# Entropy-Learned Hashing
# Constant Time Hashing with Controllable Uniformity

Brian Hentschel
Harvard University

Utku Sirin
Harvard University

Stratos Idreos
Harvard University

## ABSTRACT

Hashing is a widely used technique for creating uniformly random numbers from arbitrary data. This is required in a large range of core data-driven operations including indexing, partitioning, filters, and sketches. As such, hashing is a core component in numerous systems including relational data systems, key-value stores, compilers, and networks. Due to both the computational and data heavy nature of hashing, it is a core systems bottleneck. For example, a typical database query in the standard TPC-H benchmark may spend 50% of its total cost in hash tables. Similarly, Google spends at least 2% of its total computational cost on C++ hash tables, resulting in a massive yearly cost footprint just from one hashing operation.

We propose a new hashing method, called Entropy-Learned Hashing, which reduces the computational cost of hashing by up to an order of magnitude. We look at hashing from a pseudorandomness point of view and the key question we ask is "how much randomness is needed?" We show that state-of-the-art hash functions do too much work to perform their core task: extracting randomness from a data source to create random outputs. Entropy-Learned Hashing 1) models and estimates the randomness (entropy) of the input data, and then 2) creates data-specific hash functions that use only the parts of the data that are needed to differentiate the outputs. The resulting hash functions dramatically reduce the amount of computation needed while we prove their output is similarly uniform to that of traditional hash functions. We test Entropy-Learned Hashing across diverse and core hashing operations such as hash tables, Bloom filters, and partitioning and we demonstrate an increase in throughput in the order of 3.7x, 4.0x, and 14x respectively compared to the best in-class hash functions and implementations used at scale by Google and Meta.

## CCS CONCEPTS

• **Information systems → Point lookups**.

## KEYWORDS

hashing, hash tables, systems, Bloom filters

**ACM Reference Format:**
Brian Hentschel, Utku Sirin, and Stratos Idreos. 2022. Entropy-Learned Hashing Constant Time Hashing with Controllable Uniformity. In *Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22), June 12–17, 2022, Philadelphia, PA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3514221.3517894


*SIGMOD '22, June 12–17, 2022, Philadelphia, PA, USA*

ACM ISBN 978-1-4503-9249-5/22/06…$15.00
https://doi.org/10.1145/3514221.3517894

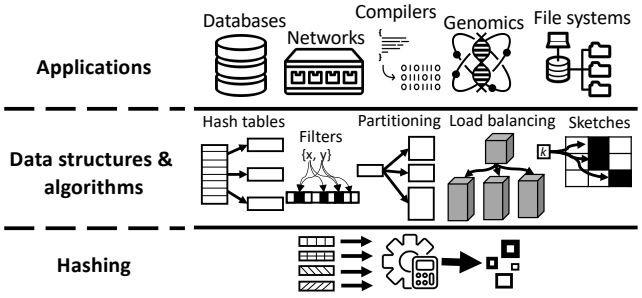



**Figure 1: Hashing is a core element for numerous fundamental components across diverse classes of systems.**

## 1 DATASET SPECIFIC HASHING

**Hashing is Central to Computer Systems.** Hashing is one of the core concepts in computer science; data structures and algorithms which use hashing exist in nearly every computer program. It's most ubiquitous use case, hash tables, is the standard way to access individual data items. They are used both for fast access to hot data in L1 cache across general purpose programs as well as for accessing colder data that lies outside of cache either in memory or on disk. For example, in relational database systems hash tables are used for joins and group by operations. Beyond hash tables, hashing is used in numerous other core parts of computer science such as filters [11], data partitioning [62], load balancing [49], and sketches [15, 30]. As a result of their many and important use cases, hashing is not only central within relational databases [65, 66] but acts as a core component of systems across compilers [3], file systems [70, 76], gaming [33], genomics [50], and more. This effect is depicted visually in Figure 1 where hashing is shown as the core design element used to build numerous fundamental operations, data structures, and algorithms (hash tables, filters, partitioning, etc.) which in turn are are core components of diverse systems.

