Coordinated Weighted Sampling for Estimating Aggregates Over Multiple Weight Assignments

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

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

Subhabrata Sen
AT&T Labs–Research
180 Park Avenue
Florham Park, NJ 07932, USA
sen@research.att.com

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 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.

VLDB '09, August 24-28, 2009, Lyon, France

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 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 *subpopulation* 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 [21, 1, 34, 12, 13]. (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 [33, 18, 7, 17]. Figure 1 (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 [36, 8, 47, 16, 32, 37]. 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 [24, 18, 19, 35]: 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 [31, 51, 6, 5, 9, 26, 27, 2, 20, 28, 13, 22, 10, 14] 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. 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: A challenge is that 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: we can tailor the sampling to be Poisson, k-mins, or order (bottom-k) sampling. In all three cases, sampling is efficient on data streams, distributed data, and metric data [9, 11, 23, 12] and there are unbiased subpopulation weight estimators that have variance that decreases linearly or faster with the sample size [9, 22, 49, 13]. Order samples [42, 46, 43, 9, 13, 40, 22], with the advantage of a fixed sample size, emerge as a better choice. Coordination loosely means that a key that is sampled under one assignment is more likely to be sampled under other assignment. 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 generic derivation of estimators, tailor it to obtain tight unbiased estimators for the min, max, and 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, 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 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 is minimized.

Empirical evaluation. We performed a comprehensive empirical evaluation using IP packet traces, movies' ratings data set (The Netflix Challenge [39]), 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 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.Section 5 presents our estimators and Section 6 provides bounds on the variance. Section 7 presents the evaluation results. Finally, Section 8 concludes the paper. Details including derivations and proofs can be found in [15].

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 [4] and order samples [45, 41, 43]. 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 order samples) was (re-)introduced in computer science as a method to support aggregations that involve multiple sets [6, 5, 9, 26, 27, 2, 13, 28, 14]. 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.

Sketches that are not samples. Sketches that are not *sample based* [36, 7, 8, 47, 16, 32, 37, 17, 25] are effective point solutions for particular metrics such as max-dominance [17] or L_1 [25] difference. Their 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[48].

Bloom filters [3, 24] 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 associating *adjusted weights* with sampled keys that are used to answer weight queries. Sample distributions are defined through *random rank assignments* [9, 43, 12, 22, 13, 14] 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 $\mathbf{F}_\mathbf{w}$ are the respective cumulative distributions). For a set J and a rank assignment r we denote by $r_i(J)$ the ith smallest rank of a key in J, we also abbreviate and write $r(J) = r_1(J)$.

- A *Poisson-τ* 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.
- An *order-k (bottom-k)* sample of J contains the k keys $i_1, \ldots, i_k$ of smallest ranks in J. The sketch $s_k(J, r)$ consists of the k pairs $(r(i_j), w(i_j))$, $j = 1, \ldots, k$, 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, $r^{(1)}, \ldots, r^{(k)}$. The sample is the set of (at most k) keys) with minimum rank values $r^{(1)}(J)$, $r^{(2)}(J)$, ..., $r^{(k)}(J)$. 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, order-k samples are k uniform draws without replacements, and Poisson-τ 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 EXP$[w]$). Equivalently, if $u \in U[0,1]$ then $-\ln(u)/w$ is an exponential random variable with parameter w. EXP$[w]$ ranks have the property that the minimum rank $r(J)$ has distribution EXP$[w(J)]$, where $w(J) = \sum_{i \in J} w(i)$ This property is useful for designing estimators and efficiently computing sketches [9, 11, 23, 12, 13]. The k-mins sample [9] 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. An order-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 [42, 29, 43].
- 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-τ sample is an IPPS sample [29] (Inclusion Probability Proportional to Size). The order-k sample is a priority sample [40, 22] (PRI).

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 $\mathsf{E}[a(i)] = w(i)$, 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. 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 Ω 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 $i \notin s$, $\mathsf{E}[a(i)] = w(i)$ ($a(i)$ is an unbiased estimator of $w(i)$). These adjusted weights are called the Horvitz-Thompson (HT) estimator [30]. For a particular Ω, the HT adjusted weights minimize VAR$[a(i)]$ for all $i \in I$. The HT adjusted weights for Poisson τ-sampling are $a(i) = w(i)/\mathbf{F}_{w(i)}(\tau)$. Poisson sampling with IPPS ranks and HT adjusted weights are known to minimize the sum $\sum_{i \in I}$ VAR$(a(i))$ of per-key variances over all AW-summaries with the same expected size.

HT on a partitioned sample space (HTP) [13]. 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 example, if s is an order-k sample 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.

For each key i we consider a partition of Ω 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 $p^i(s) = \Pr\{i \in s' \mid s' \in P^i(s)\}$ 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, $\mathsf{E}[a(i)] = w(i)$. Therefore, also over Ω we have $\mathsf{E}[a(i)] = w(i)$.

Rank Conditioning (RC**)** is an HTP method designed for an order-k sketch [13]. For each i and possible rank value τ we have an equivalence class P^i_τ containing all sketches in which the kth smallest rank value assigned to a key other than i is τ. Note that if $i \in s$ then this is the $(k+1)$st smallest rank which is included in the sketch. It is easy to see that the inclusion probability of i in a sketch in P^i_τ is $p^i_\tau = \mathbf{F}_{w(i)}(\tau)$.

Assume s contains $i_1, \ldots, i_k$ and the $(k+1)$st smallest rank value r_{k+1}. Then for key i_j, we have $s \in P^{i_j}_{r_{k+1}}$ and $a(i_j) = \frac{w(i_j)}{\mathbf{F}_{w(i_j)}(r_{k+1})}$.

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

The RC estimator for order-k samples with IPPS ranks [22] 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 $k+1$ [49]. Order sampling emerges as superior to Poisson sampling, since it matches its estimation quality per expected sample size and has the desirable property of a fixed sample size.

