Tighter Estimation using Bottom-k Sketches

Edith Cohen
AT&T Labs–Research
180 Park Avenue
Florham Park, NJ 07932, USA
edith@research.att.com

Haim Kaplan
School of Computer Science
Tel Aviv University
Tel Aviv, Israel
haimk@cs.tau.ac.il

ABSTRACT

Summaries of massive data sets support approximate query processing over the original data. A basic aggregate over a set of records is the weight of subpopulations specified as a predicate over records' attributes. *Bottom-k* sketches are a powerful summarization format of weighted items that includes priority sampling [22], and the classic weighted sampling without replacement. They can be computed efficiently for many representations of the data including distributed databases and data streams and support coordinated and all-distances sketches.

We derive novel unbiased estimators and confidence bounds for subpopulation weight. Our *rank conditioning* (RC) estimator is applicable when the total weight of the sketched set cannot be computed by the summarization algorithm without a significant use of additional resources (such as for sketches of network neighborhoods) and the tighter *subset conditioning* (SC) estimator that is applicable when the total weight is available (sketches of data streams).

Our estimators are derived using clever applications of the Horvitz-Thompson estimator (that is not directly applicable to bottom-k sketches). We develop efficient computational methods and conduct performance evaluation using a range of synthetic and real data sets. We demonstrate considerable benefits of the SC estimator on larger subpopulations (over all other estimators); of the RC estimator (over existing estimators for weighted sampling without replacement); and of our confidence bounds (over all previous approaches).

1. INTRODUCTION

Consider a weighted set (I, w) where I is a set of records, and w is a weight function assigning a weight $w(i) \geq 0$ for each $i \in I$. A basic aggregate over such sets is *subpopulation weight*. A *subpopulation weight query* specifies a subpopulation $J \subset I$ as a predicate on the values of the attributes of the records in I. The result of the query is $w(J)$, the sum of the weights of records in J. This aggregate can be used to estimate other aggregates over subpopulations such as selectivity $(w(J)/w(I))$, variance, and higher moments of $\{w(i) \mid i \in J\}$ [11].

We study probabilistic summarization algorithms, producing a small sketch of the original data, from which we can answer subpopulation weight queries approximately. The use of sketches speeds up query processing and addresses storage limitations when the original dataset can not be stored, is distributed, or resides on a slower media.

In order to support subpopulation selection with arbitrary predicates, the summary must retain content of some individual records so we can determine for each record in the sketch whether it satisfies the predicate. Two such methods are *k-mins* and *bottom-k* sketches. Bottom-k sketches are obtained by assigning a *rank*, $r(i)$, to each record $i \in I$ that is independently drawn for each i from a distribution that depends on $w(i)$. The *bottom-k sketch* contains the k records with smallest ranks [6, 29, 4, 2, 23, 15]. The distribution of the sketches is determined by the family of distributions that is used to draw the ranks. If we draw $r(i)$ from an exponential distribution with parameter $w(i)$ then we obtain sketches that are distributed as if we draw records without replacement with probability proportional to their weights (see e.g. [15]). We denote sampling without replacement with probability proportional to weight by WS, and we call a bottom-k sketch with exponential ranks a WS sketch.

To obtain a *k-mins* sketch [6] we also assign independent random ranks to records (again, the distribution of $r(i)$ depends on $w(i)$). The record of smallest rank is selected, and this is repeated k times, using k independent rank assignments. When we draw $r(i)$ from an exponential distribution with parameter $w(i)$ then k-mins sketches are equivalent to weighted sampling with replacement of k records. We denote sampling with replacement with probability proportional to weight by WSR, and we call a k-mins sketch with exponential ranks a WSR sketch.

Subpopulation weight query is a general primitive with numerous applications. A concrete example is queries over the set of all IP flows going through a router during some time period. Flow records are collected at IP routers by tools such as Cisco's NetFlow [31] (now emerging as an IETF standard). Each flow record contains the source and destination IP addresses, port and protocol numbers, and the number of packets and bytes of the flow. A router can produce a bottom-k sketch of the flows that it collects during a time period, before moving the raw data to external storage or discarding it. A network manager can then use the sketches from a single router or from multiple routers to answer various subpopulation queries. For example, esti-

VLDB '08, August 24-30, 2008, Auckland, New Zealand

mate "the bandwidth used for an application such as p2p or Web traffic" or "the bandwidth destined to a specified Autonomous System." The ability to answer such queries is critical for improving network performance, and for anomaly detection.

Bottom-k and k-mins sketches can be used to summarize a single weighted set or to summarize multiple sets that are defined over the same universe of items. When we summarize multiple sets we want the sketches of different sets to be defined using the same rank assignment to the items. That is, if two sets A and B contain some item x, the same rank value for x is used to compute the sketches of both A and B. We call sketches of multiple sets that share the rank assignment *coordinated sketches*.

A very useful feature of coordinated sketches is that they support subpopulation queries where the selection predicate includes conditions on sets' memberships. In particular we can estimate sizes of (selected subpopulations) of intersections and unions of sets. On our example of IP flow records, consider multiple routers and a sketch of the set of flows through each router. We can approximate queries like "How many flows to a particular destination pass through a particular subset of the routers". Since a particular flow can go through multiple routers, the use of coordinated sketches is critical for obtaining good estimates. Additional application domains of subpopulation queries over coordinated bottom-k sketches include document-features and market-basket datasets [4, 6, 29, 2].

Bottom-k sketches are also used in applications where the subsets are defined implicitly as neighborhoods in a metric space. An example of such an application is a peer to peer network where each node has a database of songs that it is sharing with the other peers. Each peer maintains a bottom-k sketch of the union of the sets of songs of all peers within a certain distance (e.g. hops) from it. If it keeps such sketch for every distance then the collection of these sketches is called an *all distances sketch*. It turns out that coordinated all distances sketches for all peers (and all neighborhoods) can be computed efficiently and stored compactly [6, 14, 15]. The peer can use this sketch to find out how many songs of a particular singer there are in close peers, how many songs of a particular genre are in the entire network etc. These coordinated sketches also support predicates with membership conditions such as "the number of songs by the Beatles that are both at a distance at most ρ_1 from peer v and at a distance at most ρ_2 from peer w." Subpopulation queries over all-distances sketches facilitate decaying aggregation [27, 12], kernel density estimators [33], and typicality estimation [26].

Predicates with membership conditions over coordinated bottom-k sketches of multiple sets are processed by first computing the bottom-k sketch of the union of the relevant sets from their individual sketches. For each included item in the union sketch we can determine membership information in each set and view it as attribute values of the item. Therefore, the complex selection predicate over multiple sets can be handled as a plain attribute-based selection predicate over a sketch of a single set (the union) and estimators and confidence bounds applicable to a bottom-k sketch of a single set can be applied. [1]

Our results. We develop accurate estimators and confidence intervals for subpopulation weight queries for bottom-k sketches and in particular for WS sketches.

Our estimators are of two kinds.
(1) Estimators that are based on the Maximum Likelihood (ML) principle. While biased, WS ML estimators can be computed efficiently and perform well in practice.
(2) Estimators that generalize the basic Horvitz-Thompson (HT) estimator [25].

The HT estimator assigns to each included record i an adjusted weight $a(i)$ equals to $w(i)$ divided by the probability that i is included in a sketch. Clearly, the expected adjusted weight of record i equals $w(i)$. The estimate of a subpopulation is the sum of the adjusted weights of the records in the sketch that belong to the subpopulation and is easily computed from the sketch by applying the selection predicate to included records. This is clearly unbiased.

The HT estimator minimizes the per-record variance of the adjusted weight for the particular distribution over sketches. The HT estimator, however, cannot be computed for bottom-k sketches, since the probability that a record is included in a sketch cannot be determined from the information available in the sketch alone [32, 34]. Our variant, which we refer to as HT *on a partitioned sample space* (HTP), overcomes this hurdle by applying the HT estimator on a partition of the sample space such that inclusion probability of records can be computed in each subspace. We believe that this technique of generalizing the HT estimator may be useful in other contexts. As an indication for this we point out that our derivation generalizes and simplifies one for priority sampling [22] and reveals general principles.

We derive two HT-based estimators. One (rank-conditioning (RC)) is suitable to use when the total weight of the sketched set is not known, and a tighter estimator (subset-conditioning (SC)) to be used when the total weight of the sketched set is known. Our RC estimator generalizes priority sampling, the best known estimator prior to this work. The SC estimator that uses the total weight is much tighter for larger subpopulations than any known estimator based on a sketching method that supports coordinated sketches.

From basic properties of the variance of a sum of random variables follows that the variance of our estimate of a particular subpopulation J is equal to $\sum_{i\in J}\textsc{var}[a(i)] + \sum_{i\neq j, i,j\in J}\textsc{cov}[a(i),a(j)]$, where $\textsc{cov}[a(i),a(j)]$ is the covariance of the adjusted weights of records i and j. For all RC estimators (and priority sampling), the covariances of different records are 0. Our SC estimator, however, has negative covariances of different records. Moreover, the sum of covariances is minimized. This guarantees that the variance of our estimator for any subpopulation is no larger and generally much smaller than the sum of the variances of the adjusted weights of the individual records of the subpopulation. This important property boosts accuracy, in particular for large subpopulations.

Confidence intervals are critical for many applications. We derive confidence bounds (tailored to applications where the total weight is or is not provided) and develop methods to efficiently compute them. We compare our confidence bounds with previous approaches (a bound for priority sampling [37] and known WSR estimators) and show that our confidence intervals are about 1/2 the width of the best previously known with any summarization method.

[1] In a followup work [16], we extend the estimators developed here to tighter estimators that directly work with the original set of sketches.

2. RELATED WORK

In [15] we initiated a comparison between bottom-k and k-mins sketches. The focus there was on quantifying the resources (storage, communication) required to produce these sketches. We distinguished between *explicit* and *implicit* representations of the data. Explicit representations list the occurrence of each record in each set which we sketch. These include a data stream of the items in a single set or item-set pairs if we sketch many sets (for example, item-basket associations in a market basket data, links in web pages, features in documents [5, 3, 29, 35, 2]). Bottom-k sketches can be computed much more efficiently than k-mins sketches when the data is represented explicitly [4, 29, 15].

Implicit representations are those where we summarize multiple sets that are specified succinctly: In the peer to peer network described above, all sketched sets are implicitly represented as neighborhoods in the network. (See also [6, 19, 18, 28, 27, 14].) In these applications, the summarization algorithm is applied to the succinct representation (the network itself in this example).

We also considered in [15] the information content in the sketches. We showed how to probabilistically draw a k-mins sketch from a bottom-k sketch, when using exponential ranks. The distribution induced on k-mins sketch is the same as if we were drawing them directly to begin with, using multiple rank functions. We called this process *mimicking* k-mins sketches from bottom-k sketches. Mimicking k-mins sketches from bottom-k sketches allows to apply simple estimators for k-mins sketches to bottom-k sketches. This, however, does not fully utilize the information in bottom-k sketches. In this paper we derive confidence bounds and tighter estimators that fully exploit the information in bottom-k sketches.

Beyond computation issues, the distinction between data representations is also important for estimation. In particular, we can see why it is important to develop estimators and confidence intervals for subpopulation weight queries for both scenarios, when the total weight is or is not available. With explicit representation, the summarization algorithm can compute the total weight of the records without a significant processing or communication overhead. With implicit representation, the total weight of each subset is not readily available. For example, in our p2p application computing the exact total weight of every neighborhood is much more resource consuming than obtaining all sketches. Even with explicit representation, the total weight of a union of subsets can not be retrieved from their sketches and weights. Therefore, total weight is not available when processing complex queries involving multiple subsets (such as, in our example, the total bandwidth to a destination through multiple routers).

A dominant estimator for subpopulation weight to date is based on priority sampling [22]. This estimator emerged as a clever modification of threshold sampling [21] so that a fixed size sample is produced. Estimators based on priority sampling perform similarly to estimators based on threshold sampling, and are better than estimators based on WSR. Priority sampling was also shown to minimize the sum of per-record variances [36]. We denote priority sampling by PRI, and a sketch which it produces by a PRI sketch.

Priority sampling, however, was not compared to weighted sampling without replacement, since unbiased estimators for the latter were not known. It turns out that a PRI sketch is a bottom-k sketch for a different family of rank functions. Our general framework for bottom-k sketches places both priority sampling and weighted sampling without replacement within a unified framework. In particular, we generalize the work of [22], and gives a simpler proof that their estimator for PRI sketches is unbiased.

Our evaluation of the estimators for WS sketches and PRI sketches show that their performance is similar when the total weight is not provided. We make a strong case for WS sketches by showing that 1) the confidence intervals which we develop for WS sketches are much tighter (about half the width) than confidence intervals known for PRI sketches [37], and by 2) developing estimators for WS sketches that use the total weight whose variance is 1/3 of the variance of PRI sketches. Another advantage of WS sketches is that they can often be computed more efficiently than PRI sketches.

Motivated by the work we present here, a summarization scheme which minimizes the sum of variances of sets of any fixed size has been developed [9]. This scheme is applicable for a summary of a single set and is not a bottom-k sketch. In particular, it does not support coordinated and all-distances sketches.

An orthogonal line of recent research is on algorithms for sketching unaggregated data [24, 20], where each item may be broken into many different pieces, each with its own weight. The sum of the weights of the pieces is equal to the weight of the item. As an example think of an IP flow broken into its individual packets. The problem is to get a sketch of the dataset for subpopulation queries without preaggregating the data which may require substantial resources. The papers [8, 7] study several estimators for this setup, some particularly appropriate for routers' architectures. A more recent work [10] applies the recent sampling technique of [9] to unaggregated data.

Last, we mention a vast literature on survey sampling that studies WSR (PPSWR Probability Proportional to Size With Replacement), WS (PPSWOR - PPS Without Replacement) and threshold sampling (related to IPPS - Inclusion Probability Proportion to Size), see e.g. [32, 34]. Nevertheless, our estimators and confidence intervals are original. Furthermore, we address important database issues such as efficient computation of sketches, and coordinated sketches, on massive datasets; efficient query processing; support for general subpopulation selection queries; and more.

A poster with a very preliminary sketch of the ideas we develop here has appeared in [13]. This paper (together with the appendix) supersedes the technical report [17].

3. PRELIMINARIES

Let (I, w) be a weighted set. A *rank assignment* r maps each item i to a random rank $r(i)$. The ranks of items are drawn independently using a family of distributions $\mathbf{f_w}$ $(w \geq 0)$, where the rank of an item with weight $w(i)$ is drawn from $\mathbf{f_{w(i)}}$.[2] For a subset J of items and a rank assignment r, we denote by i_j the item of jth largest rank in J. We also define $r(J) = r(i_1)$ to be the smallest rank in J according to r.

A *k-mins sketch* of a set J is the vector

$$(r^{(1)}(J), r^{(2)}(J), \ldots, r^{(k)}(J))$$

[2] We assume that we use enough bits to represent the ranks such that no two items obtain the same rank.

where $r^{(1)}, \ldots, r^{(k)}$ are k independent rank assignments. For some applications we store with $r^{(\ell)}(J)$, $1 \le \ell \le k$, additional attributes of the corresponding item such as its weight.

