

# Coordinated Weighted Sampling: Estimation of Multiple-Assignment Aggregates

Edith Cohen\*

Haim Kaplan<sup>†</sup>

Subhabrata Sen<sup>‡</sup>

## Abstract

Many data sources are naturally modeled by multiple weight assignments over a set of keys: snapshots of an evolving database at multiple points in time, measurements collected over multiple time periods, requests for resources served at multiple locations, and records with multiple numeric attributes. Over such vector-weighted data we are interested in aggregates with respect to one set of weights, such as weighted sums, and aggregates over multiple sets of weights such as the  $L_1$  difference.

Sample-based summarization is highly effective for data sets that are too large to be stored or manipulated. The summary facilitates approximate processing of queries that may be specified after the summary was generated. Current designs, however, are geared for data sets where a single *scalar* weight is associated with each key.

We develop a sampling framework based on *coordinated weighted samples* that is suited for multiple weight assignments and obtain estimators that are *orders of magnitude tighter* than previously possible. We demonstrate the power of our methods through an extensive empirical evaluation on diverse data sets ranging from IP network to stock quotes data.

## 1 Introduction

Many business-critical applications today are based on extensive use of computing and communication network resources. These systems are instrumented to collect a wide range of different types of data. Examples include performance or environmental measurements, traffic traces, routing updates, or SNMP traps in an IP network, and transaction logs, system resource (CPU, memory) usage statistics, service level end-end performance statistics in an end- service infrastructure. Retrieval of useful information from this vast amount of data is critical to a wide range of compelling applications including network and service management, troubleshooting and root cause analysis, capacity provisioning, security, and sales and marketing.

Many of these data sources produce data sets consisting of numeric vectors (*weight vectors*) associated with a set of identifiers (*keys*) or equivalently as a set of *weight assignments* over *keys*. Aggregates over the data are specified using this abstraction.

We distinguish between data sources with *co-located* or *dispersed* weights. A data source has **dispersed weights** if entries of the weight vector of each key occur in different times or locations: (i) Snapshots of a database that is modified over time (each snapshot is a weight assignment, where the weight of a key is the value of a numeric attribute in a record with this key.) (ii) measurements of a set of parameters (keys) in different time periods (weight assignments). (iii) number of requests for different objects (keys) processed at multiple servers (weight assignments). A data source has **co-located weights** when a complete weight vector is “attached” to each key: (i) Records with multiple numeric attributes such as IP flow records

---

\*AT&T Labs-Research, 180 Park Avenue, Florham Park, NJ.

<sup>†</sup>The Blavatnik School of Computer Science, Tel Aviv University.

<sup>‡</sup>AT&T Labs-Research, 180 Park Avenue, Florham Park, NJ.

generated by a statistics module at an IP router, where the attributes are the number of bytes, number of packets, and unit. (ii) Document-term datasets, where keys are documents and weight attributes are terms or features (The weight value of a term in a document can be the respective number of occurrences). (iii) Market-basket datasets, where keys are baskets and weight attributes are goods (The weight value of a good in a basket can be its multiplicity). (iv) Multiple numeric functions over one (or more) numeric measurement of a parameter. For example, for measurement  $x$  we might be interested in both first and second moments, in which case we can use the weight assignments  $x$  and  $x^2$ .

A very useful common type of query involves properties of a *sub-population* of the monitored data that are *additive* over keys. These aggregates can be broadly categorized as : (a) *Single-assignment* aggregates, defined with respect to a single attribute, such as the weighted sum or selectivity of a subpopulation of the keys. An example over IP flow records is the total bytes of all IP traffic with a certain destination Autonomous System [25, 1, 39, 16, 17]. (b) *Multiple-assignment* aggregates include similarity or divergence metrics such as the  $L_1$  difference between two weight assignments or maximum/minimum weight over a subset of assignments [38, 22, 9, 21]. Figure 2 (A) shows an example of three weight assignments over a set of keys and key-wise values for multiple-assignment aggregates including the minimum or maximum value of a key over subset of assignments and the  $L_1$  distance. The aggregate value over selected keys is the sum of key-wise values.

Multiple-assignment aggregates are used for clustering, change detection, and mining emerging patterns. Similarity over corpus of documents, according to a selected subset of features, can be used to detect near-duplicates and reduce redundancy [41, 10, 52, 20, 37, 42]. A retail merchant may want to cluster locations according to sales data for a certain type of merchandise. In IP networks, these aggregates are used for monitoring, security, and planning [28, 22, 23, 40]: An increase in the amount of distinct flows on a certain port might indicate a worm activity, increase in traffic to a certain set of destinations might indicate a flash crowd or a DDoS attack, and an increased number of flows from a certain source may indicate scanner activity. A network security application might track the increase in traffic to a customer site that originates from a certain suspicious network or geographic area.

Exact computation of such aggregates can be prohibitively resource-intensive: Data sets are often too large to be either stored for long time periods or to be collated across many locations. Computing multiple-assignment aggregates may require gleaning information across data sets from different times and locations. We therefore aim at concise summaries of the data sets, that can be computed in a scalable way and facilitate approximate query processing.

Sample-based summaries [36, 56, 7, 6, 11, 31, 32, 3, 24, 33, 17, 26, 14, 18] are more flexible than other formats: they naturally facilitate subpopulation queries by focusing on sampled keys that are members of the subpopulation and are suitable when the exact query of interest is not known beforehand or when there are multiple attributes of interest. When keys are weighted, weighted sampling, where heavier keys are more likely to be sampled, is essential for performance. Existing methods, however, are designed for one set of weights and are either not applicable or perform poorly on multiple-assignment aggregates.

## Contributions

We develop sample-based summarization framework for vector-weighted data that supports efficient approximate aggregations. The challenges differ between the dispersed and co-located models due to the particular constraints imposed on scalable summarization.

**Dispersed weights model:** Any scalable algorithm must decouple the processing of different assignments – collating dispersed-weights data to obtain explicit key/vector-weight representation is too expensive. Hence, processing of one assignment can not depend on other assignments.

We propose summaries based on *coordinated weighted samples*. The summary contains a “classic”

weighted sample taken with respect to each assignment. Coordination loosely means that a key that is sampled under one assignment is more likely to be sampled under other assignments. We can tailor the sampling to be Poisson,  $k$ -mins, or bottom- $k$  (order) sampling. In all three cases, sampling is efficient on data streams, distributed data, and metric data [11, 15, 27, 16] and there are unbiased subpopulation weight estimators that have variance that decreases linearly or faster with the sample size [11, 26, 54, 17]. Bottom- $k$  samples [47, 51, 48, 11, 17, 45, 26], with the advantage of a fixed sample size, emerge as a better choice. Our design has the following important properties:

- **Scalability:** The processing of each assignment is a simple adaptation of single-assignment weighted sampling algorithm. Coordination is achieved by using the same hash function across assignments.
- **Weighted sample for each assignment:** Our design is especially appealing for applications where sample-based summaries are already used, such as periodic (hourly) summaries of IP flow records. The use of our framework versus independent sampling in different periods facilitates support for queries on the relation of the data across time periods.
- **Tight estimators:** We provide a principled “template” derivation of estimators, tailor it to obtain tight unbiased estimators for the min, max, and range ( $L_1$ ), and bound the variance.

**Colocated weights model:** For colocated data, the full weight vector of each key is readily available to the summarization algorithm and can be easily incorporated in the summary. We discuss the shortcomings of applying previous methods to summarize this data. One approach is to sample records according to one particular weight assignment. Such a sample can be used to estimate aggregates that involve other assignments,<sup>1</sup> but estimates may have large variance and be biased. Another approach is to concurrently compute multiple weighted samples, one for each assignment. In this case, single-assignment aggregates can be computed over the respective sample but no unbiased estimators for multiple-assignment aggregates were known. Moreover, such a summary is wasteful in terms of storage as different assignments are often correlated (such as number of bytes and number of IP packets of an IP flow).

We consider summaries where the set of included keys embeds a weighted sample with respect to each assignment. The set of embedded samples can be independent or coordinated. Such a summary can be computed in a scalable way by a stream algorithm or distributively.

- We derive estimators, which we refer to as *inclusive estimators*, that utilize all keys included in the summary. An inclusive estimator of a single-assignment aggregate applied to a summary that embeds a certain weighted sample from that assignment is at least as tight, and typically significantly tighter, than an estimator directly applied to the embedded sample. Moreover, inclusive estimators are applicable to multiple-assignment aggregates, such as the min, max, and  $L_1$ .
- We show that when the embedded samples are coordinated, the number of distinct keys in the summary (in the union of the embedded samples) is minimized.

**Empirical evaluation.** We performed a comprehensive empirical evaluation using IP packet traces, movies’ ratings data set (The Netflix Challenge [44]), and stock quotes data set. These data sets and queries also demonstrate potential applications. For dispersed data we achieve *orders of magnitude* reduction in variance over previously-known estimators and estimators applied to independent weighted samples. The variance of these estimators is comparable to the variance of a weighted sum estimator of a single weight assignment.

For co-located data, we demonstrate that the size of our combined sample is significantly smaller than the sum of the sizes of independent samples one for each weight assignment. We also demonstrate that even

---

<sup>1</sup>This is standard, by multiplying the per-key estimate with an appropriate ratio [51]

for single assignment aggregates, our estimators which use the combined sample are much tighter than the estimators that use only a sample for the particular assignment.

**Organization.** The remainder of the paper is arranged as follows. Section 2 reviews related work, Section 3 presents key background concepts and Section 4 presents our sampling approach. Sections 5-7 present our estimators: Section 5 presents a template estimator which we apply to colocated summaries in Section 6 and to dispersed summaries in Section 7. Section 8 provides bounds on the variance. This is an extended version of [19].

## 2 Related work

**Sample coordination.** Sample coordination was used in survey sampling for almost four decades. *Negative coordination* in repeated surveys was used to decrease the likelihood that the same subject is surveyed (and burdened) multiple times. *Positive coordination* was used to make samples as similar as possible when parameters change in order to reduce overhead. Coordination is obtained using the PRN (Permanent Random Numbers) method for Poisson samples [5] and order (bottom- $k$ ) samples [50, 46, 48]. PRN resembles our “shared-seed” coordination method. The challenges of massive data sets, however, are different from those of survey sampling and in particular, we are not aware of previously existing unbiased estimators for multiple-assignment aggregates over coordinated weighted samples.

Coordination (of Poisson,  $k$ -mins, and bottom- $k$  samples) was (re-)introduced in computer science as a method to support aggregations that involve multiple sets [7, 6, 11, 31, 32, 3, 17, 33, 18] and as a form of locality sensitive hashing [9]. Coordination addressed the issue that independent samples of different sets over the same universe provide weak estimators for multiple-set aggregates such as intersection size or similarity. Intuitively, two large but almost identical sets are likely to have disjoint independent samples – the sampling does not retain any information on the relations between the sets.

This previous work, however, considered restricted weight models: *uniform*, where all weights are 0/1, and *global weights*, where a key has the same weight value across all assignments where its weight is strictly positive (but the weight can vary between keys). Allowing the same key to assume different positive weights in different assignments is clearly essential for our applications.

While these methods can be applied with general weights, by ignoring weight values and performing coordinated uniform sampling, resulting estimators are weak. Intuitively, uniform sampling performs poorly on weighted data because it is likely to leave out keys with dominant weights. Weighted sampling, where keys with larger weights are more likely to be represented in the sample, is essential for boundable variance of weighted aggregates.

The structure of coordinated samples for general weights turns out to be more involved than with global weights, where essentially the samples of different assignments (sets) are derived from a single “global” (random) ranking of keys. The derivation of unbiased estimators was also more challenging: global weights allow us to make simple inferences on inclusion of keys in a set when the key is not represented in the sample. These inferences facilitate the derivation of estimators but do not hold under general weights.

**Unaggregated data.** Sample-based sketches [30, 13, 12] and sketches that are not samples were also proposed for *unaggregated* data streams (the scalar weight of each key appears in multiple data points) [2]. This is a more general model with weaker estimators than “aggregated” keys. We leave for future work summarization of unaggregated data set with vector-weights.

VAROPT is a weighted sampling design [8, 14] that realizes all the advantages of other schemes but it is not clear if it can be applied with coordinated samples (even with global weights).

**Sketches that are not samples.** Sketches that are not *sample based* [41, 10, 52, 20, 37, 42, 21, 29] are effective point solutions for particular metrics such as max-dominance [21] or  $L_1$  [29] difference. Their

**weighted set**  $(I, w)$  with keys  $I = \{i_1, \dots, i_6\}$  and a **rank assignment**  $r$

key $i$ :	$i_1$	$i_2$	$i_3$	$i_4$	$i_5$	$i_6$
weight $w(i)$	20	10	12	20	10	10
$u(i) \in U[0, 1]$	0.22	0.75	0.07	0.92	0.55	0.37
$r(i) = u(i)/w(i)$	0.011	0.075	0.0583	0.046	0.055	0.037

**Poisson samples with expected size  $k = 1, 2, 3$  and AW-summaries:**

$$p(i) = \min\{1, w(i)\tau\}, k = \sum_i p(i), a(i) = w(i)/p(i)$$

$k$	sample	$\tau$	$i$ :	$i_1$	$i_2$	$i_3$	$i_4$	$i_5$	$i_6$
1	$i_1$	$\frac{1}{82}$	$p(i)$ :	0.24	0.12	0.15	0.24	0.12	0.12
			$a(i)$ :	82	0	0	0	0	0
2	$i_1$	$\frac{2}{82}$	$p(i)$ :	0.49	0.24	0.29	0.49	0.24	0.24
			$a(i)$ :	41	0	0	0	0	0
3	$i_1$	$\frac{3}{82}$	$p(i)$ :	0.73	0.37	0.44	0.73	0.37	0.37
			$a(i)$ :	27.40	0	0	0	0	0

**Bottom- $k$  samples of size  $k = 1, 2, 3$  and AW-summaries:**

$$p(i) = \min\{1, w(i)r_{k+1}\}, a(i) = w(i)/p(i)$$

$k$	sample	$r_{k+1}$	$i$ :	$i_1$	$i_2$	$i_3$	$i_4$	$i_5$	$i_6$
1	$i_1$	0.037	$p(i)$ :	0.74					
			$a(i)$ :	27.02	0	0	0	0	0
2	$i_1, i_6$	0.046	$p(i)$ :	0.92					0.46
			$a(i)$ :	21.74	0	0	0	0	21.74
3	$i_1, i_6, i_4$	0.055	$p(i)$ :	1			1		0.55
			$a(i)$ :	20.00	0	0	20.00	0	18.18

Figure 1: Example of a weighted set, a random rank assignment with IPPS ranks, Poisson and bottom- $k$  samples, and respective AW-summaries.

disadvantage is less flexibility in terms of supported aggregates and in particular, no support for aggregates over selected subpopulations of keys: we can estimate the overall  $L_1$  difference between two time periods but we can not estimate the difference restricted to a subpopulation such as flows to particular destination or certain application. There is also no mechanism to obtain “representatives” keys[53]. Lastly, even a practical implementation of [21, 29] involves constructions of stable distributions or range summable random variables (whereas for our sample-based summaries all is needed is “random-looking” hash functions).

Bloom filters [4, 28] also support estimation of similarity metrics but summary size is not tunable and grows linearly with the number of keys.

### 3 Preliminaries

A *weighted set*  $(I, w)$  consists of a set of keys  $I$  and a function  $w$  assigning a scalar weight value  $w(i) \geq 0$  to each key  $i \in I$ . We review components of sample-based summarizations of a weighted set: sample distributions, respective *sketches*, that in our context are samples with some auxiliary information, and estimators for weight queries in the form of *adjusted weights* associated with sampled keys.

Sample distributions are defined through *random rank assignments* [11, 48, 16, 26, 17, 18] that map each key  $i$  to a rank value  $r(i)$ . The rank assignment is defined with respect to a family of probability density functions  $\mathbf{f}_w$  ( $w \geq 0$ ), where each  $r(i)$  is drawn independently according to  $\mathbf{f}_{w(i)}$ . We say that  $\mathbf{f}_w$  ( $w \geq 0$ ) are *monotone* if for all  $w_1 \geq w_2$ , for all  $x$ ,  $\mathbf{F}_{w_1}(x) \geq \mathbf{F}_{w_2}(x)$  (where  $\mathbf{F}_w$  are the respective cumulative distributions). For a set  $J$  and a rank assignment  $r$  we denote by  $r_i(J)$  the  $i$ th smallest rank of a key in  $J$ , we also abbreviate and write  $r(J) = r_1(J)$ .

- A *Poisson- $\tau$*  sample of  $J$  is defined with respect to a rank assignment  $r$ . The sample is the set of keys with  $r(i) < \tau$ . The sample has *expected* size  $k = \sum_i \mathbf{F}_{w(i)}(\tau)$ . Keys have independent inclusion probabilities. The sketch includes the pairs  $(r(i), w(i))$  and may include key identifiers with attribute values.
- A *bottom- $k$  (order- $k$ )* sample of  $J$  contains the  $k$  keys  $i_1, \dots, i_k$  of smallest ranks in  $J$ . The sketch  $s_k(J, r)$  consists of the  $k$  pairs  $(r(i_j), w(i_j))$ ,  $j = 1, \dots, k$ , and  $r_{k+1}(J)$ . (If  $|J| \leq k$  we store only  $|J|$  pairs.), and may include the key identifiers  $i_j$  and additional attributes.
- A  *$k$ -mins sample* of  $J \subset I$  is produced from  $k$  independent rank assignments,  $r^{(1)}, \dots, r^{(k)}$ . The sample is the set of (at most  $k$ ) keys) with minimum rank values  $r^{(1)}(J), r^{(2)}(J), \dots, r^{(k)}(J)$ . The sketch includes the minimum rank values and, depending on the application, may include corresponding key identifiers and attribute values.

When weights of keys are uniform, a  $k$ -mins sample is the result of  $k$  uniform draws with replacement, bottom- $k$  samples are  $k$  uniform draws without replacements, and Poisson- $\tau$  samples are independent Bernoulli trials. The particular family  $\mathbf{f}_w$  matters when weights are not uniform. Two families with special properties are:

- EXP ranks:  $\mathbf{f}_w(x) = we^{-wx}$  ( $\mathbf{F}_w(x) = 1 - e^{-wx}$ ) are exponentially-distributed with parameter  $w$  (denoted by  $\text{EXP}[w]$ ). Equivalently, if  $u \in U[0, 1]$  then  $-\ln(u)/w$  is an exponential random variable with parameter  $w$ .  $\text{EXP}[w]$  ranks have the property that the minimum rank  $r(J)$  has distribution  $\text{EXP}[w(J)]$ , where  $w(J) = \sum_{i \in J} w(i)$ . (The minimum of independent exponentially distributed random variables is exponentially distributed with parameter equal to the sum of the parameters of these distributions). This property is useful for designing estimators and efficiently computing sketches [11, 15, 27, 16, 17]. The  $k$ -mins sample [11] of a set is a sample drawn *with replacement* in  $k$  draws where a key is selected with probability equal to the ratio of its weight and the total weight. A bottom- $k$  sample is the result of  $k$  such draws performed *without replacement*, where keys are selected according to the ratio of their weight and the weight of remaining keys [47, 34, 48].
- IPPS ranks:  $\mathbf{f}_w$  is the uniform distribution  $U[0, 1/w]$  ( $\mathbf{F}_w(x) = \min\{1, wx\}$ ). This is the equivalent to choosing rank value  $u/w$ , where  $u \in U[0, 1]$ . The Poisson- $\tau$  sample is an IPPS sample [34] (Inclusion Probability Proportional to Size). The bottom- $k$  sample is a priority sample [45, 26] (PRI).

Figure 1 shows an example of a weighted set with 6 keys and a respective rank assignment with IPPS ranks. The figure shows the corresponding Poisson samples of expected size  $k = 1, 2, 3$ . The value  $\tau$  is calculated according to the desired expected sample size. The sample includes all keys with rank value that is below  $\tau$ . This particular rank assignment yielded a Poisson sample of size 1 when the expected size was 1, 2, 3. The figure also shows bottom- $k$  samples of sizes  $k = 1, 2, 3$ , containing the  $k$  keys with smallest rank values.

**Adjusted weights.** A technique to obtain estimators for the weights of keys is by assigning an adjusted weight  $a(i) \geq 0$  to each key  $i$  in the sample (adjusted weight  $a(i) = 0$  is implicitly assigned to keys not in the sample). The adjusted weights are assigned such that  $\mathbb{E}[a(i)] = w(i)$ , where the expectation is over the randomized algorithm choosing the sample. We refer to the (random variable) that combines a weighted sample of  $(I, w)$  together with adjusted weights as an *adjusted-weights summary* (AW-summary) of  $(I, w)$ . An AW-summary allows us to obtain an unbiased estimate on the weight of *any* subpopulation  $J \subset I$ . The estimate  $\sum_{j \in J} a(j) = \sum_{j \in J | a(j) > 0} a(j)$  is easily computed from the summary provided that we have sufficient auxiliary information to tell for each key in the summary whether it belongs to  $J$  or not. Figure 1 shows example AW-summaries for the Poisson and bottom- $k$  samples. The set  $J = \{i_2, i_4, i_6\}$  with weight

$w(J) = w(i_2) + w(i_4) + w(i_6) = 10 + 20 + 10 = 40$  has estimate of 0 using the three Poisson AW-summaries and estimates 0, 21.74, 38.18 respectively by the three bottom- $k$  AW-summaries. Moreover, for any secondary numeric function  $h()$  over keys' attributes such that  $h(i) > 0 \implies w(i) > 0$  and any subpopulation  $J$ ,  $\sum_{j \in J | a(j) > 0} a(j)h(j)/w(j)$  is an unbiased estimate of  $\sum_{j \in J} h(j)$ .

**Horvitz-Thompson (HT).** Let  $\Omega$  be the distribution over samples such that if  $w(i) > 0$  then  $p^{(\Omega)}(i) = \Pr\{i \in s | s \in \Omega\}$  is positive. If we know  $p^{(\Omega)}(i)$  for every  $i \in s$ , we can assign to  $i \in s$  the adjusted weight  $a(i) = \frac{w(i)}{p^{(\Omega)}(i)}$ . Since  $a(i)$  is 0 when  $i \notin s$ ,  $\mathbb{E}[a(i)] = w(i)$  ( $a(i)$  is an unbiased estimator of  $w(i)$ ). These adjusted weights are called the Horvitz-Thompson (HT) estimator [35]. For a particular  $\Omega$ , the HT adjusted weights minimize  $\text{VAR}[a(i)]$  for all  $i \in I$ . The HT adjusted weights for Poisson  $\tau$ -sampling are  $a(i) = w(i)/\mathbf{F}_{w(i)}(\tau)$ . Figure 1 shows the inclusion probability  $\mathbf{F}_{w(i)}(\tau)$  and a corresponding AW-summary for the Poisson samples. Poisson sampling with IPPS ranks and HT adjusted weights are known to minimize the sum  $\sum_{i \in I} \text{VAR}(a(i))$  of per-key variances over all AW-summaries with the same expected size.

**HT on a partitioned sample space (HTP) [17].** This is a method to derive adjusted weights when we cannot determine  $\Pr\{i \in s | s \in \Omega\}$  from the information contained in the sketch  $s$  alone.

For each key  $i$  we consider a partition of  $\Omega$  into equivalence classes. For a sketch  $s$ , let  $P^i(s) \subset \Omega$  be the equivalence class of  $s$ . This partition must satisfy the following requirement: Given  $s$  such that  $i \in s$ , we can compute the conditional probability  $p^i(s) = \Pr\{i \in s' | s' \in P^i(s)\}$  from the information included in  $s$ .

