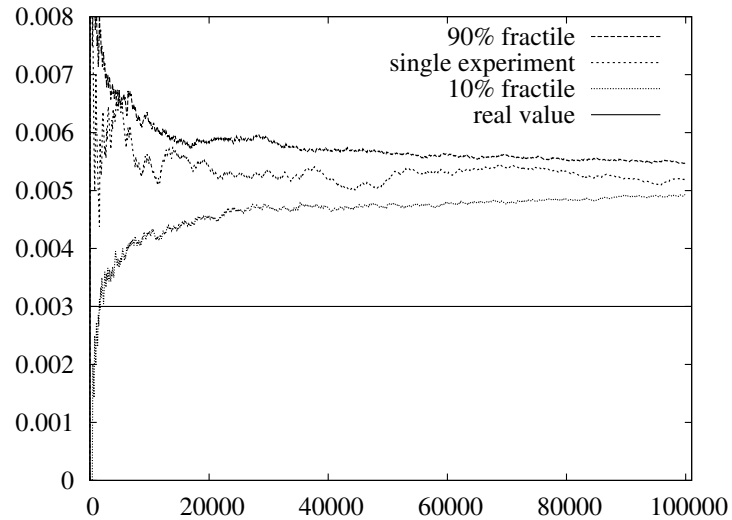
Conclusion

- Both k-mins and bottom-k generalize the idea of storing the smallest hash value.
- With limited independence
  k-mins has major problems with bias whereas bottom-k works perfectly even with 2-indepedence.

bottom-k



k-mins

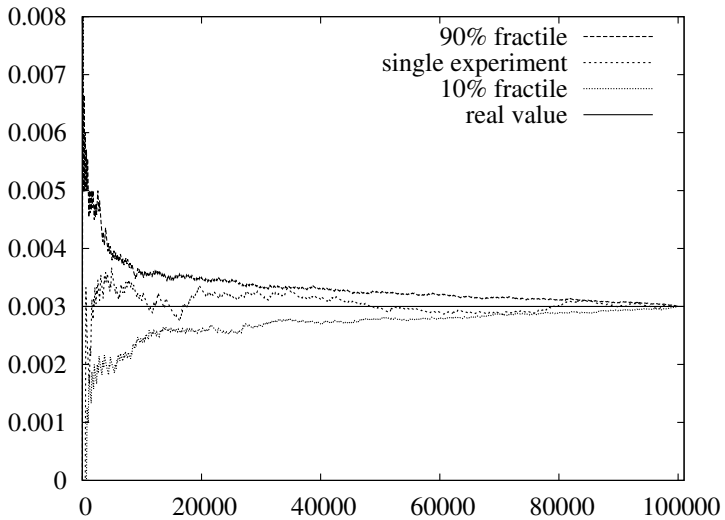


- Technically the proof is very simple.

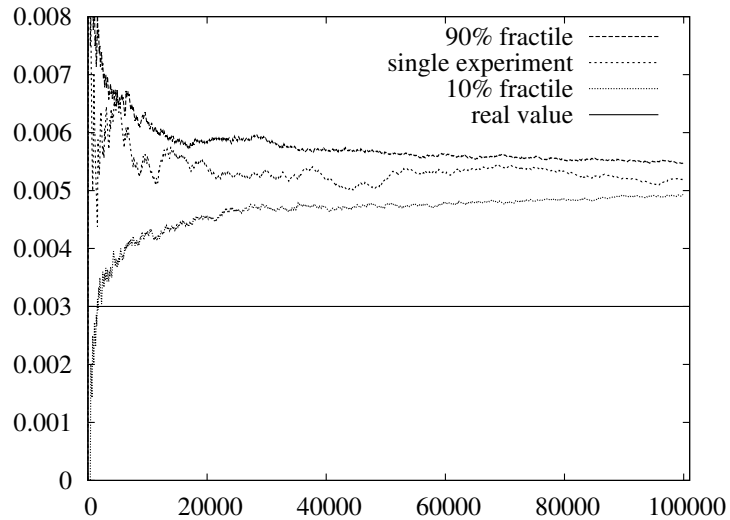
Conclusion

- Both k-mins and bottom-k generalize the idea of storing the smallest hash value.
- With limited independence
  k-mins has major problems with bias whereas bottom-k works perfectly even with 2-indepedence.

bottom-k



k-mins



- Technically the proof is very simple.
- Similar results hold for weighted sampling but proofs much much harder.

