Distance Queries from Sampled Data: Accurate and Efficient

Edith Cohen
Microsoft Research
Mountain View, CA, USA
editco@microsoft.com

ABSTRACT

[1] Distance queries are a basic tool in data analysis. They are used for detection and localization of change for the purpose of anomaly detection, monitoring, or planning. Distance queries are particularly useful when data sets such as measurements, snapshots of a system, content, traffic matrices, and activity logs are collected repeatedly.

Random sampling, which can be efficiently performed over streamed or distributed data, is an important tool for scalable data analysis. The sample constitutes an extremely flexible summary, which naturally supports domain queries and scalable estimation of statistics, which can be specified after the sample is generated. The effectiveness of a sample as a summary, however, hinges on the estimators we have.

We derive novel estimators for estimating L_p distance from sampled data. Our estimators apply with the most common weighted sampling schemes: Poisson Probability Proportional to Size (PPS) and its fixed sample size variants. They also apply when the samples of different data sets are independent or coordinated. Our estimators are admissible (Pareto optimal in terms of variance) and have compelling properties.

We study the performance of our Manhattan and Euclidean distance ($p = 1, 2$) estimators on diverse datasets, demonstrating scalability and accuracy even when a small fraction of the data is sampled. Our work, for the first time, facilitates effective distance estimation over sampled data.

1. INTRODUCTION

Data is commonly generated or collected repeatedly, where each *instance* has the form of a value assignment to a set of *keys*: Daily summaries of the number of queries containing certain keywords, activity in a social network, transmitted bytes for IP flow keys, performance parameters (delay, throughput, or loss) for IP source destination pairs, environmental measurements for sensor locations, and requests for resources. In these examples, each set of values (instance) corresponds to a particular time or location. The universe of possible key values is fixed across instances but the values of a key are different.

One of the most basic operations in data analysis are *Distance queries*, which are used to detect, measure, and localize change [7, 26]. Their applications include anomaly detection, monitoring, and planning. Formally, a difference query between instances is specified by a meta-data based selection predicate that is applied to the keys. The result is the distance between the vectors projected on the selected keys. The simple example in Figure 1 shows two instances on 7 keys, and some queries specified on this data using Euclidean and Manhattan distances.

The collection and warehousing of massive data is subject to limitations on storage, throughput, and bandwidth. Even when the data is stored in full, exact processing of queries may be slow and resource consuming. Random sampling of datasets is widely used as a means to obtain a flexible summary over which we can query the data while meeting these limitations [27, 36, 6, 5, 8, 20, 21, 3, 17, 3, 22, 12, 19, 10, 13].

data set

key h:	a	b	c	d	e	f
$v_1(h)$	5	0	4	5	8	7
$v_2(h)$	7	10	3	0	6	7

L_p-distance for $H \subset \{a, b, c, d, e, f\}$: $L_p(H) = (L_p^p(H))^{\frac{1}{p}}$
where $L_p^p(H) = \sum_{h \in H} |v_1(h) - v_2(h)|^p$

Example Queries:

$L_1(\{a, \ldots, f\}) = 20$ $\quad L_2(\{a, \ldots, f\}) = \sqrt{134} \approx 11.6$
$L_1(\{d, e, f\}) = 7$ $\quad L_2^2(\{a, e\}) = 4$

Figure 1: Distances between two instances, Table shows a data set with two instances $i \in \{1, 2\}$ and 7 keys $\{a, \ldots, f\}$ and the values $v_i(h)$ of key h in instance i. The figure also provides example distance queries, specified for a selected set H of keys.

The sampling scheme is applied to our data, and a set of random bits, and returns a small subset of the entries. Our sample includes these entries and their values and has a fraction of the size of the original dataset. The sample only includes nonzero entries, this is particularly important when the data is sparse, meaning that the vast majority of keys have associated value 0. When the values distribution is skewed, we often apply *weighted* sampling, meaning that the probability a key is sampled depends on its value – Favoring heavier keys allows for more accurate estimates of sums and other statistics.

Perhaps the most basic sampling scheme is Poisson sampling, where keys are sampled independently. Weighting is often done using Probability Proportional to Size (PPS) sampling [23], where the inclusion probability is proportional to the value. Other common sampling schemes are bottom-k (order) samples, which have the advantage over independent sampling of yielding a sample size of exactly k. Bottom-k sampling generalizes reservoir sampling and includes Priority (Sequential Poisson) sampling and weighted sampling without replacement [33, 32, 30, 11, 19, 12, 13]. These sampling schemes are very efficient to apply also when the data is streamed or distributed.

Once we have the sample, we can quickly process approximate queries posed over the original data. This is done by applying an *estimator* to the sample. We seek estimators that provide good results when a small fraction of the data is sampled. In particular, we would like them to be *admissible*, that is, optimally use the information we have in the sample, and be efficient to compute. When estimating nonnegative quantities, such as distances, we are interested in nonnegative estimators.

[1] This is a full version of a KDD 2014 paper.

Consider the basic problem of estimating, for a given selection predicate, the sum of values of keys selected by the predicate. This problem is solved well by the classic Horvitz-Thompson (HT) [24] estimator. The estimate is the sum, over sampled keys i satisfying the predicate, of the ratio v_i/p_i, where v_i is the value and p_i is the inclusion probability of i. This *inverse-probability* estimate is clearly unbiased (if $v_i > 0 \implies p_i > 0$): if the key is not sampled, the estimate is 0 and otherwise it is v_i/p_i, giving expectation v_i. The estimate is also nonnegative when values are. Moreover, the inverse-probability estimator is a UMVUE (uniform minimum variance unbiased estimator), meaning that among all unbiased and nonnegative estimators, it minimizes variance point wise. To apply this estimator, we need the inclusion probability p_i to be available to it when v_i is. This is the case with both Poisson and bottom-k sampling, when implemented correctly.

The HT estimate is a sum estimator: conceptually, we apply an estimator to each key in the universe. Keys that are not sampled have an estimate of 0. Keys that are sampled have a positive estimate. We then sum the estimators of different keys. This makes it highly suitable for domain (selection) queries, which aggregate over a subset of keys that can be specified by a predicate.

We now turn to our problem of estimating the distance between instances from their samples. We seek an estimator with similar properties to the HT estimator: a sum estimator, unbiased and nonnegative, and with optimal use of the information in the sample.

This problem turns out to be significantly more challenging. One can attempt to apply again the HT estimator: When the outcome reveals the value of the estimated quantity, the estimate is equal to the value divided by the probability of such an outcome. The estimate is 0 otherwise. Inverse probability estimates, however, are inapplicable to distance estimation over weighted samples, since they require that there is a positive probability for an outcome which reveals the exact value of the estimated quantity: the absolute difference between the values of the key in two instances. With multiple instances and weighted sampling, keys that have zero value in one instance and positive value in another have positive contribution to the distance but because zero values are never sampled, there is zero probability for determining the value from the outcome: In the Example in Figure 1, key c has value 5 in the first instance and value 0 in the second, and thus has a contribution of 5 to the L_1 distance. The key however will never be included in a weighted sample of instance 2.

When considering multiple instances, we also need to specify how their samples relate to each other. The sampling schemes we mentioned, Poisson and bottom-k, specify the distribution for a single instance. Our first requirement, since data of different instances can be distributed or collected in different times, is that the sample of one instance can not depend on values assumed in another [16, 14]. The random bits, however, can be reused (using random hash functions), to make sampling probabilities dependent. The two extremes of the joint distribution of samples of different instances are *independent sampling* (independent sets of random bits for each instance) and *coordinated sampling* (identical sets of random bits). With coordinated sampling, which is a form of locality sensitive hashing, similar instances have similar samples whereas independent samples of identical instances can be completely disjoint. Each of these two relations has unique advantages and therefore in our work here we address both: Coordination [4, 34, 31, 33, 6, 5, 8, 20, 21, 3, 12, 22, 13, 16] allows for tighter estimates of many basic queries including distinct counts (set union) [8, 20, 21, 13], quantile sums [16], Jaccard similarity [6, 5], and more recently L_1 distance [16]. The drawbacks of coordination are that it results in unbalanced "burden" where same keys tend to be sampled across instances – an issue when, for example, being sampled translates to overhead which we would like to balance across keys. Moreover, while beneficial for some queries, the variance on some queries – notably sum queries that span values from multiple instances ("total number of search queries by Californians on Monday-Wednesday", from daily summaries) – is larger than with independent sampling – an issue if our samples are primarily used for such queries.

Contributions:

We derive unbiased nonnegative estimators for L_p^p distance queries over weighted samples. These include the two important cases of Manhattan distances L_1 and Euclidean distances L_2 (which is the square root of L_2^2). Our estimators apply with either Poisson or bottom-k sampling. We also address the two important cases where the samples of different instances are independent or coordinated. Our work facilitates, for the first time, the use of weighted samples as summaries that support distance queries.

Our estimators have several compelling properties. Similarly to the HT estimator, our estimators are *sum* estimators: We estimate L_p^p as a sum, over *the selected keys*, of nonnegative unbiased estimates of $\mathrm{RG}_p = |v_1 - v_2|^p$ of the values assumed by the key (see Figure1). Our estimators are unbiased, nonnegative, and admissible (Pareto optimal), meaning that another (nonnegative and unbiased) estimator with strictly lower variance on some data must have strictly higher variance on another data. The estimate $\widehat{\mathrm{RG}}_p$ obtained for a particular key has high variance, since most likely, the key is not sampled in any instance (in which case the estimate is 0), but unbiasedness allows for diminishing relative error when more keys are selected. The distance L_p can be estimated by the pth root of our L_p^p estimate.

Our estimates $\widehat{\mathrm{RG}}_p$ can be positive also when the samples do not reveal the respective value RG_p but only partial information on it. This property turns out to be critical for obtaining admissible estimators, what is not possible, as we mentioned above, with the inverse-probability estimate. It also means, however, that the estimators we derive have to be carefully tailored to the power p.

For independently-sampled instances, we present an estimator for RG_p of two values ($p > 0$). This derivation uses a technique we presented in [14]. Our estimator, which we call the L^* estimator, is the unique symmetric and *monotone* admissible estimator, where monotonicity means that the estimate is non-decreasing with the information we can glean from the outcome.

For coordinated samples of instances, we apply our framework of monotone sampling [9]. We derive the L* and U* estimators for the RG_p functions, which are admissible, unbiased, and nonnegative. The L* estimator is monotone and has lower variance for data with small difference (range) whereas the U* estimator performs better when the range is large. This choice is important, because it allows us to customize estimation to properties of the data set. Network traffic data, for example, is likely to have larger differences and thus the U* estimator may perform better whereas the L* estimator would be preferable when differences are smaller. The L* estimator also exhibits a compelling theoretical property of being "variance competitive" [15], meaning that for all data vectors, its variance is not too far off the minimum possible variance for the vector by a nonnegative unbiased estimator. This property makes the L* estimator a good default choice.

For $p = 1, 2$, which are the important special cases of the Manhattan and the Euclidean distances, we compute closed form expressions of estimators and their variance and also obtain tighter bounds on the "competitiveness" of the L* estimator. We evaluate and compare the performance of our L_1 and L_2^2 distance estimators on queries over diverse data sets. The queries vary in the support size (number of keys in the data set satisfying the selection

predicate) and in the relative difference (difference normalized by norm). We show that in all cases, we achieve good results when a small fraction of the data is sampled. Over coordinated samples, we examine the behavior of the L* and U* estimators and also consider the *optimally competitive* estimator, which minimizes the worst-case ratio, and we compute by a program. Finally, we provide guidelines to choosing between these estimators based on properties of the data.

Roadmap: Section 2 contains necessary background and definitions. We present difference estimators for independent samples in Section 3 and for coordinated samples in Section 4. Section 7 contains an experimental evaluation.

2. PRELIMINARIES

We denote by $v_{ih} \in \mathbb{R}_{\geq 0}$ the value of key $h \in K$ in instance $i \in [r]$ and by the vector $\boldsymbol{v}(h)$, the values of key h in all instances (the column vector v_{ih}). The *exponentiated range* of a vector $\boldsymbol{v}$ is:

$$\mathrm{RG}_p(\boldsymbol{v}) = (\max(\boldsymbol{v}) - \min(\boldsymbol{v}))^p \ \ (p > 0) \tag{1}$$

where $\max(\boldsymbol{v}) \equiv \max_i v_i$ and $\min(\boldsymbol{v}) = \min_i v_i$ are the maximum and minimum entry values of the vector $\boldsymbol{v}$. We omit the subscript when $p = 1$.

We are interested in queries which specify a selected subset $H \subset K$ of keys, through a predicate on K, and return

$$L_p^p(H) = \sum_{h \in H} \mathrm{RG}_p(\boldsymbol{v}(h)) \ . \tag{2}$$

The L_p-distance of two instances ($r = 2$) is $L_p(H) \equiv (L_p^p(H))^{1/p}$.

When data is sampled, we estimate L_p^p, by summing estimates $\hat{\mathrm{RG}}_p$ for the respective single-key primitive $\mathrm{RG}_p(\boldsymbol{v}(h))$ over keys $h \in H$. We use nonnegative unbiased estimators for the primitives, which result, from linearity of expectation, in unbiased estimates for the sums. We measure error by the coefficient of variation (CV), which is the ratio of the square root of the variance to the mean. Our estimates for each key have high variance, but when inclusions of different keys are pairwise independent, variance is additive and the CV decreases with $|H|$, allowing for accurate estimates of the sum. Finally, we can estimate $L_p(H)$ by taking the pth root of the estimate for $L_p^p(H)$. This estimate is biased, but the error is small when the CV of our $L_p^p(H)$ estimate is small.