We can therefore compute for all  $i \in s$  the assignment  $a(i) = w(i)/p^i(s)$  (implicitly,  $a(i) = 0$  for  $i \notin s$ .) It is easy to see that within each equivalence class,  $\mathbb{E}[a(i)] = w(i)$ . Therefore, also over  $\Omega$  we have  $\mathbb{E}[a(i)] = w(i)$ .

**Rank Conditioning (RC) [17].** When  $s$  is a bottom- $k$  sketch of  $(I, w)$ , then  $\Pr\{i \in s | s \in \Omega\}$  generally depends on all the weights  $w(i)$  for  $i \in I$  and therefore cannot be determined from  $s$ . Therefore, we can not directly apply the HT estimator. RC is an HTP method designed for bottom- $k$  sketches. For each  $i$  and possible rank value  $\tau$  we have an equivalence class  $P_\tau^i$  containing all sketches in which the  $k$ th smallest rank value assigned to a key other than  $i$  is  $\tau$ . Note that if  $i \in s$  then  $\tau$  is the  $(k + 1)$ st smallest rank which is included in the sketch. The inclusion probability of  $i$  in a sketch in  $P_\tau^i$  is  $p_\tau^i = \mathbf{F}_{w(i)}(\tau)$  and it can be computed from the sample.

Applying RC, consider  $s$  containing keys  $i_1, \dots, i_k$  and the  $(k + 1)$ st smallest rank value  $r_{k+1}$ . Then for key  $i_j$ , we have  $s \in P_{r_{k+1}}^{i_j}$  and  $a(i_j) = \frac{w(i_j)}{\mathbf{F}_{w(i_j)}(r_{k+1})}$ . The RC estimator for bottom- $k$  samples with IPPS ranks [26] has a sum of per-key variances that is at most that of an HT estimator applied to a Poisson sample with IPPS ranks and expected size  $k + 1$  [54].

This RC method was extended in [18] to obtain estimators over coordinated bottom- $k$  sketches with global weights [18]. The RC method facilitated the derivation and analysis of unbiased estimators over bottom- $k$  samples – it provided a way to “get around” the dependence between inclusions of different keys, typically, without sacrificing accuracy with respect to estimators over Poisson samples.

Figure 1 shows the  $(k + 1)$ st smallest rank value, the conditional inclusion probability  $\mathbf{F}_{w(i)}(r_{k+1})$  and the corresponding AW-summary for each bottom- $k$  sample in the example.

We subsequently use the notation  $\Omega(i, r^{-i})$  for the probability subspace of rank assignments that contains all rank assignments  $r'$  that agree on  $r$  for all keys in  $I \setminus \{i\}$ .

**Sum of per-key variances** Different AW-summaries are compared based on their *estimation quality*. Variance is the standard metric for the quality of an estimator for a single quantity. For a subpopulation  $J$  and AW-summaries  $a()$ , the variance is  $\text{VAR}[a(J)] = \mathbb{E}[a(J)]^2 - w(J)^2$ . Since our application is for subpopulations that may not be specified a priori, the notion of a good metric is more subtle. Clearly there is no single AW-summary that dominates all others of the same size (minimizes the variance) for all subpopulations  $J$ .

The metric we use in our performance evaluation is the *sum of per-key variances*  $\Sigma V[a] \equiv \sum_{i \in I} \text{VAR}[a(i)]$  [26, 17]. For AW-summaries with *zero covariances* (for any two keys  $i, j$ ,  $\text{COV}[a(i), a(j)] = \mathbb{E}[a(i)a(j)] -$

keys:  $I = \{i_1, \dots, i_6\}$

weight assignments:  $w^{(1)}, w^{(2)}, w^{(3)}$

assignment/key	$i_1$	$i_2$	$i_3$	$i_4$	$i_5$	$i_6$
$w^{(1)}$	15	0	10	5	10	10
$w^{(2)}$	20	10	12	20	0	10
$w^{(3)}$	10	15	15	0	15	10

Example functions  $f(i_j)$

$w^{(\max\{1,2\})}$	20	10	12	20	10	10
$w^{(\max\{1,2,3\})}$	20	15	15	20	15	10
$w^{(\min\{1,2\})}$	15	0	10	0	0	10
$w^{(\min\{1,2,3\})}$	10	0	10	0	0	10
$w^{(L_1\{1,2\})}$	5	10	2	15	10	0
$w^{(L_1\{2,3\})}$	10	5	3	20	15	0

(A)

Consistent shared-seed IPPS ranks:

key:	$i_1$	$i_2$	$i_3$	$i_4$	$i_5$	$i_6$
$u$	0.22	0.75	0.07	0.92	0.55	0.37
$r^{(1)}$	0.0147	$+\infty$	0.007	0.184	0.055	0.037
$r^{(2)}$	0.011	0.075	0.0583	0.046	$+\infty$	0.037
$r^{(3)}$	0.022	0.05	0.0047	$+\infty$	0.0367	0.037

Independent IPPS ranks:

key:	$i_1$	$i_2$	$i_3$	$i_4$	$i_5$	$i_6$
$u^{(1)}$	0.22	0.75	0.07	0.92	0.55	0.37
$r^{(1)}$	0.0147	$+\infty$	0.007	0.184	0.055	0.037
$u^{(2)}$	0.47	0.58	0.71	0.84	0.25	0.32
$r^{(2)}$	0.0235	0.058	0.0592	0.042	$+\infty$	0.032
$u^{(3)}$	0.63	0.92	0.08	0.59	0.32	0.80
$r^{(3)}$	0.063	0.0613	0.0053	$+\infty$	0.0213	0.08

(B)

bottom-3 samples:

$w^{(1)}$   $i_3, i_1, i_6$   
 $w^{(2)}$   $i_1, i_6, i_4$   
 $w^{(3)}$   $i_3, i_1, i_5$

bottom-3 samples:

$w^{(1)}$   $i_3, i_1, i_6$   
 $w^{(2)}$   $i_1, i_6, i_4$   
 $w^{(3)}$   $i_3, i_5, i_2$

Figure 2: (A): Example data set with keys  $I = \{i_1, \dots, i_6\}$  and weight assignments  $w^{(1)}, w^{(2)}, w^{(3)}$  and per-key values for example aggregates. (B): random rank assignments and corresponding bottom-3 samples.

$w^{(i)}w^{(j)} = 0$ ),  $\Sigma V[a]$  also measures *average variance* over subpopulations of any given weight or size [55]. RC adjusted weights on single-assignment [17] and their extension to coordinated sketches with global weights [18] have zero covariances. HT adjusted weights for Poisson sketches have zero covariances (this is immediate from independence). When covariances are zero, the variance of  $a(J)$  for a particular subpopulation  $J$  is equal to  $\sum_{i,j \in J} \text{COV}[a(i), a(j)] = \sum_{i \in J} \text{VAR}[a(i)]$ .

Estimators for Poisson,  $k$ -mins, and bottom- $k$  sketches with EXP or IPPS ranks have  $\Sigma V[a] \leq \frac{w(J)^2}{k-2}$  (where  $k$  is the (expected) sample size) [11, 16, 26, 54]. For a subpopulation  $J$  with expected  $k'$  samples in the sketch, the variance on estimating  $w(J)$  is bounded by  $w(J)^2/(k' - 2)$ . For bottom- $k$  and Poisson sketches, the variance is smaller when the weight distribution is more skewed [16, 26].

## 4 Model and summary formats

We model the data using a set of keys  $I$  and a set  $\mathcal{W}$  of *weight assignments* over  $I$ . For each  $b \in \mathcal{W}$ ,  $w^{(b)} : I \rightarrow \mathcal{R}_{\geq 0}$  maps keys to nonnegative reals. Figure 2 shows a data set with  $I = \{i_1, \dots, i_6\}$  and  $\mathcal{W} = \{1, 2, 3\}$ . For  $i \in I$  and  $\mathcal{R} \subset \mathcal{W}$ , we use the notation  $w^{(\mathcal{R})}(i)$  for the *weight vector* with entries  $w^{(b)}(i)$  ordered by  $b \in \mathcal{R}$ .

We are interested in aggregates of the form  $\sum_{i|d(i)=1} f(i)$  where  $d$  is a selection predicate and  $f$  is a numeric function, both defined over the set of keys  $I$ .  $f(i)$  and  $d(i)$  may depend on the attribute values associated with key  $i$  and on the weight vector  $w^{(\mathcal{W})}(i)$ .

We say that the function  $f$ /predicate  $d$  is *single-assignment* if it depends on  $w^{(b)}(i)$  for a single  $b \in \mathcal{W}$ . Otherwise we say that it is *multiple-assignment*. The *relevant assignments* of  $f$  and  $d$  are those necessary for determining all keys  $i$  such that  $d(i) = 1$  and evaluating  $f(i)$  for these keys.

The *maximum* and *minimum* with respect to a set of assignments  $\mathcal{R} \subset \mathcal{W}$ , are defined by  $f(i)$  as follows:

$$w^{(\max \mathcal{R})}(i) \equiv \max_{b \in \mathcal{R}} w^{(b)}(i) \quad w^{(\min \mathcal{R})}(i) \equiv \min_{b \in \mathcal{R}} w^{(b)}(i). \quad (1)$$

The relevant assignments for  $f$  in this case are  $\mathcal{R}$ . Sums over these  $f$ 's are also known as the *max-dominance* and *min-dominance* norms [21, 22] of the selected subset. The maximum reduces to the size of set union and the minimum to the size of set intersection for the special case of global weights.

The ratio  $\sum_{i \in J} w^{(\min \mathcal{R})}(i) / \sum_{i \in J} w^{(\max \mathcal{R})}(i)$  when  $|\mathcal{R}| = 2$  is the *weighted Jaccard similarity* of the assignments  $\mathcal{R}$  on  $J$ . The *range* ( $L_1$  difference when  $|\mathcal{R}| = 2$ ) can be expressed as a sum aggregate by choosing  $f(i)$  to be

$$w^{(L_1 \mathcal{R})}(i) \equiv w^{(\max \mathcal{R})}(i) - w^{(\min \mathcal{R})}(i). \quad (2)$$



For the example in Figure 2, the max dominance norm over even keys (specified by a predicate  $d$  that is true for  $i_2, i_4, i_6$ ) and assignments  $\mathcal{R} = \{1, 2, 3\}$  is  $w^{(\max\{1,2,3\})}(i_2) + w^{(\max\{1,2,3\})}(i_4) + w^{(\max\{1,2,3\})}(i_6) = 15 + 20 + 10 = 45$ , the  $L_1$  distance between assignments  $\mathcal{R} = \{2, 3\}$  over keys  $i_1, i_2, i_3$  is  $w^{(L_1\{2,3\})}(i_1) + w^{(L_1\{2,3\})}(i_2) + w^{(L_1\{2,3\})}(i_3) = 10 + 5 + 3 = 18$ .

This classification of dispersed and colocated models differentiates the summary formats that can be computed in a scalable way: With colocated weights, each key is processed once, and samples for different assignments  $b \in \mathcal{W}$  are generated together and can be coupled. Moreover, the (full) weight vector can be easily incorporated with each key included in the final summary. With dispersed weights, any scalable summarization algorithm must decouple the sampling for different  $b \in \mathcal{W}$ . The process and result for  $b \in \mathcal{W}$  can only depend on the values  $w^{(b)}(i)$  for  $i \in I$ . The final summary is generated from the results of these disjoint processes.

**Random rank assignments for  $(I, \mathcal{W})$ .** A *random rank assignment* for  $(I, \mathcal{W})$  associates a rank value  $r^{(b)}(i)$  for each  $i \in I$  and  $b \in \mathcal{W}$ . If  $w^{(b)}(i) = 0$ ,  $r^{(b)}(i) = +\infty$ . The *rank vector* of  $i \in I$ ,  $r^{(\mathcal{W})}(i)$ , has entries  $r^{(b)}(i)$  ordered by  $b \in \mathcal{W}$ . The distribution  $\Omega$  is defined with respect to a monotone family of density functions  $\mathbf{f}_w$  ( $w \geq 0$ ) and has the following properties: (i) For all  $b$  and  $i$  such that  $w^{(b)}(i) > 0$ , the distribution of  $r^{(b)}(i)$  is  $\mathbf{f}_{w^{(b)}(i)}$ . (ii) The *rank vectors*  $r^{(\mathcal{W})}(i)$  for  $i \in I$  are independent. (iii) For all  $i \in I$ , the distribution of the rank vector  $r^{(\mathcal{W})}(i)$  depends only on the weight vector  $w^{(\mathcal{W})}(i)$ .

It follows from (i) and (ii) that for each  $b \in \mathcal{W}$ ,  $\{r^{(b)}(i) | i \in I\}$  is a random rank assignment for the weighted set  $(I, w^{(b)})$  with respect to the family  $\mathbf{f}_w$  ( $w \geq 0$ ). The distribution  $\Omega$  is specified by the mapping (iii) from weight vectors to distributions of rank vectors specifies  $\Omega$ .

**Independent or consistent ranks.** If for each key  $i$ , the entries  $r^{(b)}(i)$  ( $b \in \mathcal{W}$ ) of the rank vector of  $i$  are independent we say that the rank assignment has *independent ranks*. In this case  $\Omega$  is the product distribution of independent rank assignments  $r^{(b)}$  for  $(I, w^{(b)})$  ( $b \in \mathcal{W}$ ).

A rank assignment has *consistent ranks* if for each key  $i \in I$  and any two weight assignments  $b_1, b_2 \in \mathcal{W}$ ,

$$w^{(b_1)}(i) \geq w^{(b_2)}(i) \Rightarrow r^{(b_1)}(i) \leq r^{(b_2)}(i).$$

(in particular, if entries of the weight vector are equal then corresponding rank values are equal, that is,  $w^{(b_1)}(i) = w^{(b_2)}(i) \Rightarrow r^{(b_1)}(i) = r^{(b_2)}(i)$ .)

In the special case of global (or uniform) weights, consistency means that the entries of each rank vector are equal and distributed according to  $\mathbf{f}_{w(i)}$  for all  $b \in \mathcal{W}$  such that  $w^{(b)}(i) > 0$ . Therefore, the distribution of the rank vectors is determined uniquely by the family  $\mathbf{f}_w$  ( $w > 0$ ). This is not true for general weights. We explore the following two distributions of consistent ranks, specified by a mapping of weight vectors to probability distributions of rank vectors.

• **Shared-seed:** Independently, for each key  $i \in I$ :

- $u(i) \leftarrow U[0, 1]$  (where  $U[0, 1]$  is the uniform distribution on  $[0, 1]$ .)
- For  $b \in \mathcal{W}$ ,  $r^{(b)}(i) \leftarrow \mathbf{F}_{w^{(b)}(i)}^{-1}(u(i))$ .

That is, for  $i \in I$ ,  $r^{(b)}(i)$  ( $b \in \mathcal{W}$ ) are determined using the same “placement” ( $u(i)$ ) in  $\mathbf{F}_{w^{(b)}(i)}$ .

Consistency of this construction is an immediate consequence of the monotonicity property of  $\mathbf{f}_w$ .

Shared-seed assignment for IPPS ranks is  $r^{(b)}(i) = u(i)/w^{(b)}(i)$  and for EXP ranks, is  $r^{(b)}(i) = -\ln(1 - u(i))/w^{(b)}(i)$ .

• **Independent-differences** is specific to EXP ranks. Recall that  $\text{EXP}[w]$  denotes the exponential distribution with parameter  $w$ . Independently, for each key  $i$ :

- Let  $w^{(b_1)}(i) \leq \dots \leq w^{(b_h)}(i)$  be the entries of the weight vector of  $i$ .
- For  $j \in 1 \dots h$ ,  $d_j \leftarrow \text{EXP}[w^{(b_j)}(i) - w^{(b_{j-1})}(i)]$ , where  $w^{(0)}(i) \equiv 0$  and  $d_j$  are independent.
  - For  $j \in 1 \dots h$ ,  $r^{(b_j)}(i) \leftarrow \min_{a=1}^j d_a$ .

For these ranks consistency is immediate from the construction. Since the distribution of the minimum of independent exponential random variables is exponential with parameter that is equal to the sum of the parameters, we have that for all  $b \in \mathcal{W}$ ,  $i \in I$ ,  $r^{(b)}(i)$  is exponentially distributed with parameter  $w^{(b)}(i)$ .

**Coordinated and independent sketches.** Coordinated sketches are derived from assignments with consistent ranks and independent sketches from assignments with independent ranks.  $k$ -mins sketches: An ordered set of  $k$  rank assignments for  $(I, \mathcal{W})$  defines a set of  $|\mathcal{W}|$   $k$ -mins sketches, one for each assignment  $b \in \mathcal{W}$ . Bottom- $k$  and Poisson sketches: A single rank assignment  $r$  on  $(I, \mathcal{W})$  defines a bottom- $k$  sketch (and a Poisson  $\tau^{(b)}$ -sketch) for each  $b \in \mathcal{W}$ , (using the rank values  $\{r^{(b)}(i) | i \in I\}$ ). Figure 2 shows examples of independent and shared-seed consistent rank assignments for the example data set and the corresponding bottom-3 samples.

Independent differences ranks allow for a generalization of the estimator for unweighted Jaccard similarity [6]:

**Theorem 4.1.**  *$k$ -mins sketches derived from rank assignments with independent-differences consistent ranks have the following property: For any  $b_1, b_2 \in \mathcal{W}$ , the probability that both assignments have the same minimum-rank key is equal to the weighted Jaccard similarity of the two weight assignments.*

*Proof.* Let  $\mathcal{R} = \{b_1, b_2\}$ . Independent differences rank distribution is equivalent to drawing (independently) two exponentially distributed random variables for each key  $i$ ,  $r'(i)$  and  $r''(i)$  according to weights  $w^{(\min \mathcal{R})}(i)$  and  $w^{(L_1 \mathcal{R})}(i)$ . If  $w^{(b_h)}(i) = \min\{w^{(b_1)}(i), w^{(b_2)}(i)\}$  we set  $r^{(b_h)}(i) \leftarrow r'(i)$  and otherwise  $r^{(b_h)}(i) \leftarrow \min\{r'(i), r''(i)\}$ . This follows from the property that the minimum of independent and exponentially distributed variables is exponentially distributed with parameter equal to the sum of the parameters.

To establish the lemma, we need to show that

$$\text{PR}[\min_i r^{(b_1)}(i) = \min_i r^{(b_2)}(i)] = \text{PR}[\min_i r'(i) < \min_i r''(i)].$$

We again use the property that the minimum of independent and exponentially distributed variables is exponentially distributed with the sum of the parameters. Hence,  $\min_i r'(i)$  and  $\min_i r''(i)$  are independent and exponentially distributed with parameters  $\sum_i w^{(\min \mathcal{R})}(i)$  and  $\sum_i w^{(L_1 \mathcal{R})}(i)$ , respectively. Therefore,

$$\text{PR}[\min_i r'(i) < \min_i r''(i)] = \frac{\sum_i w^{(\min \mathcal{R})}(i)}{\sum_i w^{(\min \mathcal{R})}(i) + \sum_i w^{(L_1 \mathcal{R})}(i)} = \frac{\sum_i w^{(\min \mathcal{R})}(i)}{\sum_i w^{(\max \mathcal{R})}(i)}.$$

The last equality is correct because for two independent exponentially distributed random variables with parameters  $y_1, y_2$ , the probability that the  $i$ th is smaller is  $y_i / (y_1 + y_2)$ .  $\square$

It follows that the fraction of common keys in the two  $k$ -mins sketches is an unbiased estimator of the weighted Jaccard similarity of the two assignments.

For Poisson sketches, shared-seed consistent ranks *maximize the sharing of keys between sketches*. In fact, they are the only joint distribution over assignments that does so. We conjecture that this holds also for bottom- $k$  and  $k$ -mins sketches.

**Theorem 4.2.** *Consider all distributions of rank assignments on  $(I, \mathcal{W})$  obtained using a family  $\mathbf{F}_w$ . Shared-seed consistent ranks (is the unique rank distribution that) minimize the expected number of distinct keys in the union of the sketches for  $(I, w^{(b)})$ ,  $b \in \mathcal{W}$ .*

*Proof.* We first consider Poisson sketches. Consider Poisson- $\tau^{(b)}$  sketches ( $b \in \mathcal{R}$ ). Since the inclusion of different keys are independent, it suffices to show the claim for a single key  $i$ . Let  $p^{(b)} = \mathbf{F}_{w^{(b)}(i)}(\tau^{(b)})$ . With any distribution of rank assignments, the probability that  $i$  is included in at least one sketch for  $b \in \mathcal{R}$  is at least  $\max_{b \in \mathcal{R}} p^{(b)}$ . With shared-seed ranks, this probability equals  $\max_{b \in \mathcal{R}} p^{(b)}$ , and hence, it is minimized.

For bottom- $k$  sketches, there is no joint distribution with the property that the inclusion probability of a key in at least one sketch is equal to its maximum inclusion probability over sketches ( $b \in \mathcal{R}$ ). What we can show is that shared-seed is optimal for a key  $i$  when fixing the distribution of  $r(I \setminus \{i\})$ . The argument uses the independence of the rank vectors of different keys. Specifically, for each key  $i$ , the joint distribution of  $\tau^{(b)} = r_k^{(b)}(I \setminus \{i\})$  ( $b \in \mathcal{R}$ ) is independent of the joint distribution of  $r^{(b)}(i)$  ( $b \in \mathcal{R}$ ). Thus, conditioned on the outcome of  $r_k^{(b)}(I \setminus \{i\})$  ( $b \in \mathcal{R}$ ), the inclusion probability of  $i$  in the sketch of  $b$  is  $p^{(b)} = \mathbf{F}_{w^{(b)}(i)}(\tau^{(b)})$ . We can now reuse the argument for Poisson sketches to show that with shared-seed ranks,  $i$  is included in at least one sketch from  $\mathcal{R}$  with probability exactly  $\max_{b \in \mathcal{R}} p^{(b)}$ . Hence, this choice is optimal under this conditioning.  $\square$

**Sketches for the maximum weight.** For  $\mathcal{R} \subset \mathcal{W}$ , let  $r^{(\min \mathcal{R})}(i) = \min_{b \in \mathcal{R}} r^{(b)}(i)$ . The following holds for all consistent rank assignments:

**Lemma 4.1.** *Let  $r$  be a consistent rank assignment for  $(I, \mathcal{W})$  with respect to  $\mathbf{f}_w$  ( $w > 0$ ). Let  $\mathcal{R} \subset \mathcal{W}$ . Then  $r^{(\min \mathcal{R})}(i)$  is a rank assignment for the weighted set  $(I, w^{(\max \mathcal{R})})$  with respect to  $\mathbf{f}_w$  ( $w > 0$ ).*

*Proof.* From the definition of consistency,  $r^{(\min \mathcal{R})}(i) \equiv r^{(b)}(i)$  where  $b = \arg \max_{b \in \mathcal{R}} w^{(b)}(i)$ . Therefore, the distribution of  $r^{(\min \mathcal{R})}(i)$  is  $\mathbf{f}_{w^{(\max \mathcal{R})}(i)}$ . It remains to show that  $\{r^{(\min \mathcal{R})}(i) | i \in I\}$  are independent. This immediately follows from the definition of a rank assignment: if sets of random variables are independent (rank vectors of different keys), so are the respective maxima.  $\square$

A consequence of Lemma 4.1 is the following:

**Lemma 4.2.** *From coordinated Poisson  $\tau^{(b)}$ -/bottom- $k$ -/ $k$ -mins sketches for  $\mathcal{R} \subset \mathcal{W}$ , we can obtain a Poisson  $\min_{b \in \mathcal{R}} \tau^{(b)}$ -/bottom- $k$ -/ $k$ -mins sketch for  $(I, w^{(\max \mathcal{R})})$ .*

*Proof.*  **$k$ -mins sketches:** we take the coordinate-wise minima (and respective keys) of the  $k$ -mins sketch vectors of  $(I, w^{(b)})$ ,  $b \in \mathcal{R}$ .