**Hashing: Expensive at Scale.** Because hashing is so ubiquitous, it is a substantial portion of overall system cost. Google states that 2% of its total CPU usage and 5% of its total RAM at the company is spent on just one hash-based data structure, hash tables, in just one of the languages used, C++ [42]. Including other languages and other hash-based operations, the total CPU and memory usage spent on hashing overall is surely much higher. Meta makes similar statements, with developers stating that hash tables are such "a ubiquitous tool in computer science that even incremental improvements have large impact" [16]. Moving from large cloud infrastructure to particular applications, inside databases hash-based joins and aggregations are amongst the most expensive and used operators; as a concrete example they account for over 50% of total time on 17 of the 22 queries on the TPC-H benchmark for

Hyrise [28, 69]. Another example can be seen in compilers, where using hash tables in linking is a substantial part of program compilation costs in Visual Studio [3]. In addition to hash tables, hash-based filters are a core component of LSM-tree based key-value stores [35] and can be a core computational bottleneck depending on the workload [25, 78] and system tuning [22–24]. Similarly, hash-based sketches act as a key computational bottleneck in network switches [46]. These observations across diverse industries, systems and data structures demonstrate that: *despite numerous algorithmic and engineering advances, hash-based operations are still expensive because of the frequency and scale at which they are used.*

**Randomness vs. Performance.** To start drilling in at both the source of the problem and our solution we will next discuss the core mechanisms and trade-offs in hashing. A core component of all hash-based data structures and algorithms is the hash function itself. Hash functions having two primary goals. The first is to create uniformly random outputs for any number of input items. That is, the output should be jointly uniform as well as marginally uniform. The second is computational efficiency. While ideally both goals would be optimally achievable, they are practically at odds with each other. Thus a central question is how much randomness is needed from the hash function for the operation at hand.

**Guarantees Without Assumptions on the Data.** To get performance guarantees without assumptions on the data, all randomness needs to come from the hash function. The main way to define this property is by bounding the likelihood of collision for arbitrary input items. In universal hashing [17], one guarantees that for any two items $x, y$ and family of functions $H : \{0,1\}^n \rightarrow \{0,1\}^m$, the probability when choosing a random $h$ from $H$ of $h(x) = h(y)$ is $\leq \frac{1}{m}$. However, this is not enough randomness for many data structures [55, 57], and so an expanded idea of hash randomness is k-independence, which is that for any set of $k$ inputs $x_1, \ldots, x_k$, and $k$ outputs $y_1, \ldots, y_k$, the probability of $P(\cap_i h(x_i) = y_i) = m^{-k}$ [75]. Given this model, it becomes possible to provide guarantees with larger amounts of independence being more computationally expensive but providing better performance guarantees [55, 57, 75].

**Hashing in Practice.** In practice, systems designers avoid expensive k-independent hash functions and instead opt for hash functions which lack formal robustness guarantees but are faster to compute [7]. For instance, RocksDB uses xxHash [19], Google heavily uses CityHash, Wyhash, and FarmHash [60, 61, 73], and C++ compilers such as g++ often choose MurmurHash [1, 6].

Choosing fast hash functions without robustness guarantees is deemed 1) necessary because the computational performance of hashing is too important, and 2) appropriate because empirically the output of these fast hash functions appears as random as if it was created from a perfectly random hash function [56, 63, 64]. This phenomena is explained by pseudorandomness: If the *data itself is random enough*, then hash functions with weaker guarantees in terms of independence can be shown to perform in expectation nearly identically to those that are fully random [18, 48]. In other words, if we give up guaranteed randomness properties of hashing on *all datasets*, then we can use a fast hash function. Most hash functions perform well on most input data, and it takes careful manipulation of the input data to craft scenarios where commonly used hash functions fail.