Priority sampling of weighted items

A bottom/top-k sampling scheme for weighted items
[Duffield, Lund, Thorup SIGMETRICS'04, JACM'07]

Stream of items i,

- each with a weight w_i
- uniformly random hash $h(i) \in (0, 1)$
- *priority* $q_i = w_i/h(i)$ (assumed distinct).

Priority sampling of weighted items

A bottom/top-k sampling scheme for weighted items
[Duffield, Lund, Thorup SIGMETRICS'04, JACM'07]

Stream of items i,

- each with a weight w_i
- uniformly random hash $h(i) \in (0, 1)$
- *priority* $q_i = w_i/h(i)$ (assumed distinct).

Priority sample S of size k

- contains k items of highest priority.
- threshold τ is the $(k+1)^{th}$ priority. Then $i \in S \iff q_i > \tau$.
- weight estimate $\hat{w}_i = \max\{w_i, \tau\}$ if $i \in S$; 0 otherwise

Priority sampling of weighted items

A bottom/top-k sampling scheme for weighted items
[Duffield, Lund, Thorup SIGMETRICS'04, JACM'07]

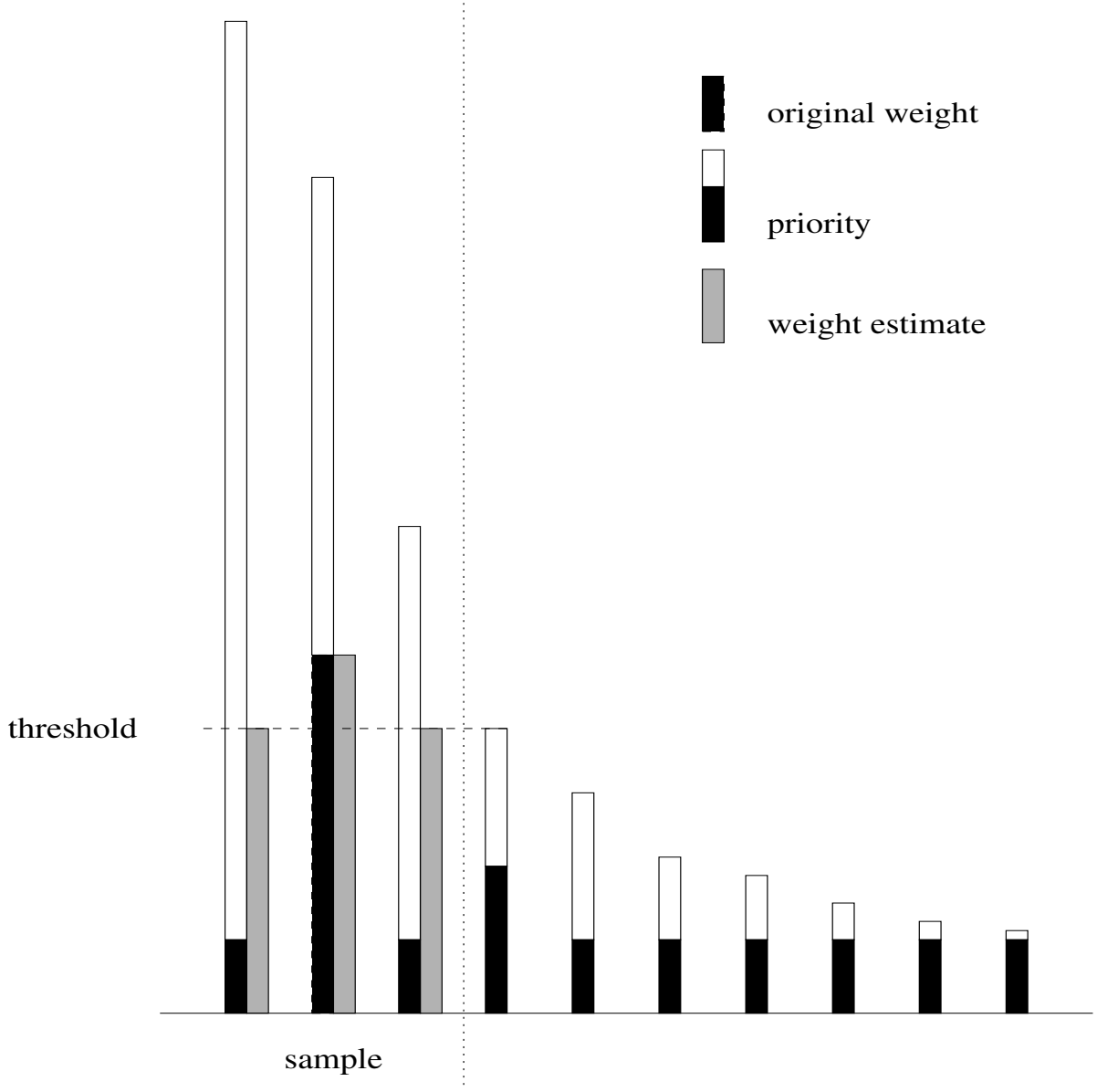
Stream of items i,

- each with a weight w_i
- uniformly random hash $h(i) \in (0, 1)$
- *priority* $q_i = w_i/h(i)$ (assumed distinct).

Priority sample S of size k (maintained in priority queue)

- contains k items of highest priority.
- threshold τ is the $(k+1)^{th}$ priority. Then $i \in S \iff q_i > \tau$.
- weight estimate $\hat{w}_i = \max\{w_i, \tau\}$ if $i \in S$; 0 otherwise

Priority sampling 3 out of 10 weighted items



Priority sampling of weighted items