A basic component in applying sum estimators is obtaining from basic sampling schemes of instances the respective estimation problems for a *single key*. We cast the basic sampling schemes discussed in the introduction in the following form, which facilitates separate treatment of each key.

The sampling of the entry v_{ih} is specified by a threshold value $\tau_{ih} \geq 0$, and random *seed* values $u_{ih} \sim U[0, 1]$ chosen uniformly at random.

$$h \text{ is sampled in instance } i \iff v_{ih} \geq \tau_{ih} u_{ih} \ . \tag{3}$$

The estimation problem for a single key h is then as follows. We know $\boldsymbol{\tau} = (\tau_{1h}, \ldots, \tau_{rh})$, the seeds $\boldsymbol{u} = (u_{1h}, \ldots, u_{rh})$, and the results of the sampling for h (the value v_{ih} in all instances i where h was sampled). We apply an estimator $\hat{\mathrm{RG}}_p$ to this information, which we refer to as the *outcome* S, to estimate $\mathrm{RG}_p(\boldsymbol{v})$. The availability of the seeds u_{ih} to the estimator turns out to be critical for estimation quality [14, 9]. We facilitate it by generating the seeds using random hash functions (pairwise independence between keys suffices for variance bounds). When we treat a single key h, we omit the reference to h from the notation.

Sampling scheme of instances. We now briefly return to sampling schemes of instances, and show how we obtain the single-key formulation (3) from them.

With Poisson PPS (Probability Proportional to Size) sampling of instance i, each key h is sampled with probability proportional to v_{ih}. An equivalent formulation is to use a global threshold value T_i, such that a key h is sampled if and only if $v_{ih} \geq u_{ih} T_i$. The expected sample size $\mathsf{E}[|S|] = \sum_{h \in K} \min\{1, v_{ih}/T_i\}$ is determined by T_i. The sampling can be easily implemented with respect to either a desired sample size or a fixed threshold value. The respective estimation problem (3) for key h has $\tau_{ih} \equiv T_i$. As an example, we can obtain a PPS Poisson sample of expected size $\mathsf{E}[|S|] = 3$ for the instances in Figure 1 using $T_1 = 29/3$ (instance 1) and $T_2 = 33/3 = 11$ (instance 2).

Priority (sequential Poisson) sampling [30, 19, 35] is performed by assigning each key a *priority* $r_{ih} = v_{ih}/u_{ih}$. The sample of instance i includes the k keys with largest priorities, the $(k+1)^{\text{th}}$ largest priority T_i, and the k^{th} largest priority T_i'.

To obtain the single-key formulation (3) for key h, we consider the sampling *conditioned* on fixing the seeds u_{ij} (and thus the priorities r_{ij}) for all $j \neq h$ [12, 19]. The *effective* threshold τ_{ih} is the k^{th} largest priority in $K \setminus \{h\}$: Key h is then sampled if and only if $v_{ih}/u_{ih} > \tau_{ih}$. When $h \in S$, $\tau_{ih} = T_i$ and when $h \notin S$, $\tau_{ih} = T_i'$.

When sampling several instances, we can make the samples *independent* when we use independent u_{ih} for all i. The samples are *coordinated* (shared-seed) if the same seed is used for the same key in all instances, that is, $\forall h \in K, \forall i \in [r], u_{ih} = u_{1h} \equiv u_h$.

Figure 2 shows Poisson PPS and priority samples, independent and coordinated, obtained for the two instances in Figure 1 from a random seed assignment u_{ih}. The figure also shows the threshold values τ_{ih}, which are available to the estimators $\hat{\mathrm{RG}}_p(S)$. As an example, the outcome for key 4, (the outcome is the input to the estimator), is as follows. When instances are independently Poisson sampled the outcome includes: $v_1 = 5$ (key 4 is sampled only in instance 1 so we know v_{14} exactly), $u_1 = 0.15$, $u_2 = 0.36$ (seeds available by applying hash functions to the key), and $\tau_1 = 29/3$, $\tau_2 = 11$ (with Poisson PPS sampling, the same threshold applies to all keys in each instance). With coordinated priority sampling of instances the outcome includes $v_1 = 5$, $u = 0.15$, $\tau_1 = 13.8$ (since key 4 is sampled in instance 1) and $\tau_2 = 30.4$ (since key 4 is not sampled in instance 2).

From here onward, we focus on estimating $\mathrm{RG}_p(\boldsymbol{v})$ for a single key from the respective outcome S. We return to sum aggregates only for the experiments in Section 7.

Estimators: For an outcome S, we denote by

$$\begin{aligned} S^* &= \{\boldsymbol{v} \mid S = S(\boldsymbol{u}, \boldsymbol{v})\} \\ &= \{\boldsymbol{z} \mid i \in S \implies z_i = v_i \ , \ i \notin S \implies z_i < \tau_i u_i\} \end{aligned} \tag{4}$$

the set of all data vectors consistent with S. We can equivalently define the outcome as the set S^* since it captures all the information available to the estimator on $\boldsymbol{v}$ and hence on $\mathrm{RG}_p(\boldsymbol{v})$. For our example in Figure 2, considering independent Poisson sampling and key 4, the set S^* includes all vectors $(5, x)$ such that $x < u_2 \tau_2 = 0.36 \cdot 11 = 3.96$. The actual data vector for key 4 is $(5, 0)$, but the outcome only partially reveals it.

We denote by $\mathcal{S}$ the set of all possible outcomes, that is, any outcome consistent with any data vector $\boldsymbol{v}$ in our domain. For data $\boldsymbol{v}$, we denote by $\mathcal{S}_{\boldsymbol{v}}$ the probability distribution over outcomes consistent with $\boldsymbol{v}$. As mentioned, we are seeking *nonnegative* estimators $\hat{\mathrm{RG}}_p(S) \geq 0$ for all $S \in \mathcal{S}$, since RG_p are nonnegative. Since we sum many estimates, we would like each estimate to be *unbiased*

	1	2	3	4	5	6
Instance 1	5	0	4	5	8	7
Instance 2	7	10	3	0	6	7

Independent sampling, PPS

seeds u_1	0.23	0.29	0.84	0.15	0.58	0.19
seeds u_2	0.81	0.17	0.48	0.36	0.15	0.49
v_1/u_1	21.7	0	4.8	33.3	13.8	36.8
v_2/u_2	8.6	58.8	6.25	0	41.7	14.3

Coordinated sampling, PPS

seeds u	0.23	0.29	0.84	0.15	0.58	0.19
v_1/u	21.7	0	4.8	33.3	13.8	36.8
v_2/u	30.4	34.5	3.6	0	10.3	36.8

Poisson samples $E[|S|] = 3$:

	τ_i	S
independent		
1	29/3	$\{1, 4, 5, 6\}$
2	11	$\{2, 5, 6\}$
coordinated $u \equiv u_1$		
1	29/3	$\{1, 4, 5, 6\}$
2	11	$\{1, 2, 5, 6\}$

priority samples $|S| = 3$:

	τ_{ih}: $h \in S, h \notin S$	S
independent		
1	13.8, 21.7	$\{4, 5, 6\}$
2	8.6, 14.3	$\{2, 5, 6\}$
coordinated $u \equiv u_1$		
1	13.8, 21.7	$\{4, 5, 6\}$
2	10.3, 30.4	$\{1, 2, 6\}$

Figure 2: Independent and coordinated samples of two instances. Poisson PPS samples of expected size 3 and priority samples of size 3 ($k = 3$).

$\mathsf{E}_{S\sim S_{\boldsymbol{v}}}[\hat{\mathrm{RG}}_p(S)] = \mathrm{RG}_p(\boldsymbol{v})$. We also seek *bounded variance* on all data $\boldsymbol{v}$, $\mathsf{E}_{S\sim S_{\boldsymbol{v}}}[\hat{\mathrm{RG}}_p(S)^2] < \infty$, and *admissibility* (Pareto variance optimality): there is no nonnegative unbiased estimator with same or lower variance on all data and strictly lower on some data. An intuitive property that is sometimes desirable is *monotonicity*: the estimate value is non decreasing with the information on the data that we can glean from the outcome $S^* \subset S'^* \implies \hat{\mathrm{RG}}_p(S) \geq \hat{\mathrm{RG}}_p(S')$.

When S^* includes vectors $\boldsymbol{v}$ such that $\mathrm{RG}_p(\boldsymbol{v}) = 0$, any unbiased and nonnegative estimator must have $\hat{\mathrm{RG}}_p(S) = 0$ (with probability 1). We therefore limit our attention to estimators satisfying this property.

We can also see that when the key h is not sampled in any instance, then S^* is consistent with $\mathrm{RG}_p(\boldsymbol{v}) = 0$, which means that $\hat{\mathrm{RG}}_p(S) = 0$ and we therefore do not need to explicitly compute the contribution of the vast majority of keys that are not sampled in at least one instance.

Finally, we will use the following definition of order optimality of estimators in our constructions. Given a partial order $\prec$ on the data domain an estimator $\hat{f}$ is $\prec$-optimal (respectively, $\prec^+$-optimal) if it is unbiased (resp., and nonnegative) for all data $\boldsymbol{v}$, and minimizes variance for $\boldsymbol{v}$ conditioned on the variance being minimized for all preceding vectors. Formally, if there is no other unbiased (resp., nonnegative) estimator that has strictly lower variance on some data $\boldsymbol{v}$ and at most the variance of $\hat{f}$ on all vectors that precede $\boldsymbol{v}$. An order optimal estimator is admissible.

3. INDEPENDENT PPS SAMPLING

We derive estimators for $\mathrm{RG}_p(\boldsymbol{v})$, where $\boldsymbol{v} = (v_1, v_2)$ (two instances $r = 2$). The estimation scheme is specified by $\boldsymbol{\tau} = (\tau_1, \tau_2)$. The outcome $S(\boldsymbol{u}, \boldsymbol{v})$ is determined by the data vector $\boldsymbol{v} = (v_1, v_2)$ and $\boldsymbol{u} = (u_1, u_2)$, where $u_i \sim U[0, 1]$ are independent. The set S^* of vectors consistent with S is (4).

We derive the L* estimator, $\hat{\mathrm{RG}}_p^{(L)}$, which is the unique symmetric, monotone, and admissible estimator. The construction adapts a framework from [14], which was used to estimate $\max\{v_1, v_2\}$: We specify an order $\prec$ on the data domain. We then formulate a set of sufficient constraints for an unbiased symmetric and order-optimal estimator $\hat{f}^{(\prec)}$ of $\mathrm{RG}_p(\boldsymbol{v})$. The constraints, however, do not incorporate nonnegativity, as this results in much more complex dependencies. But if we find a nonnegative solution $\hat{f}^{(\prec)}$, then we find an estimator with all the desired properties. We therefore hope for a good "guess" of $\prec$.

We work with $\prec$ that prioritizes smaller distances, that is, $\boldsymbol{v} \prec \boldsymbol{z}$ if and only if $\mathrm{RG}(\boldsymbol{v}) < \mathrm{RG}(\boldsymbol{z})$. With each outcome $S \in \mathcal{S}$, we associate its *determining vector* $\phi(S)$, which we define as the $\prec$-minimal vector in (the closure of) S^*. The closure is the set obtained when using a non-strict inequality in (4). The determining vector is unique for all outcomes S that are not consistent with $\mathrm{RG}_p(\boldsymbol{v}) = 0$, that is, $\mathrm{RG}_p(\boldsymbol{v}) > 0$ for all $\boldsymbol{v} \in S^*$. As we mentioned earlier, we only need to specify the estimator on these outcomes, since we only consider estimators that are 0 on outcomes consistent with $\mathrm{RG}_p(\boldsymbol{v}) = 0$. The mapping of outcomes S to the determining vector $\phi(S)$ is shown in Table 1 (Bottom).

We now formulate sufficient constraints for $\prec$-optimality. Conditioned on fixing the estimator on outcomes S such that $\phi(S) \prec \boldsymbol{v}$, the "best" we can do, in terms of minimizing variance, is to set it to a fixed value on all outcomes such that $\phi(S) = \boldsymbol{v}$. This fixed value is determined by the unbiasedness requirement. Since $\hat{f}^{(\prec)}$ is the same for all outcomes with same determining vector, we specify it as a function of the determining vector $\hat{f}^{(\prec)}(S) \equiv \hat{f}^{(\prec)}(\phi(S))$.