A *bottom-k sketch* is produced from a single rank assignment r. It is the list of the k pairs $(r(i_j), w(i_j))$, $1 \le j \le k$, sorted by increasing rank together with $r(i_{k+1})$. Depending on the application we may store with the sketch additional attributes of the items that attain the k smallest ranks. (If $|J| < k$ then the sketch contains only $|J|$ pairs.) We often abbreviate $r(i_{k+1})$ to r_{k+1}.

Bottom-k sketches must include the items' weights but can omit all rank values except r_{k+1}. The reason is that using the weights of the items with k smallest ranks and r_{k+1}, we can *redraw* the rank value of an item with weight w from the density function $\mathbf{f_w(x)}/\mathbf{F_w(r_{k+1})}$ for $0 \le x \le r_{k+1}$ and 0 elsewhere. Here $\mathbf{F_w(x)}$ is the cumulative distribution function of f_w. This is equivalent to redrawing a rank assignment from the subspace of the probability space of rank assignments where r_{k+1} and *the set* of the k smallest items is fixed. Redrawing ranks can also be viewed as redrawing a particular ordered set of bottom-k items from this subspace. Moreover, as we shall see in Section 6, if $w(J)$ is provided and we use WS sketches, we can redraw all rank values (there is no need to retain r_{k+1}). Redrawing all ranks is equivalent to obtaining a rank assignment from the probability subspace where the subset of items with k smallest ranks is the same. These properties are not only important for limiting storage requirements for the sketch but as we shall see, also facilitate the derivation of tighter estimators.

ws sketches. Clearly, the choice of which family of random rank functions to use matters only when items are weighted. Rank functions $\mathbf{f_w}$ with some useful properties are exponential distributions with parameter w [6]. The density function of this distribution is $\mathbf{f_w(x) = we^{-wx}}$, and its cumulative distribution function is $\mathbf{F_w(x) = 1 - e^{-wx}}$. Since the minimum of independent exponentially distributed random variables is exponentially distributed with parameter equal to the sum of the parameters of these distributions it follows that $r(J)$ for a subset J is exponentially distributed with parameter $w(J) = \sum_{i \in J} w(i)$. Cohen [6] used this property to obtain unbiased low-variance estimators for both the weight and the inverse weight of a set J using a k-mins sketch of J.[3]

With exponential ranks the item with the minimum rank $r(J)$ is a *weighted random sample* from J: The probability that an item $i \in J$ is the item of minimum rank is $w(i)/w(J)$. Therefore, a k-mins sketch of a subset J corresponds to a weighted random sample of size k, drawn **with replacement** from J. We call k-mins sketch using exponential ranks a WSR sketch. On the other hand, a bottom-k sketch of a subset J with exponential ranks corresponds to a weighted k-sample drawn **without replacement** from J [15]. We call such a sketch a WS sketch.

The following property of exponentially-distributed ranks is a consequence of the memoryless nature of the exponential distribution.

LEMMA 3.1. *[15] Consider a probability subspace of rank assignments over J where the k items of smallest ranks are $i_1, \ldots, i_k$ in increasing rank order. The rank differences $r_1(J), r_2(J) - r_1(J), \ldots, r_{k+1}(J) - r_k(J)$ are independent random variables, where $r_j(J) - r_{j-1}(J)$ $(j = 1, \ldots, k+1)$ is exponentially distributed with parameter $w(J) - \sum_{\ell=1}^{j-1} w(i_\ell)$. (we formally define $r_0(J) \equiv 0$.)*

WS sketches can often be computed more efficiently than other bottom-k sketches. Computing a bottom-k sketch on unaggregated data (each item appears in multiple "pieces") generally requires *pre-aggregating* the data, so that we have a list of all items and their weight, which is a costly operation when the data is distributed or in external memory. A key property of exponential ranks is that we can obtain a rank value for an item by computing *independently* a rank value for each piece, based on the weight of the piece. The rank value of the item is the minimum rank value of its pieces. The WS sketch can therefore be computed in two $O(k)$ communication rounds over distributed data or in two linear passes using $O(k)$ memory: The first pass identifies the k items with smallest rank values. The second pass is used to add up the weights of the pieces of each of these k items.

Computing a Bottom-k sketch requires processing of each item. When items are partitioned such that we have the weight of each part, WS sketches can be computed while processing only a fraction of the items. A key property is that the minimum rank value over a set of items depends only on the sum of the weights of the items. Using this property, we can quickly determine which parts contribute to the sketch and eliminate chunks of items that belong to other parts.

The same property is also useful when sketches are computed online over a stream [23]. Bottom-k sketches are produced using a priority queue that maintains the $k+1$ items with smallest ranks. We draw a rank for each item and update the queue if this rank is smaller than the largest rank in the queue. With WS sketches, we can simply draw directly from a distribution the accumulated weight of items that can be "skipped" before we obtain an item with a smaller rank value than the largest rank in the queue. The stream algorithm simply adds up the weight of items until it reaches one that is incorporated in the sketch.

PRI ***sketches.*** With priority ranks [22, 1] the rank value of an item with weight w is selected uniformly at random from $[0, 1/w]$. This is equivalent to choosing a rank value r/w, where $r \in U[0,1]$, the uniform distribution on the interval $[0,1]$. It is well known that if $r \in U[0,1]$ then $-\ln(r)/w$ is an exponential random variable with parameter w. Therefore, in contrast with priority ranks, exponential ranks correspond to using rank values $-\ln r/w$ where $r \in U[0,1]$.

PRI sketches are of interest because one can derive from them an estimator that (nearly) minimizes the sum of per-item variances: $\sum_{i \in I} \text{VAR}[a(i)]$ [36]. More precisely, Szegedy showed that the sum of per-item variances using PRI sketches of size k is no larger than the smallest sum of per-item variances attainable by an estimator that uses sketches with average size $k-1$.[4]

Review of weight estimators for wsr sketches. Recall that for a subset J, the rank values in the k-mins sketch

[3] Estimators for the inverse-weight are useful for obtaining unbiased estimates for quantities where the weight appears in the denominator such as the weight ratio of two different subsets.

[4] Szegedy's proof applies only to estimators based on adjusted weight assignments.

$r_1(J),\ldots,r_k(J)$ are k independent samples from an exponential distribution with parameter $w(J)$. The quantity $\frac{k-1}{\sum_{h=1}^{k} r_h(J)}$ is an unbiased estimator of $w(J)$. The standard deviation of this estimator is equal to $w(J)/\sqrt{k-2}$ and the average relative error is approximately $\sqrt{2/(\pi(k-2))}$ [6]. The quantity $\frac{k}{\sum_{h=1}^{k} r_h(J)}$ is the maximum likelihood estimator of $w(J)$. This estimator is a factor of $k/(k-1)$ larger than the unbiased estimator. Hence, it is obviously biased, and the bias is equal to $w(J)/(k-1)$. Since the standard deviation is about $(1/\sqrt{k})w(J)$, the bias is not significant when $k \gg 1$. The quantity $\frac{\sum_{h=1}^{k} r_h(J)}{k}$ is an unbiased estimator of the *inverse weight* $1/w(J)$. The standard deviation of this estimate is $1/(\sqrt{k}w(J))$.

4. MAXIMUM LIKELIHOOD ESTIMATORS

We apply the Maximum Likelihood (ML) principle to derive WS ML estimators. These estimators are applicable to WS sketches as our derivation exploits special properties of the exponential distribution used to produce these sketches. We show the derivation of an estimator for the total weight of the sketched set. Using the same technique in a slightly more subtle way we obtain similar estimators for the weight of a subpopulation when we do not know the total weight and when we do know the total weight. These derivations can be found in Appendix A.

Consider a set I and its bottom-k sketch s. Recall that $i_1, i_2, \ldots, i_k$ are the items in s ordered by increasing ranks. (We assume that $|k| < |I|$ as otherwise $w(I)$ is just the sum of the weights of the items in the sketch.)

Consider the rank differences, $r(i_1), r(i_2)-r(i_1), \ldots, r(i_{k+1})-r(i_k)$. From Lemma 3.1, they are independent exponentially distributed random variables in the appropriate subspace. The joint probability density function of this set of differences is therefore the product of the density functions

$$w(I)\exp(-w(I)r(i_1))(w(I)-s_1)\exp(-(w(I)-s_1)(r(i_2)-r(i_1)))\cdots$$

where $s_\ell = \sum_{j=1}^{\ell} w(i_j)$. Think about this probability density as a function of $w(I)$. The maximum likelihood estimate for $w(I)$ is the value that maximizes this function. To find the maximum, take the natural logarithm (for simplification) of the expression and look at the value which makes the derivative zero. We obtain that the maximum likelihood estimator $\tilde{w}(I)$ is the solution of the equation

$$\sum_{i=0}^{k} \frac{1}{\tilde{w}(I) - s_i} = r(i_{k+1}) \ . \qquad (1)$$

The left hand side is a monotone function, and the equation can be solved by a binary search on the range $[s_k + 1/r(i_{k+1}), s_k + (k+1)/r(i_{k+1})]$.

5. ADJUSTED WEIGHTS

In this section we introduce variants of the Horvitz-Thompson (HT) estimator [25]. Proofs can be found in Appendix B.

The idea here is to assign a positive adjusted weight $a(i)$ to each item in the sample, such that if we also set $a(i) = 0$ when i is not sampled then $\mathsf{E}[a(i)] = w(i)$. (The expectation is over the draw of the sample. Once the sample is determined the assignment of $a(i)$ is usually deterministic.) We call a sample together with the adjusted weights an *adjusted weight summary* (AW-summary) of the weighted set (I, w).

An AW-*summarization algorithm* is a probabilistic algorithm that inputs a weighted set (I, w) and returns an AW-summary of (I, w). An AW-summarization algorithm for (I, w) provides unbiased estimators for the weight of I and for the weight of subsets of I since by linearity of expectation, for any $H \subseteq I$, the sum $\sum_{i\in H} a(i)$ is an unbiased estimator of $w(H)$. Notice that if there is another weight function h defined over I then $\sum_{i\in H} h(i)a(i)/w(i)$ is an unbiased estimator of $h(J)$ for any $J \subseteq I$.

Let Ω be the probability space of rank assignments over I. Each $r \in \Omega$ has a sketch $s(r)$ associated with it. Suppose that given $s(r)$ we can compute the probability $\Pr\{i \in s(r) \mid r \in \Omega\}$ for all $i \in s(r)$ (since I is a finite set, these probabilities are strictly positive for all $i \in s(r)$). Then we can make $s(r)$ into an AW-summary using the Horvitz-Thompson (HT) estimator [25] which assigns to each $i \in s(r)$ the adjusted weight

$$a(i) = \frac{w(i)}{\Pr\{i \in s(r) \mid r \in \Omega\}} \ .$$

It is well known and easy to see that these adjusted weights are unbiased and have minimal variance *for each item* for the particular distribution over the sketches that is derived from Ω.

HT **on a partitioned sample space** (HTP) is a method to derive adjusted weights when we cannot determine $\Pr\{i \in s(r) \mid s(r) \in \Omega\}$ from the sketch $s(r)$ alone. For example if $s(r)$ is a bottom-k sketch, then the probability $\Pr\{i \in s(r) \mid r \in \Omega\}$ depends on all the weights $w(i)$ for $i \in I$.

For each item i we partition Ω into subsets $P_1^i, P_2^i \ldots$. This partition satisfies the following two requirements: (1) Given a sketch $s(r)$, we can determine the set P_j^i containing r, and (2) For every set P_j^i we can compute the conditional probability $p_j^i = \Pr\{i \in s(r) \mid r \in P_j^i\}$.

For each $i \in s(r)$, we identify the set P_j^i containing r and use the adjusted weight $a(i) = w(i)/p_j^i$ (which is the HT adjusted weight in P_j^i).[5] The expected adjusted weight of each item i within each subspace of the partition is $w(i)$ and therefore its expected adjusted weight over Ω is $w(i)$.

Rank Conditioning (RC**) adjusted weights** for bottom-k sketches are HTP adjusted weights where the partition $P_1^i, \ldots, P_\ell^i$ which we use is based on *rank conditioning.* For each possible rank value r we have a set P_r^i containing all rank assignments in which the kth rank assigned to an item other than i is r. (If $i \in s(r)$ then this is the $(k+1)$st smallest rank and otherwise its the kth smallest rank.)

The probability that i is included in a bottom-k sketch given that the rank assignment is from P_r^i is the probability that its rank value is smaller than r. For WS sketches, this probability is equal to $1 - \exp(-w(i)r)$. Assume $s(r)$ contains $i_1, \ldots, i_k$ and that the $(k+1)$st smallest rank is r_{k+1}. Then for item i_j, the rank assignment belongs to $P_{r_{k+1}}^{i_j}$, and therefore the adjusted weight of i_j is $\frac{w(i_j)}{1-\exp(-w(i_j)r_{k+1})}$. The PRI RC adjusted weight for an item i_j (obtained by a tailored derivation in [1]), is $\max\{w(i_j), 1/r_{k+1}\}$.

Variance of RC *adjusted weights*

[5] In fact all we need is the probability p_j^i. In some cases we can compute it from some parameters of P_j^i, without identifying P_j^i precisely.

LEMMA 5.1. *Consider* RC *adjusted weights and two items i, j. Then,* $\textsc{cov}[a(i), a(j)] = 0$ *(The covariance of the adjusted weight of i and the adjusted weight of j is zero.)*

As mentioned in the introduction from simple properties of the variance of a sum of random variables we have that

$$\textsc{var}[a(J)] = \sum_{i \in J} \textsc{var}[a(i)] + \sum_{i \neq j, i, j \in J} \textsc{cov}[a(i), a(j)] \ .$$

This implies for RC adjusted weights the following corollary of Lemma 5.1.

COROLLARY 5.2. *For a subset $J \subset I$,*

$$\textsc{var}[a(J)] = \sum_{j \in J} \textsc{var}[a(j)] \ .$$

Therefore, with RC adjusted weights, the variance of the weight estimate of a subpopulation is equal to the sum of the per-item variances, just like when items are selected independently. This Corollary, combined with Szegedy's result [36], shows that when we have a choice of a family of rank functions, PRI weights are the best rank functions to use when using RC adjusted weights.

Selecting a partition. The variance of the adjusted weight $a(i)$ obtained using HTP depends on the particular partition in the following way.

LEMMA 5.3. *Consider two partitions of the sample space, such that one partition is a refinement of the other, and the* AW*-summaries obtained by applying* HTP *using these partitions. For each $i \in I$, the variance of $a(i)$ using the coarser partition is at most that of the finer partition.*

It follows from Lemma 5.3 that when applying HTP, it is desirable to use the coarsest partition for which we can compute the probability p^i_j from the information in the sketch. In particular a partition that includes a single component minimizes the variance of $a(i)$ (This is the HT estimator). The RC partition yields the same adjusted weights as conditioning on the rank values of all items in $I \setminus i$, so it is in a sense also the *finest* partition we can work with. It turns out that when the total weight $w(I)$ is available we can use a coarser partition.