Given a rank assignment  $r$  for  $(I, \mathcal{W})$  then by Lemma 4.1  $r^{(\min \mathcal{R})}(i)$  is a rank assignment for  $(I, w^{(\max \mathcal{R})})$ . So by the definition of a  $k$ -mins sketch we should take the key achieving  $\min_{i \in I} r^{(\min \mathcal{R})}(i)$  to the  $k$ -mins sketch of  $(I, w^{(\max \mathcal{R})})$ , and repeat this for  $k$  different rank assignments.

Let  $j_b$  be the key such that  $r^{(b)}(j_b)$  is minimum among all  $r^{(b)}(i)$ . The lemma follows since

$$\min_{b \in \mathcal{R}} r^{(b)}(j_b) = \min_{b \in \mathcal{R}} \min_{i \in I} r^{(b)}(i) = \min_{i \in I} \min_{b \in \mathcal{R}} r^{(b)}(i) = \min_{i \in I} r^{(\min \mathcal{R})}(i).$$

**Poisson  $\tau^{(b)}$ -sketches:** we include all keys with rank value at most  $\min_{b \in \mathcal{R}} \tau^{(b)}$  in the union of the sketches.

**Bottom- $k$  sketches:** we take the  $k$  distinct keys with smallest rank values in the union of the sketches. The proof is deferred and is a consequence of Lemma 7.2.  $\square$

This property of coordinated sketches generalizes the union-sketch property of coordinated sketches for global and uniform weights, which facilitates multiple-set aggregates [11, 7, 6, 17, 18].

**Fixed number of distinct keys for colocated data.** The number of distinct keys in a set of bottom- $k$  sketches of different assignments is between  $k$  and  $|\mathcal{W}|k$ . It is closer to  $k$  with coordination of sketches and similarity of assignments. Therefore, even though  $k$ , which is the size of each embedded sample, is fixed the total size varies. A natural goal under storage constraints is to fix the number of distinct keys in the combined sample.

For a rank assignment  $r$ , we choose  $\ell$  ( $\ell \geq k$ ) to be the largest such that there are at most  $|\mathcal{W}|k$  distinct keys in the union of the bottom- $\ell$  samples taken with respect to  $r^{(b)}$  ( $b \in \mathcal{W}$ ). The total number of distinct

keys is at least  $|\mathcal{W}|(k - 1) + 1$ . Such a sample can be computed by a simple adaptation of the stream sampling algorithm for the fixed- $k$  variant. For Poisson sketches, we can similarly fix the *expected size* of the combined samples, a stream algorithm can similarly select  $\tau^{(b)}$  adaptively so that the expected size of each embedded sample is at least  $\ell \geq k$  and the expected total size is in  $[|\mathcal{W}|(k - 1) + 1, |\mathcal{W}|k]$ .

In the sequel we develop estimators over bottom- $k$  sketches. The treatment of Poisson sketches is similar and simpler. Derivations extend easily to bottom- $k^{(b)}$  sketches (with different sizes for different assignments) and to colocated data sketches with fixed number of distinct keys. We shall denote by  $S(r)$  the summary consisting of  $|\mathcal{W}|$  bottom- $k$  sketches obtained using a rank assignment  $r$ .

**Computing coordinated sketches.** Coordinated bottom- $k$  sketches can be computed by a small modification of existing bottom- $k$  sampling algorithms. If weights are colocated the computation is simple (for both shared-seed and independent-differences), as each key is processed once. For dispersed weights and shared-seed, random hash functions must be used to ensure that the same seed  $u(i)$  is used for the key  $i$  in different assignments. We apply the common practice of assuming perfect randomness of the rank assignment in the analysis. This practice is justified by a general phenomenon [49, 43], that simple heuristic hash functions and pseudo-random number generators [3] perform in practice as predicted by this simplified analysis. This phenomenon is also supported by our evaluation.

Independent-differences are not suited for dispersed weights as they require range summable universal hash functions [29, 49].

## 5 Template Estimator

Consider  $(I, \mathcal{W})$  a sample space of rank assignments  $\Omega$ , and a summary  $S$ .

We present an unbiased template estimator for sum aggregates  $\sum_{i \in I | d(i)=1} f(i)$  (where  $f$  is a numeric function  $f$   $d$  is a predicate  $d$ ). The template estimator is an adaptation of the HT estimator to our multiple-assignments setting.

The template estimator assigns adjusted  $f$ -weights  $a^{(f)}(i) \geq 0$  to the keys (we have implicit zero adjusted weights  $a^{(f)}(i) \equiv 0$  for  $i \notin S$ ). The estimate on  $\sum_{i | d(i)=1} f(i)$  is the sum  $\sum_{i \in S | d(i)=1} a^{(f)}(i)$  of the adjusted  $f$ -weights of keys in  $S$  that satisfy the predicate  $d$ .<sup>2</sup>

We subsequently apply this template to different summary types (colocated or dispersed weights), independent or coordinated distributions of rank assignments, and  $f$  and  $d$  with different dependence on the weight vector.

We start with the template estimator for Poisson sketches. Independence of key inclusions implies that there is no gain in considering estimates of  $f(i)$  that depend on keys other than  $i$ . We limit our attention to a single key  $i$ , with weight vector  $w^{(\mathcal{W})}(i)$  and ranks  $r^{\mathcal{W}}(i)$ . Let  $\Omega \equiv \Omega(i)$  denote the sample space of rank assignments for  $i$ ,  $S(r, w)$  the sample obtained when the rank assignment is  $r \in \Omega$  and the weight vector of  $i$  is  $w \equiv w^{(\mathcal{W})}(i)$ , and by  $\mathcal{S}$  the universe of all possible outcomes (in terms of key  $i$ ) over different  $r$  and  $w$ .

### Template estimator (Poisson)

---

<sup>2</sup>With sum aggregates defined this way, the same adjusted  $f$ -weights can be used for different selection predicates  $d()$ . This is natural when multiple queries share the same  $f$  but has different attribute-based selections. For example, the  $L_1$  distance of bandwidth (bytes) for IP destination between two time periods, for different subpopulations of flows (applications, destination AS, etc.) We note that nonetheless, the selection predicate  $d(i)$  is technically redundant, as  $\sum_{i | d(i)=1} f(i) = \sum_{i \in I} d(i)f(i)$ , and we can replace  $f$  and  $d$  with the weight function  $d(i)f(i)$  without a predicate.

Identify  $\mathcal{S}^* \subset \mathcal{S}$  and functions  $f(S) > 0$  and  $0 \leq p(S) \leq 1$  for all  $S \in \mathcal{S}^*$  such that the following holds:

1. for any weight vector  $w$  such that  $d(i) = \text{true}$  and  $f(i) > 0$ ,  $\text{PR}[\mathcal{S}^*|w] > 0$ .
2. for each  $S \in \mathcal{S}^*$ , for all  $w$  and  $r \in \Omega$  such that  $S(r, w) \equiv S$ .
  - $d(i) = \text{true}, f(i) \equiv f(S)$
  - $p(S) \equiv \text{PR}[\mathcal{S}^*|w] > 0$

Estimate:

- if  $S \notin \mathcal{S}^*$ ,  $a^{(f)}(i) \equiv 0$ .
- if  $S \in \mathcal{S}^*$ ,  $a^{(f)}(i) = f(S)/p(S)$ .

Where  $\text{PR}[\mathcal{S}^*|w]$  is the probability that the sample  $S$  is a member of  $\mathcal{S}^*$  when the weight vector of key  $i$  is  $w$ . The first requirement is clearly necessary. It means that any key that has a positive contribution to the aggregate must have a positive probability of being accounted for. It is not hard to see that the template estimator is unbiased. The expectation of  $a^{(f)}(i)$  on weights  $w$  is  $\text{PR}[\mathcal{S}^*|w] * f(S)/p(S) \equiv f(i)$ . This is well defined when the first requirement is satisfied, namely,  $\text{PR}[\mathcal{S}^*|w] > 0$ .

We next apply the RC method to obtain a template estimator for bottom- $k$  sketches. The template can be twicked to handle summaries with sketches with different sizes for each  $b \in \mathcal{W}$  or for colocated samples with fixed number of distinct keys.

For a key  $i$  and rank assignment  $r^{-i} : I \setminus \{i\}$ , we consider the probability space  $\Omega(i, r^{-i})$  containing all rank assignments  $r' : I$  that are identical to  $r^{-i}$  on  $I \setminus \{i\}$  (that is,  $\forall b \in \mathcal{W}, \forall j \in I \setminus \{i\}, r'^{(b)}(j) = r^{(b)}(j)$ ).

We denote by  $S(r, w)$  the sample obtained when the rank assignment is  $r$  and the weights are  $w \equiv w^{(\mathcal{W})}$  and by  $\mathcal{S}$  the universe of all possible outcomes over different  $r$  and  $w$ .

**Template RC estimator (Bottom- $k$  sketches):**

Identify, for each key  $i$ , a subset  $\mathcal{S}^*(i) \subset \mathcal{S}$  and functions  $f(S, i) > 0$  and  $0 \leq p(S, i) \leq 1$  for  $S \in \mathcal{S}^*(i)$  such that the following holds:

1. For all  $i$ , for all  $w$  such that  $d(i) = \text{true}$  and  $f(i) > 0$ , and each  $r^{-i}$ ,  $\text{PR}[\mathcal{S}^*(i)|w, r^{-i}] > 0$ .
2. for all  $w$  and  $r$  such that  $S(r, w) \equiv S$ :
  - $d(i) = \text{true}, f(i) \equiv f(S, i)$
  - $p(S, i) \equiv \text{PR}[\mathcal{S}^*(i)|w, r^{-i}]$ .

Estimate:

- if  $S \notin \mathcal{S}^*(i)$ ,  $a^{(f)}(i) \equiv 0$ .
- if  $S \in \mathcal{S}^*(i)$ ,  $a^{(f)}(i) = f(S, i)/p(S, i)$ .

Where for weights  $w$  and ranks  $r^{-i}$  for  $I \setminus \{i\}$ ,  $\text{PR}[\mathcal{S}^*(i)|w, r^{-i}]$  is the probability that a sample is in  $\mathcal{S}^*(i)$  over  $\Omega(i, r^{-i})$ .

This formulation is an instance of HTP that builds on the Rank Conditioning (RC) method [17] (see Section 3). Correctness, which is equivalent to saying that for every  $i \in I$ ,  $\mathbb{E}[a^{(f)}(i)] = f(i)$ , is immediate from HTP. The estimate is well defined when  $p(S, i) > 0$ , which follows from the first requirement.

The template is fully specified by the selection of  $\mathcal{S}^*$ . A necessary condition for inclusion of  $S$  in  $\mathcal{S}^*$  is that  $d(i)$  and  $f(i)$  can be determined from  $S$ . Ideally,  $\mathcal{S}^*$  would include all such samples, but it is not always possible, as  $\text{PR}[\mathcal{S}^*|w]$  may not be unique across all applicable  $w$  that are consistent with outcome  $S$ .

We show that the more inclusive  $\mathcal{S}^*(i)$  is, the lower the variance of  $a^{(f)}(i)$ .

To obtain tight estimators, we need to apply the template with the *most inclusive suitable selection*  $\mathcal{S}^*$ .

**Lemma 5.1.** *Consider two selections  $\mathcal{S}_1^*(i)$  and  $\mathcal{S}_2^*(i)$  such that  $\mathcal{S}_1^*(i) \subseteq \mathcal{S}_2^*(i)$ . Let  $a_1^{(f)}$  and  $a_2^{(f)}$  be the corresponding adjusted weights of  $i$ . Then for all  $w$ ,  $\text{VAR}[a_1^{(f)}(i)] \geq \text{VAR}[a_2^{(f)}(i)]$ .*

*Proof.* Fix  $w$  and  $r^{-i}$ . Let  $p_h(i, r^{-i})$  be the corresponding probabilities, in  $\Omega(i, r^{-i})$ , that  $S \in \mathcal{S}_h^*(i)$  ( $h = 1, 2$ ).

From definition, for all  $i \in I$  and  $r \in \Omega$ ,  $p_1(i, r^{-i}) \leq p_2(i, r^{-i})$ .

Observe now that it suffices to establish the relation of the variance for a particular  $\Omega(i, r^{-i})$  (since the projection of  $r$  on  $I \setminus \{i\}$  is a partition of  $\Omega$  and the adjusted weights are unbiased in each partition.) We have  $\text{VAR}_{\Omega(i, r^{-i})}[a_h^{(f)}(i)] = f(i)^2(1/p_h(i, r^{-i}) - 1)$  for  $h = 1, 2$  (variance of the HT estimator on  $\Omega(i, r^{-i})$ ).  $\square$

When adjusted weights  $a^{(f)}$  of different keys have zero covariances,  $\text{VAR}[a^{(f)}(J)] = \sum_{j \in J} \text{VAR}[a^{(f)}(j)]$ . In particular, this means that a most inclusive  $\mathcal{S}^*$  implies at most the variance for any subset  $J$ .

## 6 Colocated weights

We apply the template (Section 5) to summaries of data sets with colocated weights.

We apply the template with  $\mathcal{S}^*(i) \equiv \{S \in \mathcal{S} | i \in S\}$ , that is, all samples where  $i$  is included in the union of the single-assignment “embedded” samples. This is the most inclusive possible selection and we refer to it as the *inclusive* estimator.

A requirement for existence of a template estimator is that for all  $f, d, w$ , and  $i$ ,  $f(i)d(i) > 0$  implies that  $i$  is sampled with positive probability. This is the first requirement of the template with the inclusive selection substituted for  $\mathcal{S}^*(i)$ . If the requirement holds for any other selection (must be a proper subset of the inclusive one), it must also hold for the inclusive selection.

With all considered forms of (single-assignment) weighted sampling, a key has a positive probability to be sampled if and only if it has a positive weight. Therefore, equivalent requirement is

$$f(i)d(i) > 0 \implies w^{(\max \mathcal{W})}(i) > 0. \quad (3)$$

We show that (3) is also sufficient, that is, the inclusive estimator is defined whenever (3) holds. Recall that in the colocated model, once a key is sampled its full weight vector  $w^{(\mathcal{W})}(i)$  is available with  $S$ . Hence,  $f(i)$  and  $d(i)$  can be determined from  $S$  for all  $i \in S$ .

It remains to show how to compute  $p(S, i)$  for each  $i \in S$ . For Poisson sketches,  $i \in S$  if and only if for at least one  $b \in \mathcal{W}$ ,  $r^{(b)}(i) \leq \tau^{(b)}$ . For bottom- $k$  sketches,  $i \in S$  if and only if for at least one  $b \in \mathcal{W}$ ,  $r^{(b)}(i) \leq r_k^{(b)}(I \setminus \{i\})$ .

We provide explicit expressions for EXP and IPPS rank distributions and bottom- $k$  sketches. The derivation for Poisson is omitted, but the expressions can be obtained by substituting  $\tau^{(b)}$  for  $r_k^{(b)}(I \setminus \{i\})$ .

The probability that  $i$  is included in  $S$  over  $\Omega(i, r^{-i})$  is:

$$p(i, r^{-i}) = \text{PR}[\exists b \in \mathcal{W}, r^{(b)}(i) < r_k^{(b)}(I \setminus \{i\}) \mid r' \in \Omega(i, r^{-i})]. \quad (4)$$

To compute (4), the summary should include, for each  $b \in \mathcal{W}$ , the rank values  $r_k^{(b)}(I)$  and  $r_{k+1}^{(b)}(I)$  and for each  $i \in S(r)$  and  $b \in \mathcal{W}$ , whether  $i$  is included in the bottom- $k$  sketch of  $b$  (that is, whether  $r_k^{(b)}(i) < r_{k+1}^{(b)}(I)$ ). This information allows us to determine the values  $r_k^{(b)}(I \setminus \{i\})$  for all  $i \in I$  and  $b \in \mathcal{W}$ :

The values  $r_k^{(b)}(I \setminus \{i\})$  can be determined from  $S$  as follows: if  $i$  is included in the sketch for  $b$  then  $r_k^{(b)}(I \setminus \{i\}) = r_{k+1}^{(b)}(I)$ . Otherwise,  $r_k^{(b)}(I \setminus \{i\}) = r_k^{(b)}(I)$ . These values are constant over  $\Omega(i, r^{-i})$ . We provide explicit expressions for  $p(i, r^{-i})$  (Eq. (4)), for  $i \in S(r)$ .

**Independent ranks** (independent bottom- $k$  sketches): The probability over  $\Omega(i, r^{-i})$  that  $i$  is included in the bottom- $k$  sketch of  $b$  is  $\mathbf{F}_{w^{(b)}(i)}(r_k^{(b)}(I \setminus \{i\}))$ . It is included in  $S(r')$  if and only if it is included for at least one of  $b \in \mathcal{W}$ . Since  $r^{(b)}(i)$  are independent,

$$p(i, r^{-i}) = 1 - \prod_{b \in \mathcal{W}} (1 - \mathbf{F}_{w^{(b)}(i)}(r_k^{(b)}(I \setminus \{i\}))). \quad (5)$$

Specifically,

For EXP ranks: 
$$p(i, r^{-i}) = 1 - \exp\left(-\sum_{b \in \mathcal{W}} w^{(b)}(i) r_k^{(b)}(I \setminus \{i\})\right).$$

For IPPS ranks: 
$$p(i, r^{-i}) = 1 - \prod_{b \in \mathcal{W}} (1 - \min\{1, w^{(b)}(i) r_k^{(b)}(I \setminus \{i\})\}).$$

**Shared-seed consistent ranks** (coordinated bottom- $k$  sketches): Item  $i$  is included in the sketch of assignment  $b$  for  $r' \in \Omega(i, r^{-i})$  if and only if  $u(i) \leq \mathbf{F}_{w^{(b)}(i)}(r_k^{(b)}(I \setminus \{i\}))$ . The probability that it is included for at least one of  $b \in \mathcal{W}$  is

$$p(i, r^{-i}) = \max_{b \in \mathcal{W}} \{\mathbf{F}_{w^{(b)}(i)}(r_k^{(b)}(I \setminus \{i\}))\}. \quad (6)$$

Specifically,

For EXP ranks: 
$$p(i, r^{-i}) = 1 - \exp(-\max_{b \in \mathcal{W}} \{w^{(b)}(i) r_k^{(b)}(I \setminus \{i\})\}).$$

For IPPS ranks: 
$$p(i, r^{-i}) = \min\left\{1, \max_{b \in \mathcal{W}} \{w^{(b)}(i) r_k^{(b)}(I \setminus \{i\})\}\right\}.$$

**Independent-differences consistent ranks** (coordinated bottom- $k$  sketches): Let  $w^{(b_1)}(i) \leq \dots \leq w^{(b_h)}(i)$  be the entries of the weight vector of  $i$ . Recall that  $r^{(b_j)}(i) \leftarrow \min_{a=1}^j d_j$  where  $d_j \leftarrow \text{EXP}[w^{(b_j)}(i) - w^{(b_{j-1})}(i)]$  (we define  $w^{(0)}(i) \equiv 0$  and  $\text{EXP}[0] \equiv 0$ ).

We also define  $M_\ell = \max_{a=\ell}^h r_k^{(b_a)}(I \setminus \{i\})$  ( $\ell \in [h]$ ), and the event  $A_j$  to consist of all rank assignments such that  $j$  is the smallest index for which  $d_j \leq M_j$ . Clearly the events  $A_j$  are disjoint and  $p(i, r^{-i}) = \sum_{\ell=1}^h \text{PR}[A_\ell]$ .

The probabilities  $\text{PR}[A_\ell]$  can be computed using a linear pass on the sorted weight vector of  $i$  using the independence of  $d_\ell$ 's as follows

$$\begin{aligned} \text{PR}[A_1] &= \text{Pr}[d_1 \leq M_1] = \mathbf{F}_{w^{(b_1)}(i)}(M_1); \\ \text{PR}[A_2] &= \text{Pr}[d_1 > M_1 \wedge d_2 \leq M_2] = (1 - \mathbf{F}_{w^{(b_1)}(i)}(M_1)) \mathbf{F}_{w^{(b_2)}(i) - w^{(b_1)}(i)}(M_2); \\ &\dots \\ \text{PR}[A_\ell] &= \text{Pr}\left[\bigwedge_{a=1}^{\ell-1} (d_a > M_a) \wedge d_\ell \leq M_\ell\right] = \prod_{j=1}^{\ell-1} (1 - \mathbf{F}_{w^{(b_j)}(i) - w^{(b_{j-1})}(i)}(M_j)) \cdot \mathbf{F}_{w^{(b_\ell)}(i) - w^{(b_{\ell-1})}(i)}(M_\ell). \end{aligned}$$

**Generic consistent rank assignments (coordinated sketches)** Let  $\mathcal{R} \subset \mathcal{W}$  be the set of assignments relevant for  $f$  and  $d$ . Let  $r_k^{(\min \mathcal{R})}(I \setminus \{i\}) \equiv \min_{b \in \mathcal{R}} r_k^{(b)}(I \setminus \{i\})$ . We use a more restrictive selection

$$i \in \mathcal{S}^*(i) \iff \min_{b \in \mathcal{R}} r^{(b)}(i) \leq r_k^{(\min \mathcal{R})}(I \setminus \{i\}). \quad (7)$$

For all consistent rank assignments, we have

$$p(i, r^{-i}) = \mathbf{F}_{w^{(\max \mathcal{R})}(i)}(r_k^{(\min \mathcal{R})}(I \setminus \{i\})).$$

notation	explanation
$w^{(\mathcal{R})}(i)$ and $r^{(\mathcal{R})}(i)$	the weight and rank assignments restricted to $\mathcal{R}$ .
$r_h^{(\min \mathcal{R})}(J) = \min_{b \in \mathcal{R}} r_h^{(b)}(J)$	the smallest $h$ -smallest rank, over $\mathcal{R}$
$r_h^{(\ell^{\text{th}} \text{smallest } \mathcal{R})}(J)$	The $\ell^{\text{th}}$ smallest value in $\{r_h^{(b)}(J)   b \in \mathcal{R}\}$
$r^{(\min \mathcal{R})}(i) = \min_{b \in \mathcal{R}} r^{(b)}(i)$	The min rank value of $i$ over $\mathcal{R}$
$r^{(\max \mathcal{R})}(i) = \max_{b \in \mathcal{R}} r^{(b)}(i)$	the max rank value of $i$ over $\mathcal{R}$
$w^{(\max \mathcal{R})}(i) = \max_{b \in \mathcal{R}} w^{(b)}(i)$	Maximum weight $i$ assumes over $\mathcal{R}$
$w^{(\min \mathcal{R})}(i) = \min_{b \in \mathcal{R}} w^{(b)}(i)$	Minimum weight $i$ assumes over $\mathcal{R}$
$w^{(\ell^{\text{th}} \text{-largest } \mathcal{R})}(i)$	$\ell^{\text{th}}$ largest weight $i$ assumes over $\mathcal{R}$ (in $\{w^{(b)}(i)   b \in \mathcal{R}\}$ )
$w^{(\text{top-}\ell \mathcal{R})}(i)$	Restriction of $w^{(\mathcal{R})}(i)$ to the top- $\ell$ assignments of $i$ .
$b^{(\max \mathcal{R})}(i) = \arg \max_{b \in \mathcal{R}} w^{(b)}(i)$	The weight assignment from $\mathcal{R}$ which maximizes $i$ 's weight.
$b^{(\min \mathcal{R})}(i) = \arg \min_{b \in \mathcal{R}} w^{(b)}(i)$	The weight assignment from $\mathcal{R}$ which minimizes $i$ 's weight.
$b^{(\text{top-}\ell \mathcal{R})}(i)$	Subset of $\mathcal{R}$ containing assignments with top- $\ell$ weights for $i$
$b^{(\ell^{\text{th}} \text{-largest } \mathcal{R})}(i)$	assignment $b \in \mathcal{R}$ with $\ell^{\text{th}}$ largest weight $w^{(b)}(i)$

Table 1: Notation table.  $\mathcal{R} \subset \mathcal{W}$  is a subset of assignments,  $i \in I$ , is a key,  $J \subset I$  is a subset of keys,  $h \geq 1$  is an integer. When the dependency on  $\mathcal{R}$  is clear from context, it is omitted.