We use the notation $\mathcal{S}_0(\boldsymbol{v})$ for the set of outcomes S that are consistent with $\boldsymbol{v}$ but also consistent with a vector that precedes $\boldsymbol{v}$:

$$\mathcal{S}_0(\boldsymbol{v}) \quad = \quad \{S | \boldsymbol{v} \in S^* \wedge \phi(S) \prec \boldsymbol{v}\}$$

The contribution of the outcomes $\mathcal{S}_0(\boldsymbol{v})$ to the expectation of $\hat{f}^{(\prec)}$ when data is $\boldsymbol{v}$ is

$$f_0(\boldsymbol{v}) = \mathsf{E}_{S\sim\mathcal{S}_{\boldsymbol{v}}}[I_{S\in\mathcal{S}_0(\boldsymbol{v})}\hat{f}^{(\prec)}(S)]\,,$$

where I is the indicator function. We obtain the following sufficient constraints for a $\prec$-optimal unbiased estimator, that may not be nonnegative, but is forced to be 0 on all outcomes consistent with $\mathrm{RG}_p(\boldsymbol{v}) = 0$. For all $\boldsymbol{v}$,

$$\mathrm{PR}_{S\sim\mathcal{S}_{\boldsymbol{v}}}[\phi(S) = \boldsymbol{v}] = 0 \quad \implies \quad f_0(\boldsymbol{v}) \equiv f(\boldsymbol{v}) \tag{5}$$

$$\mathrm{PR}_{S\sim\mathcal{S}_{\boldsymbol{v}}}[\phi(S) = \boldsymbol{v}] > 0 \quad \implies \quad \hat{f}^{(\prec)}(\boldsymbol{v}) = \frac{f(\boldsymbol{v}) - f_0(\boldsymbol{v})}{\mathrm{PR}_{S\sim\mathcal{S}_{\boldsymbol{v}}}[\phi(S) = \boldsymbol{v}]} \tag{6}$$

$\boldsymbol{\phi} = (\phi_1, \phi_2)$	$\hat{\mathrm{RG}}_p^{(L)}(\boldsymbol{\phi})$
$\boldsymbol{\phi} = (0, 0)$	0
$\phi_1 \geq \phi_2 > \tau_2$	$\frac{\tau_1}{\min\{\tau_1,\phi_1\}}(\phi_1 - \phi_2)^p$
$\phi_1 \geq \phi_2 \leq \tau_2$	$\frac{p\tau_1\tau_2}{\min\{\phi_1,\tau_1\}} \int_{\max\{0,\phi_1-\tau_2\}}^{\phi_1-\phi_2} \frac{y^{p-1}}{\phi_1-y} dy +$ $+ \frac{\tau_1 \max\{0,\phi_1-\tau_2\}^p}{\min\{\phi_1,\tau_1\}}$

outcome S		$\phi(S)_1$	$\phi(S)_2$
$S = \emptyset$	:	0	0
$S = \{1\}$	:	v_1	$\min\{u_2\tau_2, v_1\}$
$S = \{2\}$	:	$\min\{u_1\tau_1, v_2\}$	v_2
$S = \{1, 2\}$	:	v_1	v_2

Table 1: Top: Estimator $\hat{\mathrm{RG}}_p^{(L)}$ for $p > 0$ over independent samples, stated as a function of the determining vector $\boldsymbol{\phi} = (\phi_1, \phi_2)$ when $\phi_1 \geq \phi_2$ (case $\phi_2 > \phi_1$ is symmetric). Bottom: mapping of outcomes to determining vectors.

We derive $\hat{\text{RG}}_p^{(L)}$ by solving the right hand side of (6) for all $\boldsymbol{v}$ such that $\text{PR}_{S\sim\mathcal{S}_{\boldsymbol{v}}}[\phi(S)=\boldsymbol{v}]>0$. The solution $\hat{\text{RG}}_p^{(L)}(\boldsymbol{\phi})$ $(p>0)$ is provided in Table 1 through a mapping of determining vectors to estimate values. The estimator is specified for $\phi_1 \geq \phi_2$, as the other case is symmetric. We can verify that for all $p>0$, the estimator $\hat{\text{RG}}_p^{(L)}$ is nonnegative, monotone (for all y, $\hat{\text{RG}}_p^{(L)}(y,x)$ is non-increasing for $x\in(0,y]$) and has finite variances (follow from $\int_0^y \hat{\text{RG}}_p^{(L)}(y,x)^2 dx<\infty$). We can also verify that condition (5) holds. Vectors $\boldsymbol{v}$ with $\text{PR}_{S\sim\mathcal{S}_{\boldsymbol{v}}}[\phi(S)=\boldsymbol{v}]=0$ are exactly those with one positive and one zero entry. We can verify that (5) is satisfied, that is, $\mathsf{E}_{S\sim\mathcal{S}_{\boldsymbol{v}}}\hat{\text{RG}}_p^{(L)}(S)=\text{RG}_p(\boldsymbol{v})$ on these vectors. Table 2 shows explicit expressions of $\hat{\text{RG}}^{(L)}$ and $\hat{\text{RG}}_2^{(L)}$.

$\boldsymbol{\phi}=(\phi_1,\phi_2)$	$\hat{\text{RG}}^{(L)}(\boldsymbol{\phi})$
$\boldsymbol{\phi}=(0,0)$	0
$\phi_1\geq\phi_2>\tau_2$	$\frac{\tau_1}{\min\{\tau_1,\phi_1\}}(\phi_1-\phi_2)$
$\phi_1\geq\phi_2\leq\tau_2$	$\frac{\tau_1\tau_2}{\min\{\tau_1,\phi_1\}}\ln\left(\frac{\min\{\phi_1,\tau_2\}}{\phi_2}\right)+\frac{\tau_1\max\{0,\phi_1-\tau_2\}}{\min\{\phi_1,\tau_1\}}$

$\boldsymbol{\phi}=(\phi_1,\phi_2)$	$\hat{\text{RG}}_2^{(L)}(\boldsymbol{\phi})$
$\boldsymbol{\phi}=(0,0)$	0
$\phi_1\geq\phi_2>\tau_2$	$\frac{\tau_1}{\min\{\tau_1,\phi_1\}}(\phi_1-\phi_2)^2$
$\phi_1\geq\phi_2\leq\tau_2$	$\frac{2\tau_1\tau_2}{\min\{\tau_1,\phi_1\}}\left(\phi_2-\min\{\phi_1,\tau_2\}+\phi_1\ln\frac{\min\{\phi_1,\tau_2\}}{\phi_2}\right)$ $+\frac{\tau_1\max\{0,\phi_1-\tau_2\}^2}{\min\{\phi_1,\tau_1\}}$

Table 2: Explicit form of estimators $\hat{\text{RG}}^{(L)}$ and $\hat{\text{RG}}_2^{(L)}$ for $r=2$ over independent samples. Estimator is stated as a function of the determining vector (ϕ_1,ϕ_2) when $\phi_1\geq\phi_2$ (case $\phi_2\geq\phi_1$ is symmetric).

We now provide the derivation. We consider vectors $\boldsymbol{v}$ in increasing $\prec$ order and solve (6) for $f^{(\prec)}$ on outcomes with determining vector $\boldsymbol{v}=(v,v-\Delta)$, where $v\geq\Delta\geq 0$.

• **Case:** $v-\Delta\geq\tau_2$. The outcomes always reveals the second entry. The determining vector is $\boldsymbol{v}$ when $u_1\tau_1\leq v$, which happens with probability $\min\{1,v/\tau_1\}$. Otherwise, the outcome is consistent with $(v-\Delta,v-\Delta)$ and the estimate is 0. We solve the equality $\Delta^p=\min\{1,v/\tau_1\}\hat{\text{RG}}_p^{(L)}$, obtaining

$$\hat{\text{RG}}_p^{(L)}(v,v-\Delta)=\frac{\tau_1}{\min\{v,\tau_1\}}\Delta^p\,. \tag{7}$$

• **Case:** $v-\Delta<\tau_2$. The determining vector is $(v,v-\Delta)$ with probability $\frac{\min\{v,\tau_1\}}{\tau_1}\frac{v-\Delta}{\tau_2}$. Otherwise, it is $(v,v-y)$ for some $y<\Delta$. We use (6) to obtain an integral equation:

$$\begin{aligned}\Delta^p &= \frac{\min\{v,\tau_1\}}{\tau_1}\frac{v-\Delta}{\tau_2}\hat{\text{RG}}_p^{(L)}(v,v-\Delta)+\\ &\quad+\frac{\min\{v,\tau_1\}}{\tau_1\tau_2}\int_{\max\{0,v-\tau_2\}}^{\Delta}\hat{\text{RG}}_p^{(L)}(v,v-y)dy\end{aligned}$$

Taking a partial derivative with respect to Δ, we obtain

$$\frac{\partial\hat{\text{RG}}_p^{(L)}(v,v-\Delta)}{\partial\Delta}=\frac{p\tau_1\tau_2}{\min\{v,\tau_1\}}\frac{\Delta^{p-1}}{v-\Delta}$$

We use the boundary value for $\Delta=\max\{0,v-\tau_2\}$:

$$\hat{\text{RG}}_p^{(L)}(v,\min\{v,\tau_2\})=\frac{\tau_1}{\min\{v,\tau_1\}}\max\{0,v-\tau_2\}^p\,,$$

and obtain the solution

$$\begin{aligned}&\hat{\text{RG}}_p^{(L)}(v,v-\Delta)= \\ &\frac{p\tau_1\tau_2}{\min\{v,\tau_1\}}\int_{\max\{0,v-\tau_2\}}^{\Delta}\frac{y^{p-1}}{v-y}dy+\frac{\tau_1\max\{0,v-\tau_2\}^p}{\min\{v,\tau_1\}}\end{aligned} \tag{8}$$

The special case $\tau_1=\tau_2=\tau$: The estimators $\hat{\text{RG}}^{(L)}$ and $\hat{\text{RG}}_2^{(L)}$ as a function of the determining vector and their variance are provided in Tables 3 and 4. For data vectors where $v_1\geq v_2\geq\tau$, $\hat{\text{RG}}^{(L)}=v_1-v_2$ and $\text{VAR}[\hat{\text{RG}}^{(L)}]=0$. If $v_1\geq\tau\geq v_2$, $\hat{\text{RG}}^{(L)}=\tau\ln\frac{\tau}{v_2}+v_1-\tau$, and $\text{VAR}[\hat{\text{RG}}^{(L)}]=-2\tau v_2\ln(\frac{\tau}{v_2})-v_2^2+(\tau)^2$. Finally, if $v_2\leq v_1\leq\tau$, $\hat{\text{RG}}^{(L)}(v_1,v_2)=\frac{(\tau)^2}{v_1}\ln\frac{v_1}{v_2}$ and

$$\begin{aligned}&\text{VAR}_{\mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}^{(L)}]= \\ &= \frac{v_1v_2}{(\tau)^2}(\frac{(\tau)^2}{v_1}\ln\frac{v_1}{v_2}-(v_1-v_2))^2+\\ &\quad+(1-\frac{v_1^2}{(\tau)^2})(v_1-v_2)^2+\\ &\quad+\frac{v_1}{(\tau)^2}\int_{v_2}^{v_1}(\frac{(\tau)^2}{v_1}\ln\frac{v_1}{y}-(v_1-v_2))^2dy\\ &= 2(\tau)^2(1-\frac{v_2}{v_1}\ln\frac{v_1}{v_2}-\frac{v_2}{v_1})-(v_1-v_2)^2\,.\end{aligned}$$

Determining vector $\phi_1\geq\phi_2$	$\hat{\text{RG}}^{(L)}(\boldsymbol{\phi})$
$\phi_1\geq\phi_2\geq\tau$	$\phi_1-\phi_2$
$\phi_1\geq\tau\geq\phi_2$	$\tau\ln\frac{\tau}{\phi_2}+\phi_1-\tau$
$\phi_2\leq\phi_1\leq\tau$	$\frac{(\tau)^2}{\phi_1}\ln\frac{\phi_1}{\phi_2}$

Data $v_1\geq v_2$	$\text{VAR}_{\mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}^{(L)}]$
$v_1\geq v_2\geq\tau$	0
$v_1\geq\tau\geq v_2$	$-2\tau v_2\ln(\frac{\tau}{v_2})-v_2^2+\tau^2$
$v_2\leq v_1\leq\tau$	$2\tau^2(1-\frac{v_2}{v_1}\ln\frac{v_1}{v_2}-\frac{v_2}{v_1})-(v_1-v_2)^2$

Table 3: $\hat{\text{RG}}^{(L)}$ and its variance for independent samples.

Similarly, for $v_2\leq v_1\leq\tau$, $\hat{\text{RG}}_2^{(L)}(v_1,v_2)=2(\tau)^2(\ln\frac{v_1}{v_2}-\frac{v_1-v_2}{v_1})$.

4. SHARED-SEED SAMPLING

We derive estimators for $\text{RG}_p(\boldsymbol{v})(p>0)$, where $\boldsymbol{v}=(v_1,\ldots,v_r)$ for $r\geq 2$. The sampling uses the same random seed u for all entries. The outcome $S(u,\boldsymbol{v})$ is determined by the data $\boldsymbol{v}$ and a *scalar* seed value $u\in(0,1]$, drawn uniformly at random: Entry i is included in S if and only if $v_i\geq\tau_i u$.