6. USING THE TOTAL WEIGHT

When the total weight is available we can use HTP estimators defined using a coarser partition of the sample space than the one used by the RC estimator. The *subset conditioning estimator* (SC), which we present in this section, partitions the space of rank assignment by the set of the $k-1$ items of smallest rank among all items other than i, in order to compute the adjusted weight of i. That is, we will show that knowing the total weight, it is possible to compute the probability that i is included in the sketch in the subspace of rank assignments where the set of $k-1$ smallest-ranked items other than i is fixed. A related estimator which also condition on the order of the ranks of these $k-1$ smallest-ranked items is called the *prefix conditioning estimator* and is presented in Appendix C. By Lemma 5.3 subset conditioning is better than prefix conditioning and rank conditioning in terms of per-item variances. A side benefit of these estimators is that they do not need r_{k+1} and thereby require one less sample.

The SC estimator has the following two important properties that RC does not have. The adjusted weights of different items have negative covariances, and $\textsc{var}[a(I)] = 0$. Two alternative ways to say this are that the sum of the adjusted weights equals the total weight of the set and the sum of the covariances of different items is "as negative as possible" and is equal to the negative of the sum of the variances of the individual items. These properties imply that the variance of the estimator on a subset is smaller than the sum of the variances of the individual items in the subset.

We now define the SC estimator precisely. In Appendix D we prove that it indeed satisfies the properties mentioned above. For a set Y of items and $\ell \geq 0$, we define

$$f(Y, \ell) = \int_{x=0}^{\infty} \ell \exp(-\ell x) \prod_{j \in Y} (1 - \exp(-w(j)x)) dx \ . \quad (2)$$

This is the probability that a random rank assignment with exponential ranks for the items in Y, and for the items in a set X such that $w(X) = \ell$, assigns the $|Y|$ smallest ranks to the items in Y and the $(|Y|+1)$st smallest rank to an item from X. For exponential ranks, this probability depends only on $w(X)$, and does not depend on how the weight of X is divided between the items. This is a critical property that allows us to compute adjusted weights with subset conditioning specifically for WS sketches.

Let s be the WS sketch. Recall that for an item i, we use the subspace with all rank assignments in which among the items in $I \setminus \{i\}$, the items in $s \setminus \{i\}$ have the $(k-1)$ smallest ranks. The probability, conditioned on this subspace, that item i is contained in the sketch is $\frac{f(s, w(I \setminus s))}{f(s \setminus \{i\}, w(I \setminus s))}$, and so the adjusted weight assigned to i is

$$a(i) = w(i) \frac{f(s \setminus \{i\}, w(I \setminus s))}{f(s, w(I \setminus s))} \ . \quad (3)$$

Note that we easily compute $w(I \setminus s)$ from the total weight and the weights of the items in s.

6.1 Computing SC adjusted weights

SC adjusted weights can be computed by numerical integration using Eq. (3). We propose an alternative method based on a Markov chain that is faster and easier to implement. The method converges to the SC adjusted weights as the number of steps grows. It can be used with any fixed number of steps and provides unbiased adjusted weights.

The key idea is to use the fact (which follows from the proof of Lemma 5.3) that the mean of the RC estimator over all rank assignments in a partition of the SC estimator is equal to the SC estimator. Let P be the set of the items in the sketch. Let Ω_P be the subspace of all rank assignments producing a sketch with items P. The subspace Ω_P is further partitioned according to the rank-order π of the items in P. Let $\Omega_{P,\pi}$ be one such subspace of Ω_P that corresponds to a permutation π of P. The subspace $\Omega_{P,\pi}$ is further partitioned according to the $k+1$ smallest rank.

We approximate SC by drawing a subspace $\Omega_{P,\pi}$ with probability $p_\pi = |\Omega_{P,\pi}|/|\Omega_P|$; then drawing the $k+1$ smallest rank according to the distribution of r_{k+1} in $\Omega_{P,\pi}$; and computing the RC adjusted weight of items in P using r_{k+1}. This process is repeated multiple times and we use the average as an unbiased estimate on the mean, which is the SC estimator.

Consider a subspace $\Omega_{P,\pi}$, and let $i_1, i_2, \ldots, i_k$ be the items ordered as in π. By Lemma 3.1 the distribution of

r_{k+1} in $\Omega_{P,\pi}$ is the sum of k independent exponential random variables with parameters $w(I)$, $w(I)-w(i_1)$,...,$w(I)-\sum_{h=1}^{k} w(i_h)$. So the adjusted weight of i_j, $j=1,\ldots,k$ is $a(i_j)=\mathsf{E}[w(i_j)/(1-\exp(-w(i_j)r_{k+1}))]$ where the expectation is over this distribution of r_{k+1}.[6]

Instead of computing the expectation, we average the RC adjusted weights $w(i_j)/(1-\exp(-w(i_j)r_{k+1}))$ over multiple draws of r_{k+1}. This average is clearly an unbiased estimator of $w(i_j)$ and its variance decreases with the number of draws. Each repetition can be implemented in $O(k)$ time (drawing and summing k random variables.).

To draw the subspace $\Omega_{P,\pi}$ we define a Markov chain over permutations of P. Starting with a permutation π defined by the original ranks, we continue to a permutation π' by applying the following process. We draw r_{k+1} as described above from the distribution of r_{k+1} in $\Omega_{P,\pi}$. We then redraw rank values for the items of P using their weights and r_{k+1} as described in Section 3. The permutation π' is obtained by reordering P according to the new rank values. This Markov chain has the following property.

LEMMA 6.1. *Let P be a (unordered) set of k items. Let p_π be the conditional probability that in a random rank assignment whose prefix consists of items of P, the order of these items in the prefix is as in π. Then p_π is the stationary distribution of the Markov chain described above.*

PROOF. Suppose we draw a permutation π of the items in P with probability p_π and then draw r_{k+1} as described above. Then this is equivalent to drawing a random rank assignment whose prefix consists of items in P and taking r_{k+1} of this assignment.

Similarly assume we draw r_{k+1} as we just described, draw ranks for items in P, and order P by these ranks. Then this is equivalent to drawing a permutation π with probability p_π. □

Our implementation is controlled by two parameters: INPERM and PERMNUM. INPERM is the number of times the rank value r_{k+1} is redrawn for a permutation π (at each step of the Markov chain). PERMNUM is the number of steps of the Markov chain (number of permutations in the sequence).

We start with the permutation $(i_1,\ldots,i_k)$ obtained in the WS sketch. We apply this Markov chain to obtain a sequence of PERMNUM permutations of $\{i_1,\ldots,i_k\}$. For each permutation π_j, $1\le j\le$ PERMNUM, we draw r_{k+1} from P_{π_j} INPERM times as described above. For each such draw we compute the RC adjusted weights for all items. The final adjusted weight is the average of the RC adjusted weights assigned to the item in the PERMNUM $*$ INPERM applications of the RC method. The total running time is $O(\text{PERMNUM}\cdot k\log k+\text{INPERM}\cdot k)$.

An important property of this process is that if we apply it for a *fixed* number of steps, and average over a fixed number of draws of r_{k+1} within each step, we still obtain unbiased estimators. Our experimental section shows that these estimators perform very well.

7. CONFIDENCE BOUNDS

We provide a general derivation of confidence bounds for bottom-k sketches, specialize it to WS sketches, and develop efficient computational techniques.

We derive bounds for subpopulation weight when the total weight is not known. These bounds use conditioning on the order of the items in the sketch. In Appendix E we use similar techniques to derive tighter confidence intervals for two other variants of the problem: the (special case of the) total weight (Appendix E.1) and subpopulations when the total weight is known (Appendix E.3). Our bounds on the total weight do not condition on the order of the items in the sketch. This conditioning weakens the bounds but simplifies the derivation and the computation for WS sketches.

A *weighted list* (Z,π) is a weighted set Z linearly ordered according to a permutation π. The linear order will be often derived from a rank assignment r to the elements in Z. In such case we may also denote a weighted list by (Z,r).

The *concatenation* $(Z^1,\pi^1)\oplus(Z^2,\pi^2)$ of two weighted lists (Z^1,π^1) and (Z^2,π^2) is a weighted list of $Z^1\cup Z^2$ where the elements in Z^1 are ordered according to π^1, the elements in Z^2 are ordered according to π^2, and all the elements of Z^1 precede all the elements of Z^2. We define $\Omega((Z,\pi))$ to be the probability subspace of rank assignments over Z such that the rank order of the items is π.

Let r be a rank assignment, s be the corresponding sketch, and ℓ be the weighted list $\ell=(J\cap s,r)$. Let $\overline{W}(\ell,r_{k+1},\delta)$ be the set of all weighted lists $h=(H,\pi)$ such that

$$\text{PR}\{r'(H)\ge r_{k+1}\mid r'\in\Omega(\ell\oplus h)\}\ge\delta\ .$$

Verbally, $\overline{W}(\ell,r_{k+1},\delta)$ consists of all weighted lists $h=(H,\pi)$ that we can concatenate to ℓ such that in at least δ fraction of the rank assignments to $(J\cap s)\cup H$ that respects the order of $\ell\oplus h$, the smallest rank in H is at least r_{k+1}.

Let $\overline{w}(\ell,r_{k+1},\delta)=\sup\{w(H)\mid(H,\pi)\in\overline{W}(\ell,r_{k+1},\delta)\}$. (If $\overline{W}(\ell,r_{k+1},\delta)=\emptyset$, then $\overline{w}(\ell,r_{k+1},\delta)=0$.)

Analogously, let $\underline{W}(\ell,r_k,\delta)$ be the set of all weighted lists $h=(H,\pi)$ such that

$$\text{PR}\{\overline{r'}(J\cap s)\le r_k\mid r'\in\Omega(\ell\oplus h)\}\ge\delta\ .$$

Let $\underline{w}(\ell,r_k,\delta)=\inf\{w(H)\mid(H,\pi)\in\underline{W}(\ell,r_k,\delta)$.
(If $\underline{W}(\ell,r_k,\delta)=\emptyset$, then $\underline{w}(\ell,r_k,\delta)=+\infty$). We prove the following (Appendix E.2)

LEMMA 7.1. *Let r be a rank assignment, s be the corresponding sketch, and ℓ be the weighted list $\ell=(J\cap s,w,r)$. Then $w(J\cap s)+\overline{w}(\ell,r_{k+1},\delta)$ is a $(1-\delta)$-confidence upper bound on $w(J)$ and $w(J\cap s)+\underline{w}(\ell,r_k,\delta)$ is a $(1-\delta)$-confidence lower bound on $w(J)$.*

7.1 Confidence bounds for ws sketches

The derivation of Lemma 7.1 incorporates "worst case" assumptions on the weight distribution of "unseen" items (items that are not included in the sketch). WS sketches have the unique property that the distribution of the ith largest rank in a weighted set, conditioned on either the set or the list of the $i-1$ items of smallest rank values, depends only on the total weight of the set (and not on

[6] The mean of RC adjusted weights over $\Omega_{P,\pi}$ are correct adjusted weights that have smaller variance than RC. Note that this is not an instance of HTP. Also note that while similar looking, this estimator is weaker than prefix conditioning: Rank assignments with the same prefix of items from $I\setminus i$, but where the item i appears in different positions in the k-prefix, can have different adjusted weights with this assignment, whereas they have the same adjusted weight with prefix conditioning. Thereby this estimator has larger variance of the adjusted weight of each item i than prefix conditioning.

the particular partition of the "unseen" weight into items). This property makes the bounds *efficient* in the respective probability subspaces (the bounds correspond to the actual quantiles of the estimator).

We provide some properties and notation that simplify the presentation of bounds derived with conditioning on the order. Consider a weighted set (I, w) and a subspace of rank assignments where the ordered set of the h items of smallest ranks is $i_1, i_2, \ldots, i_h$. Let $s_j = \sum_{\ell=1}^{j} w(i_\ell)$. For convenience we define $s_0 \equiv 0$ and $r_0 = 0$. By Lemma 3.1, for $j = 0, \ldots, h$, the rank difference $r(i_{j+1}) - r(i_j)$ is an exponential r.v. with parameter $w(I) - s_j$, and these rank differences are independent. Therefore for $j \in \{0, \ldots, h\}$, the distribution of $r(i_j)$ (also the sum of the first i rank differences) is a sum of exponential random variables.

For $0 \le x_0 \le \cdots \le x_h < t$, we use the notation $v(t, x_0, \ldots, x_h)$ for the random variable that is the sum of $h+1$ independent exponential random variables with parameters $t - x_j$ $(j = 0, \ldots, h)$. With this notation the distribution of $r(i_j)$ is $v(w(I), s_0, \ldots, s_{j-1})$. As we had seen in Lemma 7.1, our confidence bounds are computed by finding quantiles of these distributions, that is, solving equations of the form

$$\text{PR}\{v(x, s_0, \ldots, s_h) \le \tau\} = \delta \ . \qquad (4)$$

From linearity of expectation,

$$\mathsf{E}[v(t, x_0, \ldots, x_h)] = \sum_{j=0}^{h} 1/(t - x_j) \ .$$

From independence, the variance is the sum of variances of the exponential random variables and is

$$\text{VAR}[v(t, x_0, \ldots, x_h)] = \sum_{j=0}^{h} 1/(t - x_j)^2 \ .$$

Bounds for the total weight w(I). We apply a derivation in the Appendix (Lemma E.1). Note that $\text{PR}\{v(x, s_0, \ldots, s_k) \le r_{k+1}\}$ is the probability that the $k+1$ largest rank is $\le r_{k+1}$ given that the "unseen" weight is $x - s_k$. This probability increases with x. If for $x = s_k$ this probability is already $\ge 1 - \delta$ it means that if the total weight is larger than s_k the event of seeing $k+1$ largest rank $\ge r_{k+1}$ is already $\le \delta$. Therefore we can take $x = s_k$ as a $1 - \delta$ confidence upper bound. Otherwise, we take the solution of the equation

$$\text{PR}\{v(x, s_0, \ldots, s_k) \le r_{k+1}\} = 1 - \delta \ .$$

to be our $1-\delta$ confidence upper bound. A larger total weight would mean that obtaining $k+1$ smallest rank $\ge r_{k+1}$ is an event which is less than δ likely to happen.

Similarly, for a $1 - \delta$ confidence lower bound we take the solution of

$$\text{PR}\{v(x, s_0, \ldots, s_k) \le r_{k+1}\} = \delta \ .$$

Bounds for subpopulation weight (with unknown w(I)). We specialize Lemma 7.1 for WS sketches. Let J be a subpopulation. For a rank assignment, let s be the corresponding sketch and let s_h $(1 \le h \le |J \cap s|)$ be the sum of the weights of the h items of smallest rank values from J (we define $s_0 \equiv 0$). Lemma 7.1 implies that the $(1 - \delta)$-confidence upper bound on $w(J)$ is the solution of the equation

$$\text{PR}\{v(x, s_0, \ldots, s_{|J \cap s|}) \le r_{k+1}\} = 1 - \delta$$

(and is $s_{|J\cap s|}$ if there is no solution $x > s_{|J\cap s|}$.) The $(1-\delta)$-confidence lower bound is 0 if $|J \cap s| = 0$. Otherwise, let $x > s_{|J\cap s|-1}$ be the solution of

$$\text{PR}\{v(x, s_0, \ldots, s_{|J\cap s|-1}) \le r_k\} = \delta \ .$$

The lower bound is $\max\{s_{|J\cap s|}, x\}$.