It is easy to see that  $\mathcal{S}^*$  satisfies the requirements of the template estimator (Section 5) for any consistent ranks distribution. While generic and simpler than the tailored derivations for shared-seed and independent differences, this estimator is weaker because  $\mathcal{S}^*$  is less inclusive (immediate consequence of Lemma 5.1).

## 7 Dispersed weights

Let  $r$  be a rank assignment for  $(I, \mathcal{W})$ . As in the co-located model the summary  $S(r)$  contains all keys in the bottom- $k$  sketches  $s_k(I, r^{(b)})$  for  $b \in \mathcal{W}$ . But in the dispersed weights model  $w^{(b)}(i)$  (for  $i \in I, b \in \mathcal{W}$ ) is included in  $S(r)$  if and only if  $i \in s_k(I, r^{(b)})$ . Table 1 summarizes the notation we use in this section.

An *aggregation* is specified by the pair  $f, d$ . Assignments  $\mathcal{R} \subset \mathcal{W}$  are *relevant assignments* for an aggregation if  $f$  and  $d$  depend only on  $w^{(\mathcal{R})}$ . In the dispersed weights model, samples taken for assignments not in  $\mathcal{R}$  do not contain any useful information for estimating our aggregate. We are therefore interested in identifying a minimal such set  $\mathcal{R}$ .

We next characterize a class of aggregates which includes the minimum ( $f = w^{(\min \mathcal{R})}$ ), maximum ( $f = w^{(\max \mathcal{R})}$ ) and quantiles over a set  $\mathcal{R}$  of assignments (for example,  $f(i)$  being the median, in this case the 4<sup>th</sup> largest weight, of  $\{w^{(1)}(i), w^{(2)}(i), \dots, w^{(7)}(i)\}$ ).

**Definition 7.1.** We say that an aggregation  $(f, d)$  is  $\text{top-}\ell$  dependent if

$$f(i) \equiv f(w^{(\text{top-}\ell \mathcal{R})}(i), b^{(\text{top-}\ell \mathcal{R})}(i)) \quad (8)$$

$$d(i) \equiv d(w^{(\text{top-}\ell \mathcal{R})}(i), b^{(\text{top-}\ell \mathcal{R})}(i)) \quad (9)$$

$$w^{(\ell^{\text{th}} \text{-largest } \mathcal{R})}(i) = 0 \Rightarrow d(i)f(i) = 0. \quad (10)$$

Notice that functions  $f$  and  $d$  that satisfy (10) for some  $\ell$  also satisfy (10) for  $\ell' < \ell$  but do not necessarily satisfy (8) and (9) for  $\ell' < \ell$ . Of special interest are the two extreme cases: *max-dependence*, when  $\ell = 1$ , and *min-dependence*, when  $\ell = |\mathcal{R}|$ . With min-dependence, (8) and (9) are redundant (always hold) and (10) is  $w^{(\min \mathcal{R})}(i) = 0 \Rightarrow f(i)d(i) = 0$ .



The aggregations specified by  $f(i) = w^{(\min \mathcal{R})}(i)$  and any predicate  $d$  are min-dependent, but not  $\text{top-}\ell$  dependent for any  $\ell \neq |\mathcal{R}|$ . The aggregations specified by  $f(i) = w^{(\max \mathcal{R})}(i)$  and any attribute-based predicate  $d$  are max-dependent, but not  $\text{top-}\ell$  dependent for any  $\ell \neq 1$ . More generally, aggregations specified by  $f(i) = w^{(\ell^{\text{th}}\text{-largest } \mathcal{R})}(i)$  (the  $\ell^{\text{th}}$  largest weight) and attribute-based  $d$  are  $\text{top-}\ell$  dependent but not  $\text{top-}h$  dependent for any  $h \neq \ell$ .

We use our template estimator to obtain unbiased nonnegative estimators for  $\text{top-}\ell$  dependent aggregations over coordinated sketches ( $1 \leq \ell \leq |\mathcal{R}|$ ) and min-dependent aggregations over independent sketches.

We present two such estimators which we name *s-set* and *l-set* and respectively denote  $\mathcal{S}_s^*$  and  $\mathcal{S}_l^*$  the respective selections in the template estimator. We have that  $\mathcal{S}_l^* \supset \mathcal{S}_s^*$  and moreover,  $\mathcal{S}_l^*$  is a maximal set for which we can determine the  $\ell$  largest weights. Hence, from Lemma 5.1, the *l-set* estimator dominates the *s-set* estimator. On the other hand, *s-set* estimators have a simple closed universal expression for all coordinated sketches whereas we present a closed-form *l-set* estimators only for shared-seed coordinated sketches.

## 7.1 *s-set top-ℓ dependence*

*s-set top-ℓ dependence estimator (coordinated sketches):*

$$\begin{array}{l}
\forall i \in S, \\
\bullet \mathcal{R}'(i) \leftarrow \left\{ b \in \mathcal{R} \mid r^{(b)}(i) < r_k^{(\min \mathcal{R})}(I \setminus \{i\}) \right\} \\
\bullet \text{if } |\mathcal{R}'(i)| \geq \ell \quad \{ * \quad S \in \mathcal{S}_s^*(i) \quad * \} \\
\\
w^{(\text{top-}\ell \mathcal{R})}(i) \leftarrow w^{(\text{top-}\ell \mathcal{R}'(i))}(i) \\
b^{(\text{top-}\ell \mathcal{R})}(i) \leftarrow b^{(\text{top-}\ell \mathcal{R}'(i))}(i) \\
f(i) \leftarrow f(w^{(\text{top-}\ell \mathcal{R})}(i), b^{(\text{top-}\ell \mathcal{R})}(i)) \\
d(i) \leftarrow d(w^{(\text{top-}\ell \mathcal{R})}(i), b^{(\text{top-}\ell \mathcal{R})}(i)) \\
p(i, r^{-i}) \leftarrow \mathbf{F}_{w^{(\ell^{\text{th}}\text{-largest } \mathcal{R})}(i)}(r_k^{(\min \mathcal{R})}(I \setminus \{i\})) \\
a^f(i) \leftarrow \frac{f(i)}{p(i, r^{-i})} \\
\\
\text{Output } \sum_{i|S \in \mathcal{S}_s^*(i) \wedge d(i)} a^f(i)
\end{array}$$

For  $f(i) = w^{(\max \mathcal{R})}(i)$  ( $\ell = 1$ ),

$$S \in \mathcal{S}_s^* \iff \exists b \in \mathcal{R} \mid r^{(b)}(i) < r_k^{(\min \mathcal{R})}(I \setminus \{i\})$$

and we obtain the adjusted weights:

$$a^{(\max \mathcal{R})}(i) = \frac{w^{(\max \mathcal{R})}(i)}{\mathbf{F}_{w^{(\max \mathcal{R})}(i)}(r_k^{(\min \mathcal{R})}(I \setminus \{i\}))}. \quad (11)$$

**Correctness:** Clearly when rank assignments are consistent or independent, any  $\ell$ -dependent  $f$  and  $d$  satisfy the first requirement of the template estimator, namely, for any  $i$  with  $f(i)d(i) > 0$  and any fixed assignment on ranks to  $I \setminus \{i\}$ , there is positive probability that  $S \in \mathcal{S}_s^*(i)$ . The  $\ell$  largest weights of  $i$  are all positive and there is a positive probability that all respective assignments have maximum rank at most  $r_k^{(\min \mathcal{R})}(I \setminus \{i\})$ .

We now establish that the requirements of the template estimator are satisfied and that it is correctly applied.

**Lemma 7.2.** *Let  $r$  be a consistent rank assignment.*

(i) If  $S \in \mathcal{S}_s^*(i)$ ,

$$\begin{aligned} w^{(\text{top-}\ell \mathcal{R})}(i) &= w^{(\text{top-}\ell \mathcal{R}'(i))}(i), \\ b^{(\text{top-}\ell \mathcal{R})}(i) &= b^{(\text{top-}\ell \mathcal{R}'(i))}(i). \end{aligned}$$

(ii) *The computation of  $p(i, r^{-i})$  is correct.*

*Proof.* (i): We have  $S \in \mathcal{S}_s^*(i)$  if and only if  $|\mathcal{R}'(i)| \geq \ell$ . Therefore, there is some  $\ell' \geq \ell$  such that  $b^{(\ell'^{\text{th}} \text{largest } \mathcal{R})}(i) \in \mathcal{R}'(i)$ . It suffices to show that for all  $h \leq \ell'$ ,  $b^{(h^{\text{th}} \text{largest } \mathcal{R})}(i) \in \mathcal{R}'(i)$ . From consistency of ranks,  $w^{(b_1)}(i) > w^{(b_2)}(i)$  if and only if  $r^{(b_1)}(i) < r^{(b_2)}(i)$ . Therefore,  $r^{(b_1)}(i) < r_k^{(\min \mathcal{R})}(I \setminus \{i\})$  implies  $r^{(b_2)}(i) < r_k^{(\min \mathcal{R})}(I \setminus \{i\})$ . Equivalently,  $b_1 \in \mathcal{R}'(i)$  implies  $b_2 \in \mathcal{R}'(i)$ .

(ii): The value  $r_k^{(\min \mathcal{R})}(I \setminus \{i\})$  depends on the rank values of all keys other than  $i$  and is the same for all assignments  $r' \in \Omega(i, r^{-i})$ . It can always be computed from  $S(r)$  since  $r_k^{(b)}(I \setminus \{i\})$  can be determined for all  $b \in \mathcal{R}$  regardless if  $i \in s_k(I, r^{(b)})$  or not ( $r_k^{(b)}(I \setminus \{i\})$  is the  $(k+1)$ th smallest rank value if  $i \in s_k(I, r^{(b)})$  and is the  $k$ th smallest rank value otherwise). By the definition of the template estimator  $p(i, r^{-i})$  should be the probability, conditioned on  $\Omega(i, r^{-i})$ , that for at least  $\ell$  assignments  $b \in \mathcal{R}$  we have  $b \in \mathcal{R}'(i)$ , that is,  $r^{(b)}(i) < r_k^{(\min \mathcal{R})}(I \setminus \{i\})$ . Using part (i), this is equivalent to the condition that for all  $b \in b^{(\text{top-}\ell \mathcal{R})}(i)$ ,  $r^{(b)}(i) < r_k^{(\min \mathcal{R})}(I \setminus \{i\})$ . From consistency of ranks, this condition holds if and only if the  $\ell$ th smallest rank in  $r^{(\mathcal{R})}(i)$ , associated with the  $\ell^{\text{th}}$  largest weight, is less than  $r_k^{(\min \mathcal{R})}(I \setminus \{i\})$ . This probability is exactly  $\mathbf{F}_{w^{(\ell^{\text{th}} \text{largest } \mathcal{R})}(i)}(r_k^{(\min \mathcal{R})}(I \setminus \{i\}))$  and the lemma follows.  $\square$

**Lemma 7.3.** *For  $\ell = 1$  (max-dependence),  $|\{i | S \in \mathcal{S}_s^*(i)\}| \geq k - 1$  (at least  $k - 1$  keys obtain nonnegative adjusted weights).*

*Proof.* Let  $b$  be such that  $r_k^{(b)}(I)$  is minimized. each one of the  $k - 1$  smallest-rank keys in  $s_k(I, r^{(b)})$  must have  $S \in \mathcal{S}_s^*(i)$ .  $\square$

### 7.1.1 Min-dependence s-set estimator.

The s-set estimator has a particularly simple formulation when  $\ell = |\mathcal{R}|$  (min-dependence). The expression for  $p_s(i, r^{-i})$  below holds for any rank distribution and can be computed also for independent ranks.

**Min-dependence s-set estimator:**

- $S \in \mathcal{S}_s^*(i) \iff \forall b \in \mathcal{R}, r^{(b)}(i) < r_{k+1}^{(\min \mathcal{R})}(I)$
- if  $S \in \mathcal{S}_s^*(i)$ ,  $p_s(i, r^{-i}) \leftarrow \text{PR}[\forall b \in \mathcal{R}, r^{(b)}(i) < r_{k+1}^{(\min \mathcal{R})}(I) \mid r' \in \Omega(i, r^{-i})]$

We have  $S \in \mathcal{S}_s^*(i)$  if  $i$  is included in all  $|\mathcal{R}|$  sketches with rank value that is at most  $r_{k+1}^{(\min \mathcal{R})}(I)$ . For coordinated sketches, we have that for all  $b \in \mathcal{R}$ ,  $r^{(b)}(i) < r_{k+1}^{(\min \mathcal{R})}(I)$  if and only if  $r^{(\max \mathcal{R})}(i) < r_{k+1}^{(\min \mathcal{R})}(I)$ . Therefore,

$$\begin{aligned} p_s(i, r^{-i}) &= \text{Pr}[r^{(\max \mathcal{R})}(i) < r_{k+1}^{(\min \mathcal{R})}(I) \mid r' \in \Omega(i, r^{-i})] \\ &= \mathbf{F}_{w^{(\min \mathcal{R})}(i)}(r_{k+1}^{(\min \mathcal{R})}(I)) \end{aligned}$$

For independent sketches, for each  $b \in \mathcal{R}$ , the events  $r^{(b)}(i) < r_{k+1}^{(\min \mathcal{R})}(I)$  are independent. Therefore,

$$p_s(i, r^{-i}) = \prod_{b \in \mathcal{R}} \mathbf{F}_{w^{(b)}(i)}(r_{k+1}^{(\min \mathcal{R})}(I))$$

For coordinated sketches and  $f(i) = w^{(\min \mathcal{R})}(i)$ ,

$$a_s^{(\min \mathcal{R})}(i) = \frac{w^{(\min \mathcal{R})}(i)}{\mathbf{F}_{w^{(\min \mathcal{R})}(i)}(r_{k+1}^{(\min \mathcal{R})}(I))}, \quad (12)$$

(when  $S \in \mathcal{S}_s^*(i)$ ,  $a_s^{(\min \mathcal{R})}(i) = 0$  otherwise).

## 7.2 $l$ -set top- $\ell$ dependence

The  $l$ -set estimator which we now present is the tightest template estimator for  $f(i) = w^{(\ell^{th} \text{ largest } \mathcal{R})}(i)$ :  $S \in \mathcal{S}_l^*(i)$  if and only if the sample (with  $k-1$ ) contains sufficient information to determine  $w^{(\text{top-}\ell \mathcal{R})}$  and  $b^{(\text{top-}\ell \mathcal{R})}$ . The latter is a necessary condition in the template and therefore  $\mathcal{S}_l^*$  is the most inclusive possible selection (Lemma 5.1).

When  $\ell = 1$  (max-dependence), the  $l$ -set and  $s$ -set estimators are the same. When  $\ell = |\mathcal{R}|$  (min-dependence),  $S \in \mathcal{S}_l^*(i)$  when  $i$  is included in all  $|\mathcal{R}|$  bottom- $k$  sketches. For  $\ell$ -dependence,  $S \in \mathcal{S}_l^*(i)$  if the top  $\ell$  weights of  $i$  are included in  $S$  and we have upper bound on all other weights that are no more than  $w^{(\ell^{th} \text{ largest } \mathcal{R})}(i)$ . We can obtain such upper bounds when the seeds  $u^{(b)}(i)$  are readily available for all  $b \in \mathcal{R}$  and  $i \in S(r)$ . Under these conditions we can compute  $f(i)$  and  $d(i)$  when  $S \in \mathcal{S}_l^*(i)$  and we will show that we can also compute  $p(i, r^{-i})$ .

Consider sketches where the seeds  $u^{(b)}(i)$  are available for all  $i \in S(r)$  and  $b \in \mathcal{R}$ . This *known seeds* requirement holds for shared-seed coordinated sketches (since  $u^{(b)}(i) \equiv u(i)$  and  $u(i)$  is available for all  $i \in S(r)$ ) and can be explicitly made for independent sketches.

**top- $\ell$  dependence  $l$ -set estimator:**

- For  $i \in S$ :
  - $\mathcal{R}'(i) \leftarrow \{b \in \mathcal{R} \mid r^{(b)}(i) < r_{k+1}^{(b)}(I)\}$
  - If  $|\mathcal{R}'(i)| \geq \ell$  and  $\forall b \in \mathcal{R} \setminus b^{(\text{top-}\ell \mathcal{R}'(i))}(i)$ ,  $u^{(b)}(i) < \mathbf{F}_{w^{(\ell^{th} \text{ largest } \mathcal{R}'(i))}(i)}}(r_k^{(b)}(I \setminus \{i\}))$
- $l^* S \in \mathcal{S}_l^*(i) \text{ *!}$
- $w^{(\text{top-}\ell \mathcal{R})}(i) \leftarrow w^{(\text{top-}\ell \mathcal{R}'(i))}(i)$ ,
- $b^{(\text{top-}\ell \mathcal{R})}(i) \leftarrow b^{(\text{top-}\ell \mathcal{R}'(i))}(i)$ ,
- $f(i) \leftarrow f(w^{(\text{top-}\ell \mathcal{R})}(i), b^{(\text{top-}\ell \mathcal{R})}(i))$
- $d(i) \leftarrow d(w^{(\text{top-}\ell \mathcal{R})}(i), b^{(\text{top-}\ell \mathcal{R})}(i))$
- $p_\ell(i, r^{-i}) \leftarrow \text{PR}[\forall b \in b^{(\text{top-}\ell \mathcal{R})}(i), r^{(b)}(i) < r_{k+1}^{(b)}(I) \wedge$   
 $\forall b \in \mathcal{R} \setminus b^{(\text{top-}\ell \mathcal{R})}(i), u^{(b)}(i) < \mathbf{F}_{w^{(\ell^{th} \text{ largest } \mathcal{R})}(i)}}(r_k^{(b)}(I \setminus \{i\})) \mid r' \in \Omega(i, r^{-i})]$
- $a^f(i) \leftarrow \frac{f(i)}{p(i, r^{-i})}$

For shared-seed coordinated sketches

$$p_\ell(i, r^{-i}) = \min\left\{ \min_{b \in b^{(\text{top-}\ell \mathcal{R})}(i)} \mathbf{F}_{w^{(b)}(i)}(r_{k+1}^{(b)}(I)), \min_{b \notin b^{(\text{top-}\ell \mathcal{R})}} \mathbf{F}_{w^{(\ell^{th} \text{ largest } \mathcal{R})}(i)}}(r_{k+1}^{(b)}(I)) \right\} \quad (13)$$

For independent ranks with known seeds:

$$p_\ell(i, r^{-i}) = \prod_{b \in b(\text{top-}\ell \mathcal{R})(i)} \mathbf{F}_{w^{(b)}(i)}(r_{k+1}^{(b)}(I)) \prod_{b \notin b(\text{top-}\ell \mathcal{R})(i)} \mathbf{F}_{w^{(\ell^{\text{th}} \text{largest})(i)}}(r_{k+1}^{(b)}(I)). \quad (14)$$

The **Min-dependence l-set estimator** has the simpler form (The l-set estimator uses the seed values  $u_i$  only when  $l < |\mathcal{R}|$ ):

- $S \in \mathcal{S}_l^*(i) \iff \forall b \in \mathcal{R}, r^{(b)}(i) < r_{k+1}^{(b)}(I)$
- if  $S \in \mathcal{S}_l^*(i)$ ,  $p_\ell(i, r^{-i}) \leftarrow \text{PR}[\forall b \in \mathcal{R}, r^{(b)}(i) < r_{k+1}^{(b)}(I) \mid r' \in \Omega(i, r^{-i})]$

$p_\ell(i, r^{-i})$  for shared-seed consistent ranks is:

$$p_\ell(i, r^{-i}) = \min_{b \in \mathcal{R}} \mathbf{F}_{w^{(b)}(i)}(r_{k+1}^{(b)}(I)). \quad (15)$$

For EXP ranks,

$$p_\ell(i, r^{-i}) = 1 - \exp(-\min_{b \in \mathcal{R}} w^{(b)}(i) r_{k+1}^{(b)}(I))$$

and for IPSS ranks,  $p_\ell(i, r^{-i}) = \min\{1, \min_{b \in \mathcal{R}} \{w^{(b)}(i) r_{k+1}^{(b)}(I)\}\}$ .

For independent sketches:

$$p_\ell(i, r^{-i}) = \prod_{b \in \mathcal{R}} \mathbf{F}_{w^{(b)}(i)}(r_{k+1}^{(b)}(I)). \quad (16)$$

By contrasting (13) and (14) or (15) and (16) we can see that the respective inclusion probability can be exponentially smaller (in  $|\mathcal{R}|$ ) for independent sketches than with coordinated sketches. Since the variance  $\text{VAR}[a(i)]$  is proportional to  $(\frac{1}{p_\ell(i, r^{-i})} - 1)$ , we can have exponentially larger variance.

**Correctness:** It is easy to see that the first requirement of the template estimator (Section 5) is satisfied: for both consistent and independent ranks, for  $\ell$ -dependent  $f$  and  $d$ , and  $i$  such that  $f(i)d(i) > 0$  has  $\text{PR}[S \in \mathcal{S}_l^*(i)] > 0$ . Furthermore, when  $S \in \mathcal{S}_l(i)$ , the top- $\ell$  weights  $w^{(\text{top-}\ell \mathcal{R})}(i)$  are available from  $S(r)$  and therefore  $f$  and  $d$  can be evaluated. Lastly, it is easy to see that the computation of the inclusion probability is correct and that it depends only on the projection of  $r$  on  $I \setminus \{i\}$ . The value  $r_k^{(b)}(I \setminus \{i\})$  can always be determined from  $S$ : it is equal to  $r_{k+1}^{(b)}(I)$  when  $i$  is in the sketch of  $b$  and to  $r_k^{(b)}(I)$  otherwise.

Let  $a_i^{(\min \mathcal{R})}(i)$  be the adjusted weight for  $f(i) = w^{(\min \mathcal{R})}(i)$  of the l-set estimator using shared-seed consistent ranks, and let  $a_{i \text{nd}}^{(\min \mathcal{R})}(i)$  be the adjusted weight for  $f(i) = w^{(\min \mathcal{R})}(i)$  of the l-set estimator using independent ranks.

### 7.3 The Range ( $L_1$ difference).

The template estimator is not applicable to range estimates. This is because there are weight vectors (say  $w(i) = (1, 0)$ ) where the range is positive ( $w^{(L_1)} = 1$  in this case) and there is 0 probability to determine it from the sample  $S$  (intuitively, because we can never be “sure” that the second value is 0).

Fortunately, we can use the relation  $w^{(L_1 \mathcal{R})}(i) = w^{(\max \mathcal{R})}(i) - w^{(\min \mathcal{R})}(i)$  with the  $\ell$ -dependence estimators and get the estimator

$$a^{(L_1 \mathcal{R})}(i) = a^{(\max \mathcal{R})}(i) - a^{(\min \mathcal{R})}(i) \quad (17)$$

We use  $a^{(\max \mathcal{R})}(i)$  for the max-dependence l-set (same as s-set) estimator and  $a^{(\min \mathcal{R})}(i)$  when the statement applies to both the s-set and l-set min-dependence estimators. When specifics is warranted, we use  $a_s^{(L_1 \mathcal{R})}(i) = a^{(\max \mathcal{R})}(i) - a_s^{(\min \mathcal{R})}(i)$  and  $a_l^{(L_1 \mathcal{R})}(i) = a^{(\max \mathcal{R})}(i) - a_l^{(\min \mathcal{R})}(i)$ .