We apply our work on estimators for monotone sampling [9] to derive two unbiased nonnegative admissible estimators: The L* estimator $\hat{\text{RG}}_p^{(L)}$ and the U* estimator $\hat{\text{RG}}_p^{(U)}$. We present closed form expressions of estimators and variances when $\boldsymbol{\tau}$ has all entries equal (to the scalar τ). The derivations easily extend to non-uniform $\boldsymbol{\tau}$.

condition	$\hat{\text{RG}}_2^{(L)}(\boldsymbol{\phi})$
$\phi_1\geq\phi_2\geq\tau$	$(\phi_1-\phi_2)^2$
$\phi_1\geq\tau\geq\phi_2$	$\phi_1^2-(\tau)^2-2\tau(\phi_1-\phi_2)+2\tau\phi_1\ln\frac{\tau}{\phi_2}$
$\phi_2<\phi_1\leq\tau$	$2(\tau)^2(\ln\frac{\phi_1}{\phi_2}-\frac{\phi_1-\phi_2}{\phi_1})$

Data $v_1\geq v_2$	$\text{VAR}_{\mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}_2^{(L)}]$
$v_1\geq v_2\geq\tau$	0
$v_1\geq\tau\geq v_2$	$-4v_1v_2\tau(2v_1-v_2)\ln\frac{\tau}{v_2}$ $+4v_1v_2(\tau)^2+\frac{(\tau)^4}{3}+\frac{8v_2^3\tau}{3}$ $-6v_1v_2^2\tau-4v_1^2v_2^2$ $-v_2^4+4v_1v_2^3+4v_1^2(\tau)^2-2v_1(\tau)^3$
$v_2\leq v_1\leq\tau$	$\frac{2(\tau)^2}{3v_1}\left(4v_2^3+5v_1^3-9v_1v_2^2\right)-(v_1-v_2)^4$ $-4(\tau)^2(2v_1-v_2)v_2\ln(\frac{v_1}{v_2})$

Table 4: $\hat{\text{RG}}_2^{(L)}$ and its variance for independent samples.

The set of data vectors consistent with outcome $S(u, \boldsymbol{v})$ is

$$S^* = \{\boldsymbol{z} | \forall i \in [r], i \in S \implies z_i = v_i \,,\, i \notin S \implies z_i < \tau_i u\} \,.$$

Observe that the sets $S^*(u, \boldsymbol{z})$ are the same for all consistent data vectors $\boldsymbol{z} \in S^*(u, \boldsymbol{v})$. Fixing the data $\boldsymbol{v}$, the set $S^*(u, \boldsymbol{v})$ is non-decreasing with u, which means that the information on the data that we can glean from the outcome can only increase when u decreases. This makes the sampling scheme *monotone* in the randomization, which allows us to apply the estimator derivations in [9].

The lower bound function. The derivations use the *lower bound function* $\underline{\text{RG}}_p$, which maps an outcome S to the infimum of RG_p values on vectors that are consistent with the outcome:

$$\underline{\text{RG}}_p(S) = \inf_{\boldsymbol{v} \in S^*} \text{RG}_p(\boldsymbol{v}) \,.$$

For RG, the lower bound is the difference between a lower bound on the maximum entry and an upper bound on the minimum entry.

$$\underline{\text{RG}}(S) \quad = \quad \max_{i \in S} v_i - \min\{\min_{i \in S} v_i, \min_{i \notin S} \tau_i u\} \,.$$

The lower bound on RG_p is the pth power of the respective bound on RG, that is, $\underline{\text{RG}}_p(S) = \underline{\text{RG}}(S)^p$. For $S(u, \boldsymbol{v})$, we use the notation $\underline{\text{RG}}_p(S(u, \boldsymbol{v})) \equiv \underline{\text{RG}}_p(u, \boldsymbol{v})$. For all-entries-equal $\boldsymbol{\tau}$:

condition	$\lvert S\rvert$	$\underline{\text{RG}}(S)$
$u > \frac{\max(\boldsymbol{v})}{\tau}$	0	0
$\frac{\max(\boldsymbol{v})}{\tau} \geq u \geq \frac{\min(\boldsymbol{v})}{\tau}$	$1 \ldots r-1$	$\max(\boldsymbol{v}) - u\tau$
$u < \frac{\min(\boldsymbol{v})}{\tau}$	r	$\text{RG}(\boldsymbol{v})$

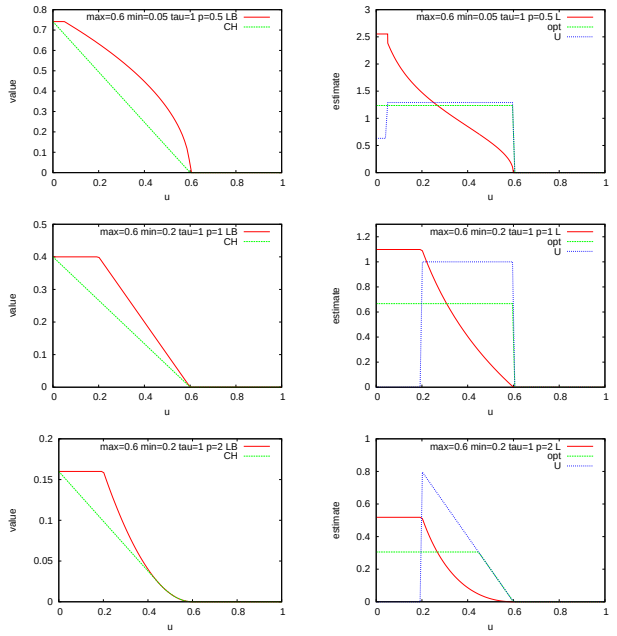


Figure 3: Left: The lower bound function and corresponding lower hull for example vectors and $p \in \{0.5, 1, 2\}$. Right: the corresponding optimal, L*, and U* estimates on outcomes consistent with the vector.

The L* estimator. The L* estimator [9] is defined for any monotone estimation problem for which an unbiased and nonnegative estimator exists. This estimator is specified as a function of the corresponding lower bound function. For RG_p we have

$$\forall S(\zeta, \boldsymbol{v}),\ \hat{\text{RG}}_p^{(L)}(S) = \frac{\underline{\text{RG}}_p(\zeta, \boldsymbol{v})}{\zeta} - \int_\zeta^1 \frac{\underline{\text{RG}}_p(u, \boldsymbol{v})}{u^2} du \,. \quad (9)$$

From [9], we know that $\hat{\text{RG}}_p^{(L)}$ has the following properties:

- It is nonnegative and unbiased.
- It is the unique (up to equivalence) admissible unbiased nonnegative monotone estimator, meaning that the estimate is non-decreasing with u.
- It is $\prec^+$-optimal with respect to the partial order $\prec$

$$\boldsymbol{v} \prec \boldsymbol{z} \iff \text{RG}_p(\boldsymbol{v}) < \text{RG}_p(\boldsymbol{z}) \,.$$

$\prec^+$-optimality with respect to this particular order means that any estimator with a strictly lower variance for a data vector must have strictly higher variance on some vector with a smaller range – this means that the L* estimator "prioritizes" data where the range (or difference when aggregated) is small.

The L* estimator has finite variances when the monotone estimation problem admits a nonnegative estimator with finite variances. It is also 4-*competitive* in terms of variance [15], meaning that for any data vector, the ratio of the expectation of the square to the minimum one possible for the data via an unbiased nonnegative estimator is at most 4

$$\forall \boldsymbol{v}, \mathsf{E}_{S \sim \mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}_p^{(L)}(S)^2] \leq 4 \mathsf{E}_{S \sim \mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}_p^{(\boldsymbol{v})}(S)^2] \,,$$

where $\hat{\text{RG}}_p^{(\boldsymbol{v})}(S)$ is a nonnegative unbiased estimator which minimizes the variance for $\boldsymbol{v}$ (We will present a construction of this estimator). Competitiveness, is a strong property that means that for *all* data vectors, the variance under the L* estimator is not too far off the minimum possible variance for that vector by a nonnegative unbiased estimator.

For our sampling scheme with all entries equal τ, we define $\max(\boldsymbol{v}) \equiv \max_i v_i$ (which is available from S whenever $|S| > 0$) and $v_{\min} = \min(\boldsymbol{v})$ if $|S| = r$ and $v_{\min} = u\tau$ otherwise, which is also always available from S, the estimator is

$$\hat{\text{RG}}_p^{(L)}(S) = \begin{cases} |S| = 0: & 0 \\ |S| \geq 1: & (\max(\boldsymbol{v}) - v_{\min})^p \max\{1, \frac{\tau}{v_{\min}}\} - \\ & \int_{\min\{1, \frac{v_{\min}}{\tau}\}}^{\min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}} \frac{(\max(\boldsymbol{v}) - x\tau)^p}{x^2} dx \end{cases} \quad (10)$$

Estimators and variance for RG and RG_2 are provided in Tables 6 and 7. We also compute a tight ratio on variance competitiveness for $p = 1, 2$:

LEMMA 4.1.

$$\forall \boldsymbol{v},\ \mathsf{E}_{S \sim \mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}^{(L)}(S)^2] \leq 2\mathsf{E}_{S \sim \mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}^{(\boldsymbol{v})}(S)^2] \quad (11)$$

$$\forall \boldsymbol{v},\ \mathsf{E}_{S \sim \mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}_2^{(L)}(S)^2] \leq 2.5\mathsf{E}_{S \sim \mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}_2^{(\boldsymbol{v})}(S)^2] \quad (12)$$

The U* estimator. The U* estimator [9] is the solution of the integral equation

$$\forall S(\zeta, \boldsymbol{v}),\ \hat{\text{RG}}_p(\zeta, \boldsymbol{v}) = \sup_{\boldsymbol{z} \in S^*} \inf_{0 \leq \eta < \zeta} \frac{\underline{\text{RG}}_p(\eta, \boldsymbol{z}) - \int_\zeta^1 \hat{\text{RG}}_p(u, \boldsymbol{v}) du}{\zeta - \eta}$$

From [9], we know that $\hat{\text{RG}}_p^{(U)}$ has the following properties:

- It is nonnegative and unbiased.
- It is $\prec^+$-optimal with respect to the partial order $\prec$

$$\boldsymbol{v} \prec \boldsymbol{z} \iff \text{RG}_p(\boldsymbol{v}) > \text{RG}_p(\boldsymbol{z}) \,.$$

This means that the U* estimator "prioritizes" data where the range (or difference when aggregated) is large. In particular, it is the nonnegative unbiased estimator with minimum variance on data with $\min(\boldsymbol{v}) = 0$.

The solution for all-entries equal τ is provided as Algorithm 1 (see Appendix A for calculation). The estimator is admissible and has finite variances for all data vectors.

The estimator $\hat{\text{RG}}_p^{(U)}$ and its variance for $p = 1, 2$ are provided in Tables 8 and 9 (See Appendix B for details).

$\boldsymbol{v}$-optimality. We say that an estimator is *$\boldsymbol{v}$-optimal* (for data $\boldsymbol{v}$), if amongst all estimators that are nonnegative and unbiased for all data, it has the minimum possible variance when the data is $\boldsymbol{v}$. In order to measure the competitiveness of our estimators, as in Lemma 4.1, we derive an expression for the *$\boldsymbol{v}$-optimal estimate values* $\hat{\text{RG}}_p^{(\boldsymbol{v})}$. It turns out that the values assumed by a $\boldsymbol{v}$-optimal estimator on outcomes consistent with $\boldsymbol{v}$ are unique (almost everywhere on $\mathcal{S}_{\boldsymbol{v}}$ [15]. Note that there is no single estimator that is $\boldsymbol{v}$-optimal for all $\boldsymbol{v}$, that is, there is no uniform minimum variance unbiased nonnegative estimator for RG_p. Therefore, the $\boldsymbol{v}$-optimal estimates for all possible values of $\boldsymbol{v}$ can not be combined into a single estimator.

We now obtain an explicit representation of $\hat{\text{RG}}_p^{(\boldsymbol{v})}$. We use the notation $H^{(\boldsymbol{v})}_{\text{RG}_p}(u)$ for the lower boundary of the convex hull (lower hull) of $\underline{\text{RG}}_p(u, \boldsymbol{v})$ and the point $(1, 0)$. This function is monotone non-increasing in u and therefore differentiable almost everywhere. We apply the following

THEOREM 4.1. *[15] A nonnegative unbiased estimator $\hat{\text{RG}}_p$ minimizes* $\text{VAR}_{S\sim\mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}_p] \iff$ almost everywhere *on* $\mathcal{S}_{\boldsymbol{v}}$

$$\hat{\text{RG}}_p^{(\boldsymbol{v})}(u) = -\frac{dH^{(\boldsymbol{v})}_{\text{RG}_p}(u)}{du} \ . \tag{13}$$

The estimates (13) are monotone non-increasing in u.

We can now specify $\hat{\text{RG}}_p^{(\boldsymbol{v})}$ for PPS sampling with all-entries-equal τ. The function $\underline{\text{RG}}_p(u, \boldsymbol{v})$ is $\max\{0, \max(\boldsymbol{v}) - \tau\}$ for $u \geq \frac{\max(\boldsymbol{v})}{\tau}$ and equal to $\text{RG}_p(\boldsymbol{v})$ for $u \leq \frac{\min(\boldsymbol{v})}{\tau}$. Therefore for $u \geq \frac{\max(\boldsymbol{v})}{\tau}$, the lower hull is $H^{(\boldsymbol{v})}_{\text{RG}_p}(u) = 0$ and $\hat{\text{RG}}_p^{(\boldsymbol{v})}(u) = 0$.

For $p \leq 1$, the function is concave for $u \in [\frac{\min(\boldsymbol{v})}{\tau}, \frac{\max(\boldsymbol{v})}{\tau}]$. The lower hull is therefore a linear function for $u \leq \frac{\max(\boldsymbol{v})}{\tau}$: when $\max(\boldsymbol{v}) \leq \tau$, $H^{(\boldsymbol{v})}_{\text{RG}_p}(u) = \text{RG}_p(\boldsymbol{v})(1 - u\frac{\tau}{\max(\boldsymbol{v})})$ and when $\max(\boldsymbol{v}) \geq \tau$, $H^{(\boldsymbol{v})}_{\text{RG}_p}(u) = \text{RG}_p(\boldsymbol{v}) - u(\text{RG}_p(\boldsymbol{v}) - (\max(\boldsymbol{v}) - \tau)^p)$. The $\boldsymbol{v}$-optimal estimates are therefore constant for $u \leq \min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}$: $\hat{\text{RG}}_p^{(\boldsymbol{v})}(u) = \text{RG}(\boldsymbol{v})\frac{\tau}{\max(\boldsymbol{v})}$ when $\max(\boldsymbol{v}) \leq \tau$, and $\hat{\text{RG}}_p^{(\boldsymbol{v})}(u) = \text{RG}_p(\boldsymbol{v}) - (\max(\boldsymbol{v}) - \tau)^p$ when $\max(\boldsymbol{v}) \geq \tau$.