**Problem Definition.** Having given the core concepts in state-of-the-art hashing, we can now restate the problem more concretely. Modern systems across diverse areas and industries utilize fast hash functions but without any guarantees. However, these fast hash functions are still not fast enough: they are still slow in that they occupy a large portion of total cost in all those systems. In this paper, we ask the following question:

*"Is it possible to improve on the speed of the best modern hash functions such that this brings significant end-to-end impact across diverse widely used hash-based operations, while at the same time maintaining and controlling their uniformity properties"?*

**The Solution: A Dataset-specific view of Hashing.** Our core intuition is to utilize the inherent randomness in the data in a controlled way. That is, if we know how random the input data is, we can use this observed randomness to create faster hash functions by doing just enough computation and data movement to create a sufficiently random output. Our key insight is that hash functions in state-of-the-art solutions are "fixed" in that they always do the same work regardless of the input. As such they end up doing more work than needed if data sources are already random enough. Our insight is to utilize such "surplus randomness" by adapting the hash function to the data to minimize computational cost.

Our solution, called *Entropy-Learned Hashing*, creates a tailored hash function for any given data source. It does so in two steps. First, it learns the amount of randomness in the data and exactly where this randomness appears. It does that by utilizing samples of past data items and queries to estimate the amount of randomness at specific subsets of bytes in data keys. In the second step, Entropy-Learned Hashing utilizes this learned randomness to choose the best subsets of bytes from the input keys that should be used in the hash function. These subsets are chosen to have just enough randomness for the task at hand. This results in hash functions that do just enough computation, while preserving the (approximate) uniformity of the hash function's output. For example, for a dataset with keys of length 120 bytes, if a consistent subset of bytes (e.g., bytes 3,7,9,12, and 15) is sufficiently random, Entropy-Learned Hashing creates a hash function that will use only this subset of bytes and as such it requires approximately only 1/24th the amount of computation.

**Constant-Time Hashing.** How much time it takes to hash input keys is now dependent on how random data is, but as we show in our experiments, most datasets possess enough randomness that just a small number of bytes of data are needed for hashing for most tasks. This divorces hashing time from key size, as adding more bytes to a key no longer adds to hash computation time. This makes hashing a sub-linear operation for Entropy-Learned Hashing, a large change with respect to traditional hash functions which have linear computational cost with key size. As a result, Entropy-Learned Hashing provides unbounded computational speedups as key sizes grow.

**Contributions.** Our contributions are as follows:

- *Entropy-Learned Hashing Formalization*: We introduce a new way to design hash functions that uses the entropy inside the data source to reduce the computation required by hash functions.
- *Optimization*: We show how to choose which bytes to hash given a collection of past queries and data items to analyze.

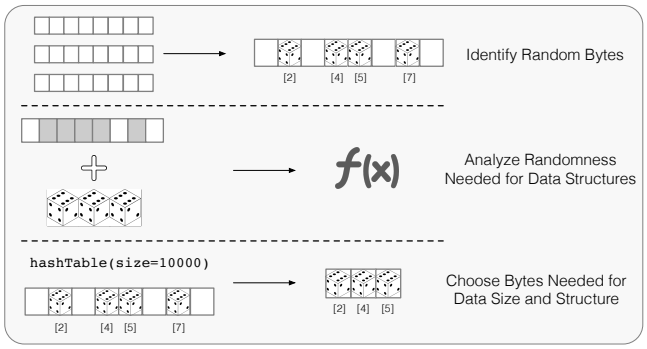



**Figure 2: The core steps in Entropy-Learned Hashing.**

- *Generalization*: We show how the entropy of partial-key hashes generalizes to data items outside the given sample of data.
- *Concrete Trade-offs*: We derive metric equations for three core use cases of Entropy-Learned Hashing: hash tables, Bloom filters, and data partitioning. This allows trading speed in hash computation for small changes in other metrics of interest such as the number of comparisons, FPR, and partition variance.
- *Computational Gains*: Comparing against state-of-the-art designs and implementations, Google's and Meta's hash tables, we show that Entropy-Learned Hashing provides higher throughput than traditional hashing. While this improvement is unbounded with respect to key size, for common medium-sized key types such as URLs, we show this improvement is up to 3.7× for hash tables, 4.0× for Bloom filters, and up to 14× for data partitioning.