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 $\mathrm{VAR}[a(J)] = \mathsf{E}[a(J)]^2 - w(J)^2$. Since our application is for arbitrary subpopulations that may not 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 J.

RC adjusted weights have *zero covariances*, that is, for any two keys i, j, $\mathrm{COV}[a(i), a(j)] = \mathsf{E}[a(i)a(j)] - w(i)w(j) = 0$ [13]. This property extends to applications of the RC method to coordinated sketches with global weights [14]. 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 $\sum_{i,j \in J} \mathrm{COV}[a(i), a(j)] = \sum_{i \in J} \mathrm{VAR}[a(i)]$. For AW-summaries with zero covariances, the *sum of per-key variances* $\Sigma V[a] \equiv \sum_{i \in I} \mathrm{VAR}[a(i)]$, also measures average variance over subpopulations of certain weight [50]. $\Sigma V[a]$ hence serves as a balanced performance metric [22, 13] and we use it in our performance evaluation.

Estimators for Poisson, k-mins, and order sketches with EXP or IPPS ranks have $\Sigma V[a] \leq \frac{\sum_{i \in I} w(i)^2}{k-2}$ (where k is the (expected) sample size) [9, 12, 22, 49]. This bound is tight when keys have uniform weights and $k \ll |I|$, but $\Sigma V[a]$ is smaller for order and Poisson sketches when the weight distribution is skewed [12, 22]. For a subpopulation J with expected k' samples in the sketch, the variance on estimating $w(J)$ is bounded by $w(J)^2/(k'-2)$.

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 $b \in \mathcal{W}$, $w^{(b)} : I \to \mathcal{R}_{\geq 0}$ maps keys to nonnegative reals. Figure 1 shows a data set with $I = \{i_1, \ldots, i_6\}$ 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 [17, 18] of the selected subset.

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 L_1 difference 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 1, 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 Ω 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 $\mathbf{f}_{\mathbf{w}^{(\mathbf{b})}(\mathbf{i})}$. (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 Ω is specified by the mapping (iii) from weight vectors to distributions of rank vectors specifies Ω.

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 Ω 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 $b_1, b_2 \in \mathcal{W}$,

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

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

order 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

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

order 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

(B)

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

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

• **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 $b \in \mathcal{W}$, $r^{(b)}(i) \leftarrow \mathbf{F}^{-1}_{w^{(b)}(i)}(u(i))$.

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 $r^{(b)}(i) = -\ln u(i)/w^{(b)}(i)$.

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

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

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 $b \in \mathcal{W}$. Order and Poisson sketches: A single rank assignment r on $(I, \mathcal{W})$ defines an order-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 1 shows examples of independent and shared-seed consistent rank assignments for the example data set and the corresponding order 3-samples.

In the sequel we mainly focus on order-k sketches. Derivations are similar (but simpler) for Poisson sketches. We shall denote by $S(r)$ the summary consisting of $|\mathcal{W}|$ order-k sketches obtained using a rank assignment r.

k-mins sketches derived from rank assignments with independent-differences consistent ranks have the following property:

THEOREM 4.1. *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.*

Therefore, the fraction of common keys in the two k-mins sketches is an unbiased estimator of the weighted Jaccard similarity. This generalizes the estimator for unweighted Jaccard similarity [5].

The following theorem shows that shared-seed consistent ranks maximizes the sharing of keys between sketches. We prove it for Poisson sketches and conjecture that it holds also for order 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 minimize the expected number of distinct keys in the union of the sketches for $(I, w^{(b)})$, $b \in \mathcal{W}$.*

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 $\mathbf{f}_w$ ($w > 0$).*

A consequence of Lemma 4.1 is the following:

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

Fixed number of distinct keys for colocated data The number of distinct keys in coordinated size-k sketches is at most $|\mathcal{W}|k$. It is smaller when weight assignments are more correlated. The size varies by the rank assignment when k is fixed. A different natural goal is instead of fixing k, to fix the number of distinct keys to be between $[|\mathcal{W}|(k-1)+1, |\mathcal{W}|k]$ distinct keys. For a rank assignment r, we define ℓ to be the largest such that there are at most $|\mathcal{W}|k$ distinct keys in the union of the order-ℓ sketches with respect to $r^{(b)}$ ($b \in \mathcal{W}$). As a result, we have varying $\ell \geq k$ but sample size in $[|\mathcal{W}|(k-1)+1, |\mathcal{W}|k]$. This sample can be computed by a simple adaptation of the stream sampling algorithm for the fixed-k variant.

Computing coordinated sketches. Coordinated order sketches can be computed by a small modification of existing order 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 [44, 38], that simple heuristic hash functions and pseudo-random number generators [2] 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 [25, 44].

5. ESTIMATORS

Consider $(I, \mathcal{W})$, a rank assignment $r \in \Omega$, and a corresponding summary $S(r)$. The input to our generic estimator is a numeric function f and a predicate d, defined for each key in I. Our estimator assigns adjusted f-weights $a^{(f)}(i)$ to a subset $S^*(r)$ of the keys included in $S(r)$. An estimate for $\sum_{i|d(i)=1} f(i)$ is obtained by summing the adjusted f-weights of keys in $S^*(r)$ that satisfy the predicate d. A handy property is that the same adjusted f-weights can be used for different selection predicates $d()$.

Recall that the probability subspace $\Omega(i,r)$ consists of all rank assignments r' such that $\forall b \in \mathcal{W}$, and $\forall j \in I \setminus \{i\}$, $r'^{(b)}(j) = r^{(b)}(j)$. Let $p(i,r)$ denote the probability that i is included in $S^*(r')$ for $r' \in \Omega(i,r)$ we apply HTP and use $a^{(f)}(i) = f(i)/p(i,r)$.

$S^*(r)$ is selected to be as inclusive as possible such that we can evaluate $d(i)$, $f(i)$, and $p(i,r)$ for all $i \in S^*(r)$ based on the information in $S(r)$.