For $p > 1$, $\underline{\text{RG}}_p(u, \boldsymbol{v})$ is convex for $u \in [\frac{\min(\boldsymbol{v})}{\tau}, \frac{\max(\boldsymbol{v})}{\tau}]$. Geometrically, the lower hull follows the lower bound function for $u > \alpha$, where α is the point where the slope of the lower bound function is equal to the slope of a line segment connecting the current point to the point $(0, \text{RG}_p(\boldsymbol{v}))$. For $u \leq \alpha$, the lower hull follows this line segment and is linear. Formally, the point α is the solution of

$$\text{RG}_p(\boldsymbol{v}) = (\max(\boldsymbol{v}) - x\tau)^{p-1}(p\tau + \max(\boldsymbol{v}) - x\tau) \ .$$

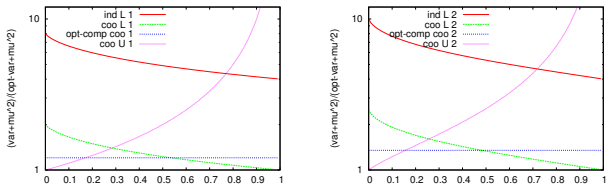
If there is no solution $\alpha \in [\frac{\min(\boldsymbol{v})}{\tau}, \min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}]$, we use $\alpha = \min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}$. The estimates for $u \in [\alpha, \min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}]$ are $\hat{\text{RG}}_p(u, \boldsymbol{v}) = -\frac{d\underline{\text{RG}}_p(u,\boldsymbol{v})}{du} = p\tau(\max(\boldsymbol{v}) - u\tau)^{p-1}$ and for $u \leq \alpha$, $\hat{\text{RG}}_p^{(\boldsymbol{v})}(u) = \frac{\text{RG}_p(\boldsymbol{v}) - (\max(\boldsymbol{v}) - \alpha\tau)^p}{\alpha}$.

Figure 3 (top) illustrates $\underline{\text{RG}}_p(u, \boldsymbol{v})$ and the corresponding lower hull $H^{(\boldsymbol{v})}_{\text{RG}_p}$ as a function of u for example vectors with $p \in \{0.5, 1, 2\}$.

Now that we expressed $\hat{\text{RG}}_p^{(\boldsymbol{v})}(S)$ on all outcomes consistent with $\boldsymbol{v}$, we can compute (for any vector $\boldsymbol{v}$), the minimum possible variance attainable for it by an unbiased nonnegative estimator:

$$\text{VAR}_{\mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}_p^{(\boldsymbol{v})}] = \int_0^1 \hat{\text{RG}}_p^{(\boldsymbol{v})}(u)^2 du - \text{RG}_p(\boldsymbol{v})^2 \ . \tag{14}$$

We use the expectation of the square to measure the "variance competitiveness" of estimators (Since the second summand in (14) is the same for all estimators). The ratio for a particular $\boldsymbol{v}$ is the expectation of the square for $\boldsymbol{v}$ divided by the optimal one which is $\mathsf{E}_{S\sim\mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}_p^{(\boldsymbol{v})}(S)^2] = \int_0^1 \hat{\text{RG}}_p^{(\boldsymbol{v})}(u)^2 du$. The competitive ratio is the maximum ratio over all $\boldsymbol{v}$.



$p = 1, \frac{\max(\boldsymbol{v})}{\tau} = 0.25$ $\qquad$ $p = 2, \frac{\max(\boldsymbol{v})}{\tau} = 0.25$

Figure 5: Ratio of the expectation of the square to the minimum possible expectation of the square for the data point (over shared-seed samples), as a function of the ratio $\min(\boldsymbol{v})/\max(\boldsymbol{v})$. Estimator $\hat{\text{RG}}_p^{(L)}$ over independent samples, and $\hat{\text{RG}}_p^{(L)}$, $\hat{\text{RG}}_p^{(U)}$, and the optimally competitive estimator over shared-seed samples. Sampling with all-entries equal τ.

The optimally competitive (OC) estimator. We used a program to compute the estimator with minimum competitive ratio. The estimator was computed for an outcome that revealed $\max(\boldsymbol{v})$ and provided an upper bound $x < \max(\boldsymbol{v})$ on $\min(\boldsymbol{v})$. The domain was discretized and the estimates were computed iteratively for decreasing x so that the estimates satisfy a certain ratio c. We then performed a search to find the minimum c for which the computation is successful.

Choosing between the L*, U*, and OC estimators. Figure 3 shows the $\boldsymbol{v}$-optimal estimates and the L* and U* estimators for example vectors, illustrating the monotonicity of L* and how the estimators relate to each other. The estimators and their variances depend only on $\boldsymbol{\tau}$ and the maximum and minimum entry values $\max(\boldsymbol{v})$ and $\min(\boldsymbol{v})$. We study the variance for all-entries-equal $\boldsymbol{\tau}$ and $\max(\boldsymbol{v}) \leq \tau$.

Figure 5 shows the expectation of the square of the L* estimator over independent samples and of the L*, U*, and OC estimators over shared-seed samples. This is as a function of the ratio of the minimum to the maximum value in the data vector. The expectation of the square plotted is the ratio to the minimum possible expectation of the square over coordinated samples (the $\boldsymbol{v}$-optimum). Recall that the $\boldsymbol{v}$-optimum is not simultaneously attainable for all vectors and is used only as a reference for variance competitiveness. We can see that the L* estimator is nearly optimal when the ratio is large and that the U* estimator is nearly optimal when the ratio is small. The OC estimator outperforms both in the mid range. For the L* estimator, the ratio is always at most 2 (for $p = 1$) and 2.5 (for $p = 2$) from the optimum whereas the U* estimator can have large ratios. The OC estimator has optimal worst-case ratios of 1.204 (for $p = 1$) and 1.35 (for $p = 2$), but the ratio is the same across the range for all data vectors with $\min(\boldsymbol{v}) < \max(\boldsymbol{v}) \leq \tau$.

We study the variance as a function of $\frac{\min(\boldsymbol{v})}{\max(\boldsymbol{v})}$. The variance is 0 when $\text{RG}(\boldsymbol{v}) = 0$ (ratio is 1). Otherwise, it is lower for $\hat{\text{RG}}_p^{(U)}$ when the ratio is sufficiently small. The threshold point is ϕ_p which

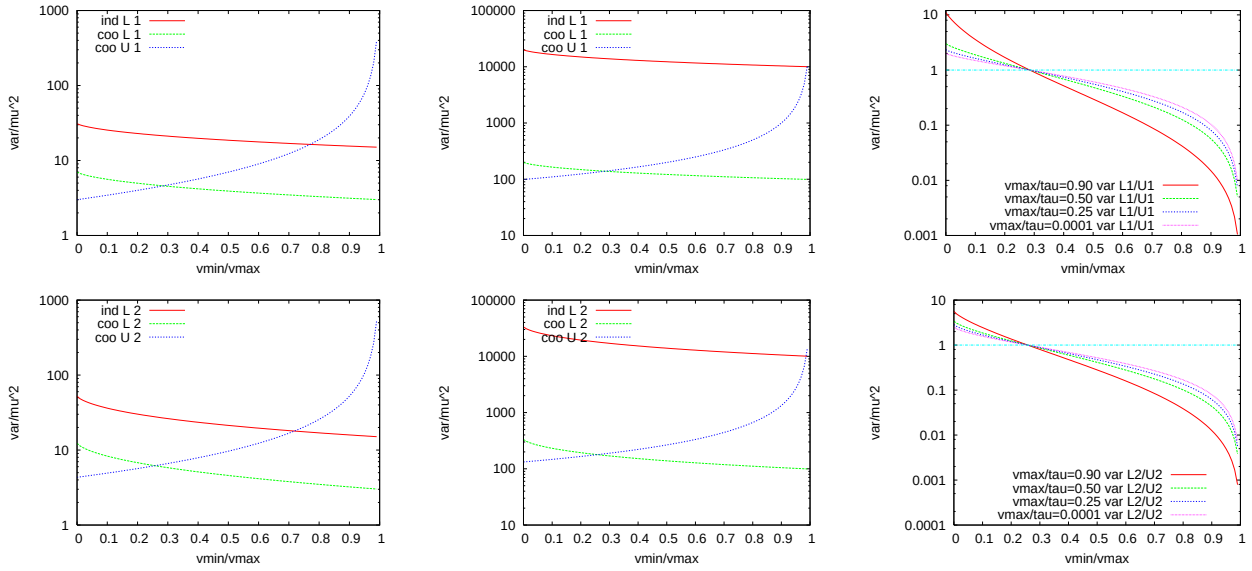


Figure 4: Variance (normalized by square of expectation) of $\hat{\text{RG}}_p^{(L)}$ estimator over independent samples and of $\hat{\text{RG}}_p^{(L)}$ and $\hat{\text{RG}}_p^{(U)}$ over shared-seed samples. Sampling with all-entries equal τ. (A): data with $\max(\boldsymbol{v}) = 0.25\tau$. (B): data with $\max(\boldsymbol{v}) = 0.01\tau$. (C): ratio $\text{VAR}[\hat{\text{RG}}_p^{(L)}]/\text{VAR}[\hat{\text{RG}}_p^{(U)}]$ for shared-seed sampling, selected ratios $\max(\boldsymbol{v})/\tau$. Sweeping $\min(\boldsymbol{v})$. Top shows $p = 1$, bottom is $p = 2$.

satisfies

$$\text{VAR}_{\mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}_p^{(U)}] < \text{VAR}_{\mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}_p^{(L)}] \iff \frac{\min(\boldsymbol{v})}{\max(\boldsymbol{v})} < \phi_p \ .$$

For $p = 1$, $\phi_1 \approx 0.285$ (is the solution of the equality $(1-x)/(2x) = \ln(1/x)$). For $p = 2$, $\phi_2 \approx 0.258$.

This suggests selecting an estimator according to expected characteristics of the data. If typically $\text{RG}(\mathbf{v}) > (1-\phi_p)\max(\mathbf{v})$, we choose $\hat{\text{RG}}_p^{(U)}$. If typically $\text{RG}(\mathbf{v}) < (1-\phi_p)\max(\mathbf{v})$ we choose $\hat{\text{RG}}_p^{(L)}$, and otherwise, we choose the OC estimator.

The variance of the L* estimator over independent samples and of the L* and U* estimators over shared-seed samples is illustrated in Figure 4. The figure also illustrates the relation between the variance of the shared-seed L* and U* estimators. When $\max(\boldsymbol{v})/\tau \ll 1$ (which we expect to be a prevailing scenario), $\text{VAR}[\hat{\text{RG}}^{(L)}]$ is nearly at most 2 times $\text{VAR}[\hat{\text{RG}}^{(U)}]$ but as $\min(\boldsymbol{v}) \to \max(\boldsymbol{v})$, the ratio $\text{VAR}[\hat{\text{RG}}^{(U)}]/\text{VAR}[\hat{\text{RG}}^{(L)}]$ is not bounded. When $\max(\boldsymbol{v})/\tau$ is close to 1, the variance of the U* estimator is close to 0, and $\text{VAR}[\hat{\text{RG}}^{(L)}]/\text{VAR}[\hat{\text{RG}}^{(U)}]$ is not bounded. Interestingly, for $p = 2$, the variance of the U* estimator is always at least $\frac{4}{3}\text{RG}_4(\boldsymbol{v})$, and thus, using (12), the variance of the L* estimator is at most 4.4 times the variance of the $\boldsymbol{v}$-optimal (and thus of the U* estimator).

5. EXTENSIONS

The *one-sided distance*, which isolates the growth or decline components, is defined as $L_{p+}^p(H) = \sum_{h\in H} \text{RG}_{p+}(\boldsymbol{v}(h))$, where $\text{RG}_{p+}(\boldsymbol{v}) = \max\{0, v_1 - v_2\}^p$. We can use any of our RG_p estimators $\hat{\text{RG}}_p(S)$ to estimate RG_{p+} as follows: If S^* includes $\boldsymbol{v}$ such that $v_1 \leq v_2$ then $\hat{\text{RG}}_{p+}(S) = 0$. Otherwise, $\hat{\text{RG}}_{p+}(S) = \hat{\text{RG}}_p(S)$. We can symetrically define L_{p-}^p and RG_{p-} and $\hat{\text{RG}}_{p-}$.

Our derivations can be extended to other sampling schemes. One such extension is to weighted sampling without replacement (PPSWR), which is bottom-k sampling with priorities $r_{ih} = -\frac{\ln(1-u_{ih})}{v_{ih}}$ [33, 32, 11, 12].

6. RELATED WORK

Prior to our work, the only distance estimator we are aware of which obtained good estimate with small fraction of data sampled is for L_1 over coordinated samples. This estimator uses the relation $|v_1 - v_2| = \max\{v_1, v_2\} - \min\{v_1, v_2\}$ to obtain an indirect estimate as the difference of two inverse probability estimates for the maximum and minimum [16]. Our U* estimator for $p = 1$ is a strengthening of this L_1 estimator.