A bottom/top-k sampling scheme for weighted items
[Duffield, Lund, Thorup SIGMETRICS'04, JACM'07]

Stream of items i,

- each with a weight w_i
- uniformly random hash $h(i) \in (0, 1)$
- *priority* $q_i = w_i/h(i)$ (assumed distinct).

Priority sample S of size k (maintained in priority queue)

- contains k items of highest priority.
- threshold τ is the $(k+1)^{th}$ priority. Then $i \in S \iff q_i > \tau$.
- weight estimate $\hat{w}_i = \max\{w_i, \tau\}$ if $i \in S$; 0 otherwise

Priority sampling of weighted items

A bottom/top-k sampling scheme for weighted items
[Duffield, Lund, Thorup SIGMETRICS'04, JACM'07]

Stream of items i,

- each with a weight w_i
- uniformly random hash $h(i) \in (0, 1)$
- *priority* $q_i = w_i/h(i)$ (assumed distinct).

Priority sample S of size k (maintained in priority queue)

- contains k items of highest priority.
- threshold τ is the $(k+1)^{th}$ priority. Then $i \in S \iff q_i > \tau$.
- weight estimate $\hat{w}_i = \max\{w_i, \tau\}$ if $i \in S$; 0 otherwise

Theorem [DLT'04] Priority sampling is unbiased: $\mathbf{E}[\hat{w}_i] = w_i$.

Priority sampling of weighted items

A bottom/top-k sampling scheme for weighted items
[Duffield, Lund, Thorup SIGMETRICS'04, JACM'07]

Stream of items i,

- each with a weight w_i
- uniformly random hash $h(i) \in (0,1)$
- *priority* $q_i = w_i/h(i)$ (assumed distinct).

Priority sample S of size k (maintained in priority queue)