$a^{(L_1 \mathcal{R})}(i)$  is obviously an unbiased estimate of  $w^{(L_1 \mathcal{R})}$ , because it is the difference of two unbiased estimators. We show that for a consistent  $r$ , it is also nonnegative.

We use the notation  $p^{(\max \mathcal{R})}(i, r^{-i})$ ,  $p_s^{(\min \mathcal{R})}(i, r^{-i})$ , and  $p_l^{(\min \mathcal{R})}(i, r^{-i})$  for the respective inclusion probabilities (and  $p^{(\min \mathcal{R})}$  when the statement applies to both the respective s-set and l-set estimators).

We first establish the following lemma.

**Lemma 7.4.** *For consistent  $r$  with IPPS or EXP ranks,  $\mathcal{R}$ ,  $k$ , and  $i \in I$ .*

$$\frac{p^{(\max \mathcal{R})}(i, r^{-i})}{p^{(\min \mathcal{R})}(i, r^{-i})} \leq \frac{w^{(\max \mathcal{R})}(i)}{w^{(\min \mathcal{R})}(i)}.$$

*Proof.* Since,  $p_s^{(\min \mathcal{R})}(i, r^{-i}) \leq p_l^{(\min \mathcal{R})}(i, r^{-i})$ , it suffices to establish the inequality for  $p_s^{(\min \mathcal{R})}(i, r^{-i})$ .

For IPPS ranks it suffices to show that for any  $\tau$

$$\frac{\min\{1, \tau w^{(\max \mathcal{R})}(i)\}}{\min\{1, \tau w^{(\min \mathcal{R})}(i)\}} \leq \frac{w^{(\max \mathcal{R})}(i)}{w^{(\min \mathcal{R})}(i)}.$$

This is clear in case the numerator of the left hand side is  $\tau w^{(\max \mathcal{R})}(i)$  and the denominator is  $\tau w^{(\min \mathcal{R})}(i)$ . Otherwise, since  $\tau w^{(\max \mathcal{R})}(i) \geq \tau w^{(\min \mathcal{R})}(i)$  the numerator is  $1 < \tau w^{(\max \mathcal{R})}(i)$  so the inequality must also hold.

For EXP ranks, we need to show that for any  $\tau$ ,

$$\frac{1 - \exp(-\tau w^{(\max \mathcal{R})}(i))}{1 - \exp(-\tau w^{(\min \mathcal{R})}(i))} \leq \frac{w^{(\max \mathcal{R})}(i)}{w^{(\min \mathcal{R})}(i)}.$$

Taking  $r = w^{(\max \mathcal{R})}(i)/w^{(\min \mathcal{R})}(i)$  and  $\gamma = \exp(-\tau w^{(\min \mathcal{R})}(i))$ , this follows since for any  $r \geq 1$  and  $0 < \gamma < 1$ ,  $\frac{1-\gamma^r}{1-\gamma} \leq r$ .  $\square$

**Lemma 7.5.** *For consistent  $r$  with IPPS or EXP ranks,  $\forall i \in I$ ,  $a^{(L_1 \mathcal{R})}(i) \geq 0$ .*

*Proof.* It suffices to show that  $a^{(\min \mathcal{R})}(i) \leq a^{(\max \mathcal{R})}(i)$ . We first observe that  $a^{(\min \mathcal{R})}(i) > 0$  implies  $a^{(\max \mathcal{R})}(i) > 0$ . If  $a^{(\max \mathcal{R})}(i) > 0$  and  $a^{(\min \mathcal{R})}(i) = 0$  we are done. Otherwise, the claim follows using Lemma 7.4.  $\square$

## 8 Variance properties

### 8.1 Covariances

We conjecture that the RC estimators we presented have zero covariances. This conjecture is consistent with empirical observations and with properties of related RC estimators [17, 18]. With zero covariances, the variance  $\text{VAR}[a^{(f)}(J)]$  is the sum over  $i \in J$  of the per-key variances  $\text{VAR}[a^{(f)}(i)]$ . Hence, if two adjusted-weights estimators  $a_1$  and  $a_2$  have  $\text{VAR}[a_1(i)] \geq \text{VAR}[a_2(i)]$  for all  $i \in I$ , then the relations holds for all  $J \subset I$ .

**Conjecture 8.1.** *All our estimators for colocated or dispersed summaries have zero covariances: For all  $i \neq j \in I$ ,  $E[a^{(f)}(i)a^{(f)}(j)] = f(i)f(j)$ .*

## 8.2 Variance bounds

We use the notation  $t_k^{(f)}(i)$  for the RC  $f$ -adjusted weights assigned by an RC estimator applied to a bottom- $k$  sketch of  $(I, f)$ . We also write  $t_k^{(w^{(b)})}(i)$  as  $t_k^{(b)}(i)$  for short.

We measure the variance of an adjusted weight assignment  $a$  using  $\Sigma V[a] = \sum_{i \in I} \text{VAR}[a(i)]$ . To establish variance relation between two estimators, it suffices to establish it for each key  $i$ . Furthermore, if the estimators are defined with respect to the same distribution of rank assignments then it suffices to establish variance relation with respect to some  $\Omega(i, r^{-i})$ . (Since these subspaces partition  $\Omega$  and our estimators are unbiased on each subspace).

The variance of adjusted  $f$ -weights  $a^{(f)}(i)$  for  $i \in I$  are

$$\text{VAR}_{\Omega(i, r^{-i})}[a^{(f)}(i)] = f(i)^2 \left( \frac{1}{p(i, r^{-i})} - 1 \right). \quad (18)$$

**Colocated single-assignment estimators.** We show that our single-assignment inclusive estimators for colocated summaries (independent or coordinated) dominate plain RC estimators based on a single bottom- $k$  sketch.

**Lemma 8.2.** *For  $b \in \mathcal{W}$  and  $i \in I$ , let  $a^{(b)}(i)$  be the adjusted weights for co-located summaries computed by our estimator (using  $S^*(r) \equiv S(r)$  and inclusion probabilities (4)). Then,  $\text{VAR}[a^{(b)}(i)] \leq \text{VAR}[t_k^{(b)}(i)]$ .*

*Proof.* Consider applying the generic estimator with  $S^*(r)$  containing all keys  $i$  with  $r^{(b)}(i) < r_{k+1}^{(b)}(I)$ . This estimator assigns to  $i$  an adjusted weight of 0 if  $r^{(b)}(i) > r_{k+1}^{(b)}(I)$  and an adjusted weight of  $w^{(b)}(i)/\mathbf{F}_{w^{(b)}(i)}(r_{k+1}^{(b)}(I))$  otherwise. This is the same adjusted weights as assigned by the RC bottom- $k$  estimator if we apply it using the rank assignment to  $(I, w^{(b)})$  obtained by restricting  $r$  (that is using  $r^{(b)}$ ). The lemma now follows from Lemma 5.1.

A direct proof: It suffices to establish the variance relation for a particular subspace  $\Omega(i, r^{-i})$  considering the restriction of  $r^{(b)}$  of  $r$  as the rank assignment for  $(I, w^{(b)})$ . In  $\Omega(i, r^{-i})$ , the variance of  $t_k^{(b)}(i)$  is

$$\text{VAR}_{\Omega(i, r^{-i})}[t_k^{(b)}(i)] = w^{(b)}(i)^2 (1/\mathbf{F}_{w^{(b)}(i)}(r_{k+1}^{(b)}(I)) - 1).$$

From (4),  $p(i, r^{-i}) \geq \mathbf{F}_{w^{(b)}(i)}(r_{k+1}^{(b)}(I))$ . Therefore,

$$\text{VAR}_{\Omega(i, r^{-i})}[a^{(b)}(i)] = w^{(b)}(i)^2 (1/p(i, r^{-i}) - 1) \leq \text{VAR}_{\Omega(i, r^{-i})}[t_k^{(b)}(i)].$$

□

**Approximation quality of multiple-assignment estimators.** The quality of the estimate depends on the relation between  $f$  and the weight assignment(s) with respect to which the weighted sampling is performed. We refer to these assignments as *primary*. Variance is minimized when  $f(i)$  are the primary weights but often  $f$  must be *secondary*:  $f$  may not be known at the time of sampling, the number of different functions  $f$  that are of interest can be large – to estimate all pairwise similarities we need  $\binom{|W|}{2}$  different “weight-assignments”. For dispersed weights, even if known apriori, weighted samples with respect to some multiple-assignment  $f$  cannot, generally, be computed in a scalable way. We bound the variance of our min, max, and  $L_1$  estimators.

**Colocated min, max, and  $L_1$  estimators.** We bound the variance of inclusive estimators for min, max, and  $L_1$  using the variance of inclusive estimators for the respective primary weight assignments.

**Lemma 8.3.** For  $f \in \{\max \mathcal{R}, \min \mathcal{R}, L_1 \mathcal{R}\}$ , let  $a^{(f)}(i)$  be the adjusted  $w^{(f)}$ -weights for co-located summaries computed by our estimator (using  $S^*(r) \equiv S(r)$  and inclusion probabilities (4)).

$$\begin{aligned}\text{VAR}[a^{(\min \mathcal{R})}(i)] &= \min_{b \in \mathcal{R}} \text{VAR}[a^{(b)}(i)], \\ \text{VAR}[a^{(\max \mathcal{R})}(i)] &= \max_{b \in \mathcal{R}} \text{VAR}[a^{(b)}(i)], \\ \text{VAR}[a^{(L_1 \mathcal{R})}(i)] &\leq \text{VAR}[a^{(\max \mathcal{R})}(i)].\end{aligned}$$

*Proof.* It suffices to establish this relation in a subspace  $\Omega(i, r^{-i})$ . All inclusive estimators share the same inclusion probabilities  $p(i, r^{-i})$  and the variance is as in Equation (18). The proof is immediate from the definitions, substituting  $w^{(\min \mathcal{R})}(i) = \min_{b \in \mathcal{R}} w^{(b)}(i)$ ,  $w^{(\max \mathcal{R})}(i) = \max_{b \in \mathcal{R}} w^{(b)}(i)$ , and  $w^{(L_1 \mathcal{R})}(i) = w^{(\max \mathcal{R})}(i) - w^{(\min \mathcal{R})}(i)$ .  $\square$

The following relations are an immediate corollary of Lemma 8.3:

$$\begin{aligned}\Sigma V[a^{(\min \mathcal{R})}] &\leq \min_{b \in \mathcal{R}} \Sigma V[a^{(b)}], \quad \Sigma V[a^{(\max \mathcal{R})}] \leq \max_{b \in \mathcal{R}} \Sigma V[a^{(b)}], \\ \Sigma V[a^{(L_1 \mathcal{R})}] &\leq \Sigma V[a^{(\max \mathcal{R})}] \leq \max_{b \in \mathcal{R}} \Sigma V[a^{(b)}].\end{aligned}$$

**Relative variance bound for max:** For both the dispersed and the colocated models, we show that the variance of the max estimator is at most that of an estimator applied to a weighted sample taken with max being the primary weight. More precisely,  $a^{(\max \mathcal{R})}(i)$  has at most the variance of an RC estimator applied to a bottom- $k$  sketch of  $(I, w^{(\max \mathcal{R})})$  (obtained with respect to the same  $\mathbf{f}_w$  ( $w > 0$ )). Hence, the relative variance bounds of single-assignment bottom- $k$  sketch estimators are applicable [16, 17, 26].

**Lemma 8.4.** Let  $t_k^{(\max \mathcal{R})}(i)$  be the adjusted weights of the RC estimator applied to a bottom- $k$  sketch of  $(I, w^{(\max \mathcal{R})})$ . For any  $i \in I$ ,  $\text{VAR}[a^{(\max \mathcal{R})}(i)] \leq \text{VAR}[t_k^{(\max \mathcal{R})}(i)]$ .

*Proof.* By Lemma 4.1 for consistent ranks  $r^{(\min \mathcal{R})} = \min_{b \in \mathcal{R}} r^{(b)}(i)$  is a valid rank assignment for  $(I, w^{(\max \mathcal{R})})$  (using the same rank distributions). So it follows that RC adjusted weights with respect to  $w^{(\max \mathcal{R})}$  can be stated as a redundant application of the template estimator where  $S \in \mathcal{S}_1^*(i)$  when  $i$  is one of the  $k$  least ranked keys with respect to  $r^{(\min \mathcal{R})}$ .

From Lemma 7.3, the selection  $\mathcal{S}^*$  is such that  $S \in \mathcal{S}^*(i)$  for the  $k - 1$  least ranked keys (or more) with respect to  $r^{(\min \mathcal{R})}$ . Hence,  $\mathcal{S}_1^* \subset \mathcal{S}^*$ . Applying Lemma 5.1, we obtain that for all  $i \in I$ ,  $\text{VAR}[a^{(\max \mathcal{R})}(i)] \leq \text{VAR}[t_k^{(\max \mathcal{R})}(i)]$ .  $\square$

**Dispersed model min and  $L_1$  estimators.** We bound the absolute variance of our  $w^{(\min \mathcal{R})}$  estimator in terms of the variance of  $w^{(b)}$ -estimators for  $b \in \mathcal{R}$ . Let  $t_k^{(b)}$  be RC adjusted  $w^{(b)}$ -weights using the bottom- $k$  sketch with ranks  $r^{(b)}$ .

**Lemma 8.5.** For shared-seed consistent  $r$ , for all  $i \in I$ ,

$$\text{VAR}[a_\ell^{(\min \mathcal{R})}(i)] \leq \max_{b \in \mathcal{R}} \text{VAR}[t_k^{(b)}(i)]$$

*Proof.* Fixing  $i$  and  $\Omega(i, r^{-i})$ , for shared seed  $p_\ell^{(\min \mathcal{R})}(i, r^{-i}) = \min_{b \in \mathcal{R}} \mathbf{F}_{w^{(b)}(i)}(r_{k+1}^{(b)}(I))$ .

Let  $b'$  be such that  $p_\ell^{(\min \mathcal{R})}(i, r^{-i}) = \mathbf{F}_{w^{(b')}(i)}(r_{k+1}^{(b')}(I))$ . We have that

$$w^{(\min \mathcal{R})}(i)^2 \left( \frac{1}{p_\ell^{(\min \mathcal{R})}(i, r^{-i})} - 1 \right) \leq w^{(b')}(i)^2 \left( \frac{1}{\mathbf{F}_{w^{(b')}(i)}(r_{k+1}^{(b')}(I))} - 1 \right).$$

From this follows that

$$w^{(\min \mathcal{R})}(i)^2 \left( \frac{1}{p^{(\min \mathcal{R})}(i, r^{-i})} - 1 \right) \leq \max_{b \in \mathcal{R}} w^{(b)}(i)^2 \left( \frac{1}{\mathbf{F}_{w^{(b)}(i)}(r_{k+1}^{(b)}(I))} - 1 \right).$$

which is equivalent to the statement of the lemma.  $\square$

It follows from this lemma that there exists a  $b \in \mathcal{R}$  such that  $\Sigma V[a^{(\min \mathcal{R})}] \leq \Sigma_{b \in \mathcal{R}} \Sigma V[a^{(b)}]$ .

**Lemma 8.6.** *For consistent  $r$ , for all  $i \in I$ ,*

$$\text{VAR}[a^{(L_1 \mathcal{R})}(i)] \leq \text{VAR}[a^{(\min \mathcal{R})}(i)] + \text{VAR}[a^{(\max \mathcal{R})}(i)]$$

*Proof.* Fixing  $i$  and  $\Omega(i, r^{-i})$ , With probability  $p^{(\min \mathcal{R})}(i, r^{-i})$ ,  $a^{(L_1 \mathcal{R})} = w^{(\max \mathcal{R})}(i)/p^{(\max \mathcal{R})}(i, r^{-i}) - w^{(\min \mathcal{R})}(i)/p^{(\min \mathcal{R})}(i, r^{-i})$ . With probability  $p^{(\max \mathcal{R})}(i, r^{-i}) - p^{(\min \mathcal{R})}(i, r^{-i})$ ,  $a^{(L_1 \mathcal{R})} = w^{(\max \mathcal{R})}(i)/p^{(\max \mathcal{R})}(i, r^{-i})$ .

We have  $\text{VAR}[a^{(L_1 \mathcal{R})}(i)] = \mathbb{E}[a^{(L_1 \mathcal{R})}(i)^2] - (w^{(\max \mathcal{R})}(i) - w^{(\min \mathcal{R})}(i))^2$ . Substituting in the above we obtain

$$\begin{aligned} \text{VAR}[a^{(L_1 \mathcal{R})}(i)] &= \\ & w^{(\max \mathcal{R})}(i)^2 \left( \frac{1}{p^{(\max \mathcal{R})}(i, r^{-i})} - 1 \right) + w^{(\min \mathcal{R})}(i)^2 \left( \frac{1}{p^{(\min \mathcal{R})}(i, r^{-i})} - 1 \right) \\ & - 2w^{(\max \mathcal{R})}(i)w^{(\min \mathcal{R})}(i) \left( \frac{1}{p^{(\max \mathcal{R})}(i, r^{-i})} - 1 \right) \\ & = \text{VAR}[a^{(\max \mathcal{R})}(i)] + \text{VAR}[a^{(\min \mathcal{R})}(i)] \\ & - 2w^{(\max \mathcal{R})}(i)w^{(\min \mathcal{R})}(i) \left( \frac{1}{p^{(\max \mathcal{R})}(i, r^{-i})} - 1 \right) \end{aligned}$$

$\square$

In particular we get  $\Sigma V[a^{(L_1 \mathcal{R})}] \leq \Sigma V[a^{(\min \mathcal{R})}] + \Sigma V[a^{(\max \mathcal{R})}]$ .

## 9 Evaluation

We evaluate the performance of our estimators on summaries, of independent and coordinated sketches, produced for the colocated and the dispersed data models.

### 9.1 Datasets

• **IP dataset1:** A set of about  $9.2 \times 10^6$  IP packets from a gateway router. For each packet we have source and destination IP addresses (srcIP and destIP), source and destination ports (srcPort and destPort), protocol, and total bytes.

◇ *Colocated data:* Packets were aggregated by each of the following keys.

**keys:** 4tuples (srcIP, destIP, srcPort and destPort) ( $1.09 \times 10^6$  distinct keys).

**Weight assignments:** number of bytes ( $4.25 \times 10^9$  total), number of packets ( $9.2 \times 10^6$  total), and uniform (weight of 1 for each key).

**keys:** destIP ( $3.76 \times 10^4$  unique destinations) **Weight assignments:** number of bytes ( $4.25 \times 10^9$  total), number of packets ( $9.2 \times 10^6$  total), number of distinct 4-tuples ( $1.09 \times 10^6$  total), and uniform (weight of 1 for each key).



◊ *Dispersed data*: We partitioned the packet stream into two consecutive sets with the same number of packets ( $4.6 \times 10^6$ ) in each. We refer to the first set as *period1* and to the second set as *period2*. For each period packets were aggregated by keys. As keys we used the destIP or a pair consisting of both the srcIP and the destIP. We considered three attributes for each key, namely, total number of bytes, number of packets, or the number of distinct 4tuples with that key. For each attribute we got two weight assignments  $w^{(1)}$  and  $w^{(2)}$  one for each period. (See Table 2).

key, weight	$\sum_i w^{(1)}(i)$	$\sum_i w^{(2)}(i)$	$\sum_i w^{(\max\{1,2\})}(i)$	$\sum_i w^{(\min\{1,2\})}(i)$	$\sum_i w^{(L_1\{1,2\})}(i)$
destIP, 4tuple	$5.42 \times 10^5$	$5.54 \times 10^5$	$7.47 \times 10^5$	$3.49 \times 10^5$	$3.98 \times 10^5$
destIP, bytes	$2.08 \times 10^9$	$2.17 \times 10^9$	$3.26 \times 10^9$	$9.96 \times 10^8$	$2.26 \times 10^9$
srcIP+destIP, packets	$4.61 \times 10^6$	$4.61 \times 10^6$	$7.61 \times 10^6$	$1.61 \times 10^6$	$6.00 \times 10^6$
srcIP+destIP, bytes	$2.08 \times 10^9$	$2.17 \times 10^9$	$3.49 \times 10^9$	$7.65 \times 10^8$	$2.72 \times 10^9$

Table 2: IP dataset1

• **IP dataset2**: IP packet trace from an IP router during August 1, 2008. Packets were partitioned to one hour time periods.

◊ *Colocated data*: **keys**: destIP or 4tuples. **weight assignments**: bytes, packets, IP flows, and uniform.

We used the packet stream of Hour3 which has  $1.73 \times 10^5$  distinct destIPs,  $1.87 \times 10^{10}$  total bytes,  $4.93 \times 10^7$  packets,  $1.30 \times 10^6$  distinct flows, and  $0.94 \times 10^5$  distinct 4tuples.

◊ *Dispersed data*: The packets in each hour were aggregated into different weight assignments. We used keys that are destIP or 4tuples and weights that are corresponding bytes. We thus obtained a weight assignment  $w^{(h)}$  for each hour.

The following table summarizes some properties of the data for the first 4 hours and for the sets of hours  $\mathcal{R} = \{1, 2\}$  and  $\mathcal{R} = \{1, 2, 3, 4\}$ . The table lists the number of distinct keys (destIP or 4tuples) and total bytes  $\sum_i w^{(h)}(i)$  for each hour or set of hours.

hours	1	2	3	4	{1, 2}	{1, 2, 3, 4}
destIP ( $\times 10^5$ )	2.17	2.96	1.73	1.76	3.41	3.61
4tuples ( $\times 10^6$ )	1.05	1.17	0.94	0.99	2.10	3.74
bytes ( $\times 10^{10}$ )	2.00	1.84	1.87	1.81	3.84	7.52

The following table lists, for destIP and 4tuple keys, the sums  $\sum_i w^{(\min \mathcal{R})}(i)$ ,  $\sum_i w^{(\max \mathcal{R})}(i)$ , and  $\sum_i w^{(L_1 \mathcal{R})}(i)$ , for  $\mathcal{R} = \{1, 2\}$  and  $\mathcal{R} = \{1, 2, 3, 4\}$ .

key	destIP		4tuple	
	{1, 2}	{1, 2, 3, 4}	{1, 2}	{1, 2, 3, 4}
min ( $\times 10^{10}$ )	1.51	1.33	0.86	0.82
max ( $\times 10^{10}$ )	2.33	3.02	2.99	4.92
$L_1$ ( $\times 10^{10}$ )	0.83	1.69	2.13	4.11

• **Netflix Prize Data [44]**. The dataset contains dated ratings of  $1.77 \times 10^4$  movies. We used all 2005 ratings ( $5.33 \times 10^7$ ). Each key corresponds to a movie and we used 12 weight assignments  $b \in \{1 \dots 12\}$  that corresponds to months. The weight  $w^{(b)}(i)$  is the number of ratings of movie  $i$  in month  $b$ . (See Table 3 for more details.)

months	1	2	3	4	5	6	7	8	9	10	11	12	1,2	1-6	1-12
movies ( $\times 10^4$ )	1.54	1.58	1.61	1.64	1.66	1.68	1.70	1.73	1.73	1.77	1.73	1.73	1.60	1.71	1.77
ratings ( $\times 10^6$ )	4.70	4.10	4.31	4.16	4.39	5.30	4.95	5.26	4.91	5.16	3.61	2.41	8.80	27.0	53.3
min ( $\times 10^6$ )													3.72	2.97	1.68
max ( $\times 10^6$ )													5.08	6.79	7.95
$L_1$ ( $\times 10^6$ )													1.35	3.82	6.27

Table 3: Netflix data set. Distinct movies (number of movies with at least one rating) and total number of ratings for each month (1, ..., 12) in 2005 and for periods  $\mathcal{R} = \{1, 2\}$ ,  $\mathcal{R} = \{1, \dots, 6\}$ , and  $\mathcal{R} = \{1, \dots, 12\}$ . For these periods, we also show  $\sum_i w^{(\min \mathcal{R})}(i)$ ,  $\sum_i w^{(\max \mathcal{R})}(i)$ , and  $\sum_i w^{(L_1 \mathcal{R})}(i)$ .