We propose two methods of solving these equations: (i) applying the *normal approximation* to the respective sum of exponentials distribution or (ii) the *quantile method* which we developed.

Normal approximation. We apply the normal approximation to the quantiles of a sum of exponentials distribution. For $\delta \ll 0.5$, let α be the Z-value that corresponds to confidence level $1 - \delta$. The approximate δ-quantile of $v(x, s_0, \ldots, s_h)$ is $\mathsf{E}[v(x, s_0, \ldots, s_h)] - \alpha\sqrt{\text{VAR}[v(x, s_0, \ldots, s_h)]}$ and the approximate $(1-\delta)$-quantile is $\mathsf{E}[v(x, s_0, \ldots, s_h)] + \alpha\sqrt{\text{VAR}[v(x, s_0, \ldots, s_h)]}$.

To approximately solve $\text{PR}\{v(x, s_0, \ldots, s_h) \le \tau\} = \delta$ (x such that τ is the δ-quantile of $v(x, s_0, \ldots, s_h)$), we solve the equation

$$\mathsf{E}[v(x, s_0, \ldots, s_h)] - \alpha\sqrt{\text{VAR}[v(x, s_0, \ldots, s_h)]} = \tau \ .$$

To approximately solving $\text{PR}\{v(x, s_0, \ldots, s_h) \le \tau\} = 1 - \delta$, we solve

$$\mathsf{E}[v(x, s_0, \ldots, s_h)] + \alpha\sqrt{\text{VAR}[v(x, s_0, \ldots, s_h)]} = \tau \ .$$

We solve these equations (to the desired approximation level) by searching over values of $x > s_h$ using standard numerical methods. The function $\mathsf{E}[v(x)] + \alpha\sqrt{\text{VAR}[v(x)]}$ is monotonic decreasing in the range $x > s_h$. The function $\mathsf{E}[v(x)] - \alpha\sqrt{\text{VAR}[v(x)]}$ is decreasing or bitonic (first increasing then decreasing) depending on the value of α.

The quantile method. Let $D^{(x)}$ be a parametric family of probability distributions defined over the same domain, such that the parameter x varies continuously in some interval.

Let τ be a value in the union of the domains of $\{D^{(x)}\}$ such that the probability $\text{PR}\{y \preceq \tau \mid y \in D^{(x)}\}$ is increasing with x. So the value of x which solves the equation $\text{PR}\{y \preceq \tau \mid y \in D^{(x)}\} = \delta$ $(Q_\delta(D^{(x)}) = \tau)$ is unique.

We assume the following two "black box" ingredients. The first ingredient is a tool for drawing a *monotone parametric sample*. A monotone parametric sample is a function s such that for every x, $s(x)$ is a sample from $D^{(x)}$, and if $x \ge y$ then $s(x) \preceq s(y)$. We say that monotone parametric samples s^1 and s^2 are *independent* if for every x, $s^1(x)$ and $s^2(x)$ are independent draws from $D^{(x)}$.

The second ingredient is a solver of equations of the form $s(x) = \tau$ for a parametric sample $s(x)$. We assume that the parametric sampling process is such that there is always a solution.

We define a distribution $\overline{D}^{(\tau)}$ such that a sample from $\overline{D}^{(\tau)}$ is obtained by drawing a monotone parametric sample $s(x)$ and returning the solution of $s(x) = \tau$. (parametric samples of different samples from $\overline{D}^{(\tau)}$ are independent.) The two black box ingredients allow us to draw samples from $\overline{D}^{(\tau)}$. Our interest in $\overline{D}^{(\tau)}$ is due to the following lemma.

LEMMA 7.2. *For any δ, the solution of $Q_\delta(D^{(x)}) = \tau$ is the δ-quantile of $\overline{D}^{(\tau)}$.*

PROOF. Consider the distribution $D^{(z)}$ such that $Q_\delta(D^{(z)}) = \tau$. Consider a parametric sample s. From the monotonicity of s we have that the solution to $s(x) = \tau$ is $\geq z$ if and only if $s(z) \geq \tau$. Similarly we have that the solution to $s(x) = \tau$ is $\leq z$ if and only if $s(z) \leq \tau$. Since $s^1(z)$, $s^2(z)$ are independent the lemma follows. □

The quantile method for approximately solving equations of the form $\text{PR}\{y \preceq \tau \mid y \in D^{(x)}\} = \delta$ draws multiple samples from $\overline{D}^{(\tau)}$ and returns the δ-quantile of the set of samples. We apply the quantile method to approximately solve equations of the form of Eq. (4). The family of distributions that we consider is $D^{(x)} = v(x, s_0, \ldots, s_h)$. This family has the monotonicity property with respect to any $\tau > 0$. A parametric sample $s(x)$ from $v(x, s_0, \ldots, s_h)$ is obtained by drawing $h+1$ independent random variables $v_0, \ldots, v_h$ from $U[0,1]$. The parametric sample is $s(x) = \sum_{j=0}^{h} -\ln v_h/(x - s_j)$ and is a monotone decreasing function of x. A sample from $\overline{D}^{(\tau)}$ is then the solution of the equation $\sum_{j=0}^{h} -\ln v_h/(x-s_j) = \tau$. Since $s(x)$ is monotone, the solution can be found using standard search.

7.2 Confidence bounds for wsr sketches

The WSR estimator on the total weight is the average of the k minimum ranks which are independent exponential random variables with (the same) parameter $w(I)$. (This is a Gamma distribution.) We used the normal approximation to this distribution in order to compute WSR confidence bounds. The expectation of the sum is $k/w(I)$ and the variance is $k/w(I)^2$. The confidence bounds are the δ and $1-\delta$ quantiles of $\overline{r}$. Let α be the Z-value that corresponds to confidence level $1-\delta$ in the standard normal distribution. By applying the normal approximation, the approximate upper bound is the solution of $k/w(I) + \alpha\sqrt{k/w(I)^2} = k\overline{r}$, and the approximate lower bound is the solution of $k/w(I) - \alpha\sqrt{k/w(I)^2} = k\overline{r}$. Therefore, the approximate bounds are $(1 \pm \alpha/\sqrt{k})/\overline{r}$.

7.3 Confidence bounds for priority sketches

We review Thorup's confidence bounds for PRI sketches [37], which we implemented and included in our evaluation. We denote $p_\tau(i) = \text{PR}\{r(i) < \tau\}$. The number of items in $J \cap s$ with $p_\tau(i) < 1$ is used to bound $\sum_{i \in J | p_\tau(i) < 1} p_\tau(i)$ (the expectation of the sum of independent Poisson trials). These bounds are then used to obtain bounds on the weight $\sum_{i \in J | p_\tau(i) < 1} w(i)$, exploiting the correspondence (specific for PRI sketches) between $\sum_{i \in J | p_\tau(i) < 1} p_\tau(i)$ and $\sum_{i \in J | p_\tau(i) < 1} w(i)$: For PRI sketches, $p_\tau(i) = \min\{1, w(i)\tau\}$. If $w(i)\tau \geq 1$ then $p_\tau(i) = 1$ (item is included in the sketch) and if $w(i)\tau < 1$ then $p_\tau(i) = w(i)\tau$. Therefore, $p_\tau(i) < 1$ if and only if $p_\tau(i) = w(i)\tau$ and

$$\sum_{i \in J | p_\tau(i) < 1} w(i) = \tau^{-1} \sum_{i \in J | p_\tau(i) < 1} p_\tau(i) \ .$$

For $n' \geq 0$, define $\overline{n}_\delta(n')$ (respectively, $\underline{n}_\delta(n')$) to be the infimum (respectively, supremum) over all μ, such that for all sets of independent Poisson trials with sum of expectations μ, the sum is less than δ likely to be at most n' (respectively, at least n'). If $n' = |\{i \in J \cap s | w(i)\tau < 1\}|$, then $\underline{n}_\delta(n')$ and $\overline{n}_\delta(n')$ are $(1-\delta)$-confidence bounds on $\sum_{i \in J \cap s | w(i)\tau < 1} p_\tau(i)$. Since

$$w(J) = \sum_{i \in J \cap s | w(i)\tau \geq 1} w(i) + \tau^{-1} \sum_{i \in J \cap s | w(i)\tau < 1} p_\tau(i) \ ,$$

we obtain $(1-\delta)$-confidence upper and lower bounds on $w(J)$ by substituting $\overline{n}_\delta(J)$ and $\underline{n}_\delta(J)$ for $\sum_{i \in J \cap s | w(i)\tau < 1} p_\tau(i)$ in this formula, respectively.

Chernoff bounds provide an upper bound on $\overline{n}_\delta(n')$ of $-\ln\delta$ if $n' = 0$ and the solution of $\exp(n' - x)(x/n')^{n'} = \delta$ otherwise; and a lower bound on $\underline{n}_\delta(n') \leq n'$ that is the solution of $\exp(n' - x)(x/n')^{n'} = \delta$ and 0 if there is no solution.

Thorup's approach is not effective for WS sketches: a bound on the sum $\sum_{i \in J} p_\tau(i)$ does not provide a corresponding good bound on the sum of the weights of items in J. In particular, $w(i)$ can be arbitrarily large when $p(i)$ approaches 1, which precludes good upper bounds.

In contrast to our WS bounds, Thorup's PRI bounds are inefficient. One source of slack is the use of Chernoff bounds rather than exactly computing $\overline{n}_\delta(n')$ and $\underline{n}_\delta(n')$. Other sources of slack are due to the fact that the actual distribution of the sum of independent Poisson trials depends on how they are distributed. In particular, variance is higher when there are more items with smaller $p(i)$'s. An inherent source of slack is that the derivation must make "worst case" assumptions on the distribution of "unseen" items, whereas the actual variance of the estimator is lower when the weight outside the sketch is attributed to a smaller number of larger items. Another source of slack is that the derivation does not utilize the (available) weights of the items in $J \cap s$ with $w(i)\tau < 1$ and extends the worst-case assumptions to the weight attributed to these items. This discussion suggests that it might be possible to tighten these PRI bounds. Alternative PRI bounds can also be derived by specializing our general derivation to PRI sketches. The derivation and evaluation of these bounds is outside the scope of this paper.

8. EXPERIMENTAL EVALUATION

Data sets. Our evaluation included synthetic distributions that allowed us to understand performance dependence on the skew (Pareto power parameter) and real-world data sets that provided natural selection predicates for subpopulations:

- Our synthetic data sets were obtained by independently drawing $n \in \{1 \times 10^3, 2 \times 10^4\}$ items from each of uniform or Pareto distributions with power parameters $\alpha \in \{1, 1.2, 2\}$. We select subpopulations (a partition into subpopulation), using a *group size* parameter g. Items are ordered by their weights and sequentially partitioned into n/g groups (subpopulations) each consisting of g items. This partition corresponds to subpopulations of similar-size items.

- Netflix Prize [30] data set contains approximately 1×10^8 dated ratings of 17,770 movies by users. Each movie corresponds to a record with weight equal to the number of ratings. We used a natural grouping of movies into subpopulations according to ranges of movie release years (same year, 2 years, 5 years, and decades).

- Two IP packet traces of 4.2×10^9 packets (*campus*) and 4.7×10^9 packets (*peering*). Items corresponded to destination IP address and application (determined by port and protocol) pairs (7593 and 41217 distinct items). The weight of each item was the total bytes of associated packets. We used natural partition into subpopulations, based on the application type (such as web, mail, p2p, and more).

Total weight. We compare estimators and confidence bounds on the total weight $w(I)$.

Estimators. We evaluate the maximum likelihood WS estimator (WS ML) (Section 4), the rank conditioning WS estimator (WS RC) (Section 5), the rank conditioning PRI estimator (PRI RC) [1](Section 5) , and the WSR estimator [6] (Section 3).

Figure 1 (left) shows the (absolute value) of the relative error, averaged over 1000 runs, as a function of k. We can see that all three bottom-k based estimators outperform the WSR estimator, demonstrating the advantage of the added information when sampling "without replacement" over sampling "with replacement" (see also [15]). The advantage of these estimators grows with the skew. The quality of the estimate is similar among the bottom-k estimators (WS ML, WS RC, and PRI RC). The maximum likelihood estimator (WS ML), which is biased, has worse performance for very small values of k where the bias is more significant. PRI RC has a slight advantage especially if the distribution is more skewed. This is because, in this setting, with unknown $w(I)$, PRI RC is a nearly optimal adjusted-weight based estimator [36].

Confidence bounds. We compare the Chernoff based PRI confidence bounds from [37] (Section 7.3) and the WS (Section 7.1) and WSR (Section 7.2) confidence bounds we derived. We apply the normal approximation with the stricter (but easier to compute) conditioning on the order for the WS confidence bounds and the normal approximation for the WSR confidence bounds (see Section 7.1). The 95%-confidence upper and lower bounds and the 90% confidence interval (*the width*, which is the difference between the upper and lower bounds), averaged over 1000 runs, are shown in Figure 1 (middle and right). We can see that the WS confidence bounds are tighter, and often significantly so, than the PRI confidence bounds. In fact PRI confidence bounds were worse than the WSR-based bounds on less-skewed distributions (including the uniform distributions). This perhaps surprising behavior is explained by the efficiency of our WS and WSR bounds and the inefficiency of the bounds in [37] (see discussion in Section 7.3).

The normal approximation provided fairly accurate confidence bounds for the total weight. The WS and WSR bounds were evidently more efficient, with actual errors closely corresponding to the desired confidence level. E.g., for the 90% confidence interval on all Pareto distributions, across values of k, the highest error rate was 12%. The true weight was within the WS confidence bounds on average in 90.5%, 90.2%, 90% of the time for the different values of α. The corresponding in-bounds rates for WSR were 90.6%, 90.3%, and 90.0%, and for PRI 99.2%, 99.1%, and 98.9%. (The high in-bounds rate for the PRI bounds reflects the inefficiency of these bounds).

Subpopulation weight. **Estimators.** We implemented an approximate version of WS SC using the Markov chain and averaging method (Section 6.1). We showed that this approximation provides unbiased estimators that are better than the WS RC estimator (better per-item variances and negative covariances for different items), but attains zero sum of covariances only at the limit. Here we quantify the benefit of WS SC over WS RC and its dependence on the size of the subpopulation. We also evaluate the quality of approximate WS SC as a function of the parameters INPERM, and PERMNUM, and compare WS SC to the PRI RC estimator.

It is always possible to design AW-summaries that artificially favor a particular subpopulation. Therefore, to obtain a meaningful comparison, we consider all subpopulations defined by a *partition* of the items. For such a partition, we compute the sum, over subpopulations, of the square error of the estimator (square of the difference between the adjusted weight and the true weight of the subpopulation), averaged over multiple runs. This sum corresponds to the sum of the variances of the estimator over the subpopulations. We considered parameterized partitions by the group size g for the Pareto distributions. For the Netflix data, we partitioned the movies according to ranges of release years. For the IP packets data, we used a fixed partition according to application type of the IP flow.