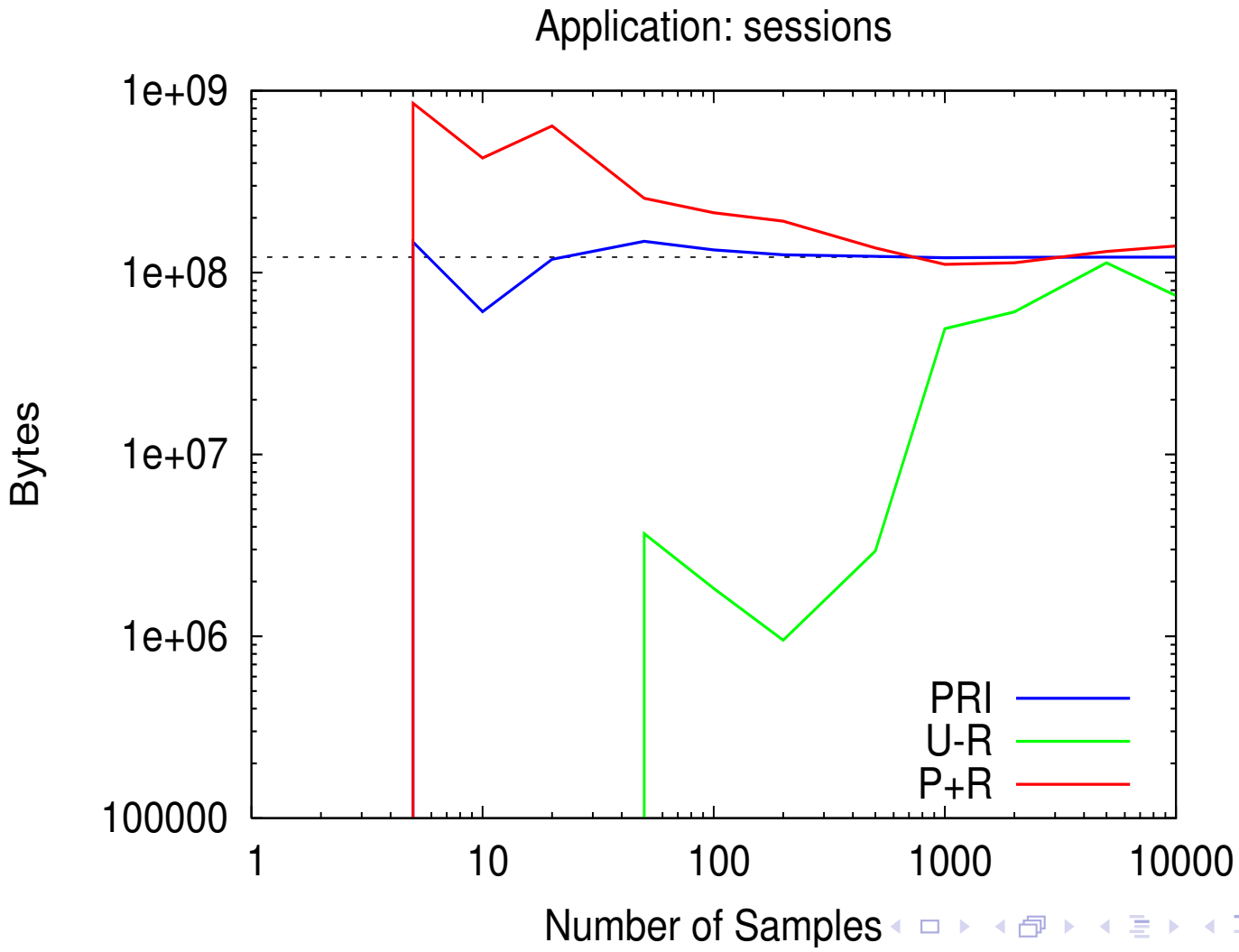
- contains k items of highest priority.
- threshold τ is the $(k+1)^{th}$ priority. Then $i \in S \iff q_i > \tau$.
- weight estimate $\hat{w}_i = \max\{w_i, \tau\}$ if $i \in S$; 0 otherwise

Theorem [DLT'04] Priority sampling is unbiased: $\mathbf{E}[\hat{w}_i] = w_i$.

Theorem [Szegey STOC'06] For any set of input weights, with one extra sample, priority sampling has smaller variance sum $\sum_i \mathrm{Var}[\hat{w}_i]$ than all other unbiased sampling schemes.