5.1 Colocated Weights

The summary $S(r)$ contains all keys $i \in I$ such that for at least one $b \in \mathcal{W}$, $r^{(b)}(i) \leq r^{(b)}_{k+1}(I)$ and the full weight vector $w^{(\mathcal{W})}(i)$ for each included key. Hence, any f and d can be evaluated for all $i \in S(r)$.

We use the generic estimator with $S^*(r) \equiv S(r)$ and refer to this as *inclusive* estimators. (We use the term inclusive since they use all keys in the union of the order-k samples.) Inclusive estimators are applicable when f and d satisfy the condition $f(i)d(i) > 0 \implies w^{(\max_{\mathcal{W}})}(i) > 0$ for all $i \in I$, which simply means that any key with a positive contribution to the aggregate has a positive probability of being sampled. The probability that i is included in $S(r')$ for $r' \in \Omega(i,r)$ is

$$p(i,r) = \textsc{pr}[\exists b \in \mathcal{W}, r'^{(b)}(i) < r^{(b)}_k(I \setminus \{i\}) | r' \in \Omega(i,r)] \ . \quad (3)$$

To compute (3), the summary should include, for each $b \in \mathcal{W}$, the rank values $r^{(b)}_k(I)$ and $r^{(b)}_{k+1}(I)$ and for each $i \in S(r)$ and $b \in \mathcal{W}$, whether i is included in the order-k sketch of b (that is, whether $r^{(b)}(i) < r^{(b)}_{k+1}(I)$). This information allows us to determine the values $r^{(b)}_k(I \setminus \{i\})$ for all $i \in I$ and $b \in \mathcal{W}$: if i is included in the sketch for b then $r^{(b)}_k(I \setminus \{i\}) = r^{(b)}_{k+1}(I)$. Otherwise, it is $r^{(b)}_k(I)$.

We provide explicit expressions for $p(i,r)$ (Eq. (3)), for $i \in S(r)$, for the rank distributions which we consider. Since we can evaluate $p(i,r)$, $f(i)$, and $d(i)$ for all $i \in S(r)$, we can indeed apply the generic estimator with $S^*(r) \equiv S(r)$.

Independent ranks (independent order-k sketches): The probability over $\Omega(i,r)$ that i is included in the order-k sketch of b is $\mathbf{F}_{w^{(b)}(i)}(r^{(b)}_k(I \setminus \{i\}))$. 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,

$$p(i,r) = 1 - \prod_{b \in \mathcal{W}} (1 - \mathbf{F}_{w^{(b)}(i)}(r^{(b)}_k(I \setminus \{i\}))) \ . \quad (4)$$

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

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

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

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

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

We also define $M_\ell = \max_{a=\ell}^{h} r^{(b_a)}_k(I \setminus \{i\})$ ($\ell \in [h]$), 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 $p(i,r) = \sum_{\ell=1}^{h} \textsc{pr}[A_\ell]$.

The probabilities $\textsc{pr}[A_\ell]$ can be computed using a linear pass on the sorted weight vector of i.

5.2 Dispersed weights

Let r be a rank assignment for $(I, \mathcal{W})$. The summary $S(r)$ is the set of order-k sketches $s_k(I, r^{(b)})$ for $b \in \mathcal{W}$. 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 $i \in s_k(I, r^{(b)})$.

For $\mathcal{R} \subset \mathcal{W}$ and $i \in I$, let $w^{(\max_{\mathcal{R}})}(i) = \max_{b \in \mathcal{R}} w^{(b)}(i)$, $b^{(\max_{\mathcal{R}})}(i) = \arg\max_{b \in \mathcal{R}} w^{(b)}(i)$ (the weight assignment from $\mathcal{R}$ which maximizes i's weight), and $r^{(\min_{\mathcal{R}})}(i) = \min_{b \in \mathcal{R}} r^{(b)}(i)$ (the smallest rank value that i assumes for $b \in \mathcal{R}$). If r is consistent then $r^{(\min_{\mathcal{R}})}(i) = r^{b^{(\max_{\mathcal{R}})}(i)}(i)$ (smallest rank value for i is assumed on the assignment with largest weight). Similarly, $w^{(\min_{\mathcal{R}})}(i) = \min_{b \in \mathcal{R}} w^{(b)}(i)$, $b^{(\min_{\mathcal{R}})}(i) = \arg\min_{b \in \mathcal{R}} w^{(b)}(i)$, and $r^{(\max_{\mathcal{R}})}(i) = \max_{b \in \mathcal{R}} r^{(b)}(i)$. When the dependency on $\mathcal{R}$ is clear from context, it is omitted.

We also use $r^{(\min_{\mathcal{R}})}_{k+1}(I) = \min_{b \in \mathcal{R}} r^{(b)}_{k+1}(I)$ and denote the weight and rank vectors of $i \in I$ by $r^{(\mathcal{R})}(i)$ and $w^{(\mathcal{R})}(i)$.

We apply the generic derivation using the following guidelines:

(1) If f can be expressed as a linear combination of the form $f(i) = f_1(i) + f_2(i) + \ldots$, we estimate each summand f_j separately. This allows for weaker conditions in the generic derivation, resulting in more inclusive sets of applicable samples and tighter estimates. In some cases it is necessary to express f as a linear combination in order to facilitate estimation, as there are $f = f_1 + f_2$ such that the generic estimator is not applicable to f but is applicable to f_1 and f_2.

(2) We determine a set $\mathcal{R} \subset \mathcal{W}$ of *relevant assignments* for f and d. The set $S^*(r)$ of applicable samples is a subset of $\bigcup_{b \in \mathcal{R}} s_k(I, r^{(b)})$.

(3) We consider the dependence of f and d on the weight vector $w^{(\mathcal{R})}$. We derive estimators for two families of f and d's that include the cases where f is $w^{(\min \mathcal{R})}$, $w^{(\max \mathcal{R})}$, or $w^{(L_1 \mathcal{R})}$ which we used in our empirical evaluation. Our methodology is applicable to other interesting f's such as quantiles over assignments.