The paper is curated to be self-contained with the most critical material and we also accompany it with an online appendix with detailed proofs and numerous additional experiments [34].

## 2 OVERVIEW & MODELING

We now move on with a detailed description of Entropy-Learned Hashing which will span the next three sections. In this section, we start with a more detailed overview as well as laying out the basics for notation and modeling which we use throughout the paper.

**Overview.** The goal of Entropy-Learned Hashing is to learn how much randomness is needed and to produce a hash function which does just enough work by controlling the input given to the hash function. To achieve this goal, Entropy-Learned Hashing looks for bytes which are highly random on input objects and passes just enough of these bytes to create a highly random output. Stated more formally, Entropy-Learned Hashing consists of creating a hash function $H'$ which is the composition of 1) a partial-key function $L$ which maps a key $x$ to any subkey of $x$ (including potentially the full key $x$), and 2) $H$, a traditional hash function. Our focus is on designing $L$, and $H$ can be any of the many well-engineered hash functions for full-keys.

In order to create the partial-key function $L$, Entropy-Learned Hashing uses three steps as shown in Figure 2. First, it analyzes the data source $x$ and identifies which bytes are highly random, and how much entropy can be expected from a choice of $L$ (Section 3). Second, it reasons about how $L$ affects data structure metrics (Section 4). Finally, it uses runtime information, such as the size of the desired Bloom filter or hash table or the number of partitions in partitioning to choose which bytes to use in $L$ (Section 5).

| Notation | Definition (filter, hash table, or load balancer) |
|---|---|
| $X, x$ | key stored in the filter or hash table |
| $H, h$ | hash function for filter or hash table |
| $Y, y$ | query key in filter or hash table |
| $m$ | size of filter (in bits), table (in slots), or # bins |
| $n$ | number of keys in filter or table |
| $K$ | set of keys |
| $S_{\|L}$ | multi-set of partial keys. Equal to $(K_{\|L}, z)$ |
| $K_{\|L}$ | Set of all partial keys. |
| $z$ | maps each key $x \in K_{\|L}$ to $\|L^{-1}(x)\|$. $z_x$ is used as shorthand for $z(x)$ throughout. |
| **Notation** | **Definition (hash table only)** |
| $\alpha$ | fill of hash table: $\frac{n}{m}$ |
| $P'$ | number of comparisons to find non-existing key |
| $P$ | average # of comparisons to retrieve a key in the dataset |

**Table 1: Notation used throughout the paper.**

**Notation.** The notation for all variables used is given in Table 1. Capital letters refer to either random variables or sets whereas lower case variables refer to fixed quantities. The new notation is because keys entered into $H$ are no longer unique. The set of keys $K$ contained in a hash-based data structure is broken down into the multi-set $S_{|L} = (K_{|L}, z)$. Here, $K_L$ is the set of all partial-keys (outputs of $L$ applied to keys in $K$), and $z$ maps each key in $K_L$ to the cardinality of its pre-image in $K$. For instance, if $L$ takes the first two characters of an input and $K = \{\text{dog, dot, cat, fan}\}$, then $K_{|L} = \{\text{"do", "ca", "fa"}\}$, $z(\text{"ca"}) = 1$, and $z(\text{"do"}) = 2$.

**Hash Function Model.** We assume that $H$ is ideally random, i.e. that for any distinct inputs $x_1, \ldots, x_n$, output range $[m] = \{1, \ldots, m\}$, and outputs $a_1, \ldots, a_n \in [m]$, we have

$$\mathbb{P}(H(x_1) = a_1, \ldots, H(x_n) = a_n) = \prod_{i=1}^{n} \mathbb{P}(H(x_i) = a_i) \quad = (\frac{1}{m})^n$$

We do not use k-independent hashing; as noted before and as shown again in our experiments, hash functions tend to perform empirically like their perfectly random counterparts. Moreover, most proofs using k-independent hashing give big-O guarantees but drop constant factors [48, 55, 57]. These constant factors are of significant importance for high performance hash functions.