Distance estimation over *unweighted* coordinated [28] or independent [14] sampling is a much simpler problem. With unweighted sampling, the inclusion probability of positive entries is independent of their weight. When carefully implemented, we can use inverse-probability estimates, which as discussed in the introduction, do not work with weighted sampling. An unweighted sample, however, is much less informative for its size when the data is skewed, as "heavy hitters" can be easily missed out. The estimators we develop here are applicable with, and take advantage, of weighted sampling.

Distance estimation of vectors (each instance in our terminology is presented as a vector of key values) was extensively studied using *linear sketches*, which are random linear projections, e.g. [25, 2, 1, 18]. Random projections have the property that the difference vector of sketches is the sketch of the difference of the two vectors. This means they are tailored for distance-like queries which aggregate over functions of the coordinate-wise differences. In particular, with linear sketches we can accurately estimate distances that are very small relative to the input vectors norms, whereas even with weighted sampling, accuracy depends on the relation of the distance to the vectors norms. A significant disadvantage of linear sketches, however, is lack of flexibility: Unlike samples, they do not support domain (selection) queries that are specified after the summary structure is computed and can only estimate distance between the full vectors. Moreover, each sketch is tailored for a specific metric, such as L_2^2. In particular, linear sketches are not suitable at all for estimating one-sided distances. Lastly, linear

sketches have size which often depends (poly) logarithmically on the number of keys. To summarize, the two techniques, sampling and linear sketches, have different advantages. Sampling provides much greater flexibility in terms of supported queries and often admits a smaller summary structure.

7. EXPERIMENTAL EVALUATION

We selected several datasets that have the form of values assigned to a set of keys, on two instances, and natural selection predicates. We consider the L_1 and L_2^2 distances on keys satisfying these predicates. Properties of the data sets with selections are summarized below and in Table 5.

• **destIP** (IP packet traces): keys: (anonymized) IP destination addresses. value: the number of IP flows to this destination IP. Instances: two consecutive time periods. Selection: all IP destination addresses in a certain subnetwork.

• **Server** (WWW activity logs): Keys: (anonymized) source IP address and Web site pairs. value: the number of HTTP requests issued to the Web site from this address. Instances: two consecutive time periods. Selection: A particular Web site.

• **Surnames and OSPD8:** keys: all words (terms). value: the number of occurrences of the term in English books digitized by Google and published within the time period [29]. Instances: the years 2007 and 2008. Surnames selection: the 18.5×10^3 most common surnames in the US. OSPD8 selection: the 7.5×10^4 8 letter words that appear in the Official Scrabble Players Dictionary (OSPD).

Each of the two instances were Poisson PPS [23] sampled (see Section 2) with different sampling threshold T, to obtain a range of sample sizes. We used both coordinated (shared-seed) and independent sampling of the two instances.

We study the quality of the L_p^p estimates obtained from our RG_p estimators. We estimate $L_p^p = \sum_{h\in H} \mathrm{RG}_p(\boldsymbol{v}(h))$ as the sum over selected keys H of RG_p estimates: $\hat{L}_p^p = \sum_{h\in H} \hat{\mathrm{RG}}_p(\boldsymbol{v}(h))$. We consider the estimator $\hat{\mathrm{RG}}_p{}^{(L)}$ for independent samples (Section 3) and the estimators $\hat{\mathrm{RG}}_p{}^{(L)}$ and $\hat{\mathrm{RG}}_p{}^{(U)}$ for coordinated samples (Section 4). To apply the estimators, we apply the selection predicate to sampled keys to identify all the ones satisfying the predicate. The estimators are then computed for keys that are sampled in at least one instance (the estimate is 0 for keys that are not sampled in any instance and do not need to be explicitly computed).

Since all our RG_p estimators are unbiased and nonnegative, so is the corresponding sum estimate $\hat{L}_p^p$. The variance is additive and is $\sum_{h\in H} \mathrm{VAR}_{\mathcal{S}_{\boldsymbol{v}(h)}}[\hat{\mathrm{RG}}_p]$. We measure the performance of the estimators using the variance normalized by the square of the expectation, which is the squared coefficient of variation $\mathrm{CV}^2(\hat{L}_p^p) = \frac{\sum_{h\in H} \mathrm{VAR}_{\mathcal{S}_{\boldsymbol{v}(h)}}[\hat{\mathrm{RG}}_p]}{(\sum_{h\in H} \mathrm{RG}_p(\boldsymbol{v}(h)))^2}$. This is the same as the normalized mean squared error (MSE) for our unbiased estimators.

Figure 6 shows the CV^2 of our L_p^p estimators ($p = 1, 2$) as a function of the sampled fraction of the dataset. We can see qualitatively, that all estimators, even over independent samples, are satisfactory, in that the CV is small for a sample that is a small fraction of the full data set. The estimator $\hat{\mathrm{RG}}_p^{(L)}$ over coordinated shared-seed samples outperforms, by orders of magnitude, the estimator $\hat{\mathrm{RG}}_p^{(L)}$ over independent samples. The gap widens for more aggressive sampling (higher T).

On the IP flows and WWW logs data, there are significant differences on the values of keys between instances: the L_1 distance is a large fraction of the total sum of values $\sum_{h\in H}\sum_{i\in[2]} v_i(h)$. Therefore, for the destIP and Server selections, $\hat{\mathrm{RG}}_p^{(U)}$ outperforms $\hat{\mathrm{RG}}_p^{(L)}$ on shared-seed samples. On the term count data there is typically a small difference between instances. We can see that for the Surnames and OSPD8 selections, $\hat{\mathrm{RG}}_p^{(L)}$ outperforms $\hat{\mathrm{RG}}_p^{(U)}$ on shared seed samples. These trends are more pronounced for the Euclidean distance ($p = 2$). In this case, on Surnames and OSPD8, $\hat{\mathrm{RG}}_p^{(L)}$ over independent samples outperform $\hat{\mathrm{RG}}_p^{(U)}$ over shared-seed samples. We can see that we can significantly improve accuracy by tailoring the selection of the estimator to properties of the data. The performance of the U* estimator, however, can significantly diverge for very similar instances whereas the competitive L* estimator is guaranteed not to be too far off. Therefore, when there is no prior knowledge on the difference, we suggest using the L* estimator.

The datasets also differ in the symmetry of change. The change is more symmetric in the IP flows and WWW logs data $L_{p+} \approx L_{p-}$ whereas there is a general growth trend $L_{p+} \gg L_{p-}$ in the term count data. Estimator performance on one-sided distances (not shown) is similar to the corresponding distance estimators.

8. CONCLUSION

Distance queries are essential for monitoring, planning, and anomaly and change detection. Random sampling is an important tool for retaining the ability to query data under resource limitations. We provide the first satisfactory solution for estimating L_p distance from sampled data sets. Our solution is comprehensive, covering common sampling schemes. It is supported by rigorous analysis and novel techniques. Our estimators scale well with data size and we demonstrated that accurate estimates are obtained for queries with small support size.

Acknowledgement. The author is grateful to Haim Kaplan for many comments, helpful feedback, and suggesting the use of the ngrams data.

	$\hat{\mathrm{RG}}^{(L)}$
$\lvert S\rvert = 0$	0
$\lvert S\rvert \geq 1$	$\max\{\max(\boldsymbol{v})-\tau, 0\} - \max\{\min(\boldsymbol{v})-\tau, 0\} + \tau \ln \frac{\min\{\max(\boldsymbol{v}),\tau\}}{\min\{v_{\min},\tau\}}$

Condition	$\mathrm{VAR}_{\mathcal{S}_{\boldsymbol{v}}}[\hat{\mathrm{RG}}^{(L)}]$
$\min(\boldsymbol{v}) \geq \tau$	0
$\max(\boldsymbol{v}) \leq \tau,\ \min(\boldsymbol{v}) = 0$	$2\mathrm{RG}(\boldsymbol{v})\tau - \mathrm{RG}(\boldsymbol{v})^2$
$\max(\boldsymbol{v}) \leq \tau,\ \min(\boldsymbol{v}) > 0$	$2\mathrm{RG}(\boldsymbol{v})\tau - \mathrm{RG}(\boldsymbol{v})^2 - 2\tau \min(\boldsymbol{v}) \ln(\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})})$
$0 < \min(\boldsymbol{v}) \leq \tau \leq \max(\boldsymbol{v})$	$(\tau)^2 - \min(\boldsymbol{v})^2 - 2\tau \min(\boldsymbol{v}) \ln(\frac{\tau}{\min(\boldsymbol{v})})$
$0 = \min(\boldsymbol{v}),\ \tau \leq \max(\boldsymbol{v})$	$(\tau)^2 - \min(\boldsymbol{v})^2$

Table 6: $\hat{\mathrm{RG}}^{(L)}$ and variance for shared-seed sampling.

9. REFERENCES

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58:137–147, 1999.

[2] B. Bahmani, A. Goel, and R. Shinde. efficient distributed locality sensitive hashing. In *CIKM*. ACM, 2012.

[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] 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.

data	# keys	p1%	p2%	$\sum_{ih} v_{ih}$	p1%	p2%	$L_1/\sum_{ih} v_{ih}$	$L_{1+}/\sum_{ih} v_{ih}$	$L_{1-}/\sum_{ih} v_{ih}$
destIP	3.8×10^4	65%	65%	1.1×10^6	49%	51%	0.36	0.19	0.18
Server	2.7×10^5	53%	56%	2.9×10^6	50%	50%	0.75	0.38	0.37
Surnames	1.9×10^4	100%	100%	8.9×10^7	48.6%	51.4%	0.094	0.0617	0.0327
OSPD8	7.5×10^5	99%	99%	1.57×10^{10}	46.8%	53.2%	0.0826	0.0727	0.0099

Table 5: Datasets with subset selections. Table shows total number of distinct keys H satisfying selection predicate that had a positive value in at least one of the two instances, corresponding percentage in each instance, total sum of values $\sum_{ih} v_{ih}$, and fraction (shown as percentage), sum in each instance $i = 1, 2$: $\frac{\sum_h v_{ih}}{\sum_{jh} v_{jh}}$, and normalized L_1, L_{1+} and L_{1-} distances.

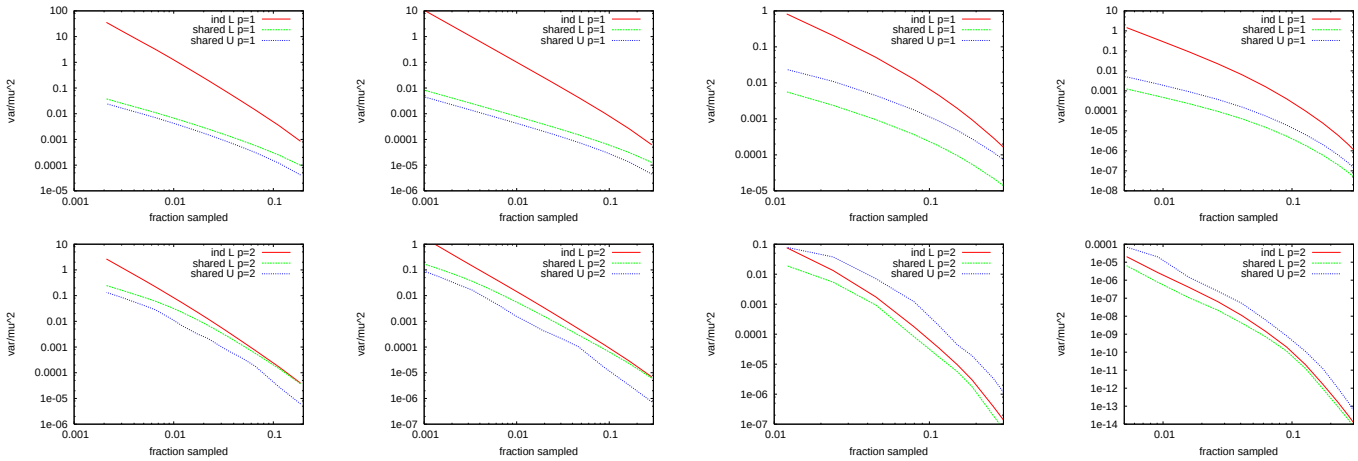


Figure 6: Queries (left to right): destIP, Server, Surnames, OSPD8. Plot shows CV^2 of L_p^p estimate for fraction of sampled items from the query support. Top shows $p = 1$ (L_1) bottom shows $p = 2$ (L_2^2).