We say that f and d are *min-dependent* if

$$w^{(\min_{\mathcal{R}})}(i) = 0 \Rightarrow f(i)d(i) = 0 \ .$$

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	5.42×10^5	5.54×10^5	7.47×10^5	3.49×10^5	3.98×10^5
destIP, bytes	2.08×10^9	2.17×10^9	3.26×10^9	9.96×10^8	2.26×10^9
srcIP+destIP, packets	4.61×10^6	4.61×10^6	7.61×10^6	1.61×10^6	6.00×10^6
srcIP+destIP, bytes	2.08×10^9	2.17×10^9	3.49×10^9	7.65×10^8	2.72×10^9

Table 1: IP dataset1

months	1	2	3	4	5	6	7	8	9	10	11	12	1,2	1-6	1-12
distinct 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 2: Netflix data set. Distinct movies (number of movies with at least one rating) and total number of ratings for each month ($1, \ldots, 12$) in 2005 and for periods $\mathcal{R} = \{1, 2\}$, $\mathcal{R} = \{1, \ldots, 6\}$, and $\mathcal{R} = \{1, \ldots, 12\}$. For these periods, we also show $\sum_i w^{(\min_\mathcal{R})}(i)$, $\sum_i w^{(\max_\mathcal{R})}(i)$, and $\sum_i w^{(L_1\mathcal{R})}(i)$.

It is easy to see that $f(i) = w^{(\min_\mathcal{R})}(i)$ and any predicate d are min-dependent, but $f(i) = w^{(\max_\mathcal{R})}(i)$ and any d which selects items i for which $w^{(\max_\mathcal{R})}(i) > 0$ is not. We derive estimators for all min-dependent f, d for both coordinated and independent sketches.

We say that f and d are *max-dependent* if

$$\begin{aligned} f(i) &\equiv f(w^{(\max_\mathcal{R})}(i), b^{(\max_\mathcal{R})}(i)) \\ d(i) &\equiv d(w^{(\max_\mathcal{R})}(i), b^{(\max_\mathcal{R})}(i)) \\ w^{(\max_\mathcal{R})}(i) = 0 &\Rightarrow d(i)f(i) = 0\,. \end{aligned}$$

In particular, $f(i) = w^{(\max_\mathcal{R})}(i)$ and any attribute-based predicate d are max-dependent. We derive estimators for max-dependent f and d for coordinated sketches. We also argue that it is not possible to obtain unbiased nonnegative estimates for $f(i) = w^{(\max_\mathcal{R})}(i)$ over independent sketches.

5.2.1 Max-dependence

Max-dependence estimator (coordinated sketches):

- $S^*(r) \leftarrow \{i \mid \exists b \in \mathcal{R}, r^{(b)}(i) < r^{(\min_\mathcal{R})}_{k+1}(I)\}$
- For $i \in S^*(r)$:

$$\begin{aligned} w^{(\max_\mathcal{R})}(i) &\leftarrow \max\{w^{(b)}(i) \mid b \in \mathcal{R}, i \in s_k(I, r^{(b)})\} \\ b^{(\max_\mathcal{R})}(i) &\leftarrow \arg \max_{b \in \mathcal{R} \mid i \in s_k(I, r^{(b)})} w^{(b)}(i) \\ p(i,r) &\leftarrow \mathbf{F}_{w^{(\max_\mathcal{R})}(i)}(r^{(\min_\mathcal{R})}_{k+1}(I)) \\ a^f(i) &\leftarrow \frac{f(w^{(\max_\mathcal{R})}(i), b^{(\max_\mathcal{R})}(i))}{p(i,r)} \end{aligned}$$

- Output $\sum_{i \in S^*(r) \mid d(w^{(\max_\mathcal{R})}(i), b^{(\max_\mathcal{R})}(i))} a^f(i)$

As a special case for $f(i) = w^{(\max \mathcal{R})}(i)$ and $i \in S^*(r)$ we obtain the adjusted weights:

$$a^{(\max_\mathcal{R})}(i) = \frac{w^{(\max_\mathcal{R})}(i)}{\mathbf{F}_{w^{(\max_\mathcal{R})}(i)}(r^{(\min_\mathcal{R})}_{k+1}(I))} \tag{6}$$

5.2.2 Min-dependence

Min-dependence l-set estimator:

- $S^*_\ell(r) \leftarrow \{i \mid \bigwedge_{b \in \mathcal{R}} r^{(b)}(i) < r^{(b)}_{k+1}(I)\}$
- $\forall i \in S^*_\ell(r)$,

$p_\ell(i,r) \leftarrow \text{PR}[\forall b \in \mathcal{R}, r'^{(b)}(i) < r^{(b)}_{k+1}(I) \mid r' \in \Omega(i,r)]$

$S^*_\ell(r)$ is the set of keys that are included in all $|\mathcal{R}|$ order-k sketches.

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

$$p_\ell(i,r) = \min_{b \in \mathcal{R}} \mathbf{F}_{w^{(b)}(i)}(r^{(b)}_{k+1}(I)) \tag{7}$$

For EXP ranks,

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

and for IPPS ranks, $p_\ell(i,r) = \min\{1, \min_{b \in \mathcal{R}}\{w^{(b)}(i) r^{(b)}_{k+1}(I)\}\}$. For independent-differences consistent ranks, $p_\ell(i,r)$ is expressed as a simultaneous bound on all prefix-sums of a set of independent exponentially-distributed random variables.

For independent ranks:

$$p_\ell(i,r) = \prod_{b \in \mathcal{R}} \mathbf{F}_{w^{(b)}(i)}(r^{(b)}_{k+1}(I))\,. \tag{8}$$

By contrasting (7) and (8) 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 $(\frac{1}{p_\ell(i,r)} - 1)$, we can have exponentially larger variance.