• **Stocks data:** Data set contains daily data for about 8.9K ticker symbols, for October 2008 (23 trading days). Daily data of each ticker had 5 price attributes (open, high, low, close, adjusted\_close) and volume traded. Table 4 lists totals of these weights for each trading day.

	1	2	3	4	5	6	7	8	9	10	11	12
open	1.81	1.80	1.75	1.68	1.65	1.55	1.56	1.42	1.50	1.61	1.54	1.47
high	1.85	1.83	1.81	1.72	1.70	1.63	1.61	1.54	1.61	1.67	1.57	1.53
low	1.78	1.73	1.70	1.57	1.57	1.50	1.45	1.33	1.46	1.52	1.45	1.40
close	1.82	1.75	1.72	1.65	1.59	1.56	1.48	1.46	1.58	1.57	1.47	1.50
adj_close	1.81	1.74	1.72	1.64	1.58	1.55	1.47	1.45	1.57	1.56	1.46	1.49
volume	1.52	1.66	1.82	2.26	1.96	2.44	2.10	3.14	1.93	2.22	1.80	2.27

	13	14	15	16	17	18	19	20	21	22	23
open	1.48	1.52	1.52	1.48	1.45	1.37	1.38	1.38	1.42	1.46	1.47
high	1.57	1.57	1.56	1.52	1.49	1.44	1.43	1.45	1.49	1.50	1.54
low	1.44	1.49	1.49	1.42	1.38	1.34	1.34	1.33	1.39	1.42	1.44
close	1.50	1.55	1.51	1.45	1.44	1.40	1.36	1.42	1.44	1.48	1.51
adj_close	1.50	1.54	1.51	1.44	1.43	1.39	1.36	1.42	1.43	1.47	1.50
volume	1.84	1.42	1.43	1.73	2.05	1.84	1.55	1.99	1.96	1.71	1.75

Table 4: Daily totals for 23 trading days in October, 2008. Prices (open, high, low, close, adjusted\_close) are  $\times 10^5$ . Volumes are in multiples of  $10^{10}$ .

The ticker prices are highly correlated both in terms of same attribute over different days and the different price attributes in a given day. The correlation is much stronger than for the volume attribute or weight assignments used in the IP datasets. At least 93% of stocks had positive volume each day and virtually all had positive (high, low, close, adjusted\_close) prices for the duration. This contrasts the IP datasets, where it is much more likely for keys (destIPs or 4tuples) to have zero weights in subsequent assignments.

◊ *Colocated data:* **keys:** ticker symbols; **weight assignments:** the six numeric attributes: open, high, low, close, adjusted\_close, and volume in a given trading day.

◊ *Dispersed data:* **keys:** ticker symbols; **weight assignments:** daily (high or volume) values for each trading days.

For multiple-assignment aggregates evaluation, we used the first 2,5,10,15,23 trading days of October:  $\mathcal{R} = \{1, 2\}$  (October 1-2),  $\mathcal{R} = \{1, \dots, 5\}$  (October 1-7),  $\mathcal{R} = \{1, \dots, 10\}$  (October 1-14),  $\mathcal{R} = \{1, \dots, 15\}$  (October 1-21),  $\mathcal{R} = \{1, \dots, 23\}$  (October 1-31). The following table lists  $\sum_i w^{(\min \mathcal{R})}(i)$ ,  $\sum_i w^{(\max \mathcal{R})}(i)$ , and  $\sum_i w^{(L_1 \mathcal{R})}(i)$  for these sets of trading days.

	high ( $\times 10^9$ )					volume ( $\times 10^{10}$ )				
	1-2	1-5	1-10	1-15	1-23	1-2	1-5	1-10	1-15	1-23
min	1.82	1.67	1.48	1.44	1.33	1.34	1.33	1.30	1.15	1.13
max	1.87	1.89	1.92	1.92	1.94	1.80	2.54	3.50	3.59	3.77
$L_1$	0.05	0.22	0.44	0.49	0.61	0.41	1.20	2.20	2.43	2.64

## 9.2 Dispersed data.

We evaluate our  $w^{(\min \mathcal{R})}$ ,  $w^{(\max \mathcal{R})}$ , and  $w^{(L_1 \mathcal{R})}$  estimators as defined in Section 7:  $a^{(\max \mathcal{R})}$ ,  $a_s^{(\min \mathcal{R})}$ ,  $a_l^{(\min \mathcal{R})}$ ,  $a_s^{(L_1 \mathcal{R})}$ , and  $a_l^{(L_1 \mathcal{R})}$  for coordinated sketches and  $a_{ind}^{(\min \mathcal{R})}$  for independent sketches.

We used shared-seed coordinated sketches and show results for the IPPS ranks (see Section 3). Results for EXP ranks were similar.

We measure performance using the absolute  $\Sigma V[a^{(f)}]$  and normalized  $n\Sigma V[a^{(f)}] \equiv \Sigma V[a^{(f)}]/(\sum_{i \in I} f(i))^2$  sums of per-key variances (as discussed in Section 3), which we approximate by averaging square errors over multiple (25-200) runs of the sampling algorithm.

**Coordinated versus Independent sketches.** We compare the  $w^{(\min \mathcal{R})}$  estimators  $a_\ell^{(\min \mathcal{R})}$  (coordinated sketches) and  $a_{ind}^{(\min \mathcal{R})}$  (independent sketches).<sup>3</sup>

<sup>3</sup>We consider  $w^{(\min \mathcal{R})}$  because the estimator is applicable for independent sketches with unknown seeds. There are no nonnegative estimators for  $w^{(\max \mathcal{R})}$  and  $w^{(L_1 \mathcal{R})}$  estimators for independent sketches with unknown seeds.

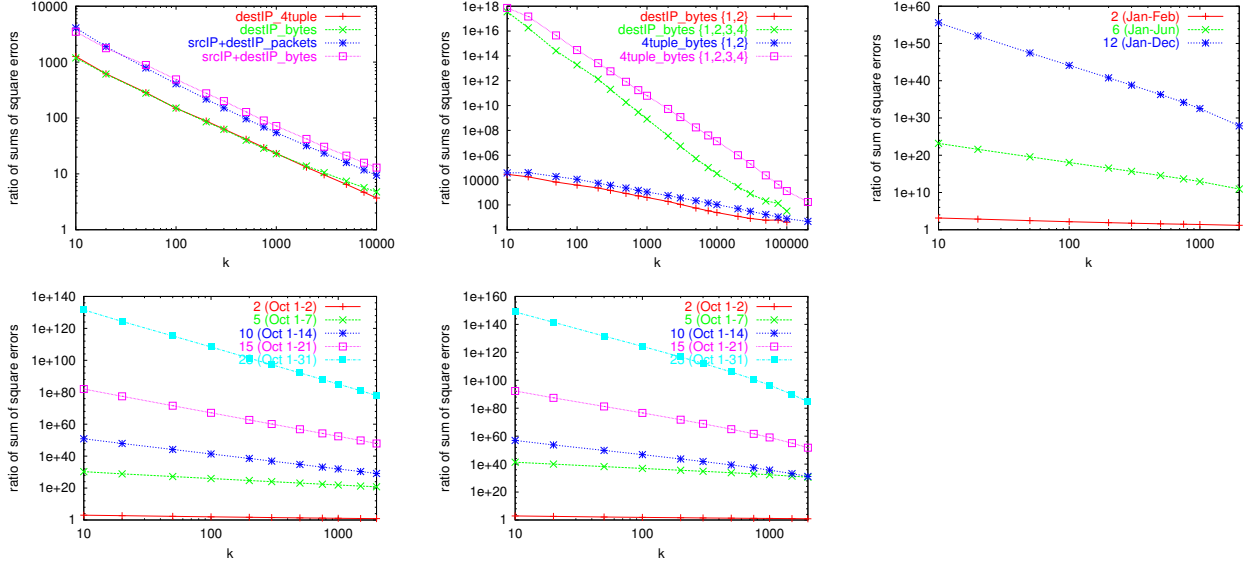


Figure 3: Top: IP dataset1 (left), IP dataset2 (middle), Netflix data set (right). Bottom: stocks dataset high values (left), stocks dataset volume values (middle). Ratio of  $w^{(\min \mathcal{R})}$  estimators for independent and coordinated sketches  $\Sigma V[a_{ind}^{(\min \mathcal{R})}] / \Sigma V[a_{\ell}^{(\min \mathcal{R})}]$ .

Figure 3 shows the ratio  $\Sigma V[a_{ind}^{(\min \mathcal{R})}] / \Sigma V[a_{\ell}^{(\min \mathcal{R})}]$  as a function of  $k$  for our datasets. Across data sets, the variance of the independent-sketches estimator is significantly larger, up to many orders of magnitude, than the variance of coordinated-sketches estimators. The ratio decreases with  $k$  but remains significant even when the sample size exceeds 10% of the number of keys.

The ratio increases with the number of weight assignments: On the Netflix data set, the ratio is 1-3 orders of magnitude for 2 months (assignments) and 10-40 orders of magnitude for 6-12 months (assignments). On IP dataset 2, the gap is 1-5 orders of magnitude for 2 assignments (hours) and 2-18 orders of magnitude for 4 assignments. On the stocks data set, the gap is 1-3 orders of magnitude for 2 assignments and reaches 150 orders of magnitude. This agrees with the exponential decrease of the inclusion probability with the number of assignments for independent sketches (see Section 7.2). These ratios demonstrate the estimation power provided by coordination.

**Weighted versus unweighted coordinated sketches.** We compare the performance of our estimators to known estimators applicable to unweighted coordinated sketches (coordinated sketches for uniform and global weights [18]). To apply these methods, all positive weights were replaced by unit weights. Because of the skewed nature of the weight distribution, the “unweighted” estimators performed poorly with variance being orders of magnitude larger.

**Variance of multiple-assignment estimators.** We relate the variance of our  $w^{(\min \mathcal{R})}$ ,  $w^{(\max \mathcal{R})}$ , and  $w^{(L_1 \mathcal{R})}$  and the variance of the optimal single-assignment estimators  $a^{(b)}$  for the respective individual weight assignments  $w^{(b)}$  ( $b \in \mathcal{R}$ ).<sup>4</sup> Because the variance of  $a_{ind}^{(\min \mathcal{R})}$  was typically many orders of magnitude worse, we include it only when it fit in the scale of the plot. The single-assignment estimators  $a^{(b)}$  are identical for independent and coordinated sketches (constructed with the same  $k$  and rank functions family), and hence are shown once.

Figures 4, 5, 6 and 7 show,  $\Sigma V[a_{\ell}^{(\min \mathcal{R})}]$ ,  $\Sigma V[a_{\ell}^{(\max \mathcal{R})}]$ , and  $\Sigma V[a_{\ell}^{(L_1 \mathcal{R})}]$  and  $\Sigma V[a^{(b)}]$  for  $b \in \mathcal{R}$  are within an order of magnitude. On our datasets,  $n\Sigma V[a^{(b)}]$  and  $n\Sigma V[a_{\ell}^{(\max \mathcal{R})}]$  are clustered together with

<sup>4</sup>For IPPS ranks,  $a^{(b)}$  are essentially optimal as they minimize  $\Sigma V[a^{(b)}]$  (and  $n\Sigma V[a^{(b)}]$ ) modulo a single sample [26, 54].

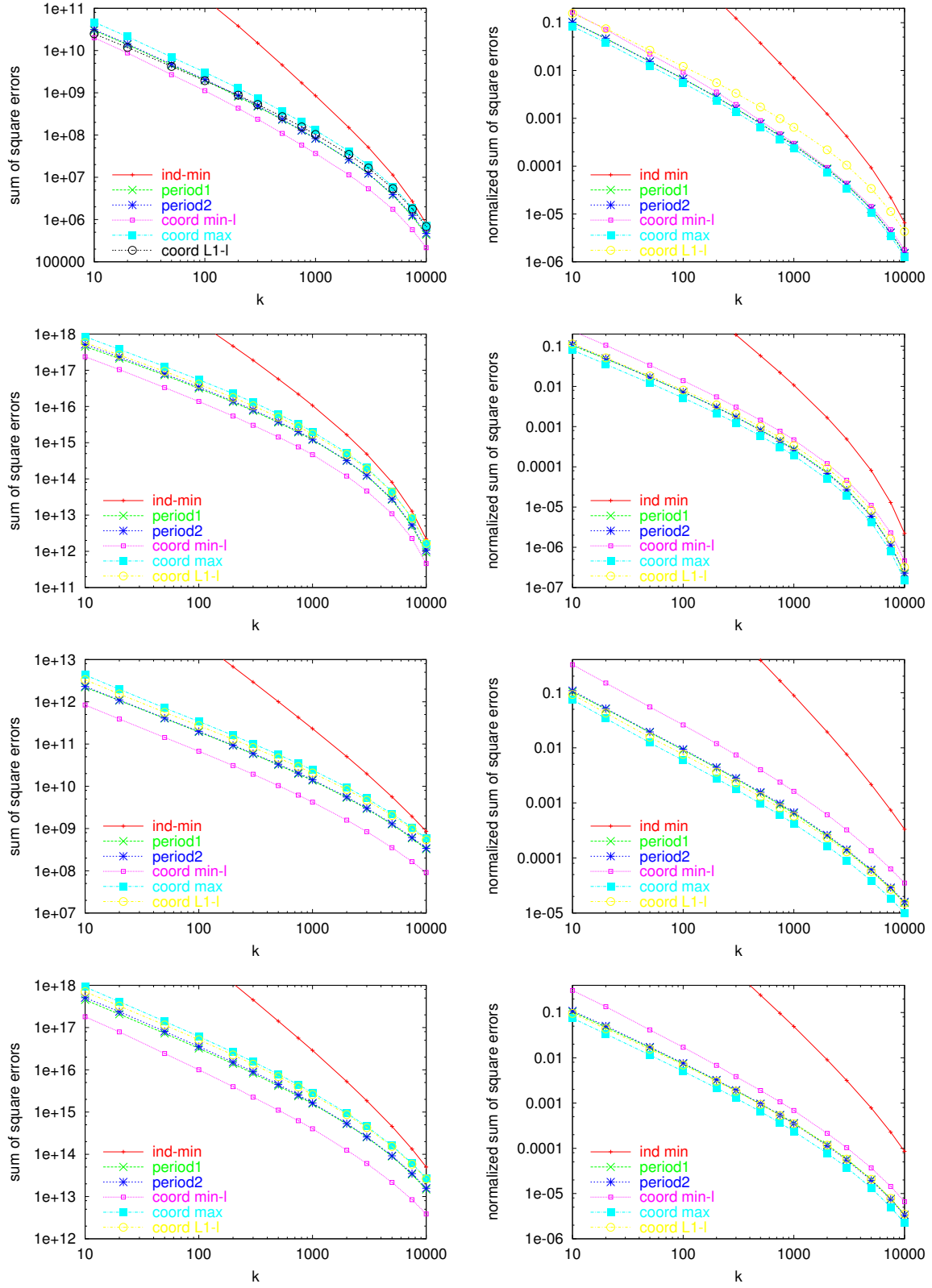


Figure 4: IP dataset1. Sum of square errors. left: absolute, right: normalized. Top: key=destIP weight=4tuple\_count, second row: key=destIP weight=bytes. Third row: key=srcIP+destIP, weight=packets. Fourth row: key=srcIP+destIP, weight=bytes

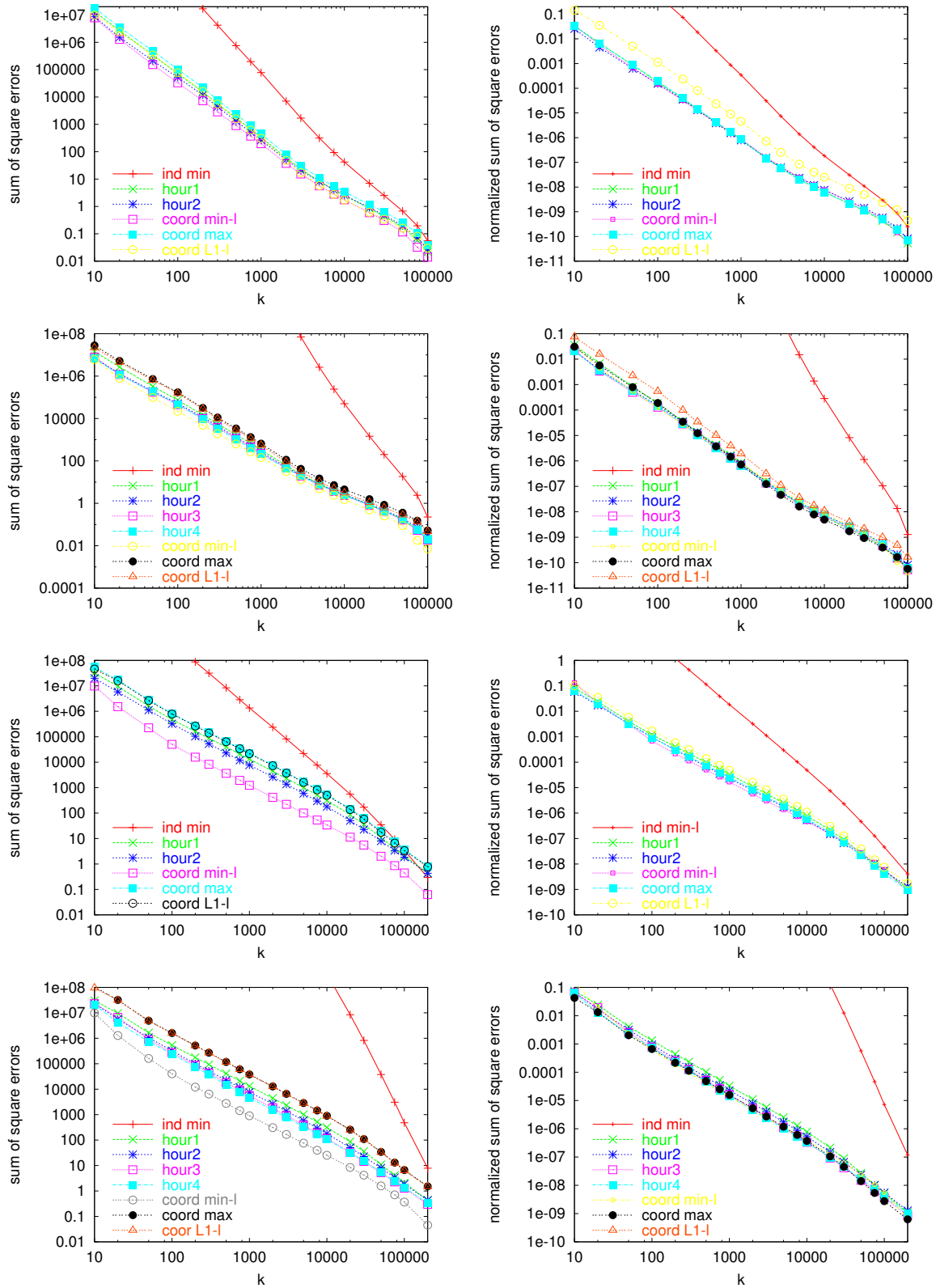


Figure 5: IP dataset2, sum of square errors. Left: absolute, Right: normalized. Top to bottom: key=destIP weight=bytes hours= {1, 2}; key=destIP weight=bytes hours= {1, 2, 3, 4}; key=4tuple weight=bytes hours= {1, 2}; key=4tuple weight=bytes hours= {1, 2, 3, 4}.

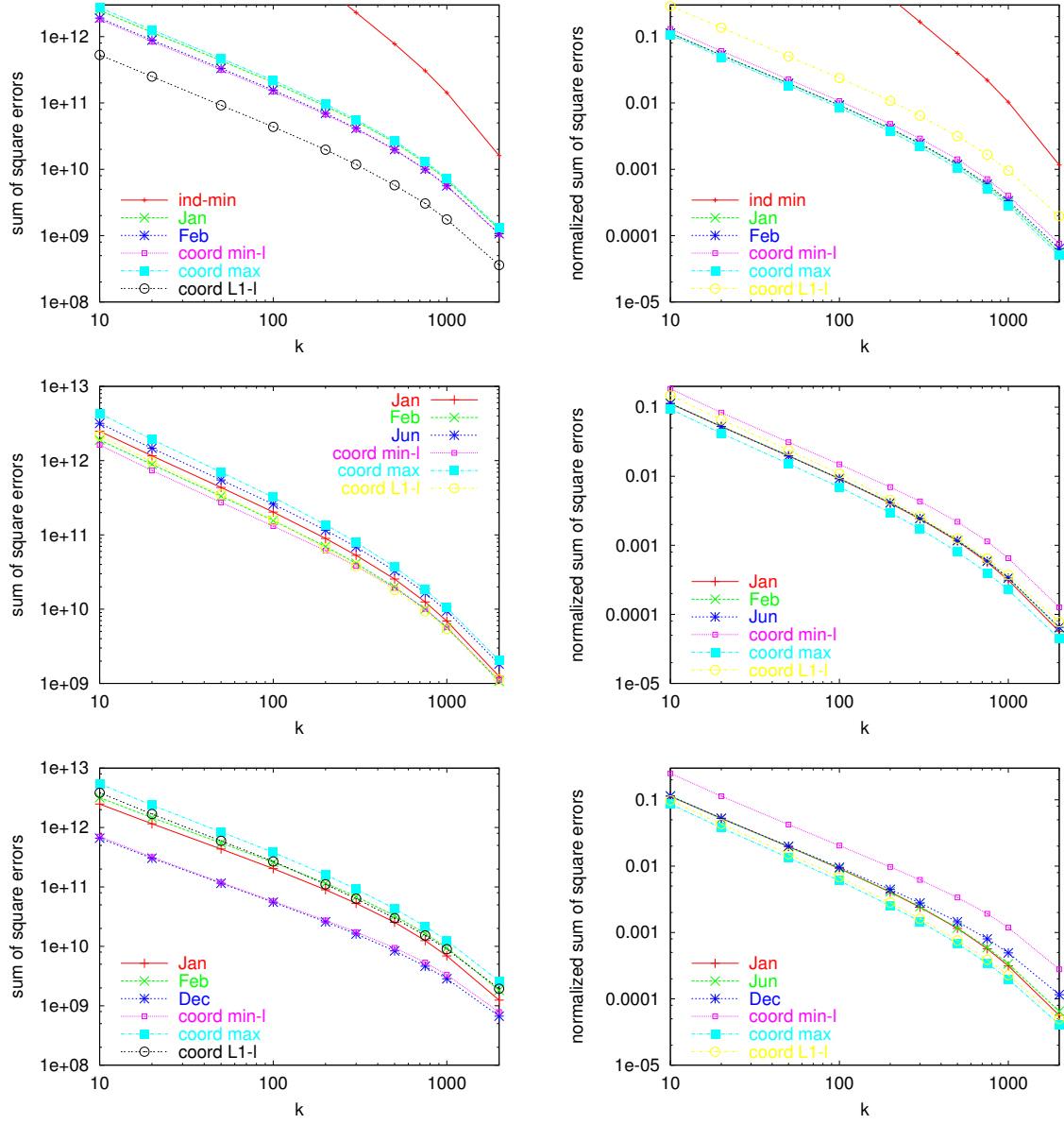


Figure 6: Netflix data set  $\mathcal{R} = \{1, 2\}$  (top),  $\mathcal{R} = \{1, \dots, 6\}$  (middle),  $\mathcal{R} = \{1, \dots, 12\}$  (bottom).  $\Sigma V$  (left) and  $n\Sigma V$  (right).