	$\hat{\mathrm{RG}}_2^{(L)}$
$\lvert S\rvert = 0$	0
$\lvert S\rvert \geq 1$	$\max\{\max(\boldsymbol{v}), \tau\}^2 - \max\{\min(\boldsymbol{v}), \tau\}^2$ $-2\max\{\min(\boldsymbol{v}), \tau\}(\max(\boldsymbol{v}) - v_{\min})$ $+2\tau \max(\boldsymbol{v}) \ln \frac{\min\{\max(\boldsymbol{v}), \tau\}}{\min\{v_{\min}, \tau\}}$

Condition	$\mathrm{VAR}_{\mathcal{S}_{\boldsymbol{v}}}[\hat{\mathrm{RG}}_2^{(L)}]$
$\min(\boldsymbol{v}) \geq \tau$	0
$\max(\boldsymbol{v}) \leq \tau$	$-4\tau \max(\boldsymbol{v}) \min(\boldsymbol{v}) \ln(\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})})(2\max(\boldsymbol{v}) - \min(\boldsymbol{v}))$ $-(\max(\boldsymbol{v}) - \min(\boldsymbol{v}))^4$ $+\frac{2\tau}{3}(5\max(\boldsymbol{v})^3 + 4\min(\boldsymbol{v})^3 - 9\max(\boldsymbol{v})\min(\boldsymbol{v})^2)$
$\min(\boldsymbol{v}) \leq \tau$ $\wedge$ $\max(\boldsymbol{v}) \geq \tau$	$4\max(\boldsymbol{v})\min(\boldsymbol{v})\tau(\min(\boldsymbol{v}) - 2\max(\boldsymbol{v}))\ln\frac{\tau}{\min(\boldsymbol{v})}$ $+4\max(\boldsymbol{v})\min(\boldsymbol{v})(\tau)^2 + \frac{(\tau)^4}{3} + \frac{8\min(\boldsymbol{v})^3\tau}{3}$ $-6\max(\boldsymbol{v})\min(\boldsymbol{v})^2\tau - 4\max(\boldsymbol{v})^2\min(\boldsymbol{v})^2$ $-\min(\boldsymbol{v})^4 + 4\max(\boldsymbol{v})\min(\boldsymbol{v})^3 + 4\max(\boldsymbol{v})^2(\tau)^2$ $-2\max(\boldsymbol{v})(\tau)^3$

Table 7: $\hat{\mathrm{RG}}_2^{(L)}$ and variance for shared-seed sampling

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

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

[7] S. S. Chawathe, A. Rajaraman, H. Garcia-Molina, and J. Widom. Change detection in hierarchically structured information. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, 1996.

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

[9] E. Cohen. Estimation for monotone sampling: Competitiveness and customization. In *PODC*. ACM, 2014. full version `http://arxiv.org/abs/1212.0243`.

[10] E. Cohen, N. Duffield, C. Lund, M. Thorup, and H. Kaplan. Efficient stream sampling for variance-optimal estimation of subset sums. *SIAM J. Comput.*, 40(5), 2011.

[11] E. Cohen and H. Kaplan. Summarizing data using bottom-k sketches. In *ACM PODC*, 2007.

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

[13] E. Cohen and H. Kaplan. Leveraging discarded samples for tighter estimation of multiple-set aggregates. In *ACM SIGMETRICS*, 2009.

[14] E. Cohen and H. Kaplan. Get the most out of your sample: Optimal unbiased estimators using partial information. In *Proc. of the 2011 ACM Symp. on Principles of Database Systems (PODS 2011)*. ACM, 2011. full version: `http://arxiv.org/abs/1203.4903`.

[15] E. Cohen and H. Kaplan. What you can do with coordinated samples. In *The 17th. International Workshop on Randomization and Computation (RANDOM)*, 2013. full version: `http://arxiv.org/abs/1206.5637`.

[16] 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. full version: `http://arxiv.org/abs/0906.4560`.

[17] 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.

[18] W. Dong, M. Charikar, and K. Li. Asymmetric distance estimation with sketches for similarity search in high-dimensional spaces. In *SIGIR*, 2008.

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

[20] 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.

Algorithm 1 $\hat{\text{RG}}_p^{(U)}(S)$

if $|S| = 0$ **then return** 0 ▷ from hereafter $|S| > 0$
$m \leftarrow \max_{i \in S} v_i$ ▷ $m = \max(\boldsymbol{v})$
if $|S| < r$ **then** $n \leftarrow 0$
else $n \leftarrow \min_{i \in S} v_i$ ▷ $n = \underline{\min}(S)$
if $n \geq \tau$ **then return** $(m-n)^p$ ▷ case: $\min(\boldsymbol{v}) \geq \tau$
if $p \leq 1$ **then** ▷ case: $\min(\boldsymbol{v}) \leq \tau$ and $p \leq 1$
if n=0 **then return** $m^p \frac{\tau}{\min\{m,\tau\}}$
else return $\frac{\tau}{n}\left((m-n)^p - \frac{\min\{m,\tau\}-n}{\min\{m,\tau\}} m^p\right)$
if $m \leq \tau$ **then** ▷ case: $\max(\boldsymbol{v}) \leq \tau, p > 1$
if $\zeta\tau > n$ **then return** $p\tau(m-\zeta\tau)^{p-1}$
else return 0
▷ case: $n < \tau < \max(\boldsymbol{v})$ and $p > 1$
$\eta_0 \leftarrow \frac{p\tau - m}{(p-1)\tau}$
if $\eta_0 \in (0,1)$ **then** ▷ subcase: $\eta_0 \in (0,1)$
if $\zeta \geq \max\{\eta_0, n/\tau\}$ **then return** $\frac{(m-\eta_0\tau)^p}{1-\eta_0}$
if $n/\tau < \zeta < \eta_0$ **then return** $p\tau(m-\zeta\tau)^{p-1}$
if $\zeta \leq n/\tau \leq \eta_0$ **then return** 0
if $\zeta \leq n/\tau \geq \eta_0$ **then**
return $\frac{\tau(m-n)^p}{n} - \frac{(\tau-n)(m-\eta_0\tau)^p}{n(1-\eta_0)}$
else ▷ subcase: $\eta_0 \notin (0,1)$
if $\zeta\tau > n$ **then return** m^p
else return $\frac{\tau}{n}(m-n)^p - m^p\left(\frac{\tau}{n} - 1\right)$

[21] 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.

[22] 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.

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

[24] 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.

[25] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. 30th Annual ACM Symposium on Theory of Computing*, pages 604–613. ACM, 1998.

[26] D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *VLDB*, 2004.

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

[28] P. Li, , K. W. Church, and T. Hastie. Conditional random sampling: A sketch-based sampling technique for sparse data. In *NIPS*, 2006.

[29] J. B. Michel, Y. K. Shen, A. Presser Aiden, A. Veres, M. K. Gray, W. Brockman, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011. http://www.sciencemag.org/content/331/6014/176.

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

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

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

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

condition	$\hat{\text{RG}}^{(U)}$
$\|S\| = 0$	0
$1 \leq \|S\| \leq r-1$	$\max\{\tau, \max(\boldsymbol{v})\}$
$\|S\| = r$	$\max\{\max(\boldsymbol{v}), \tau\} - \max\{\min(\boldsymbol{v}), \tau\}$

condition on $\boldsymbol{v}$	$\text{VAR}[\hat{\text{RG}}^{(U)} \mid \boldsymbol{v}]$
$\min(\boldsymbol{v}) \geq \tau$	0
$\max(\boldsymbol{v}) \leq \tau$	$\text{RG}(\boldsymbol{v})(\tau - \text{RG}(\boldsymbol{v}))$
$\min(\boldsymbol{v}) < \tau < \max(\boldsymbol{v})$	$\min(\boldsymbol{v})(\tau - \min(\boldsymbol{v}))$

Table 8: $\hat{\text{RG}}^{(U)}$ **and variance for share-seed sampling.**

[34] 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.

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

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

APPENDIX

A. DERIVATION OF $\hat{\text{RG}}_P^{(U)}$

If $\zeta\tau > \max(\boldsymbol{v})$ then $\underline{\text{RG}}_p(\zeta, \boldsymbol{v}) = 0$ and $\hat{\text{RG}}_p^{(U)}(\zeta, \boldsymbol{v}) = 0$. Otherwise, when $\min(\boldsymbol{v}) < \zeta\tau \leq \max(\boldsymbol{v})$, noting that the supremum is obtained by a vector $\boldsymbol{v}'$ with maximum entry $\max(\boldsymbol{v})$ and minimum entry 0,

$$\hat{\text{RG}}_p^{(U)}(\zeta, \boldsymbol{v}) = \tag{15}$$
$$= \inf_{0 \leq \eta < \zeta} \frac{(\max(\boldsymbol{v}) - \eta\tau)^p - \int_\zeta^{\min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}} \hat{\text{RG}}_p{}^{(U)}(u, \boldsymbol{v}) du}{\zeta - \eta}$$

If $\zeta\tau \leq \min(\boldsymbol{v})$,

$$\hat{\text{RG}}_p^{(U)}(\zeta, \boldsymbol{v}) = \frac{\text{RG}_p(\boldsymbol{v}) - \int_\zeta^{\min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}} \hat{\text{RG}}_p{}^{(U)}(u, \boldsymbol{v}) du}{\zeta}$$
$$= \frac{\text{RG}_p(\boldsymbol{v}) - \int_{\min\{1, \frac{\min(\boldsymbol{v})}{\tau}\}}^{\min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}} \hat{\text{RG}}_p{}^{(U)}(u, \boldsymbol{v}) du}{\min\{1, \min(\boldsymbol{v})/\tau\}} \tag{16}$$

If $\min(\boldsymbol{v}) \geq \tau$, $|S| = r$, $\hat{\text{RG}}_p^{(U)} = \hat{\text{RG}}_p^{(L)} \equiv \text{RG}_p(\boldsymbol{v})$. If $\max(\boldsymbol{v}) \leq \tau$, $\int_\zeta^1 \hat{\text{RG}}_p(u, \boldsymbol{v}) du = \underline{\text{RG}}_p(\zeta, \boldsymbol{v})$ and the infimum is the derivative of the lower bound function, and thus, $\hat{\text{RG}}_p(\zeta, \boldsymbol{v}) = p\tau(\max(\boldsymbol{v}) - \zeta\tau)^{p-1}$.

$\|S\|$	$\hat{\text{RG}}_p^{(U)}$
$0, r:$	0
$1 \ldots r-1:$	$p\tau(\max(\boldsymbol{v}) - \zeta\tau)^{p-1}$

(17)

We now consider the case $\min(\boldsymbol{v}) \leq \tau \leq \max(\boldsymbol{v})$, solving (15) for $\zeta > \min(\boldsymbol{v})/\tau$. For $\zeta = 1$ we obtain the equation

$$\hat{\text{RG}}_p^{(U)}(1, \boldsymbol{v}) = \inf_{0 \leq \eta < \zeta} \frac{(\max(\boldsymbol{v}) - \eta\tau)^p}{1 - \eta} .$$

When $p = 1$, the derivative is positive and the infimum is $\max(\boldsymbol{v})$. We obtain that $\hat{\text{RG}}^{(U)}(\zeta, \boldsymbol{v}) = \max(\boldsymbol{v})$ for $\zeta \geq \min(\boldsymbol{v})/\tau$. Using (16), $\hat{\text{RG}}^{(U)}(\zeta, \boldsymbol{v}) = \max(\boldsymbol{v}) - \tau$ when $\zeta \leq \frac{\min(\boldsymbol{v})}{\tau}$.

For $p \neq 1$, we need to find the value where $h(\eta) = \frac{(\max(\boldsymbol{v}) - \eta\tau)^p}{1-\eta}$ is minimized. The derivative is

$$\frac{\partial h(\eta)}{\partial \eta} = \frac{(\max(\boldsymbol{v}) - \eta\tau)^{p-1}}{1 - \eta}\left(-\tau p + \frac{\max(\boldsymbol{v}) - \eta\tau}{1 - \eta}\right) .$$

The derivative is 0 at

$$\eta_0 = \frac{p\tau - \max(\boldsymbol{v})}{\tau(p-1)} .$$

condition on $S(\zeta, \boldsymbol{v})$	$\hat{\text{RG}}_2^{(U)}(S)$
$\frac{\max(\boldsymbol{v})}{\tau} \geq 2, \zeta > \frac{\min(\boldsymbol{v})}{\tau}$	$\max(\boldsymbol{v})^2$
$\frac{\max(\boldsymbol{v})}{\tau} \geq 2, \zeta \leq \frac{\min(\boldsymbol{v})}{\tau}$	$\max(\boldsymbol{v})^2 - 2\tau\max(\boldsymbol{v}) + \min(\boldsymbol{v})\tau$
$\frac{\max(\boldsymbol{v})}{\tau} \leq 1, \zeta \in (\frac{\min(\boldsymbol{v})}{\tau}, \frac{\max(\boldsymbol{v})}{\tau}]$	$2\tau(\max(\boldsymbol{v}) - \zeta\tau)$
$\frac{\max(\boldsymbol{v})}{\tau} \leq 1, \zeta \leq \frac{\min(\boldsymbol{v})}{\tau}$	0
$\frac{\max(\boldsymbol{v})}{\tau} \leq 1, \zeta \geq \frac{\max(\boldsymbol{v})}{\tau}$	0
$\frac{\max(\boldsymbol{v})}{\tau} \in [1,2], \zeta > 2 - \frac{\max(\boldsymbol{v})}{\tau}, \zeta > \frac{\min(\boldsymbol{v})}{\tau}$	$4\tau(\max(\boldsymbol{v}) - \tau)$
$\frac{\max(\boldsymbol{v})}{\tau} \in [1,2], \zeta < 2 - \frac{\max(\boldsymbol{v})}{\tau}, \zeta > \frac{\min(\boldsymbol{v})}{\tau}$	$2\tau(\max(\boldsymbol{v}) - \zeta\tau)$
$\frac{\max(\boldsymbol{v})}{\tau} \in [1,2], \zeta < 2 - \frac{\max(\boldsymbol{v})}{\tau}, \zeta < \frac{\min(\boldsymbol{v})}{\tau}$	0
$\frac{\max(\boldsymbol{v})}{\tau} \in [1,2], \zeta \leq \frac{\min(\boldsymbol{v})}{\tau} > 2 - \frac{\max(\boldsymbol{v})}{\tau}$	$\frac{\tau}{\min(\boldsymbol{v})}\text{RG}_2(\boldsymbol{v}) - 4\tau(\max(\boldsymbol{v}) - \tau)(\frac{\tau}{\min(\boldsymbol{v})} - 1)$