Let $a^{(\min_\mathcal{R})}_\ell(i)$ be the adjusted weight for $f(i) = w^{(\min_\mathcal{R})}(i)$ of the l-set estimator using shared-seed consistent ranks, and let $a^{(\min_\mathcal{R})}_{ind}(i)$ be the adjusted weight for $f(i) = w^{(\min_\mathcal{R})}(i)$ of the l-set estimator using independent ranks.

We can also use a smaller set of samples as follows.

Min-dependence s-set estimator:

- $S^*_s(r) \leftarrow \{i \mid \bigwedge_{b \in \mathcal{R}} r^{(b)}(i) < r^{(\min_\mathcal{R})}_{k+1}(I)\}$
- $\forall i \in S^*_s(r)$,

$p_s(i,r) \leftarrow \text{PR}[\forall b \in \mathcal{R}, r'^{(b)}(i) < r^{(\min_\mathcal{R})}_{k+1}(I) \mid r' \in \Omega(i,r)]$

$S^*_s(r)$ is the set of keys that are included in all $|\mathcal{R}|$ sketches with rank value at most $r^{(\min_\mathcal{R})}_{k+1}(I)$. The advantage of the s-set estimator is that for coordinated sketches the inclusion probabilities have a simpler formula which is easier to compute namely

$$p_s(i,r) = \mathbf{F}_{w^{(\min_\mathcal{R})}(i)}(r^{(\min_\mathcal{R})}_{k+1}(I))\,.$$

The s-set estimator can be used with independent ranks but there is no advantage in doing so.

As a special case, we obtain adjusted weights for $f(i) = w^{(\min_\mathcal{R})}(i)$ by

$$a^{(\min_\mathcal{R})}_s(i) = \frac{w^{(\min_\mathcal{R})}(i)}{\mathbf{F}_{w^{(\min \mathcal{R})}(i)}(r^{(\min_\mathcal{R})}_{k+1}(I))}\,, \tag{9}$$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
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	1.48	1.52	1.52	1.48	1.45	1.37	1.38	1.38	1.42	1.46	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	1.57	1.57	1.56	1.52	1.49	1.44	1.43	1.45	1.49	1.50	1.54
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	1.44	1.49	1.49	1.42	1.38	1.34	1.34	1.33	1.39	1.42	1.44
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	1.50	1.55	1.51	1.45	1.44	1.40	1.36	1.42	1.44	1.48	1.51
adj_close	1.81	1.74	1.72	1.64	1.58	1.55	1.47	1.45	1.57	1.56	1.46	1.49	1.50	1.54	1.51	1.44	1.43	1.39	1.36	1.42	1.43	1.47	1.50
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	1.84	1.42	1.43	1.73	2.05	1.84	1.55	1.99	1.96	1.71	1.75

Table 3: Daily totals for 23 trading days in October, 2008. Prices (open, high, low, close, adjusted_close) are $\times 10^5$. Volumes are in$\times 10^{10}$.

for every $i \in S_s^*(r)$, and $a_s^{(\min_{\mathcal{R}})}(i) = 0$ otherwise.

s-set versus l-set estimators. The l-set estimators have lower variance than the s-set estimators:

LEMMA 5.1. *For any weight function f and $i \in I$,*

$$\text{VAR}[a_l^{(f)}(i)] \leq \text{VAR}[a_s^{(f)}(i)]$$

5.3 L_1 difference.

For a consistent r, we define the $w^{(L_1\mathcal{R})}$ adjusted weights

$$a_s^{(L_1\mathcal{R})}(i) = a^{(\max_{\mathcal{R}})}(i) - a_s^{(\min_{\mathcal{R}})}(i) \quad (10)$$

$$a_\ell^{(L_1\mathcal{R})}(i) = a^{(\max_{\mathcal{R}})}(i) - a_\ell^{(\min_{\mathcal{R}})}(i). \quad (11)$$

We use the notation $p^{(\max_{\mathcal{R}})}(i,r)$, $p_s^{(\min_{\mathcal{R}})}(i,r)$, and $p_\ell^{(\min_{\mathcal{R}})}(i,r)$ for the respective inclusion probabilities. We use the notation $a^{(\min_{\mathcal{R}})}$, $a^{(L_1\mathcal{R})}$, $p^{(\min_{\mathcal{R}})}$ when the statement applies to both the respective s-set and l-set estimators.

We show that for coordinated sketches, our $w^{(L_1\mathcal{R})}$ adjusted weights are "well behaved," in the sense that they are nonnegative.

LEMMA 5.2. *For consistent r with* IPPS *or* EXP *ranks, $\forall i \in I$, $a^{(L_1\mathcal{R})}(i) \geq 0$.*

6. VARIANCE PROPERTIES

We conjecture that the estimators we presented have zero covariances. That is, for all $i \neq j \in I$, $\mathsf{E}[a^{(f)}(i)a^{(f)}(j)] = f(i)f(j)$. This conjecture is consistent with empirical observations and with properties of related RC estimators [13, 14]. 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 $J \subset I$.

We use the notation $t_k^{(f)}(i)$ for the RC f-adjusted weights assigned by an RC estimators applied to a order-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 $\Omega(i,r)$. (Since these subspaces partition Ω and our estimators are unbiased on each subspace).

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

$$\text{VAR}_{\Omega(i,r)}[a^{(f)}(i)] = f(i)^2 \left(\frac{1}{p(i,r)} - 1\right). \quad (12)$$

Colocated single-assignment estimators. We show that our single-assignment inclusive estimators for co-located summaries (independent or coordinated) dominate plain RC estimators based on a single order-k sketch.