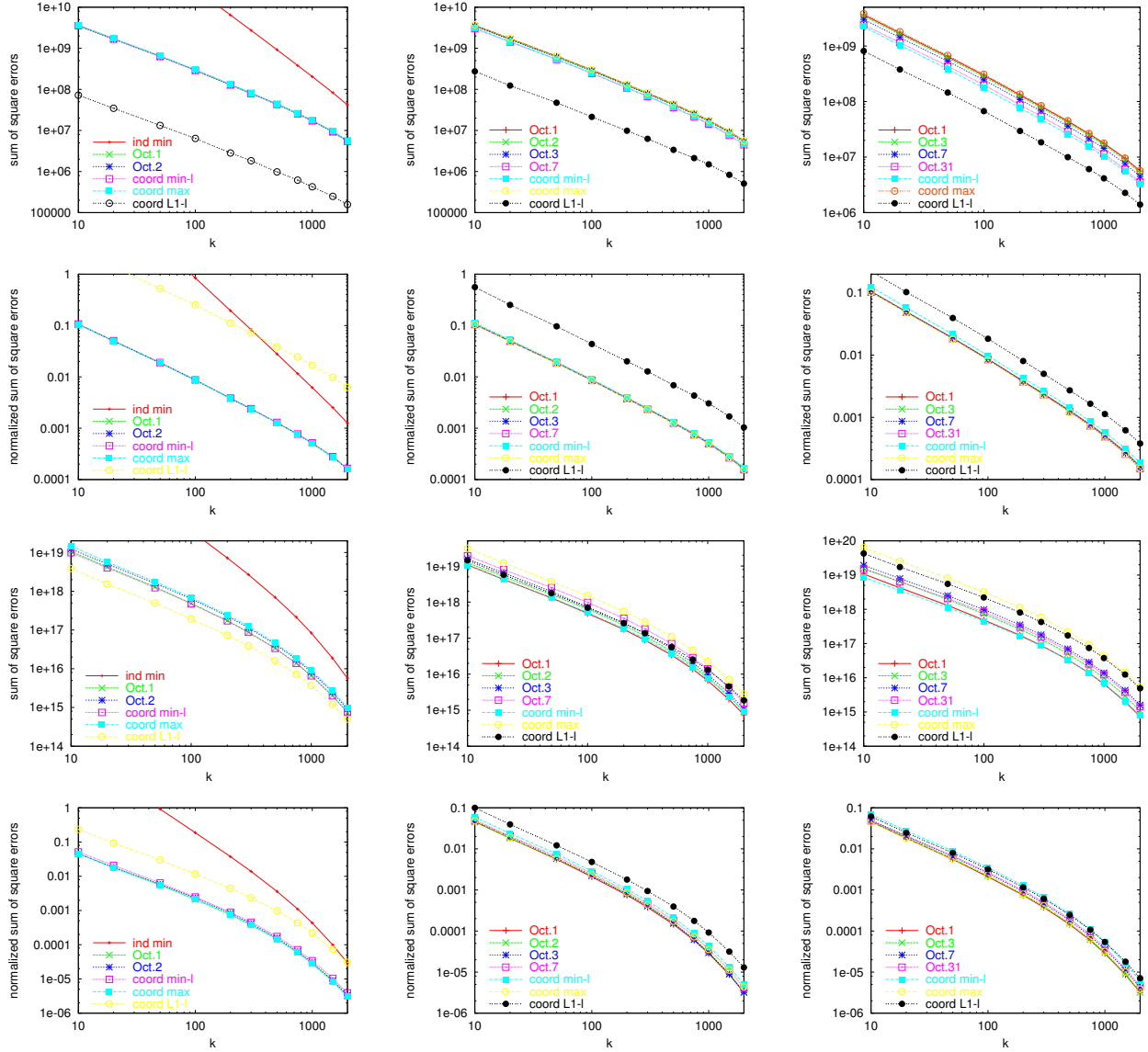


Figure 7: Stock dataset. Left:  $\mathcal{R} = \{1, 2\}$  (October 1-2, 2008), Middle:  $\mathcal{R} = \{1, \dots, 5\}$  (trading days in October 1-7, 2008), Right:  $\mathcal{R} = \{1, \dots, 23\}$  (all trading days in October, 2008). Upper two rows are  $\Sigma V$  and  $n\Sigma V$  with “high” weights. lower two rows are “volume” weights.

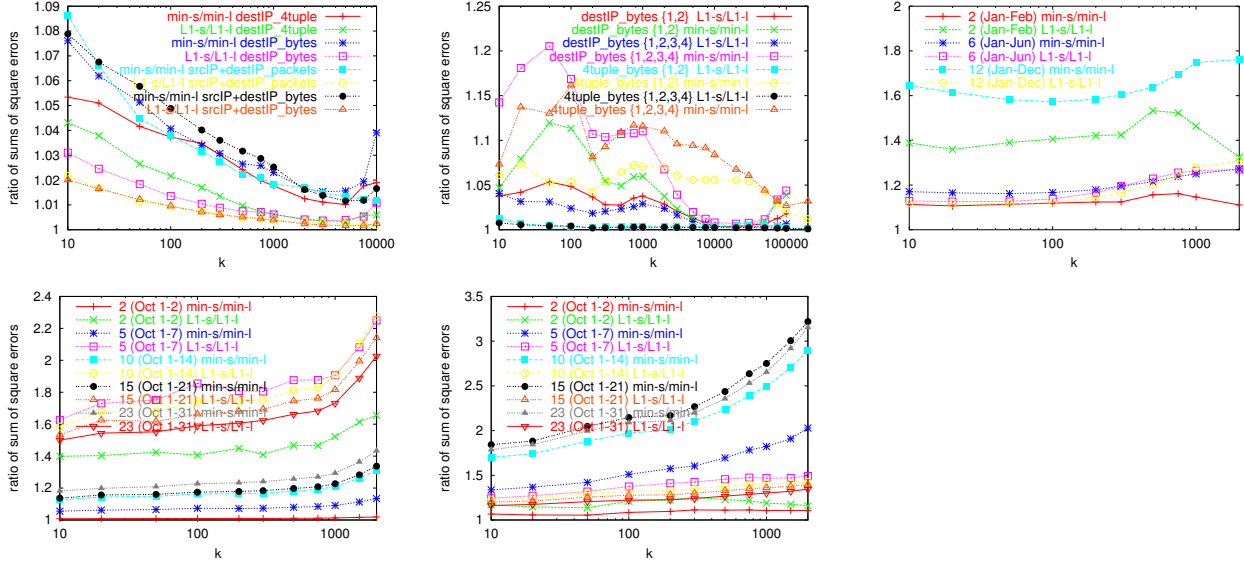


Figure 8: Top: IP dataset1 (left), IP dataset2 (middle), Netflix data set (right). Bottom: stocks dataset high-values (left), stocks dataset volume values (middle).  $\Sigma V$  ratio of s-set and l-set estimators for  $w^{(\min \mathcal{R})}$  and  $w^{(L_1 \mathcal{R})}$ ,  $\Sigma V[a_s^{(\min \mathcal{R})}]/\Sigma V[a_l^{(\min \mathcal{R})}]$  and  $\Sigma V[a_s^{(L_1 \mathcal{R})}]/\Sigma V[a_l^{(L_1 \mathcal{R})}]$ .

$kn\Sigma V \ll 1$  (and decreases with  $k$ ) (theory says  $(k-2)n\Sigma V \leq 1$ .) We also observed that  $n\Sigma V[a_l^{(L_1 \mathcal{R})}]$  and  $n\Sigma V[a_l^{(\min \mathcal{R})}]$  are typically close to  $n\Sigma V[a^{(b)}]$ . We observe the empirical relations  $\Sigma V[a_l^{(\min \mathcal{R})}] < \Sigma V[a_l^{(\max \mathcal{R})}]$  (with larger gap when the  $L_1$  difference is very small),  $\Sigma V[a_l^{(L_1 \mathcal{R})}] < \Sigma V[a_l^{(\max \mathcal{R})}]$ , and  $\Sigma V[a_l^{(\min \mathcal{R})}] < \min_{b \in \mathcal{R}} \Sigma V[a^{(b)}]$ . Empirically, the variance of our multi-assignment estimators with respect to single-assignment weights is significantly lower than the worst-case analytic bounds in Section 8 (Lemma 8.5 and 8.6). For normalized (relative) variances, we observe the “reversed” relations  $n\Sigma V[a_l^{(\min \mathcal{R})}] > n\Sigma V[a_l^{(\max \mathcal{R})}]$ ,  $n\Sigma V[a_l^{(L_1 \mathcal{R})}] > n\Sigma V[a_l^{(\max \mathcal{R})}]$ , and  $n\Sigma V[a_l^{(\min \mathcal{R})}] > \max_{b \in \mathcal{R}} n\Sigma V[a^{(b)}]$  which are explained by smaller normalization factors for  $w^{(\min \mathcal{R})}$  and  $w^{(L_1 \mathcal{R})}$ .

**S-set versus L-set estimators.** Figure 8 quantifies the advantage of the stronger l-set estimators over the s-set estimators for coordinated sketches. The advantage highly varies between datasets: 15%-80% for the Netflix dataset, 0%-9% for IP dataset1, 0%-20% for IP dataset2, and 0%-300% on the Stocks data set.

### 9.3 Colocated data

We computed shared-seed coordinated and independent sketches and show results for IPPS ranks (see Section 3). Results for EXP ranks were similar.

We consider the following  $w^{(b)}$ -weights estimators.  $a_c^{(b)}$ : the shared-seed coordinated sketches inclusive estimator (Section 6, Eq. (6)).  $a_i^{(b)}$ : the independent sketches inclusive estimator in (Section 6, Eq. (5)).  $a_p^{(b)}$ : the plain bottom- $k$  sketch RC estimator ([26] for IPPS ranks). Among all keys of the combined sketch this estimator uses only the keys which are part of the bottom- $k$  sketch of  $b$ .

We study the benefit of our inclusive estimators by comparing them to plain estimators. Since plain estimators can not be used effectively for multiple assignment aggregates, we focus on single-assignment aggregates.

**Inclusive versus plain estimators.** The plain estimators we used are optimal for individual bottom- $k$  sketches and the benefit of inclusive estimators comes from utilizing keys that were sampled for “other”



weight assignments. We computed the ratios

$$\Sigma V[a_i^{(b)}]/\Sigma V[a_p^{(b)}] \text{ and } \Sigma V[a_c^{(b)}]/\Sigma V[a_p^{(b)}]$$

as a function of  $k$ . As Figures 9, 10 and 11 show, the ratios vary between 0.05 to 0.9 on our datasets and shows a significant benefit for inclusive estimators. Our inclusive estimators are considerably more accurate with both coordinated and independent sketches. With independent sketches the benefit of the inclusive estimators is larger than with coordinate sketches since the independent sketches contain many more distinct keys for a given  $k$ .

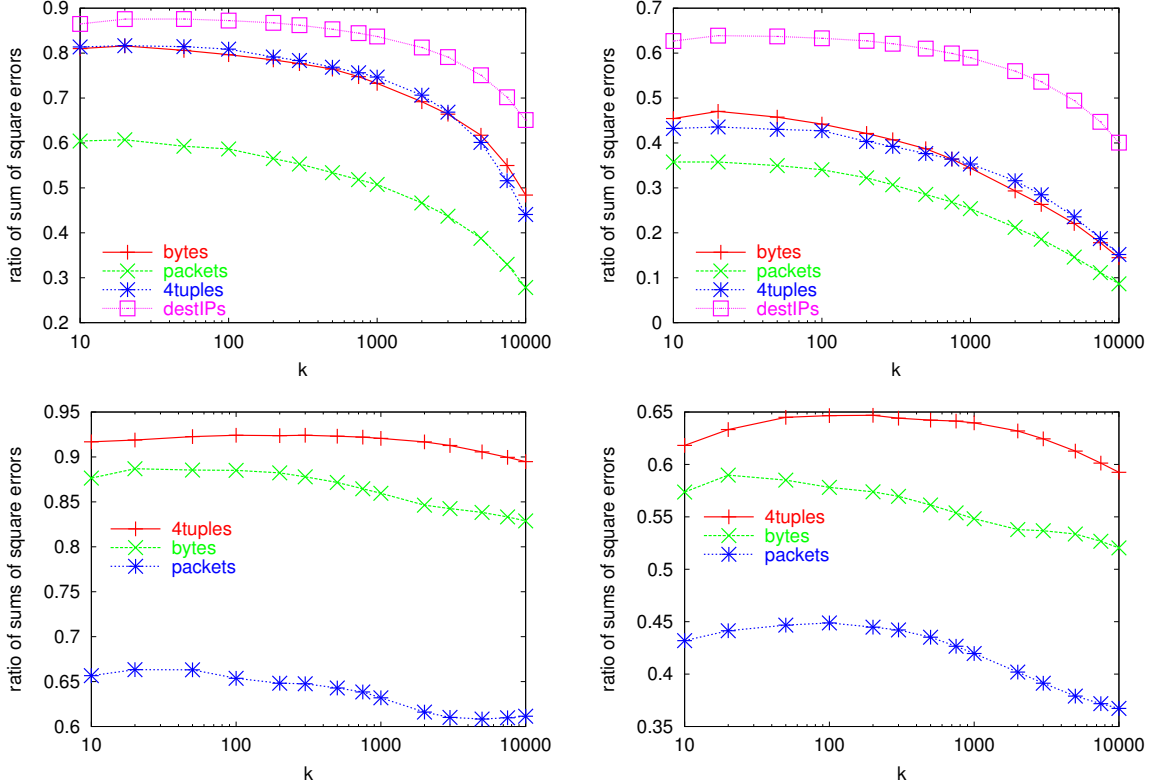


Figure 9: IP dataset1: Top: key=destIP. Bottom: key=4tuple. Left:  $\Sigma V[a_c^{(b)}]/\Sigma V[a_p^{(b)}]$ . Right:  $\Sigma V[a_i^{(b)}]/\Sigma V[a_p^{(b)}]$ .

**Variance versus storage.** For a fixed  $k$ , the plain estimator is in fact identical for independent and coordinated bottom- $k$  sketches. Independent bottom- $k$  sketches, however, tend to be larger than coordinated bottom- $k$  sketches. Here we compare the performance relative to the *combined sample size*, which is the number of distinct keys in the combined sample. We therefore use the notation  $a_{p,i}^{(b)}$  for the plain estimator applied to independent sketches and  $a_{p,c}^{(b)}$  for the plain estimator applied to coordinated sketches.

We compare summaries (coordinated and independent) and estimators (inclusive and plain) based on the tradeoff of variance versus summary size (number of distinct keys). Figures 12, 13, 14, and 15 show the normalized sums of variances, for inclusive and plain estimators  $n\Sigma V[a_i^{(b)}]$ ,  $n\Sigma V[a_c^{(b)}]$ ,  $n\Sigma V_{p,c}[a^{(b)}]$ ,  $n\Sigma V[a_{p,i}^{(b)}]$ , as a function of the combined sample size. For a fixed sketch size, plain estimators perform worse for independent sketches than for coordinated sketches. This happens since an independent sketch of

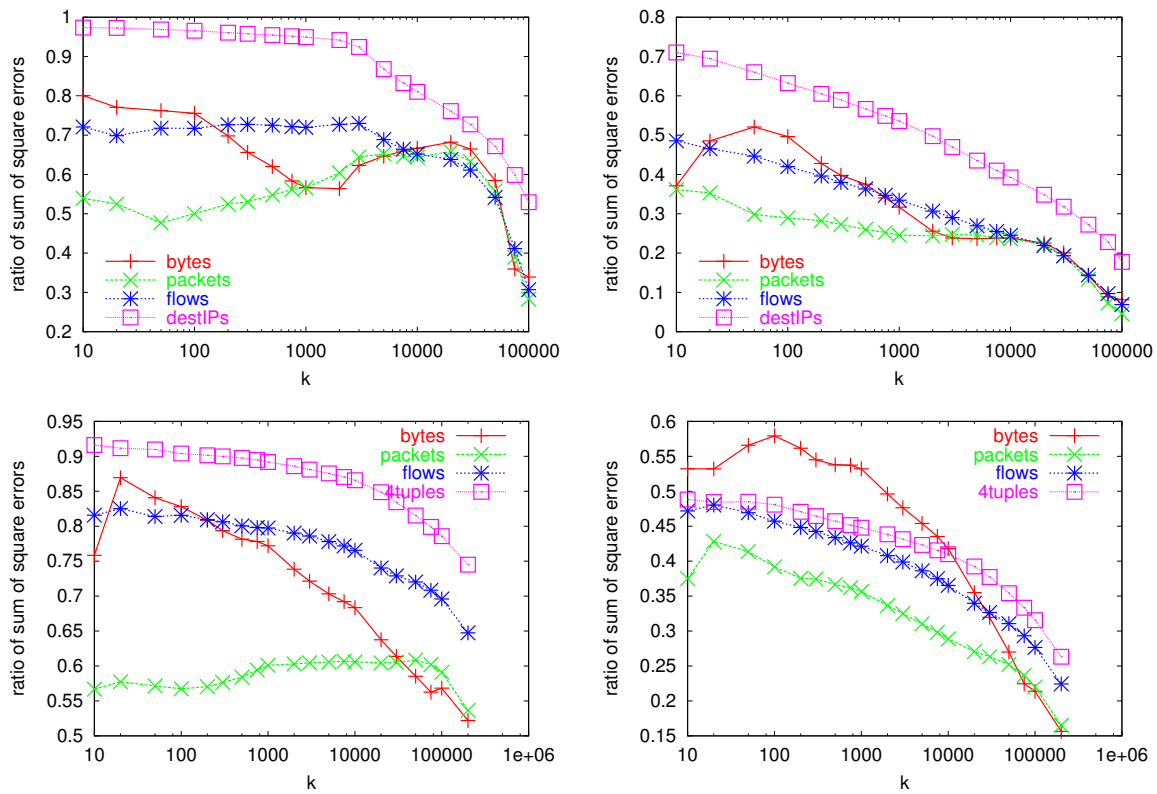


Figure 10: IP dataset2: Top: key=destIP. Bottom: key=4tuple. Left:  $\Sigma V[a_c^{(b)}]/\Sigma V[a_p^{(b)}]$ . Right:  $\Sigma V[a_i^{(b)}]/\Sigma V[a_p^{(b)}]$ .

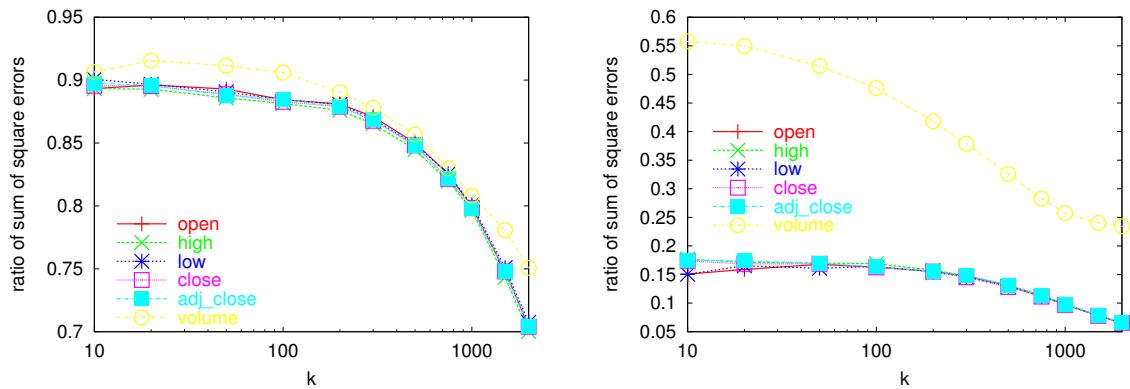


Figure 11: Stocks dataset: October 1, 2008. Left:  $\Sigma V[a_c^{(b)}]/\Sigma V[a_p^{(b)}]$ . Right:  $\Sigma V[a_i^{(b)}]/\Sigma V[a_p^{(b)}]$ .

some fixed size contains a smaller sketch for each weight assignment than a coordinated sketch of the same size. In other words the “ $k$ ” which we use to get an independent sketch of some fixed size is smaller than the “ $k$ ” which we use to get a coordinated sketch of the same size. Inclusive estimators for independent and coordinated sketches of the same size had similar variance. (Note however that for a given union size, we get weaker confidence bounds with independent samples than with coordinated samples, simply because we are guaranteed fewer samples with respect to each particular assignment.)

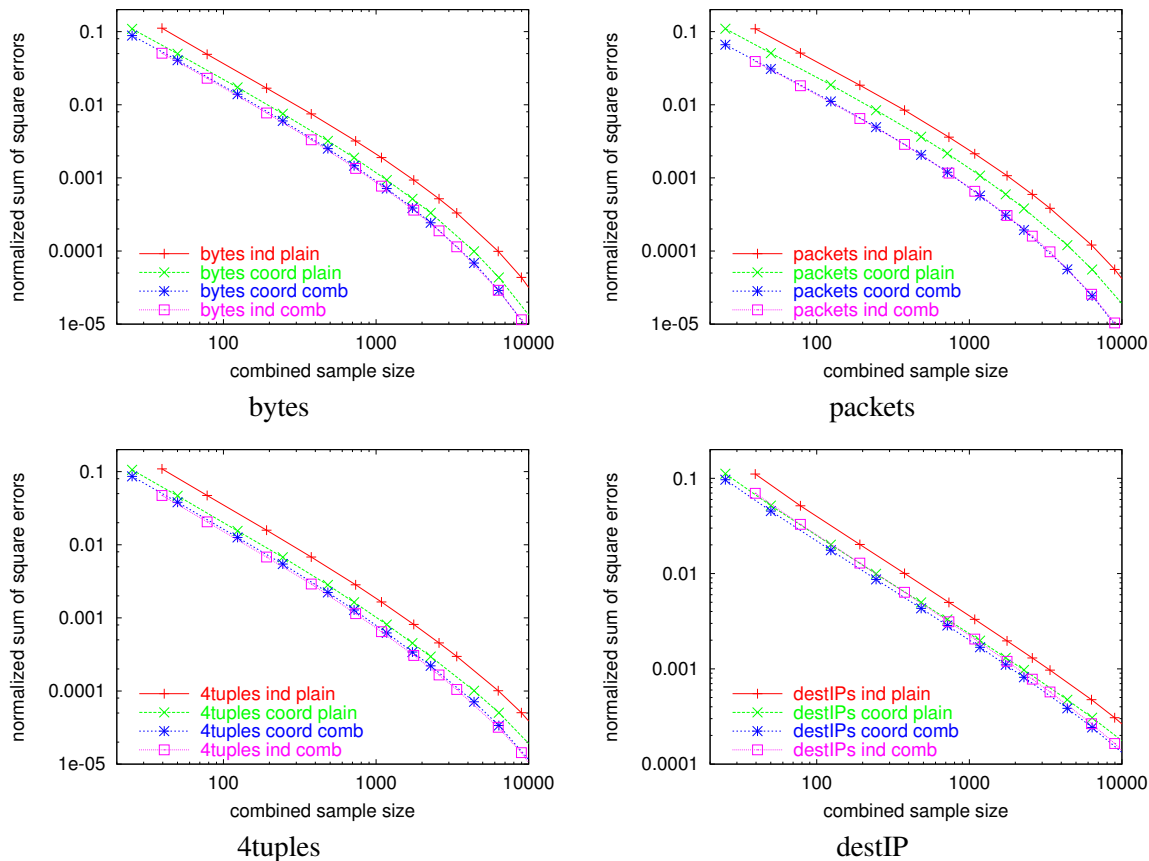


Figure 12: IP dataset1: key=destIP.  $n\Sigma V[a_i^{(b)}]$ ,  $n\Sigma V[a_c^{(b)}]$ ,  $n\Sigma V[a_{p,i}^{(b)}]$ ,  $n\Sigma V[a_{p,c}^{(b)}]$  as a function of (combined) sample size.