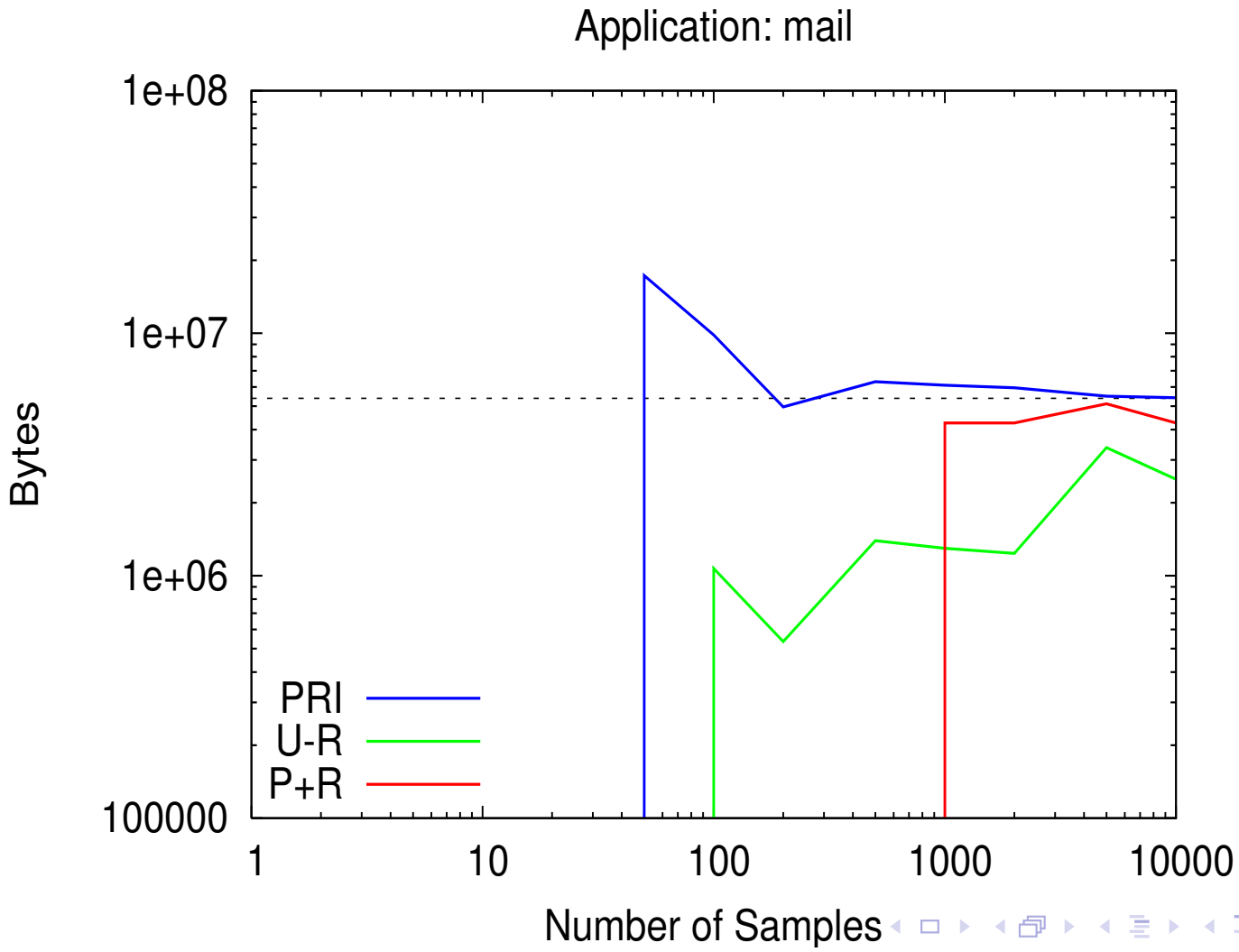
Sampled Internet traffic. Estimate traffic classes

- priority sampling
- uniform sampling without replacement.
- probability proportional to size with replacement



Sampled Internet traffic. Estimate traffic classes

- priority sampling
- uniform sampling without replacement.
- probability proportional to size with replacement



Priority sampling works with minimal independence

- Theorems a bit hard to state. Essential result is that priority sampling gives good concentration even with 2-indepence.

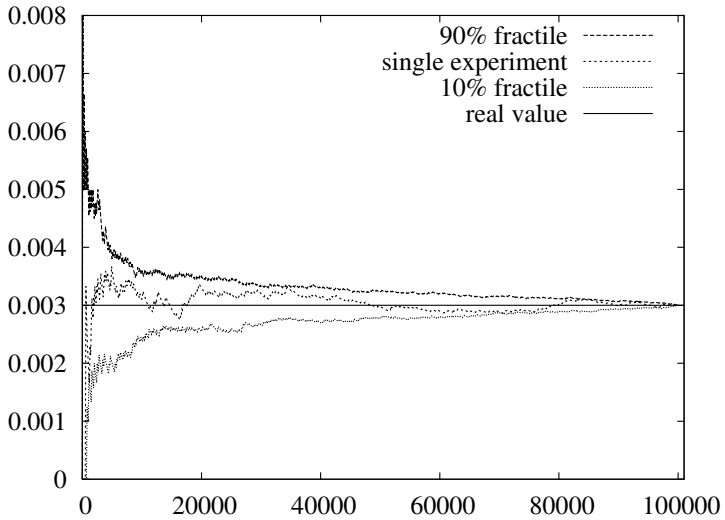
Priority sampling works with minimal independence

- Theorems a bit hard to state. Essential result is that priority sampling gives good concentration even with 2-indepence.
- Thus priority sampling trivial to implement including the hash function.

Priority sampling works with minimal independence

- Theorems a bit hard to state. Essential result is that priority sampling gives good concentration even with 2-indepence.
- Thus priority sampling trivial to implement including the hash function.
- With his mathematical analysis of linear probing, Knuth [1963] started the field of trying to understand simple "magical" algorithms: if they really work, or if they are just waiting to blow up. This is an area full of surprises, and an area where mathematics can really help the outside world.

bottom-k



k-mins