LEMMA 6.1. *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 (3)). Then, $\text{VAR}[a^{(b)}(i)] \leq \text{VAR}[t^{(b)}(i)]$.*

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 $\binom{|\mathcal{W}|}{2}$ 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 6.2. *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 (3)).*

$$\text{VAR}[a^{(\min_{\mathcal{R}})}(i)] = \min_{b \in \mathcal{R}} \text{VAR}[a^{(b)}(i)],$$

$$\text{VAR}[a^{(\max_{\mathcal{R}})}(i)] = \max_{b \in \mathcal{R}} \text{VAR}[a^{(b)}(i)],$$

$$\text{VAR}[a^{(L_1\mathcal{R})}(i)] \leq \text{VAR}[a^{(\max_{\mathcal{R}})}(i)].$$

The following relations are an immediate corollary of Lemma 6.2:

$$\Sigma V[a^{(\min_{\mathcal{R}})}] \leq \min_{b \in \mathcal{R}} \Sigma V[a^{(b)}], \quad \Sigma V[a^{(\max_{\mathcal{R}})}] \leq \max_{b \in \mathcal{R}} \Sigma V[a^{(b)}],$$

$$\Sigma V[a^{(L_1\mathcal{R})}] \leq \Sigma V[a^{(\max_{\mathcal{R}})}] \leq \max_{b \in \mathcal{R}} \Sigma V[a^{(b)}].$$

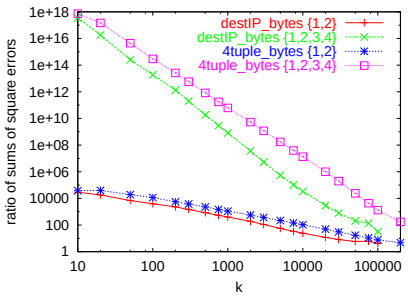
Relative variance bound for $\max$: For both the dispersed and the colocated models, we show that the variance of the $\max$ estimator is at most that of an estimator applied to a weighted sample taken with $\max$ being the primary weight. More precisely, $a^{(\max_{\mathcal{R}})}(i)$ has at most the variance of an RC estimator applied to an order-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 order-k sketch estimators are applicable [12, 13, 22].

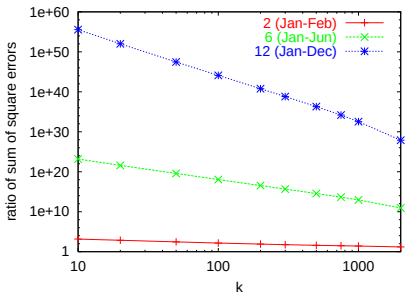
LEMMA 6.3. *Let $t_k^{(\max_{\mathcal{R}})}(i)$ be the adjusted weights of the* RC *estimator applied to an order-k sketch of $(I, w^{(\max_{\mathcal{R}})})$. For any $i \in I$, $\text{VAR}[a^{(\max_{\mathcal{R}})}(i)] \leq \text{VAR}[t_k^{(\max_{\mathcal{R}})}(i)]$.*

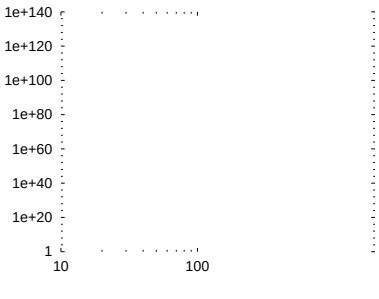
Dispersed model $\min$ and L_1 estimators. We bound the absolute variance of our $w^{(\min_{\mathcal{R}})}$ estimator in terms of the variance of $w^{(b)}$-estimators for $b \in \mathcal{R}$. Let $t_k^{(b)}$ be RC adjusted $w^{(b)}$-weights using the order-k sketch with ranks $r^{(b)}$.

LEMMA 6.4. *For shared-seed consistent r, for all $i \in I$,*

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


destIP_bytes {1,2}
destIP_bytes {1,2,3,4}
4tuple_bytes {1,2}
4tuple_bytes {1,2,3,4}
ratio of sums of square errors
k
1e+18
1e+16
1e+14
1e+12
1e+10
1e+08
1e+06
10000
100
1
10
100
1000
10000
100000


2 (Jan-Feb)
6 (Jan-Jun)
12 (Jan-Dec)
ratio of sum of square errors
k
1e+60
1e+50
1e+40
1e+30
1e+20
1e+10
1
10
100
1000


1e+140
1e+120
1e+100
1e+80
1e+60
1e+40
1e+20
1
10
100

$\mathcal{R} = \{1, \ldots, 5\}$ (October 1-7), $\mathcal{R} = \{1, \ldots, 10\}$ (October 1-14), $\mathcal{R} = \{1, \ldots, 15\}$ (October 1-21), $\mathcal{R} = \{1, \ldots, 23\}$ (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 ($\times 10^5$)					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

7.2 Dispersed data.

We evaluate our $w^{(\min_{\mathcal{R}})}$, $w^{(\max_{\mathcal{R}})}$, and $w^{(L_1 \mathcal{R})}$ estimators as defined in Section 5.2: $a^{(\max_{\mathcal{R}})}$, $a_s^{(\min_{\mathcal{R}})}$, $a_l^{(\min_{\mathcal{R}})}$, $a_s^{(L_1 \mathcal{R})}$, 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).

Figure 2 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 assignments (months) and 10-40 orders of magnitude for 6-12 months. 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 5.2.2). 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 [14]). 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 (plots are omitted).

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 $w^{(b)}$ ($b \in \mathcal{R}$). 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.