condition on $\boldsymbol{v}$	$\text{VAR}_{\mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}_2^{(U)}]$
$\min(\boldsymbol{v}) \geq \tau$	0
$\max(\boldsymbol{v}) \leq \tau$	$\text{RG}_3(\boldsymbol{v})(\frac{4}{3}\tau - \text{RG}(\boldsymbol{v}))$
$\frac{\max(\boldsymbol{v})}{\tau} \in [1,2], \frac{\min(\boldsymbol{v})}{\tau} \geq 2 - \frac{\max(\boldsymbol{v})}{\tau}$	$\frac{2(\tau - \min(\boldsymbol{v}))}{\tau}\left(\text{RG}_2(\boldsymbol{v})) - 4\tau(\max(\boldsymbol{v}) - \tau)\right)^2$
$\frac{\max(\boldsymbol{v})}{\tau} \in [1,2], \frac{\min(\boldsymbol{v})}{\tau} < 2 - \frac{\max(\boldsymbol{v})}{\tau}$	$\frac{\max(\boldsymbol{v}) - \tau}{\tau}\left(4\tau(\max(\boldsymbol{v}) - \tau) - \text{RG}_2(\boldsymbol{v})\right)^2 + \frac{\min(\boldsymbol{v})}{\tau}\text{RG}_4(\boldsymbol{v})$ $+ \frac{(\text{RG}_2(\boldsymbol{v}) + 4(\tau)^2 - 4\max(\boldsymbol{v})\tau)^3 - \text{RG}_3(\boldsymbol{v})(\text{RG}(\boldsymbol{v}) - 2\tau)^3}{6(\tau)^2}$
$\frac{\max(\boldsymbol{v})}{\tau} \geq 2$	$(2\max(\boldsymbol{v}) - \min(\boldsymbol{v}))^2 \min(\boldsymbol{v})(\tau - \min(\boldsymbol{v}))$

Table 9: Estimator $\hat{\text{RG}}_2^{(U)}$ and its variance for shared-seed sampling.

If η_0 is outside $(0,1)$, the infimum is obtained at $\eta = 0$ and the estimate is $\hat{\text{RG}}^{(U)}(\zeta, \boldsymbol{v}) = \max(\boldsymbol{v})^p$ for $\zeta \geq \min(\boldsymbol{v})/\tau$ and, using (16),

$$\hat{\text{RG}}^{(U)}(\zeta, \boldsymbol{v}) = \frac{\tau}{\min(\boldsymbol{v})}\text{RG}_p(\boldsymbol{v}) - \max(\boldsymbol{v})^p(\frac{\tau}{\min(\boldsymbol{v})} - 1)$$

for $\zeta < \min(\boldsymbol{v})/\tau$.

Otherwise, if $\eta_0 \in (0,1)$, the infimum is achieved at η_0. Using (15), the estimate is

$$\begin{aligned}\hat{\text{RG}}_p(\zeta, \boldsymbol{v}) &= \frac{(\max(\boldsymbol{v}) - \eta_0\tau)^p}{1 - \eta_0} \\ &= \frac{\tau(p-1)(\max(\boldsymbol{v})\frac{p-2}{p-1} - \tau\frac{p}{p-1})^p}{\max(\boldsymbol{v}) - \tau}\end{aligned}$$

for $\zeta \in [\max\{\eta_0, \frac{\min(\boldsymbol{v})}{\tau}\}, 1]$ and $\hat{\text{RG}}_p(\zeta, \boldsymbol{v}) = p\tau(\max(\boldsymbol{v}) - \zeta\tau)^{p-1}$ for $\zeta \in (\frac{\min(\boldsymbol{v})}{\tau}, \eta_0)$. Using (16), when $\zeta \leq \frac{\min(\boldsymbol{v})}{\tau}$, then $\hat{\text{RG}}_p(\zeta, \boldsymbol{v}) = 0$ when $\frac{\min(\boldsymbol{v})}{\tau} < \eta_0$ and

$$\hat{\text{RG}}_p(\zeta, \boldsymbol{v}) = \frac{\text{RG}_p(\boldsymbol{v}) - (\frac{\min(\boldsymbol{v})}{\tau} - \eta_0)\frac{(\max(\boldsymbol{v}) - \eta_0\tau)^p}{1 - \eta_0}}{1 - \frac{\min(\boldsymbol{v})}{\tau}}$$

when $\frac{\min(\boldsymbol{v})}{\tau} \geq \eta_0$.

B. VARIANCE OF $\hat{\text{RG}}^{(U)}$ AND $\hat{\text{RG}}_2^{(U)}$

The estimators $\hat{\text{RG}}^{(U)}$ and $\hat{\text{RG}}_2^{(U)}$, provided in Tables 8 and 9, are obtained by substituting $p = 1$ and $p = 2$ respectively in Algorithm 1. We calculate the variance of these estimators.

Variance of $\hat{\text{RG}}^{(U)}$: When $\max(\boldsymbol{v}) \leq \tau$, we have $\hat{\text{RG}}^{(U)} = \tau$ for $\zeta \in (\frac{\min(\boldsymbol{v})}{\tau}, \frac{\max(\boldsymbol{v})}{\tau}]$ and $\hat{\text{RG}}^{(U)} = 0$ otherwise. Hence,

$$\begin{aligned}&\text{VAR}[\hat{\text{RG}}^{(U)}] \\ &= (\text{RG}(\boldsymbol{v})^2(1 - \text{RG}(\boldsymbol{v})/\tau) + (\tau - \text{RG}(\boldsymbol{v}))^2\text{RG}(\boldsymbol{v})/\tau \\ &= \text{RG}(\boldsymbol{v})\tau - \text{RG}(\boldsymbol{v})^2\end{aligned}$$

Variance of $\hat{\text{RG}}_2^{(U)}$: When $\max(\boldsymbol{v}) > \tau$, we have $\eta_0 = 2 - \frac{\max \boldsymbol{v}}{\tau}$. Thus $\eta_0 \in (0,1) \iff \frac{\max(\boldsymbol{v})}{\tau} \in (1,2)$. We use

$$\int (\text{RG}_2(\boldsymbol{v}) - 2\tau\max(\boldsymbol{v}) + 2(\tau)^2 u)^2 du = \frac{(\text{RG}_2(\boldsymbol{v}) - 2\tau\max(\boldsymbol{v}) + 2(\tau)^2 u)^3}{6(\tau)^2}. \quad (18)$$

We start with the case $\max(\boldsymbol{v}) \leq \tau$. Remaining cases are omitted due to lack of space.

$$\begin{aligned}\text{VAR}_{\mathcal{S}_{\boldsymbol{v}}}[\hat{\text{RG}}_2^{(U)}] &= (1 - \frac{\text{RG}(\boldsymbol{v})}{\tau})\text{RG}_4(\boldsymbol{v}) \\ &\quad + \frac{(\text{RG}_2(\boldsymbol{v}) - 2\tau\max(\boldsymbol{v}) + 2(\tau)^2 u)^3}{6(\tau)^2}\Big|_{\frac{\min(\boldsymbol{v})}{\tau}}^{\frac{\max(\boldsymbol{v})}{\tau}} \\ &= (1 - \frac{\text{RG}(\boldsymbol{v})}{\tau})\text{RG}_4(\boldsymbol{v}) + \frac{\text{RG}_6(\boldsymbol{v})}{6(\tau)^2} - \frac{\text{RG}_3(\boldsymbol{v})(\text{RG}(\boldsymbol{v}) - 2\tau)^3}{6(\tau)^2} \\ &= \text{RG}_3(\boldsymbol{v})(\frac{4}{3}\tau - \text{RG}(\boldsymbol{v}))\end{aligned}$$

C. VARIANCE OF $\hat{\text{RG}}^{(L)}$ AND $\hat{\text{RG}}_2^{(L)}$

The estimators $\hat{\text{RG}}^{(L)}$ and $\hat{\text{RG}}_2^{(L)}$, provided in Tables 6 and 7 are obtained using (10). We calculate their variance.

Variance of $\hat{\text{RG}}^{(L)}$: When $\max(\boldsymbol{v}) \leq \tau$, we have $\hat{\text{RG}}^{(L)} = \tau\ln(\frac{\max(\boldsymbol{v})}{\tau\zeta})$ when $1 \leq |S| \leq r - 1$ and $\hat{\text{RG}}^{(L)} = \tau\ln(\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})})$ when $|S| = r$. The

variance is

$$\begin{aligned}
&\textsc{var}[\hat{\textsc{rg}}^{(L)}|\boldsymbol{v}] \\
&= (1-\frac{\max(\boldsymbol{v})}{\tau})\textsc{rg}(\boldsymbol{v})^2 + \\
&\int_{\frac{\min(\boldsymbol{v})}{\tau}}^{\frac{\max(\boldsymbol{v})}{\tau}} (\textsc{rg}(\boldsymbol{v}) - \tau\ln(\frac{\max(\boldsymbol{v})}{\tau y}))^2 dy + \\
&\frac{\min(\boldsymbol{v})}{\tau}(\textsc{rg}(\boldsymbol{v}) - \tau\ln(\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})})^2 \\
&= -2\tau\min(\boldsymbol{v})\ln(\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})}) + 2\textsc{rg}(\boldsymbol{v})\tau - (\textsc{rg}(\boldsymbol{v}))^2
\end{aligned}$$

When $\min(\boldsymbol{v}) \leq \tau \leq \max(\boldsymbol{v})$, $\hat{\textsc{rg}}^{(L)} = \max(\boldsymbol{v}) - \tau + \tau\ln(\frac{1}{\zeta})$ when $1 \leq |S| \leq r-1$ and $\hat{\textsc{rg}}^{(L)} = \max(\boldsymbol{v}) - \tau + \tau\ln(\frac{\tau}{\min(\boldsymbol{v})})$ when $|S| = r$. The variance is $\textsc{var}[\hat{\textsc{rg}}^{(L)}|\boldsymbol{v}] = (\tau)^2 - \min(\boldsymbol{v})^2 - 2\tau\min(\boldsymbol{v})\ln(\frac{\tau}{\min(\boldsymbol{v})})$.

Variance of $\hat{\textsc{rg}}_2^{(L)}$:

If $\min(\boldsymbol{v}) < \tau \leq \max(\boldsymbol{v})$, $\hat{\textsc{rg}}_2^{(L)} = \max(\boldsymbol{v})^2 - (\tau)^2 + 2\tau(u\tau - \max(\boldsymbol{v}) + \max(\boldsymbol{v})\ln\frac{1}{u}$ when $|S| \in [r-1]$ and $\hat{\textsc{rg}}_2^{(L)} = \max(\boldsymbol{v})^2 - (\tau)^2 + 2\tau(\min(\boldsymbol{v}) - \max(\boldsymbol{v}) + \max(\boldsymbol{v})\ln\frac{\tau}{\max(\boldsymbol{v})}$ when $|S| = r$. The variance is

$$\begin{aligned}
\textsc{var}[\hat{\textsc{rg}}_2^{(L)}|\boldsymbol{v}] &= 4\max(\boldsymbol{v})\min(\boldsymbol{v})\tau(\min(\boldsymbol{v}) - 2\max(\boldsymbol{v}))\ln\frac{\tau}{\min(\boldsymbol{v})} \\
&+4\max(\boldsymbol{v})\min(\boldsymbol{v})(\tau)^2 + \frac{(\tau)^4}{3} - 6\max(\boldsymbol{v})\min(\boldsymbol{v})^2\tau - \min(\boldsymbol{v})^4 \\
&-4\max(\boldsymbol{v})^2\min(\boldsymbol{v})^2 + 4\max(\boldsymbol{v})\min(\boldsymbol{v})^3 + 4\max(\boldsymbol{v})^2(\tau)^2 \\
&-2\max(\boldsymbol{v})(\tau)^3 + \frac{8\min(\boldsymbol{v})^3\tau}{3}
\end{aligned}$$

If $\max(\boldsymbol{v}) < \tau$, $\hat{\textsc{rg}}_2^{(L)} = 2\tau(u\tau - \max(\boldsymbol{v}) + \max(\boldsymbol{v})\ln\frac{\max(\boldsymbol{v})}{u\tau})$ when $|S| \in [r-1]$ and $\hat{\textsc{rg}}_2^{(L)} = 2\tau(\min(\boldsymbol{v}) - \max(\boldsymbol{v}) + \max(\boldsymbol{v})\ln\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})})$ when $|S| = r$. The variance is

$$\begin{aligned}
\textsc{var}[\hat{\textsc{rg}}_2^{(L)}|\boldsymbol{v}] &= -4\tau\max(\boldsymbol{v})\min(\boldsymbol{v})\ln(\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})})(2\max(\boldsymbol{v}) - \min(\boldsymbol{v})) + \\
&+\frac{2\tau}{3}(5\max(\boldsymbol{v})^3 - 9\max(\boldsymbol{v})\min(\boldsymbol{v})^2 + 4\min(\boldsymbol{v})^3) - \textsc{rg}_4(\boldsymbol{v})
\end{aligned}$$