To evaluate how the quality of the estimators varies with subpopulation size, we sweep the parameter g for the Pareto distributions. The RC estimators have zero covariances, and therefore, the sum of square errors should remain constant when sweeping g. The WS SC estimator has negative covariances and therefore we expect the sum to decrease as a function of g. For $g = 1$, this sum corresponds to the sum of the variances of the items which should be the same for the WS estimators and smaller for the PRI estimator. The sum of square errors, as a function of g, is constant for the RC estimators, but decreases with the WS SC estimator. For $g = n$, we obtain the variance of the sum of the adjusted weights, which should be 0 for the WS SC estimator (but not for the approximate versions).

Representative results are shown in Figure 3 (Pareto distributions) and in Figure 2 (the Netflix and IP packets data sets).

We observed that for $g = 1$, the PRI RC estimator (that obtains the minimum sum of per-item variances by a sketch of size $k + 1$) performs slightly better than the WS RC estimator when the data is more skewed (smaller α). The performance of WS SC is close and better for small values of k (it uses one fewer sample). For $g > 1$, the WS SC estimator outperforms both RC estimators and has significantly smaller variance for larger subpopulations. We similarly observe that on the Netflix and IP packet data, PRI RC was slightly better than WS RC and WS SC on smaller subpopulations and WS SC was significantly better than the RC estimators on larger subpopulations (25%-50% smaller variance).

We conclude that in applications when $w(I)$ is provided, the WS SC estimator is a considerably better choice than the RC estimators. Our results also demonstrate that the metric of the sum of per-item variances, that PRI RC is nearly optimal [36] with respect to it, is not a sufficient notion of optimality for subpopulation weight estimators. It must be augmented with comparison of covariances, making their sum as small as possible.

Figure 4 compares different choices of the parameters INPERM, and PERMNUM for the approximate (Markov chain based) WS SC estimator. We denote each such choice as a pair (INPERM, PERMNUM). We compare estimators with parameters $(400, 1)$, $(20, 20)$, $(1, 400)$, and $(5, 2)$. We conclude the following: (i) A lot of the benefit of WS SC on moderate-size subsets is obtained for small values: $(5, 2)$ performs nearly as well as the variants that use more steps and iterations. (ii) There is a considerable benefit of redrawing within a

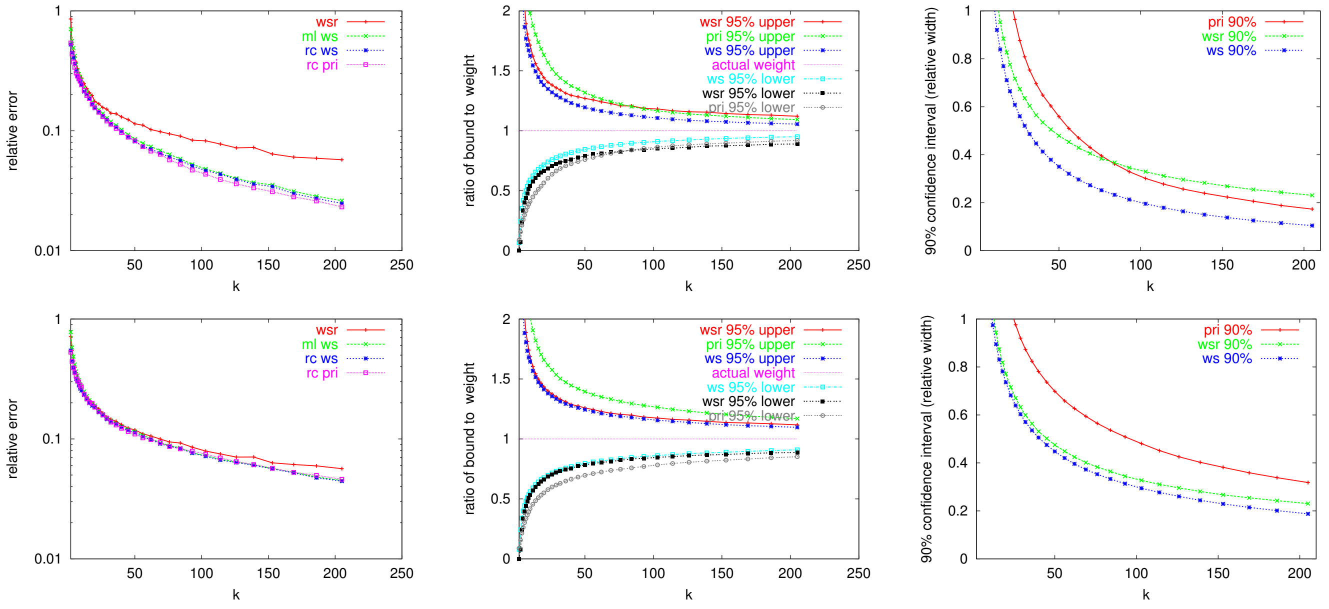


Figure 1: Left: Absolute value of the relative error of the estimator of $w(I)$ averaged over 1000 repetitions. Middle: 95% confidence upper and lower bounds for estimating $w(I)$. Right: width of 90% confidence interval for estimating $w(I)$. We show results for Pareto distributions with $n = 1 \times 10^3$, $\alpha = 1$ (top row) and $\alpha = 2$ (bottom row).

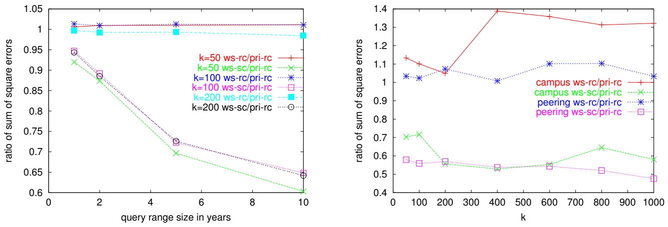


Figure 2: Ratio of the sum of square errors (averaged over 200 repetitions) over a partition, of WS RC to PRI RC and of WS SC with INPERM = 20 and PERMNUM = 20 to PRI RC. Left: The netflix request stream, groupings corresponded to movies released in the same year or range of years (same year, 2 years, 5 years, and decade), we sweep the size of the range and show $k = 50, 100, 200$. Right: IP packet streams, where items correspond to destination IP and application-type pairs. We sweep the summary size k.

permutation: $(400, 1)$ that iterates within a single permutation performs well. (iii) Larger subsets, however, benefit from larger PERMNUM: $(1, 400)$ performs better than $(20, 20)$ which in turn is better than $(400, 1)$.

Confidence bounds. We evaluate confidence bounds on subpopulation weight using the PRI Chernoff-based bounds [37] (PRI) (see Section 7.3), and the WS bounds that use $w(I)$ (WS $+w(I)$) (see Appendix E.3) or do not use $w(I)$ (WS $-w(I)$) (see Section 7.1). The WS bounds are computed using the quantile method with 200 draws from the appropriate distribution.

We consider the relative and square error of the bounds

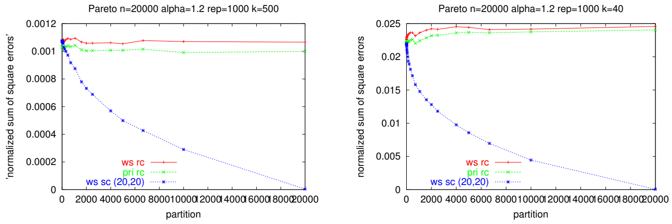


Figure 3: Normalized sum of square errors (averaged over 1000 repetitions) over a partition as a function of group size for $k = 500$ and $k = 40$ and $\alpha = 1.2$.

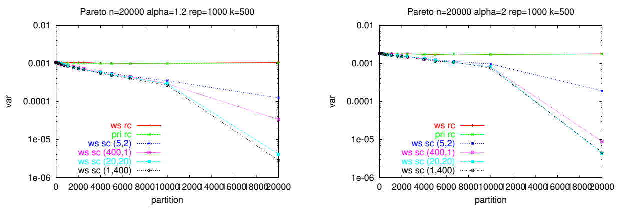
and the width of the confidence interval. The confidence bounds, intervals, and square errors, were normalized using the weight of the corresponding subpopulation. For each distribution, value of k, and partition g, the normalized bounds were averaged across 500 repetitions and across all subpopulations. Across these distributions, the WS $+w(I)$ confidence bounds are tighter (more so for larger g) than WS $-w(I)$ and both are significantly tighter than the PRI confidence bounds. Representative results are shown in Figure 5.

9. REFERENCES

[1] N. Alon, N. Duffield, M. Thorup, and C. Lund. Estimating arbitrary subset sums with few probes. In *Proceedings of the 24th ACM Symposium on Principles of Database Systems*, pages 317–325, 2005.

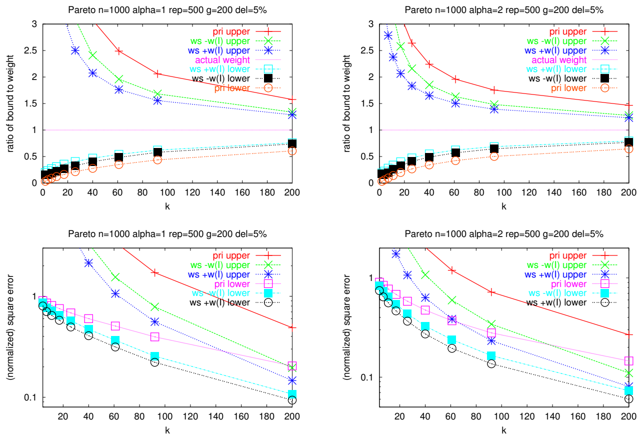
[2] K. S. Beyer, P. J. Haas, B. Reinwald, Y. Sismanis, and R. Gemulla. On synopses for distinct-value estimation under multiset operations. In *SIGMOD*, pages 199–210. ACM, 2007.

[3] K. Bharat and A. Z. Broder. Mirror, mirror on the web: A study of host pairs with replicated content. In *Proceedings of the 8th International World Wide Web Conference (WWW)*, pages 501–512, 1999.



$\alpha = 1.2$ $\alpha = 2$

Figure 4: Sum of variances in a partition for $k = 500$ as a function of group size for different combinations of INPERM and PERMNUM.



Pareto $n = 1000$, $\alpha = 1$ Pareto $n = 1000$,$\alpha = 2$

Figure 5: Subpopulation 95% confidence bounds (top) and (normalized) squared error of the 95% confidence bounds (bottom) for $g = 200$.

[4] A. Z. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*, pages 21–29. ACM, 1997.

[5] A. Z. Broder. Identifying and filtering near-duplicate documents. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, volume 1848 of *LLNCS*, pages 1–10. Springer, 2000.

[6] E. Cohen. Size-estimation framework with applications to transitive closure and reachability. *J. Comput. System Sci.*, 55:441–453, 1997.

[7] E. Cohen, N. Duffield, H. Kaplan, C. Lund, and M. Thorup. Algorithms and estimators for accurate summarization of Internet traffic. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (IMC)*, 2007.

[8] E. Cohen, N. Duffield, H. Kaplan, C. Lund, and M. Thorup. Sketching unaggregated data streams for subpopulation-size queries. In *Proc. of the 2007 ACM Symp. on Principles of Database Systems (PODS 2007)*. ACM, 2007.

[9] E. Cohen, N. Duffield, H. Kaplan, C. Lund, and M. Thorup. Variance optimal sampling based estimation of subset sums. Technical Report cs.DS/0803.0473, Computing Research Repository (CoRR), 2008.

[10] E. Cohen, N. Duffield, C. Lund, M. Thorup, and H. Kaplan. Summarization framework for unaggregated data. Submitted, 2008.

[11] E. Cohen and H. Kaplan. Efficient estimation algorithms for neighborhood variance and other moments. In *Proc. 15th ACM-SIAM Symposium on Discrete Algorithms*. ACM-SIAM, 2004.

[12] E. Cohen and H. Kaplan. Spatially-decaying aggregation over a network: model and algorithms. In *SIGMOD*. ACM, 2004.

[13] E. Cohen and H. Kaplan. Bottom-k sketches: Better and more efficient estimation of aggregates. In *Proceedings of the ACM SIGMETRICS'07 Conference*, 2007. poster.

[14] E. Cohen and H. Kaplan. Spatially-decaying aggregation over a network: model and algorithms. *J. Comput. System Sci.*, 73:265–288, 2007.

[15] E. Cohen and H. Kaplan. Summarizing data using bottom-k sketches. In *Proceedings of the ACM PODC'07 Conference*, 2007.

[16] E. Cohen and H. Kaplan. Estimating aggregates over multiple subsets. Manuscript, 2008.

[17] E. Cohen and H. Kaplan. Sketch-based estimation of subpopulation-weight. Technical Report 802.3448, CORR, 2008.

[18] E. Cohen and M. Strauss. Maintaining time-decaying stream aggregates. In *Proc. of the 2003 ACM Symp. on Principles of Database Systems (PODS 2003)*. ACM, 2003.

[19] E. Cohen, Y.-M. Wang, and G. Suri. When piecewise determinism is almost true. In *Proc. Pacific Rim International Symposium on Fault-Tolerant Systems*, pages 66–71, Dec. 1995.

[20] T. Dasu, T. Johnson, S. Muthukrishnan, and V. Shkapenyuk. Mining database structure; or, how to build a data quality browser. In *SIGMOD Conference*, pages 240–251, 2002.

[21] N. Duffield, M. Thorup, and C. Lund. Learn more, sample less, control of volume and variance in network measurement. *IEEE Transactions in Information Theory*, 51(5):1756–1775, 2005.

[22] N. Duffield, M. Thorup, and C. Lund. Priority sampling for estimating arbitrary subset sums. *J. Assoc. Comput. Mach.*, 54(6), 2007.

[23] P. S. Efraimidis and P. G. Spirakis. Weighted random sampling with a reservoir. *Inf. Process. Lett.*, 97(5):181–185, 2006.

[24] P. Gibbons and Y. Matias. New sampling-based summary statistics for improving approximate query answers. In *SIGMOD*. ACM, 1998.

[25] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[26] M. Hua, J. Pei, A. W. C. Fu, X. Lin, and H.-F. Leung. Efficiently answering top-k typicality queries on large databases. In *Proceedings of the 33rd VLDB Conference*, 2007.

[27] H. Kaplan and M. Sharir. Randomized incremental constructions of three-dimensional convex hulls and planar voronoi diagrams, and approximate range counting. In *SODA '06: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 484–493, New York, NY, USA, 2006. ACM Press.

[28] D. Mosk-Aoyama and D. Shah. Computing separable functions via gossip. In *Proceedings of the ACM PODC'06 Conference*, 2006.

[29] R. Motwani, E. Cohen, M. Datar, S. Fujiware, A. Gronis, P. Indyk, J. Ullman, and C. Yang. Finding interesting associations without support pruning. *IEEE Transactions on Knowledge and Data Engineering*, 13:64–78, 2001.

[30] The Netflix Prize. `http://www.netflixprize.com/`.

[31] Cisco NetFlow. `http://www.cisco.com/warp/public/732/Tech/netflow`.

[32] S. Sampath. *Sampling Theory and Methods*. CRC press, 2000.

[33] D. W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley & Sons, New York, 1992.

[34] R. Singh and N. S. Mangat. *Elements of survey sampling*. Springer-Verlag, New York, 1996.