Across all datasets (Figure 3 shows selected plots), $\Sigma V[a_l^{(\min_{\mathcal{R}})}]$, $\Sigma V[a_l^{(\max_{\mathcal{R}})}]$, and $\Sigma V[a_l^{(L_1 \mathcal{R})}]$ and $\Sigma V[a^{(b)}]$ for $b \in \mathcal{R}$ are within an order of magnitude. On our datasets ($n\Sigma V$ not shown), $n\Sigma V[a^{(b)}]$ and $n\Sigma V[a_l^{(\max_{\mathcal{R}})}]$ are clustered together with $kn\Sigma V \ll 1$ (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 $\Sigma V[a_\ell^{(\min_{\mathcal{R}})}] < \Sigma V[a_\ell^{(\max_{\mathcal{R}})}]$ (with larger gap when the L_1 difference is very small), $\Sigma V[a_\ell^{(L_1 \mathcal{R})}] < \Sigma V[a_\ell^{(\max_{\mathcal{R}})}]$, and $\Sigma V[a_\ell^{(\min_{\mathcal{R}})}] < \min_{b \in \mathcal{R}} \Sigma V[a^{(b)}]$. 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 6 (Lemma 6.4 and 6.5). For normalized (relative) variances, we observe the "reversed" relations $n\Sigma V[a_\ell^{(\min_{\mathcal{R}})}] > n\Sigma V[a_\ell^{(\max_{\mathcal{R}})}]$, $n\Sigma V[a_\ell^{(L_1 \mathcal{R})}] > n\Sigma V[a_\ell^{(\max_{\mathcal{R}})}]$, and $n\Sigma V[a_\ell^{(\min_{\mathcal{R}})}] > \max_{b \in \mathcal{R}} n\Sigma V[a^{(b)}]$ which are explained by smaller normalization factors for $w^{(\min_{\mathcal{R}})}$ and $w^{(L_1 \mathcal{R})}$.

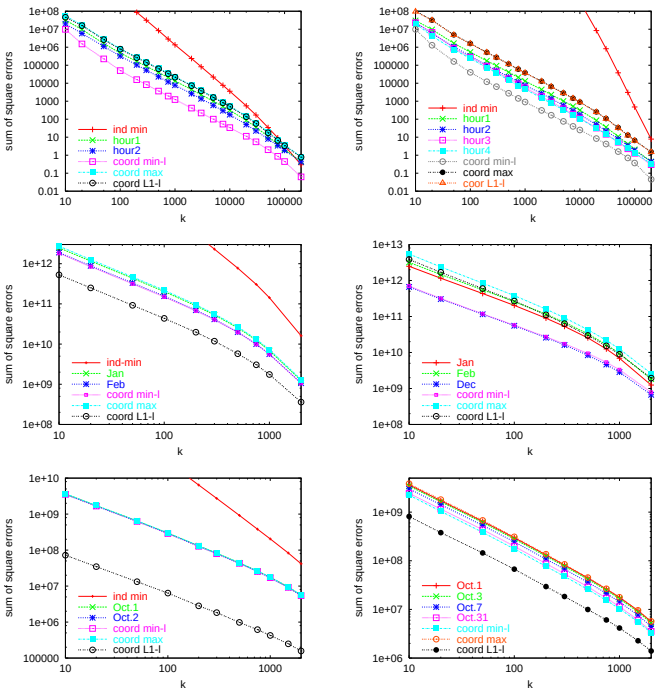


Figure 3: Top row: IP dataset2 key=4tuple weight=bytes hours= $\{1, 2\}$; IP dataset2 key=4tuple weight=bytes hours= $\{1, 2, 3, 4\}$. Middle row: Netflix data set $\mathcal{R} = \{1, 2\}$, $\mathcal{R} = \{1, \ldots, 12\}$. Bottom row: Stock dataset, high values: $\mathcal{R} = \{1, 2\}$ (October 1-2, 2008), $\mathcal{R} = \{1, \ldots, 23\}$ (all trading days in October, 2008).

S-set versus L-set estimators. To understand the advantage of the stronger l-set estimators over the s-set estimators, we studied the ratios $\Sigma V[a_s^{(\min_{\mathcal{R}})}]/\Sigma V[a_l^{(\min_{\mathcal{R}})}]$ and $\Sigma V[a_s^{(L_1 \mathcal{R})}]/\Sigma V[a_l^{(L_1 \mathcal{R})}]$ as a function of k. 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.

7.3 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 5.1, Eq. 5). $a_i^{(b)}$: the independent sketches inclusive estimator in (Section 5.1, Eq. 4). $a_p^{(b)}$: the plain order-k sketch RC estimator ([22] for IPPS ranks). Among all keys of the combined sketch this estimator uses only the keys which are part of the order-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) weights.

Inclusive versus plain estimators. The plain estimators we used are optimal for individual order-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. These ratios vary between 0.05 to 0.9 on our datasets and shows a significant benefit for inclusive estimators (see Figure 4). 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.

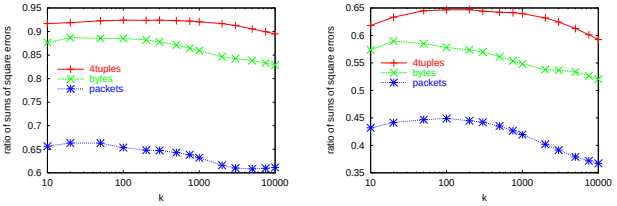


Figure 4: Inclusive versus plain estimators. IP dataset1, key=4tuple. Left: $\Sigma V[a_c^{(b)}]/\Sigma V[a_p^{(b)}]$ (coordinated sketches). Right: $\Sigma V[a_i^{(b)}]/\Sigma V[a_p^{(b)}]$ (independent sketches).

Variance versus storage. For a fixed k, the plain estimator is in fact identical for independent and coordinated order-k sketches. Independent order-k sketches, however, tend to be larger than coordinated order-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). We considered the normalized sums of variances, for inclusive and plain estimators $n\Sigma V[a_i^{(b)}]$, $n\Sigma V[a_c^{(b)}]$, $n\Sigma V[a_{p,c}^{(b)}]$, $n\Sigma V[a_{p,i}^{(b)}]$, as a function of the combined sample size (see Figure 5). For a fixed sketch size, plain estimators perform worse for independent sketches than for coordinated sketches. This happens since an independent sketch of 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.)