**Sharing index.** The *sharing index*,  $|S|/(k * |\mathcal{W}|)$  of a colocated summary  $S$  is the ratio of the expected number of distinct keys in  $S$  and the product of  $k$  and the number of weight assignments  $|\mathcal{W}|$ . The sharing index falls in the interval  $[1/|\mathcal{W}|, 1]$  and is lower (better) when more keys are shared between samples of different assignments. More keys are shared when coordination is used and when weight assignments are more similar – when all assignments are identical and coordinated sketches are used then the index is exactly  $1/|\mathcal{W}|$ .

Figure 17 plots the sharing index for coordinated and independent bottom- $k$  sketches, as a function of  $k$  for different data sets. Coordinated sketches minimize the sharing index (Theorem 4.2). On our datasets, the index lies in  $[0.25, 0.68]$  for coordinated sketches and in  $[0.4, 1]$  for independent sketches. The sharing index decreases when  $k$  becomes a larger fraction of keys, both for independent and coordinated sketches – simply because it is more likely that a key is included in a sample of another assignment. For independent sketches, the sharing index is above 0.85 for smaller values of  $k$  and can be considerably higher than with

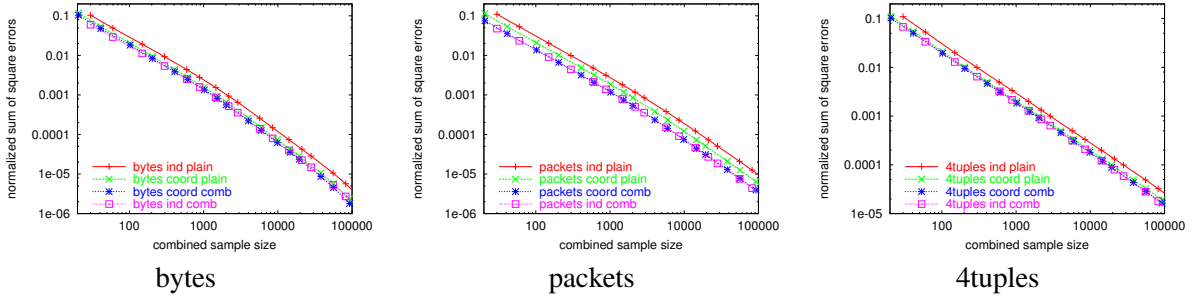


Figure 13: IP dataset1: key=4tuples.  $n\Sigma V[a_i^{(b)}]$ ,  $n\Sigma V[a_c^{(b)}]$ ,  $n\Sigma V[a_{p,i}^{(b)}]$ ,  $n\Sigma V[a_{p,c}^{(b)}]$  as a function of (combined) sample size.

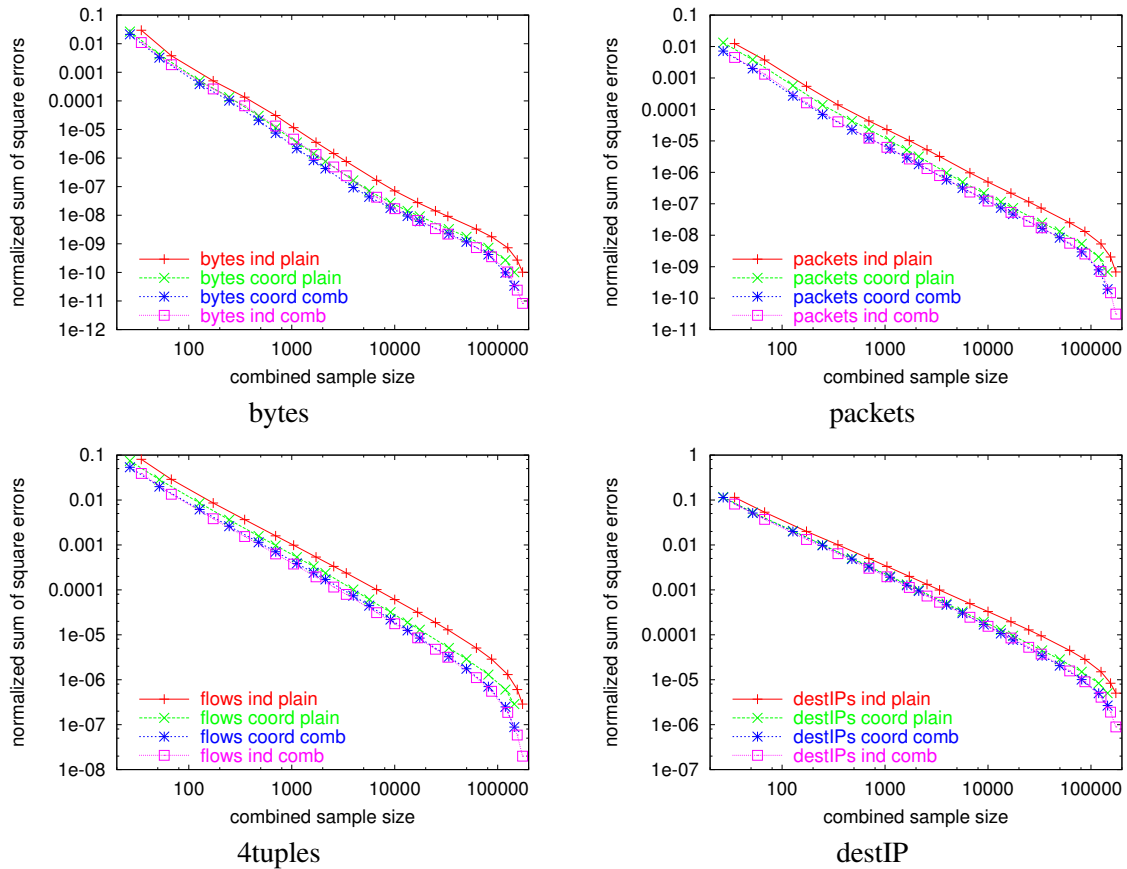


Figure 14: IP dataset2: key=destIP hour3.  $n\Sigma V[a_i^{(b)}]$ ,  $n\Sigma V[a_c^{(b)}]$ ,  $n\Sigma V[a_{p,i}^{(b)}]$ ,  $n\Sigma V[a_{p,c}^{(b)}]$  as a function of (combined) sample size.

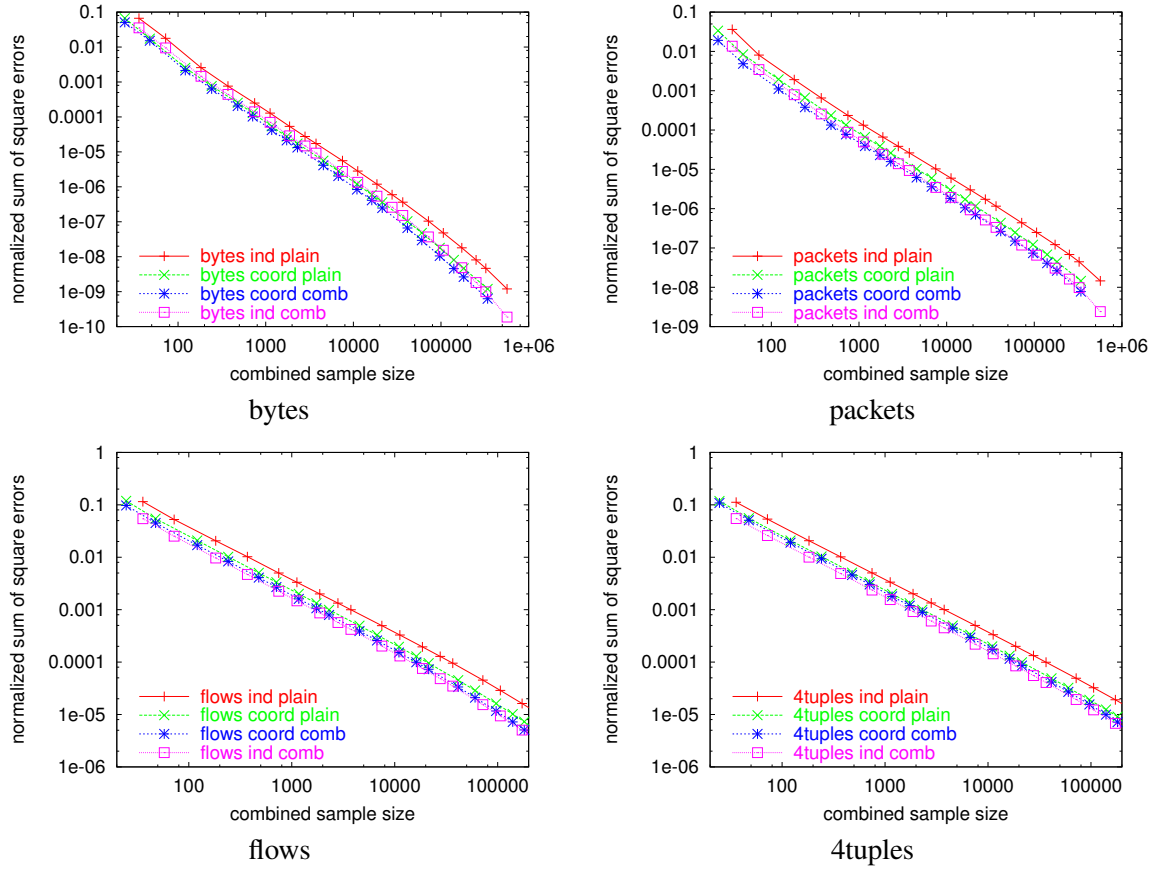


Figure 15: IP dataset2: key=4tuple hour3.  $n\Sigma V[a_i^{(b)}]$ ,  $n\Sigma V[a_c^{(b)}]$ ,  $n\Sigma V[a_{p,i}^{(b)}]$ ,  $n\Sigma V[a_{p,c}^{(b)}]$  as a function of (combined) sample size.

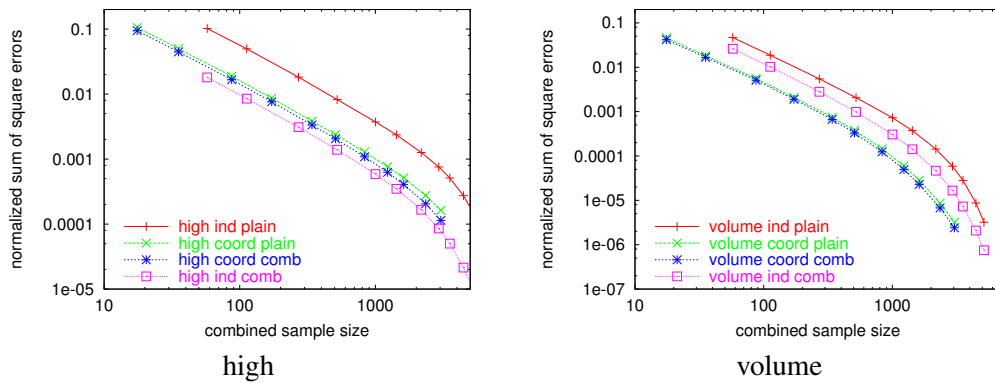


Figure 16: Stocks dataset.  $n\Sigma V[a_i^{(b)}]$ ,  $n\Sigma V[a_c^{(b)}]$ ,  $n\Sigma V[a_{p,i}^{(b)}]$ ,  $n\Sigma V[a_{p,c}^{(b)}]$  as a function of (combined) sample size.

coordinated sketches.

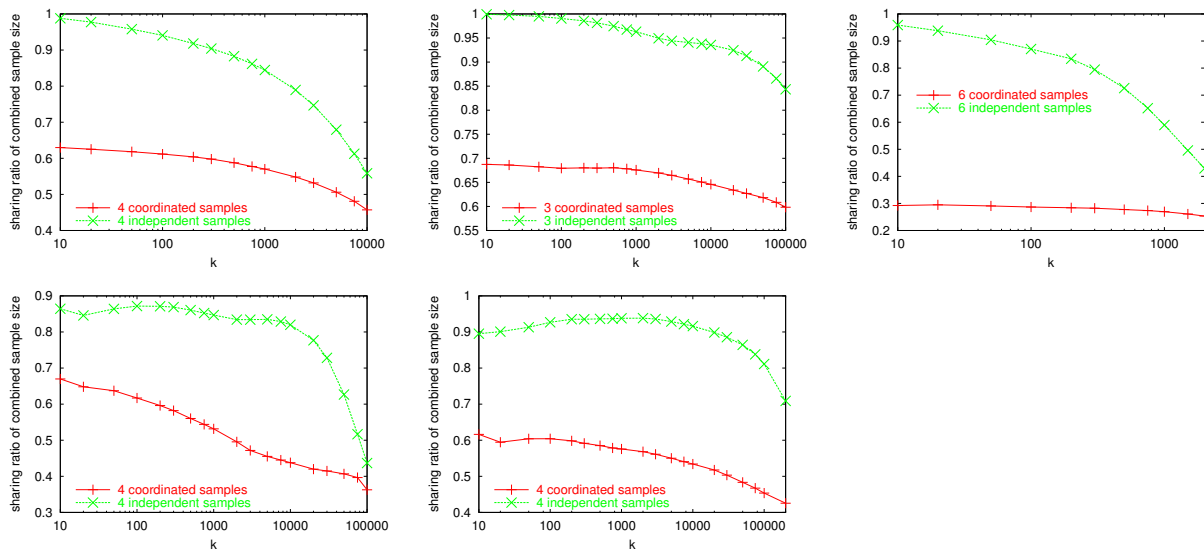


Figure 17: Sharing index of coordinated and independent sketches. Left: IP dataset1, key=destIP (4 weight assignments: bytes, packets, 4tuples, IPdests). Middle: IP dataset1, key=4tuple (3 weight assignments: bytes, packets, 4tuples). Right: Stocks dataset (6 weight assignments). Bottom: IP dataset2, key=destIP (left) IP dataset2, key=4tuple (middle)

## 10 Conclusion

We motivate and study the problem of summarizing data sets modeled as keys with *vector* weights. We identify two models for these data sets, *dispersed* (such as measurements from different times or locations) and *collocated* (records with multiple numeric attributes), that differ in the constraints they impose on scalable summarization. We then develop a sampling framework and accurate estimators for common aggregates, including aggregations over subpopulations that are specified a posteriori.

Our estimators over coordinated weighted samples for single-assignment and multiple-assignment aggregates including weighted sums and the  $L_1$  difference, max, and min improve over previous methods by orders of magnitude. Previous methods include independent weighted samples from each assignment, which poorly supports multiple-assignment aggregates, and coordinated samples for uniform weights (each item has the same weight in all assignments where its weight is nonzero), which perform poorly when, as is often the case, weight values are skewed. For collocated data sets, our coordinated weighted samples achieve optimal summary size while guaranteeing embedded weighted samples of certain sizes with respect to each individual assignment. We derive estimators for single-assignment and multiple-assignment aggregates over both independent or coordinated samples that are significantly tighter than existing ones.

As part of ongoing work, we are applying our sampling and estimation framework to the challenging problem of detection of network problems. We are also exploring the system aspects of deploying our approach within the network monitoring infrastructure in a large ISP.

## 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] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58:137–147, 1999.
- [3] 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.
- [4] B. Bloom. Space/time tradeoffs in in hash coding with allowable errors. *Communications of the ACM*, 13:422–426, 1970.
- [5] K. R. W. Brewer, L. J. Early, and S. F. Joyce. Selecting several samples from a single population. *Australian Journal of Statistics*, 14(3):231–239, 1972.
- [6] A. Z. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*, pages 21–29. ACM, 1997.
- [7] A. Z. Broder. Identifying and filtering near-duplicate documents. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, volume 1848 of *LNCS*, pages 1–10. Springer, 2000.
- [8] M. T. Chao. A general purpose unequal probability sampling plan. *Biometrika*, 69(3):653–656, 1982.
- [9] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. 34th Annual ACM Symposium on Theory of Computing*. ACM, 2002.
- [10] A. Chowdhury, O. Frieder, D. Grossman, and M. C. McCabe. Collection statistics for fast duplicate document detection. *ACM Transactions on Information Systems*, 20(1):171–191, 2002.
- [11] E. Cohen. Size-estimation framework with applications to transitive closure and reachability. *J. Comput. System Sci.*, 55:441–453, 1997.
- [12] 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.
- [13] 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.
- [14] E. Cohen, N. Duffield, H. Kaplan, C. Lund, and M. Thorup. Stream sampling for variance-optimal estimation of subset sums. In *Proc. 20th ACM-SIAM Symposium on Discrete Algorithms*. ACM-SIAM, 2009.
- [15] E. Cohen and H. Kaplan. Spatially-decaying aggregation over a network: model and algorithms. *J. Comput. System Sci.*, 73:265–288, 2007.
- [16] E. Cohen and H. Kaplan. Summarizing data using bottom-k sketches. In *Proceedings of the ACM PODC’07 Conference*, 2007.

- [17] E. Cohen and H. Kaplan. Tighter estimation using bottom-k sketches. In *Proceedings of the 34th VLDB Conference*, 2008.
- [18] E. Cohen and H. Kaplan. Leveraging discarded samples for tighter estimation of multiple-set aggregates. In *Proceedings of the ACM SIGMETRICS'09 Conference*, 2009.
- [19] E. Cohen, H. Kaplan, and S. Sen. Coordinated weighted sampling for estimating aggregates over multiple weight assignments. *Proceedings of the VLDB Endowment*, 2(1–2), 2009.
- [20] J. G. Conrad and C. P. Schriber. Constructing a text corpus for inexact duplicate detection. In *SIGIR 2004*, pages 582–583, 2004.
- [21] G. Cormode and S. Muthukrishnan. Estimating dominance norms of multiple data streams. In *Proceedings of the 11th European Symposium on Algorithms*, pages 148–161. Springer-Verlag, 2003.
- [22] G. Cormode and S. Muthukrishnan. What's new: finding significant differences in network data streams. *IEEE/ACM Transactions on Networking*, 13(6):1219–1232, 2005.
- [23] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *KDD'08*. ACM, 2008.
- [24] T. Dasu, T. Johnson, S. Muthukrishnan, and V. Shkapenyuk. Mining database structure; or, how to build a data quality browser. In *Proc. SIGMOD Conference*, pages 240–251, 2002.
- [25] N. Duffield, C. Lund, and M. Thorup. Estimating flow distributions from sampled flow statistics. In *Proceedings of the ACM SIGCOMM'03 Conference*, pages 325–336, 2003.
- [26] N. Duffield, M. Thorup, and C. Lund. Priority sampling for estimating arbitrary subset sums. *J. Assoc. Comput. Mach.*, 54(6), 2007.
- [27] P. S. Efraimidis and P. G. Spirakis. Weighted random sampling with a reservoir. *Inf. Process. Lett.*, 97(5):181–185, 2006.
- [28] L. Fan, P. Cao, J. Almeida, and A. Z. Broder. Summary cache: a scalable wide-area Web cache sharing protocol. *IEEE/ACM Transactions on Networking*, 8(3):281–293, 2000.
- [29] J. Feigenbaum, S. Kannan, M. Strauss, and M. Viswanathan. An approximate L1-difference algorithm for massive data streams. In *Proc. 40th IEEE Annual Symposium on Foundations of Computer Science*, pages 501–511. IEEE, 1999.
- [30] P. Gibbons and Y. Matias. New sampling-based summary statistics for improving approximate query answers. In *SIGMOD*. ACM, 1998.
- [31] P. Gibbons and S. Tirthapura. Estimating simple functions on the union of data streams. In *Proceedings of the 13th Annual ACM Symposium on Parallel Algorithms and Architectures*. ACM, 2001.
- [32] P. B. Gibbons. Distinct sampling for highly-accurate answers to distinct values queries and event reports. In *International Conference on Very Large Databases (VLDB)*, pages 541–550, 2001.
- [33] M. Hadjieleftheriou, X. Yu, N. Koudas, and D. Srivastava. Hashed samples: Selectivity estimators for set similarity selection queries. In *Proceedings of the 34th VLDB Conference*, 2008.
- [34] J. Hájek. *Sampling from a finite population*. Marcel Dekker, New York, 1981.



- [35] 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.
- [36] D.E. Knuth. *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*. Addison-Wesley, 1969.
- [37] A. Kolcz, A. Chowdhury, and J. Alsepector. Improved robustness of signature-based near-replica detection via lexicon randomization. In *SIGKDD 2004*, pages 605–610, 2004.
- [38] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen. Sketch-based change detection: Methods, evaluation, and applications. In *In Internet Measurement Conference*, pages 234–247. ACM Press, 2003.
- [39] A. Kumar, M. Sung, J. Xu, and E. W. Zegura. A data streaming algorithm for estimating subpopulation flow size distribution. *ACM SIGMETRICS Performance Evaluation Review*, 33, 2005.
- [40] G. Maier, R. Sommer, H. Dreger, A. Feldmann, V. Paxson, and F. Schneider. Enriching network security analysis with time travel. In *SIGCOMM' 08*. ACM, 2008.
- [41] U. Manber. Finding similar files in a large file system. In *Usenix Conference*, pages 1–10, 1994.
- [42] G. S. Manku, A. Jain, and A. D. Sarma. Detecting near-duplicates for web crawling. In *Proceedings of the 16th International World Wide Web Conference (WWW)*, 2007.
- [43] M. Mitzenmacher and S. Vadhan. Why simple hash functions work: exploiting the entropy in a data stream. In *Proc. 19th ACM-SIAM Symposium on Discrete Algorithms*, pages 746–755. ACM-SIAM, 2008.
- [44] The Netflix Prize. <http://www.netflixprize.com/>.
- [45] E. Ohlsson. Sequential poisson sampling. *J. Official Statistics*, 14(2):149–162, 1998.
- [46] E. Ohlsson. Coordination of pps samples over time. In *The 2nd International Conference on Establishment Surveys*, pages 255–264. American Statistical Association, 2000.
- [47] B. Rosén. Asymptotic theory for successive sampling with varying probabilities without replacement, I. *The Annals of Mathematical Statistics*, 43(2):373–397, 1972.
- [48] B. Rosén. Asymptotic theory for order sampling. *J. Statistical Planning and Inference*, 62(2):135–158, 1997.
- [49] F. Rusu and A. Dobra. Fast range-summable random variables for efficient aggregate estimation. In *Proc. of the 2006 ACM SIGMOD Int. Conference on Management of Data*, pages 193–204. ACM, 1990.
- [50] P. J. Saavedra. Fixed sample size pps approximations with a permanent random number. In *Proc. of the Section on Survey Research Methods, Alexandria VA*, pages 697–700. American Statistical Association, 1995.
- [51] C-E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer, 1992.
- [52] S. Schleimer, D. Wilkerson, and A. Aiken. Winnowing: local algorithms for document fingerprinting. In *Proceedings of the ACM SIGMOD*, 2003.

- [53] R. Schweller, A. Gupta, E. Parsons, and Y. Chen. Reversible sketches for efficient and accurate change detection over network data streams. In *in ACM SIGCOMM IMC*, pages 207–212. ACM Press, 2004.
- [54] M. Szegedy. The DLT priority sampling is essentially optimal. In *Proc. 38th Annual ACM Symposium on Theory of Computing*. ACM, 2006.
- [55] M. Szegedy and M. Thorup. On the variance of subset sum estimation. In *Proc. 15th ESA*, 2007.
- [56] J.S. Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, 1985.