[35] N. T. Spring and D. Wetherall. A protocol-independent technique for eliminating redundant network traffic. In *Proceedings of the ACM SIGCOMM'00 Conference*. ACM, 2000.

[36] M. Szegedy. The DLT priority sampling is essentially optimal. In *Proc. 38th Annual ACM Symposium on Theory of Computing*. ACM, 2006.

[37] M. Thorup. Confidence intervals for priority sampling. In *ACM SIGMETRICS Performance Evaluation Review*, 2006.

APPENDIX

A. ML ESTIMATORS FOR WS SKETCHES

A general observation for our ML estimators is that tighter estimators can be obtained by redrawing the rank values of the items $i_1, \ldots, i_k$ (see Section 3) and taking the expectation (or average over multiple draws) of the solution of Eq. (1) over the corresponding permutations of the first k items.

Estimating a subpopulation weight.

We derive maximum likelihood subpopulation weight estimators that use and do not use the total weight $w(I)$. Let $J \subset I$ be a subpopulation. Let $j_1, \ldots, j_a$ be the items in s that are in $I \setminus J$.[7] Let $r'_1, \ldots, r'_a$ be their respective rank values and let $s'_i = \sum_{h<i} w(j_h)$ $(i = 1, \ldots, a)$. Define $s'_0 \equiv 0$. Let $i_1, i_2, \ldots, i_c$ be the items in $J \cap s$. Let $r_1, \ldots, r_c$ be their respective rank values and let $s_i = \sum_{h<i} w(i_h)$ $(i = 1, \ldots, c)$. Define $s_0 \equiv 0$.

WS ML **subpopulation weight estimator that does not use $w(I)$:** Consider rank assignments such that rank values in $I \setminus J$ are fixed and the order of ranks of the items in J is fixed. The probability density of the observed ranks of the first k items in J is that of seeing the same rank differences (probability density is $(w(J) - s_i)\exp(-(w(J) - s_i)(r_{i+1} - r_i)$ for the ith difference) and of the rank difference between the $c+1$ and c smallest ranks in J being at least $\tau - r_c$ (where τ is the $(k+1)$st smallest rank in the sketch), which is $\exp(-(w(J) - s_c)(\tau - r_c))$. Rank differences are independent, and therefore, the probability density as a function of $w(J)$ is the product of the above densities. The maximum likelihood estimator for $w(J)$ is the value that maximizes this probability. If $c = 0$, the expression $\exp(-w(J)\tau)$ is maximized for $w(J) = 0$. Otherwise, by taking the natural logarithm and deriving we find that the value of $w(J)$ that maximizes the probability density is the solution of $\sum_{h=0}^{c-1} \frac{1}{\tilde{w}(J) - s_h} = \tau$. As with the estimator of the total weight, we can obtain a tighter estimator by redrawing the rank values.

WS ML **subpopulation weight estimator that uses $w(I)$:** We compute the probability density, as a function of $w(J)$, of the event that we obtain the sketch s with these ranks given that the prefix of sampled items from $I \setminus J$ is $j_1, \ldots, j_a$ and the prefix of sampled items from J is $i_1, \ldots, i_c$. We take the natural logarithms of the joint probability density and derive with respect to $w(J)$. If $c = 0$, the derivative is positive and the probability density is maximized for $w(J) = 0$. If $a = 0$, the derivative is negative and the probability density is maximized for $w(J) = w(I)$. Otherwise, if $a > 0$ and $c > 0$, the probability density is maximized for $\tilde{w}(J)$ that is the solution of

$$\sum_{h=0}^{c-1} \frac{1}{\tilde{w}(J) - s_h} - \sum_{h=0}^{a-1} \frac{1}{(w(I) - \tilde{w}(J)) - s'_h} = 0 \ .$$

The equation is easy to solve numerically, because the left hand side is a monotone decreasing function of $w(J)$.

B. ADJUSTED WEIGHTS

LEMMA B.1. *(Lemma 5.1) Consider* RC *adjusted weights and two items i and j. Then,* $\textsc{cov}[a(i), a(j)] = 0$.

PROOF. It suffices to show that $\mathsf{E}[a(i)a(j)] = w(i)w(j)$. Consider a partition of the sample space of all rank assignments according to the $(k-1)$th smallest rank of an item in $I \setminus \{i, j\}$.[8] Consider a subset in the partition and let r_{k-1} denote the value of the $(k-1)$th smallest rank of an item in $I \setminus \{i, j\}$ for rank assignments in this subset. We show that in this subset $\mathsf{E}[a(i)a(j)] = w(i)w(j)$. The product $a(i)a(j)$ is positive in this subset only when $r(i) < r_{k-1}$ and $r(j) < r_{k-1}$, which (since rank assignments are independent) happens with probability $\textsc{pr}\{r(i) < r_{k-1}\}\textsc{pr}\{r(j) < r_{k-1}\}$. In this case the kth smallest rank in $I \setminus \{i\}$ and $I \setminus \{j\}$ is r_{k-1} and therefore, $a(i) = \frac{w(i)}{\textsc{pr}\{r(i) < r_{k-1}\}}$, and $a(j) = \frac{w(j)}{\textsc{pr}\{r(j) < r_{k-1}\}}$. It follows that

$$\begin{aligned}&\mathsf{E}[a(i)a(j)] = \\ &\textsc{pr}\{r(i) < r_{k-1}\}\textsc{pr}\{r(j) < r_{k-1}\}\frac{w(i)}{\textsc{pr}\{r(i) < r_{k-1}\}}\frac{w(j)}{\textsc{pr}\{r(j) < r_{k-1}\}} \\ &= w(i)w(j) \ .\end{aligned}$$

□

We can extend the proof of Lemma 5.1 to show that for any subset $J \subset I$, $\mathsf{E}[\prod_{i \in I} a(i)] = \prod_{i \in I} w(i)$.

LEMMA B.2. *(Lemma 5.3) Consider two partitions of the sample space, such that one partition is a refinement of the other, and the* AW*-summaries obtained by applying* HTP *using these partitions. For each $i \in I$, the variance of $a(i)$ using the coarser partition is at most that of the finer partition.*

PROOF. We use the following simple property of the variance. Consider two random variables A_1 and A_2 over a probability space Ω. Suppose that there is a partition $\{B_j\}$ of Ω such that for every B_j, and for every $s \in B_j$, $A_2(s) = \mathsf{E}[A_1(s) | s \in B_j]$. Then $\textsc{var}[A_2] \leq \textsc{var}[A_1]$.

Let P^i_j be the sets in the fine partition, and let C^i_ℓ be the sets in the coarse partition such that $C^i_\ell = \bigcup_t P^i_{\ell t}$. Let $\overline{P}^i_j$ be the subset containing all $s \in P^i_j$ such that $i \in s$. Similarly, let $\overline{C}^i_\ell$ be the subset containing all $s \in C^i_\ell$ such that $i \in s$. Let $a(i, s)$ be the adjusted weight of i in a sketch s according to the partition P^i_j, and let $\overline{a}(i, s)$ be the adjusted weight of i in a sketch s according to the partition C^i_ℓ. We will show that for $s \in \overline{C}^i_\ell$ such that $i \in s$, $\overline{a}(i, s) = \mathsf{E}_{s' \in \overline{C}^i_\ell}[a(i, s')]$. From this and the property of the variance stated above the lemma follows. We remove the superscript i from the sets P^i_j, C^i_ℓ, $\overline{P}^i_j$, and $\overline{C}^i_\ell$ in the rest of the proof.

Let $p_j = \textsc{pr}(s \in \overline{P}_j \mid s \in P_j)$ and $\overline{p}_\ell = \textsc{pr}(s \in \overline{C}_j \mid s \in C_\ell)$. Now,

$$\begin{aligned}\mathsf{E}_{s' \in C^i_\ell}[a(i, s')] &= \frac{\sum_t \textsc{pr}(s \in \overline{P}_{\ell_t})\frac{w(i)}{p_{\ell_t}}}{\textsc{pr}(s \in \overline{C}_\ell)} \\ &= \frac{\sum_t \textsc{pr}(s \in P_{\ell_t})p_{\ell_t}\frac{w(i)}{p_{\ell_t}}}{\textsc{pr}(s \in C_\ell)\overline{p}_\ell} \\ &= \frac{w(i)\sum_t \textsc{pr}(s \in P_{\ell_t})}{\textsc{pr}(s \in C_\ell)\overline{p}_\ell} \\ &= \frac{w(i)}{\overline{p}_\ell} = \overline{a}(i, s) \ .\end{aligned}$$

□

[7] We assume that using meta attributes of items in the sketch we can decide which among them are in J.

[8] We can use a finer partitions in which all the ranks in $I \setminus \{i, j\}$ are fixed.

C. PREFIX CONDITIONING ESTIMATOR

For an item $i \in s$ we partition the sample space according to the sequence (prefix) of $k-1$ items with smallest ranks drawn from $I \setminus \{i\}$. That is if $i \notin s$, then s belongs to the partition associated with the $k-1$ items in s of smallest ranks. If $i \in s$, then s belongs to the partition associated with the sequence of $k-1$ items in $s \setminus \{i\}$.

We assign adjusted weights as follows. Consider a sketch s and $i \in s$. Let P be the set of sketches with the same prefix of $k-1$ items from $I \setminus \{i\}$ as in s. We compute the probability $\textsc{pr}\{i \in s \mid s \in P\}$, that is, the probability that i is in a sketch from P. We compute the probability of i occurring in each of the positions $j \in 1, \ldots, k$ and the probability that it does not occur at all. We use the notation $\textsc{pfx}_J(j_1, \ldots, j_k)$ for the event that the first k items drawn by weighted sampling without replacement from a subset J are $j_1, \ldots, j_k$.

We denote by i_ℓ $(1 \le \ell \le k-1)$ the ℓth item in $s \setminus \{i\}$. For each $j = 1, \ldots, k$, the probability e_j that i appears in the jth position in a sketch from P is

$$p(i \rightarrow j \cap s \in P) = \textsc{pr}\{\textsc{pfx}_I(i_1, i_2, i_{j-1}, i, i_j, i_{k-1})\} =$$
$$\frac{w(i_1)}{w(I)} \frac{w(i_2)}{w(I) - w(i_1)} \frac{w(i_{j-1})}{w(I) - \sum_{m=1}^{j-2} w(i_m)} \frac{w(i)}{w(I) - \sum_{m=1}^{j-1} w(i_m)}$$
$$\frac{w(i_j)}{w(I) - \sum_{m=1}^{j-1} w(i_m) - w(i)} \cdots \frac{w(i_{k-1})}{w(I) - \sum_{m=1}^{k-2} w(i_m) - w(i)} .$$

The probability that the sketch is from P but i does not appear in it (technically, appears in a position $k+1$ or beyond) is

$$p(i \notin s \cap s \in P) = \textsc{pr}\{ \bigcup_{\ell \in \{I \setminus s\}} \textsc{pfx}(i_1, i_2, \ldots, i_{k-1}, \ell)\} =$$
$$\frac{w(i_1)}{w(I)} \frac{w(i_2)}{w(I) - w(i_1)} \cdots \frac{w(i_{k-1})}{w(I) - \sum_{m=1}^{k-2} w(i_m)}$$
$$\frac{w(I) - w(i) - \sum_{m=1}^{k-1} w(i_m)}{w(I) - \sum_{m=1}^{k-1} w(i_m)} .$$

Therefore,

$$\textsc{pr}\{i \in s \mid s \in P\} = \frac{\sum_{j=1}^{k} p(i \rightarrow j \cap s \in P)}{\sum_{j=1}^{k} p(i \rightarrow j \cap s \in P) + p(i \notin s \cap s \in P)} .$$

The computation of the prefix conditioning adjusted weights is quadratic in k for each item i. RC adjusted weights, on the other hand, can be computed in constant number of operations per item.

D. SUBSET CONDITIONING

Lemma D.1. *Let s be a* WS *sketch of I and let $a(i)$ be* SC *adjusted weights. Then, $\sum_{i \in s} a(i) = w(I)$.*

Proof. Observe that for any sketch s, $i \in s$, and $\ell \ge 0$

$$f(s, \ell) = f(s \setminus \{i\}, \ell) - f(s \setminus \{i\}, \ell + w(i)) \frac{\ell}{\ell + w(i)} . \quad (5)$$

This relation follows by manipulating Eq. (2), or by the following argument: Let $X = I \setminus s$ and $w(X) = \ell$. The probability that the items with smallest ranks in $s \cup X$ are the items in s is equal to the probability that the $|s|-1$ items of smallest ranks in $(s \setminus \{i\}) \cup X$ are $s \setminus \{i\}$ minus the probability that the $|s|-1$ items of smallest ranks in $s \cup X$ are $s \setminus \{i\}$ and the $|s|$th smallest rank is from $X \setminus \{i\}$. This latter probability is equal to

$$f(s \setminus \{i\}, w(X \cup \{i\})) \frac{\ell}{\ell + w(i)} .$$

Using Equation (5) we obtain that

$$\begin{aligned}
\sum_{i \in s} a(i) &= \\
&= \frac{\sum_{i \in s} w(i) f(s \setminus \{i\}, w(I \setminus s))}{f(s, w(I \setminus s))} \\
&= \frac{\sum_{i \in s} w(i)(f(s, w(I \setminus s)) + f(s \setminus \{i\}, w(I \setminus \{s \setminus \{i\}\})) \frac{w(I \setminus s)}{w(i) + w(I \setminus s)})}{f(s, w(I \setminus s))} \\
&= \sum_{i \in s} w(i) + w(I \setminus s) \sum_{i \in s} \frac{\frac{w(i)}{w(i) + w(I \setminus s)} f(s \setminus \{i\}, w(I \setminus \{s \setminus \{i\}\}))}{f(s, w(I \setminus s))} \\
&= w(I) .
\end{aligned}$$

To verify the last equality, observe that

$$\frac{w(i)}{w(i) + w(I \setminus s)} f(s \setminus \{i\}, w(I \setminus \{s \setminus \{i\}\}))$$

is the probability that the first $|s|-1$ items drawn from I are $s \setminus \{i\}$ and the $|s|$th item is i. These are disjoint events and their union is the event that the first $|s|$ items drawn from I are s. The probability of this union is $f(s, w(I \setminus s))$. □

Lemma D.2. *Consider* SC *adjusted weights of two items $i \ne j$. Then,* $\textsc{cov}[a(i), a(j)] < 0$.