Sharing ratio. The *sharing ratio*, $|S|/(k * |\mathcal{W}|)$ 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 ratio measures the combined sketch size needed so that we include an order-k sketch for all weight assignments. We computed the sharing ratio for coordinated and independent order-k sketches as a function of k (see Figure 6). Coordinated sketches minimize the sharing ratio (Theorem 4.2). On our datasets, the ratio varies between 0.25-0.68 for coordinated sketches and 0.4-1 for independent sketches. The sharing ratio 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 ratio is above 0.85 for smaller values of k and can be considerably higher than with coordinated sketches. Coordinated sketches have lower (better) sharing ratio when weight assignments are more correlated.

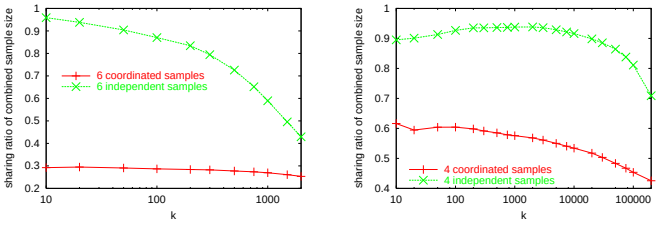


Figure 6: Sharing ratio of coordinated and independent sketches. Left: Stocks dataset (6 weight assignments). Right: IP dataset2, key=4tuple.

8. 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.

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. For colocated 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.

9. REFERENCES

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

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

[3] B. Bloom. Space/time tradeoffs in in hash coding with allowable errors. *Communications of the ACM*, 13:422–426, 1970.

[4] 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.

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

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

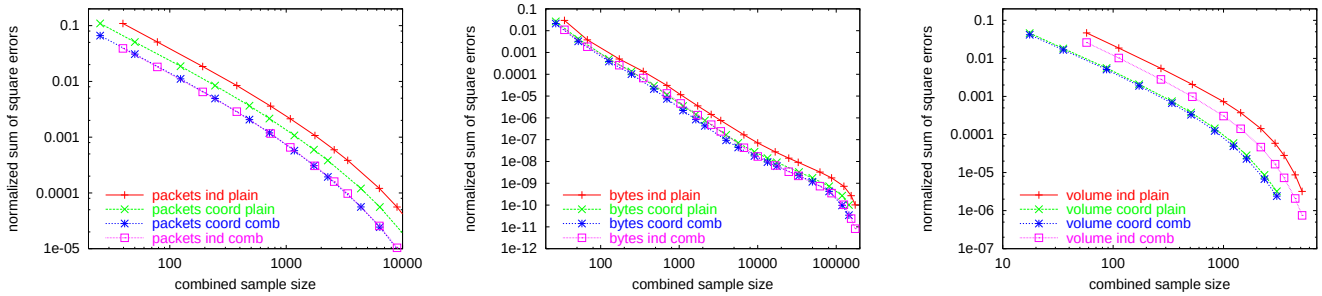


Figure 5: $n\Sigma V[a_c^{(b)}]$, $n\Sigma V[a_{p,i}^{(b)}]$, $n\Sigma V[a_{p,c}^{(b)}]$, $n\Sigma V[a_i^{(b)}]$ **as a function of combined sample size. Left: IP dataset1 key=destIP, attribute= number of packets. Middle: IP dataset2 hour3: key=destIP, attribute = number of bytes; Right: Stocks dataset, volume.**

[7] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. 34th Annual ACM Symposium on Theory of Computing*. ACM, 2002.

[8] 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.

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

[10] 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.

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

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

[13] E. Cohen and H. Kaplan. Tighter estimation using bottom-k sketches. In *Proceedings of the 34th VLDB Conference*, 2008.

[14] E. Cohen and H. Kaplan. Leveraging discarded samples for tighter estimation of multiple-set aggregates. In *Proceedings of the ACM SIGMETRICS'09 Conference*, 2009.

[15] E. Cohen, H. Kaplan, and S. Sen. Coordinated weighted sampling for estimating aggregates over multiple weight assignments. Technical Report cs.DS/0906.4560, Computing Research Repository (CoRR), 2009.

[16] J. G. Conrad and C. P. Schriber. Constructing a text corpus for inexact duplicate detection. In *SIGIR 2004*, pages 582–583, 2004.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] J. Hájek. *Sampling from a finite population*. Marcel Dekker, New York, 1981.

[30] 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.

[31] D. Knuth. *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*. Addison-Wesley, 1969.

[32] A. Kolcz, A. Chowdhury, and J. Alspector. Improved robustness of signature-based near-replica detection via lexicon randomization. In *SIGKDD 2004*, pages 605–610, 2004.

[33] 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.

[34] 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.

[35] 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.

[36] U. Manber. Finding similar files in a large file system. In *Usenix Conference*, pages 1–10, 1994.

[37] G. S. Manku, A. Jain, and A. D. Sarma. Detecting nearduplicates for web crawling. In *Proceedings of the 16th International World Wide Web Conference (WWW)*, 2007.

[38] 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.

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

[40] E. Ohlsson. Sequential poisson sampling. *J. Official Statistics*, 14(2):149–162, 1998.

[41] E. Ohlsson. Coordination of pps samples over time. In *The 2nd International Conference on Establishment Surveys*, pages 255–264. American Statistical Association, 2000.

[42] B. Rosén. Asymptotic theory for successive sampling with varying probabilities without replacement, I. *The Annals of Mathematical Statistics*, 43(2):373–397, 1972.

[43] B. Rosén. Asymptotic theory for order sampling. *J. Statistical Planning and Inference*, 62(2):135–158, 1997.

[44] F. Rusu and A. Dobra. Fast range-summable random variables for efficeint aggregate estimation. In *Proc. of the 2006 ACM SIGMOD Int. Conference on Management of Data*, pages 193–204. ACM, 1990.

[45] 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.

[46] C.-E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer, 1992.

[47] S. Schleimer, D. Wilkerson, and A. Aiken. Winnowing: local algorithms for document fingerprinting. In *Proceedings of the ACM SIGMOD*, 2003.

[48] 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.

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

[50] M. Szegedy and M. Thorup. On the variance of subset sum estimation. In *Proc. 15th ESA*, 2007.

[51] J. Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, 1985.