Proof. Consider a partition of rank assignments according to the items in $I \setminus \{i, j\}$ that have the $k-2$ smallest ranks. Consider a part in this partition and denote this set of $k-2$ items by c. We compute the expectation of $a(i)a(j)$ conditioned on this part. Let $\ell = w(I) - w(c) - w(i) - w(j)$. The probability of this part is $f(c, \ell)$, the probability that $a(i)a(j) > 0$ in c is equal to $f(c \cup \{i, j\}, \ell)$. Therefore, the conditional probability is $\frac{f(c \cup \{i,j\}, \ell)}{f(c, \ell)}$. In this case, the adjusted weight assigned to i is set according to items $c \cup \{j\}$ having the $(k-1)$ smallest ranks in $I \setminus \{i\}$. Therefore, this weight is $a(i) = w(i) \frac{f(c \cup \{j\}, \ell)}{f(c \cup \{i,j\}, \ell)}$. Symmetrically for j, $a(j) = w(j) \frac{f(c \cup \{i\}, \ell)}{f(c \cup \{i,j\}, \ell)}$. We therefore obtain that $\mathsf{E}[a(i)a(j)]$ conditioned on this part is

$$w(i) w(j) \frac{f(c \cup \{j\}, \ell) f(c \cup \{i\}, \ell)}{f(c \cup \{i,j\}, \ell) f(c, \ell)} .$$

It suffices to show that

$$\frac{f(c \cup \{j\}, \ell) f(c \cup \{i\}, \ell)}{f(c \cup \{i,j\}, \ell) f(c, \ell)} \le 1 .$$

To show that, we apply Eq. (5) and substitute in the numerator $f(c \cup \{j\}, \ell) = f(c, \ell) - f(c, \ell + w(j)) \frac{\ell}{\ell + w(j)}$ and in the denominator $f(c \cup \{i, j\}, \ell) = f(c \cup \{i\}, \ell) - f(c \cup \{i\}, \ell + w(j)) \frac{\ell}{\ell + w(j)}$ The numerator being at most the denominator therefore follows from the immediate inequality

$$f(c, \ell) f(c \cup \{i\}, \ell + w(j)) \le f(c, \ell + w(j)) f(c \cup \{i\}, \ell) .$$

□

Lemma D.3. *Consider* WS *sketches of a weighted set (I, w) and subpopulation $J \subset I$. The* SC *estimator for the weight of J has smaller variance than the* RC *estimator for the weight of J.*

PROOF. By Lemma 5.1 the variance of the RC estimator for J is $\sum_{j\in J} \textsc{var}_{\mathrm{RC}}[a(j)]$. So using Lemma 5.3 we obtain that $\sum_{j\in J} \textsc{var}_{\mathrm{SC}}[a(j)]$ is no larger than the variance of the RC estimator for J. Finally since

$$\textsc{var}_{\mathrm{SC}}[\sum_{j\in J} a(j)] = \sum_{j\in J} \textsc{var}_{\mathrm{SC}}[a(j)] + \sum_{i\neq j, i,j\in J} \textsc{cov}_{\mathrm{SC}}[a(i),a(j)] \ ,$$

and Lemma D.2 that implies that the second term is negative the lemma follows. □

E. CONFIDENCE BOUNDS

We provide derivations omitted from Section 7.

E.1 Total weight

Let r be a rank assignment of a weighted set $Z = (H, w)$. Recall that for $H' \subseteq H$, $r(H')$ is the minimum rank of an item in H'. In this section it will be useful to denote by $\overline{r}(H')$ the maximum rank of an item in H'. We define $r(\emptyset) = +\infty$ and $\overline{r}(\emptyset) = 0$. For a distribution D over a totally ordered set (by $\prec$) and $0 < \alpha < 1$, we denote by $Q_\alpha(D)$ the α-quantile of D. That is, $\textsc{pr}_{y\in D}\{y \preceq Q_\alpha(D)\} \leq \alpha$ and $\textsc{pr}_{y\in D}\{y \succeq Q_\alpha(D)\} \geq 1-\alpha$. [9]

For two weighted sets $Z_1 = (H_1, w_1)$ and $Z_2 = (H_2, w_2)$, let $\Omega(Z_1, Z_2)$ be the probability subspace that contains all rank assignments r over $Z_1 \cup Z_2$ such that $\overline{r}(H_1) < r(H_2)$.[10]

Let (I, w) be a weighted set, let r be a rank assignment for (I, w), and let $s = s(r)$ be the bottom-k sketch that corresponds to r (we also use s as the set of k items with smallest ranks). Let $\overline{W}((s,w), r_{k+1}, \delta)$ be the set containing all weighted sets $Z' = (H, w')$ such that $\textsc{pr}\{r'(H) \geq r_{k+1} \mid r' \in \Omega((s,w), Z')\} \geq \delta$. Define $\overline{w}((s,w), r_{k+1}, \delta)$ as follows. If $\overline{W}((s,w), r_{k+1}, \delta) = \emptyset$, then $\overline{w}((s,w), r_{k+1}, \delta) = 0$. Otherwise, let $\overline{w}((s,w), r_{k+1}, \delta) = \sup\{w'(H) \mid (H, w') \in \overline{W}((s,w), r_{k+1}, \delta)\}$. (This supremum is well defined for "reasonable" families of rank functions, otherwise, we allow it to be $+\infty$)

Let $\underline{W}((s,w), r_{k+1}, \delta)$ be the set of all weighted sets $Z' = (H, w')$ such that $\textsc{pr}\{r'(H) \leq r_{k+1} \mid r' \in \Omega((s,w), Z')\} \geq \delta$. Define $\underline{w}((s,w), r_{k+1}, \delta)$ as follows. We have $\underline{W}((s,w), r_{k+1}, \delta) \neq \emptyset$ for "reasonable" families of rank functions, but if it is empty, we define $\underline{w}((s,w), r_{k+1}, \delta) = +\infty$. Otherwise, let $\underline{w}((s,w), r_{k+1}, \delta) = \inf\{w'(H) | (H, w') \in \underline{W}((s,w), r_{k+1}, \delta)\}$. (This infimum is well defined since weighted sets have nonnegative weights.)

LEMMA E.1. *Let r be a rank assignment for the weighted set (I, w), and let s be the bottom-k sketch that corresponds to r Then $w(s) + \overline{w}((s,w), r_{k+1}, \delta)$ is a $(1-\delta)$-confidence upper bound on $w(I)$, and $w(s) + \underline{w}((s,w), r_{k+1}, \delta)$ is a $(1-\delta)$-confidence lower bound on $w(I)$.*

PROOF. We prove (1). The proof of (2) is analogous.

We show that in each subspace $\Omega((s,w), (I \setminus s, w))$ of rank assignments our bound is correct with probability $1-\delta$. Since these subspaces, specified by $s \subset I$ of size $|s| = k$, form a partition of the rank assignments over (I, w), the lemma follows.

Let D_{k+1} be the distribution of the $(k+1)$st smallest rank over rank assignments in $\Omega((s,w), (I \setminus s, w))$ (the smallest rank in $I \setminus s$). Assume that r is a rank assignment in $\Omega((s,w), (I \setminus s, w))$. We show that if $r_{k+1} \leq Q_{1-\delta}(D_{k+1})$ then our upper bound is correct. Since by the definition of a quantile $r_{k+1} \leq Q_{1-\delta}(D_{k+1})$ with probability $\geq (1-\delta)$ in $\Omega((s,w), (I \setminus s, w))$, it follows that our bound is correct with probability $\geq (1-\delta)$ in $\Omega((s,w), (I \setminus s, w))$.

If $r_{k+1} \leq Q_{1-\delta}(D_{k+1})$ then

$$\begin{aligned} \textsc{pr}\{r'(I \setminus s) \geq r_{k+1} \mid r' \in \Omega((s,w), (I \setminus s, w))\} &\geq \\ \textsc{pr}\{r'(I \setminus s) \geq Q_{1-\delta}(D_{k+1}) \mid r' \in \Omega((s,w), (I \setminus s, w))\} &\geq \delta \ . \end{aligned}$$

So we obtain that $(I \setminus s, w) \in \overline{W}((s,w), r_{k+1}, \delta)$ and therefore $w(I \setminus s) \leq \overline{w}((s,w), r_{k+1}, \delta)$. □

WS sketches: We apply Lemma E.1 for WS sketches as follows. For a weighted set (s, w), $|s| = k$, and $\ell \geq 0$, consider a weighted set U of weight $w(s) + \ell$ containing (s, w). Let y be the $(k+1)$th smallest rank value, over rank assignments over U such that the k items with smallest rank values are the elements of s. The probability density function of y is (see Section 6 and Eq. (2))

$$D(\ell, y) = \frac{\exp(-\ell y)\prod_{j\in s}(1-\exp(-w(i_j)y))}{\int_{x=0}^{\infty}\exp(-\ell x)\prod_{j\in s}(1-\exp(-w(i_j)x))dx} \qquad (6)$$

Let r_{k+1} be the observed $k+1$ smallest rank. The $(1-\delta)$-confidence upper bound is $w(s)$ plus the value of ℓ that solves the equation $\int_0^{r_{k+1}} D(\ell, y)dy = 1-\delta$. The function $\int_0^{r_{k+1}} D(\ell, y)$ is an increasing function of ℓ (the probability of the $(k+1)$st smallest rank being at most r_{k+1} is increasing with ℓ.) If $\int_0^{r_{k+1}} D(0, y)dy > 1-\delta$, then there is no solution and the upper bound is $w(s)$.

The lower bound is $w(s)$ plus the value of ℓ that solves the equation $\int_0^{r_{k+1}} D(\ell, y)dy = \delta$. If there is no solution ($\int_0^{r_{k+1}} D(0, y)dy > \delta$), then the lower bound is $w(s)$.

Remark. Lemma E.1 also holds for an ordered variant, where we consider rank assignments r (and corresponding subspaces) where the items in s appear in the same order as in r. WS bounds with this variant are provided in Section 7.1.

E.2 Subpopulation weight

The derivation of confidence bounds on the weight of a subpopulation $J \subset I$ (Section 7) is more subtle than the one for the total weight: The number of items from J that we see in the sketch can vary between 0 and k and we do not know if the $(k+1)$th smallest rank belongs to an item in J or in $I \setminus J$. Here we use the definition of $\Omega()$ given in Section 7. That is $\Omega((Z, \pi))$ is the probability subspace of rank assignments over a weight **list** Z such that the rank order of the items is π. We provide the proof of Lemma 7.1:

LEMMA E.2. *(Lemma 7.1) Let r be a rank assignment, s be the corresponding sketch, and ℓ be the weighted list $\ell = (J \cap s, w, r)$. Then $w(J \cap s) + \overline{w}(\ell, r_{k+1}, \delta)$ is a $(1-\delta)$-confidence upper bound on $w(J)$ and $w(J \cap s) + \underline{w}(\ell, r_k, \delta)$ is a $(1-\delta)$-confidence lower bound on $w(J)$.*

PROOF. We partition the space of rank assignments over (I, w) according to the ranks of items in $I \setminus J$ and the order of the ranks of the items in J. We show that the confidence bounds hold within each subspace. Fix one such subspace Φ of rank assignments. Let π denote the order of the items in J, and let $a(i)$ denote the rank of each item $i \in I \setminus$

[9] Note that the distributions we are dealing with may obtain some discrete values with positive probabilities. In such case $\textsc{pr}_{y\in D}\{y \prec Q_\alpha(D)\}$ may be strictly smaller than $\textsc{pr}_{y\in D}\{y \preceq Q_\alpha(D)\}$.

[10] Note that we use a different definition of $\Omega()$ in this section, in Section E.2, and in Section E.3.

J, which are fixed for rank assignments in Φ. Note that there is bijection between rank assignments in $\Omega((J,\pi))$ and rank assignments in Φ obtained by augmenting the rank assignment in $\Omega((J,\pi))$ with the ranks $a(j)$ for items $j \in I \setminus J$. We show that the statement of the lemma holds for rank assignments in Φ.

Let D_{k+1} be the distribution of r_{k+1} for $r \in \Phi$ and let D_k be the distribution of r_k for $r \in \Phi$. Over rank assignments in Φ we have $\text{PR}\{r_{k+1} \le Q_{1-\delta}(D_{k+1})\} \ge 1-\delta$ and $\text{PR}\{r_k \ge Q_\delta(D_k)\} \ge 1-\delta$. Specifically we show that

- The upper bound is correct for rank assignments $r \in \Phi$ such that $r_{k+1} \le Q_{1-\delta}(D_{k+1})$. Therefore, it is correct with probability at least $(1-\delta)$.
- The lower bound is correct for rank assignments $r \in \Phi$ such that $r_k \ge Q_\delta(D_k)$. Therefore, it is correct with probability at least $(1-\delta)$.

Fix a rank assignment $r \in \Phi$. Let s be the items in the sketch defined by r. Let $\ell = (J \cap s, r)$ and $\ell^{(c)} = (J \setminus s, r)$ be the weighted lists of the items in $J \cap s$ or $J \setminus s$, respectively, It is easy to check that another rank assignment $r' \in \Phi$ has $r'_{k+1} \ge r_{k+1}$ if and only if $r'(J \setminus s) \ge r_{k+1}$.[11] So if r is such that $r_{k+1} \le Q_{1-\delta}(D_{k+1})$ then

$$\begin{aligned}\text{PR}_{r'\in\Phi}\{r'(J \setminus s) \ge r_{k+1}\} &= \text{PR}_{r'\in\Phi}\{r'_{k+1} \ge r_{k+1}\} \\ &\ge \text{PR}_{r'\in\Phi}\{r'_{k+1} \ge Q_{1-\delta}(D_{k+1})\} \\ &\ge \delta\ .\end{aligned}$$

Now notice that drawing $r' \in \Phi$ is the same as drawing $r \in \Omega((J,\pi))$, and $\Omega((J,\pi))$ is the same as $\Omega(\ell \oplus \ell^{(c)})$. Therefore, from the definition of $\overline{W}(\ell, r_{k+1}, \delta)$ follows that $\ell^{(c)} \in \overline{W}(\ell, r_{k+1}, \delta)$, and hence $w(J \setminus s) \le \overline{w}(\ell, r_{k+1}, \delta)$ and the upper bounds holds.

Analogously, a rank assignment $r' \in \Phi$ has $r'_k \le r_k$ if and only if $\overline{r'}(J \cap s) \le r_k$. So if $r \in \Phi$ such that $r_k \ge Q_\delta(D_k)$

$$\begin{aligned}\text{PR}_{r'\in\Phi}\{\overline{r'}(J \cap s) \le r_k\} &= \text{PR}_{r'\in\Phi}\{r'_k \le r_k\} \\ &\ge \text{PR}_{r'\in\Phi}\{r'_k \le Q_\delta(D_k)\} \\ &\ge \delta\end{aligned}$$

Therefore, $\ell^{(c)} \in \underline{W}(\ell, r_k, \delta)$, and hence $w(J \setminus s) \ge \underline{w}(\ell, r_k, \delta)$ and the lower bound holds. □

E.3 Subpopulation weight using w(I)

We derive tighter confidence intervals that use the total weight $w(I)$. For weighted lists $h^1 = (H^1, \pi^1)$ and $h^2 = (H^2, \pi^2)$ we define here the probability space $\Omega(h^1, h^2)$ of all rank assignments r to $H^1 \cup H^2$ such that the order induced by the ranks on H^1 is π^1 and the order induced by the ranks on H^2 is π^2. (Here we have no requirement of the order between an item from H^1 and an item from H^2.) For $r \in \Omega(h^1, h^2)$ we define $c(r)$ to be the number of items amongst those with k smallest ranks that are in H^1 (equivalently, it is i such that $r_i(H^1) < r_{k-i+1}(H^2)$ and $r_{k-i}(H^2) < r_{i+1}(H^1)$). We also define $d(r)$ to be the difference between the largest rank values of items in H^2 and H^1 that are amongst the k least ranked items. That is

$$d(r) = r_{k-c(r)}(H^2) - r_{c(r)}(H^1)\ .$$

We denote by $(c_1, d_1) \preceq (c_2, d_2)$ the reverse lexicographic order. That is $(c_1, d_1) \preceq (c_2, d_2)$ if $c_1 > c_2$ or if $c_1 = c_2$ and $d_1 \ge d_2$. Note that if we keep $w(H_1) + w(H_2)$ fixed then as we increase $w(H_1)$ and decrease $w(H_2)$

$$\text{PR}\{(c(r'), d(r')) \preceq \tau) \mid r' \in \Omega(h^1, h^2)\}\ , \tag{7}$$

for any fixed pair $\tau = (c, d)$, increases.

Let r be a rank assignment, and let s be the sketch corresponding to r. Let $\Delta = \overline{r}((I \setminus J) \cap s) - \overline{r}(J \cap s)$, and let $\ell_1 = (J \cap s, r)$ and $\ell_2 = ((I \setminus J) \cap s, r)$.

Let $\overline{W}(\ell_1, \ell_2, \Delta, \delta)$ be the set of all pairs (h_1, h_2) of weighted lists $h_1 = (H_1, \pi_1)$ and $h_2 = (H_2, \pi_2)$ such that $w(H_1) + w(H_2) = w(I) - w(s)$ and

$$\text{PR}\{(c(r'), d(r')) \succeq (|J \cap s|, \Delta) \mid r' \in \Omega(\ell_1 \oplus h_1, \ell_2 \oplus h_2)\} \ge \delta\ , \tag{8}$$

or alternatively,

$$\text{PR}\{(c(r'), d(r')) \preceq (|J \cap s|, \Delta) \mid r' \in \Omega(\ell_1 \oplus h_1, \ell_2 \oplus h_2)\} \le 1 - \delta\ . \tag{9}$$

Clearly as we increase $w(H_1)$ and decrease $w(H_2)$ the probability of the event on the left hand side of Equation (9) increases. To get an upper bound $\overline{w}(\ell_1, \ell_2, \Delta, \delta)$ on how large can the "unseen" part of J be, we set $\overline{w}(\ell_1, \ell_2, \Delta, \delta) = 0$ if $\overline{W}(\ell_1, \ell_2, \Delta, \delta) = \emptyset$, and otherwise, $\overline{w}(\ell_1, \ell_2, \Delta, \delta) = \sup\{w(H_1) \mid (h_1, h_2) \in \overline{W}(\ell_1, \ell_2, \Delta, \delta)\}$.

Let $\underline{W}(\ell_1, \ell_2, \Delta, \delta)$ be the set of all pairs (h_1, h_2) of weighted lists $h_1 = (H_1, \pi_1)$ and $h_2 = (H_2, \pi_2)$ such that $w(H_1) + w(H_2) = w(I) - w(s)$ and

$$\text{PR}\{(c(r'), d(r')) \preceq (|J \cap s|, \Delta) \mid r' \in \Omega(\ell_1 \oplus h_1, \ell_2 \oplus h_2)\} \ge \delta\ . \tag{10}$$

If $\underline{W}(\ell_1, \ell_2, \Delta, \delta) = \emptyset$, then $\underline{w}(\ell_1, \ell_2, \Delta, \delta) = w(I) - w(s)$. Otherwise,
$\underline{w}(\ell_1, \ell_2, \Delta, \delta) = \inf\{w(H_1) \mid (h_1, h_2) \in \underline{W}(\ell_1, \ell_2, \Delta, \delta)\}$.

LEMMA E.3. *Let r be a rank assignment, s be the corresponding sketch, let $\Delta = \overline{r}((I \setminus J) \cap s) - \overline{r}(J \cap s)$, and let $\ell_1 = (J \cap s, r)$ and $\ell_2 = ((I \setminus J) \cap s, r)$. Then $w(J \cap s) + \overline{w}(\ell_1, \ell_2, \Delta, \delta)$ is a $(1-\delta)$-confidence upper bound on $w(J)$, and $w(J \cap s) + \underline{w}(\ell_1, \ell_2, \Delta, \delta)$ is a $(1-\delta)$-confidence lower bound on $w(J)$.*

PROOF. The lower bound on $w(J)$ is equal to $w(I)$ minus a $(1-\delta)$-confidence upper bound, $w((I \setminus J) \cap s) + \overline{w}(\ell_2, \ell_1, -\Delta, \delta)$ on $w(I \setminus J)$. Therefore it suffices to prove the upper bound.

We show that the bound holds with probability at least $(1-\delta)$ in the subspace of rank assignments over (I, w) where the rank order of the items in J and the rank order of the items in $I \setminus J$ are fixed. These subspaces are a partition of the space of rank assignments and therefore the lemma follows. Consider a subspace $\Phi = \Omega(\ell'_1, \ell'_2)$ where $\ell'_1 = (J, \pi_1)$ is a weighted list of J, and $\ell'_2 = (I \setminus J, \pi_2)$ is a weighted list of $I \setminus J$. We show that the bound holds for $1 - \delta$ fraction of the rank assignments in Φ.

Let D be the distribution over the pairs $(c(r), d(r))$ for $r \in \Phi$. We define the quantile $Q_{1-\delta}(D)$ with respect to the lexicographic order over the pairs.

We show that the upper bound is correct for all $r \in \Phi$ such that $(c(r), d(r)) \preceq Q_{1-\delta}(D)$. Therefore, it holds with probability at least $1 - \delta$ in Φ.

Let $r \in \Phi$ such that $(c(r), d(r)) \preceq Q_{1-\delta}(D)$. Let s be the corresponding sketch, $\ell_1 = (J \cap s, r)$, $\ell_2 = ((I \setminus J) \cap s, r)$, $\ell_1^{(c)} = (J \setminus s, r)$, $\ell_2^{(c)} = ((I \setminus J) \setminus s, r)$. By definition, $c(r) =$

[11] Note that the statement with strict inequalities does not necessarily hold.

$|J \cap s|$, $\Delta = d(r) = \overline{r}((I \setminus J) \cap s) - \overline{r}(J \cap s)$, $\ell_1' = \ell_1 \oplus \ell_1^{(c)}$, and $\ell_2' = \ell_2 \oplus \ell_2^{(c)}$. It follows that

$$\begin{aligned} \text{PR}\{(c(r'), d(r')) &\succeq (|J \cap s|, \Delta) \mid r' \in \Phi\} \geq \\ \text{PR}\{(c(r'), d(r')) &\succeq Q_{1-\delta}(D) \mid r' \in \Phi\} \geq \delta \ . \end{aligned}$$

Therefore, $(\ell_1^{(c)}, \ell_2^{(c)}) \in \overline{W}(\ell_1, \ell_2, \Delta, \delta)$, and hence,

$$w(J \setminus s) \leq \overline{w}(\ell_1, \ell_2, \Delta, \delta) \ .$$

□

Subpopulation weight using w(I) for WS *sketches.* We specialize the conditions in Lemma E.3 to WS sketches. Consider the distribution of $(c(r), d(r))$ for $r \in \Omega(\ell_1 \oplus h_1, \ell_2 \oplus h_2)$. We shall refer to items of h_1 as items of J and to items of h_2 as items of $I \setminus J$. This distribution in general depends on the decomposition of the weighted lists h_1 and h_2 into items. However we show that for WS sketches the probability

$$\text{PR}\{(c(r), d(r)) \preceq (|J \cap s|, \Delta) \mid r \in \Omega(\ell_1 \oplus h_1, \ell_2 \oplus h_2)\} \qquad (11)$$

depends on ℓ_1, ℓ_2 and $w(H^1)$ (which also determines $w(H^2)$ since $w(H^1) + w(H^2)$ is fixed.)

Indeed, note that $(c(r), d(r)) \preceq (|J \cap s|, \Delta)$ is equivalent to the following condition

$$\begin{array}{l} (r(H_2) > \overline{r}(J \cap s)) \wedge \\ \left(\begin{array}{l} (r(H_1) < \overline{r}(s \cap (I \setminus J))) \vee \\ (r(H_1) > \overline{r}(s \cap (I \setminus J)) \wedge (\overline{r}((I \setminus J) \cap s) - \overline{r}(J \cap s) > \Delta)) \end{array} \right) \end{array} . \qquad (12)$$

The first line guarantees that we have at least $|J \cap s|$ items of J among the k items of smallest ranks. If the second line holds then we have strictly more than $|J \cap s|$ items of J among the k items of smallest ranks. If the third line holds then we have exactly $|J \cap s|$ items of J among the k items of smallest ranks and $\overline{r}((I \setminus J) \cap s) - \overline{r}(J \cap s) > \Delta$.

Now the last observation to make is that the predicate of Equation (12) depends only on the rank values of the $|J \cap s|$ and $|J \cap s| + 1$ smallest ranks in J and of the $|(I \setminus J) \cap s|$ and $|(I \setminus J) \cap s| + 1$ smallest ranks in $I \setminus J$. For WS sketches, the distribution of these ranks is determined by the weighted lists ℓ_1, ℓ_2 and $w(H^1)$.

So we pick a weighted list h_1 with a single item of weight $x - w(J \cap s)$, and a weighted list h_2 with a single item of weight $w(I) - x - w((I \setminus J) \cap s)$, and let $D^{(x)}$ be the distribution of $(c(r), d(r))$ for $r \in \Omega(\ell_1 \oplus h_1, \ell_2 \oplus h_2)$. To emphasis the dependency of r on x we shall denote by $r^{(x)}$ a rank assignment drawn from $\Omega(\ell_1 \oplus h_1, \ell_2 \oplus h_2)$ where $w(H_1) = x$.

Since the largest rank in $J \cap s$ and the smallest rank of an item in H_1 decrease with x, and the largest rank in $(I \setminus J) \cap s$ and the smallest rank in H_2 increase with x, it follows that $\text{PR}\{y \preceq \tau \mid y \in D^{(x)}\}$ is increasing with x for $\tau = (|J \cap s|, \Delta)$ so we can apply the quantile method.

Obviously, $w(J \setminus s) \in [0, w(I) - w(s)]$. Therefore, we can truncate the bounds to be in this range. So the upper bound on $w(J \setminus s)$ is the minimum of $w(I) - w(s)$ and x such that $Q_{1-\delta}(D^{(x)}) = (|J \cap s|, \Delta)$. If there is no solution then the upper bound is 0. (Similarly we get an upper bound on $w((I \setminus J) \setminus s)$.) The upper bound on $w(J)$ is $w(J \cap s)$ plus the upper bound on $w(J \setminus s)$.

We apply the quantile method (Section 7.1) to solve the equations

$$Q_{1-\delta}(D^{(x)}) = (|J \cap s|, \Delta) \ ,$$

and

$$Q_\delta(D^{(x)}) = (|J \cap s|, \Delta) \ .$$

Computing the range (L, U).

- If $i' = 0$, let $U = w(I) - w(s)$. Otherwise ($i' > 0$), U is the solution of $\sum_{h=0}^{i} \frac{-\ln v_h}{x - s_h} - \sum_{h=0}^{i'-1} \frac{-\ln v_h'}{w(I) - x - s_h'} = 0$. (There is always a solution $U \in (s_i, w(I) - s_{i'-1}')$.)
- If $i = 0$, let $L = 0$. Otherwise ($i > 0$), L is the solution of $\sum_{h=0}^{i-1} \frac{-\ln v_h}{x - s_h} - \sum_{h=0}^{i'} \frac{-\ln v_h'}{w(I) - x - s_h'} = 0$. (There is always a solution $L \in (s_{i-1}, w(I) - s_{i'}')$.)

Search for $x \in (L, U)$ such that $d(x) = \Delta$.

- If $i = 0$ (we must have $\Delta > 0$) we set M to be the solution of $\sum_{h=1}^{i'-1} \frac{-\ln v_h'}{w(I) - x - s_h} = \Delta$ in the range (L, U). If there is no solution, we set $M \leftarrow L$.
- If $i' = 0$ (we must have $\Delta < 0$), we set M to be the solution of $\sum_{h=0}^{i-1} \frac{-\ln v_h}{x - s_h} = -\Delta$ in the range (L, U). If there is no solution, we set $M \leftarrow U$.
- Otherwise, if $i > 0$ and $i' > 0$, we set M to be the solution of $\sum_{h=0}^{i-1} \frac{-\ln v_h}{x - s_h} - \sum_{h=0}^{i'-1} \frac{-\ln v_h'}{w(I) - x - s_h} = \Delta$. There must be a solution in the range (L, U).

Truncating the solution.

- We can have $L \in (s_{i-1}, s_i)$ and hence possibly $M < s_i$. In this case we set $M = s_i$. Similarly, we can have $U \in (w(I) - s_{i'}, w(I) - s_{i'-1})$ and hence possibly $M > w(I) - s_{i'}$. In this case we set $M = w(I) - s_{i'}$.
- We return M.

Figure 6: Solver for $s(x) = \tau$ for subpopulation weight with known $w(I)$. Here $i = |J \cap s|$ and $i' = k - i = |(I \setminus J) \cap s|$.

The first black box ingredient that we need for the quantile method is drawing a monotone parametric sample $s(x)$ from $D^{(x)}$. Let s_i ($i \in (0, 1, \ldots, |J \cap s|)$) be the sum of the weights of the first i items from J in ℓ_1. Let s_i' ($i \in (0, 1, \ldots, k - |J \cap s|)$) be the respective sums for $I \setminus J$. We draw a rank assignment $r^{(x)} \in \Omega(\ell_1 \oplus h_1, \ell_2 \oplus h_2)$ as follows. We draw $k+2$ independent random variables $v_0, \ldots, v_{|J \cap s|}, v_0', \ldots, v_{k-|J \cap s|}'$ from $U[0, 1]$. We let the jth rank difference between items from J be $-\ln(v_j)/(x - s_j)$, and the jth rank difference between items from $(I \setminus J)$ be $-\ln(v_j')/(x - s_j')$. These rank differences determine $\overline{r}(J \cap s)$ and $r(H_1)$ (sums of $|J \cap s|$ and $|J \cap s| + 1$ first rank differences from J, respectively), and $\overline{r}((I \setminus J) \cap s)$ and $r(H_2)$ (sums of $|(I \setminus J) \cap s|$ and $|(I \setminus J) \cap s| + 1$ first rank differences from $I \setminus J$, respectively). Then $s(x)$ is the pair $(c(r^{(x)}), d(r^{(x)}))$.

The second black box ingredient is solving the equation $s(x) = \tau$. Let $i = |J \cap s|$ and let $i' = k - i = |(I \setminus J) \cap s|$ as before. The solver has three phases: We first compute the range (L, U) of values of x such that the first coordinate of the pair $s(x)$ is equal to $|J \cap s|$. That is, the rank assignment r has exactly $|J \cap s|$ items from J among the first k items. Let $d(r^{(x)}) = r_{i'}(I \setminus J) - r_i(J)$ denote the second coordinate in the pair $s(x)$. In the second phase we look for a value $x \in (L, U)$ (if there is one) such that $d(r^{(x)}) = \Delta$ (the second coordinate of $s(x)$ is equal to Δ). The function $d(r^{(x)})$ is monotone increasing in this range, which simplifies numeric solution. The third phase is truncating the solution to be in $[0, w(I) - w(s)]$. See Figure 6.