**Source Model.** Conditioned on $L$ we assume that the partial-keys $L(X)$ are i.i.d. distributed because the main metrics for hash-based algorithms tend to be order-independent. For instance, whether keys are ordered $x_1, \ldots, x_n$ or in the reverse order $x_n, \ldots, x_1$, the slots filled in a linear probing hash table or the length of the linked lists in a separate chaining hash table are identical. Similar statements hold for the false positive rate of Bloom filters and the partitions produced by partitioning. Thus, even if the original source has a temporal nature that might be better modelled by a Markovian assumption, the marginal distribution over time is more important.

## 3 CREATING PARTIAL-KEY FUNCTIONS

The first step is to create the partial-key function $L$ which needs knowledge about the data we expect. In the case of fixed datasets, such as read-only indexes like those used in the levels of LSM-based key-value stores [54], this is the actual dataset. With updates, we need a sample of past data and queries.

**Metric for Partial-Key Hash Functions.** Partial-key functions have two metrics. The first is the number of bytes in their output, with fewer being better so that subsequent hash computation is faster. The second is the Rényi Entropy of order 2 of their output, also known as the collision entropy. For a given discrete random variable $X$, its Rényi Entropy of order 2 is $H_2(X) = -\log \sum_{i=1}^{n} p_i^2$ where $p_i$ is the probability that $X$ takes on the $i$th symbol in an alphabet $\mathcal{A} = \{s_1, \ldots s_n\}$. It draws its name from the fact that if $X_1, X_2$ are drawn i.i.d. from the same distribution as $X$, then $H_2(X) = -\log_2 \mathbb{P}(X_1 = X_2)$. We use collision probability to refer to $\mathcal{P}(X) = \mathbb{P}(X_1 = X_2)$ and mean Rényi Entropy of order 2 whenever we use the term entropy. For Entropy-Learned Hashing, Rényi Entropy tells us how likely collisions are to occur. The following lemma will be useful in our analysis:

Lemma 1. *Given $n$ i.i.d. samples from a distribution $X$, the number of observed collisions over the number of 2-combinations is an unbiased estimator of the collision probability for $X$. That is, if $n_i$ is the number of times a symbol $s_i$ appears in the sample, then we have*

$$\mathbb{E}[\sum_i \frac{n_i^{\underline{2}}}{2}] = \frac{n^{\underline{2}}}{2}\mathcal{P}(X)$$

*where $x^{\underline{2}} = x(x-1)$ is the 2nd falling power. Equivalently,*

$$\mathbb{E}[\sum_i \frac{n_i^{\underline{2}}}{2}] = \frac{n^{\underline{2}}}{2}2^{-H_2(X)}$$

Proof. There are $\binom{n}{2}$ possible 2-combinations in $n$ samples, each of which can produce a collision. The probability of collision is $2^{-H_2(X)}$ and so the expected number of collisions is $\binom{n}{2}\mathcal{P}(X)$. □

**Optimization: Selecting the Bytes to Hash.** The goal is to optimize our two metrics on our optimization set, which is either the fixed dataset or a training set of a sample of prior data items. Since our two metrics are at odds, the goal is to find an optimal Pareto frontier establishing for each $k = 1, 2, 3, \ldots$, what set of $k$ bytes from our full-key input produces the most entropy.

Insight into this problem, as well as potential solutions, can be found by analyzing the similar problem for maximizing Shannon entropy (equivalently, Rényi entropy of order 1). In particular, for Shannon entropy selecting the best subset of size $k$ of random variables from amongst $n$ random variables is known to be NP-hard [39], suggesting that an optimal solution for Rényi entropy is likely computationally difficult. However, the greedy algorithm, described in detail below, is known to provide a $1-\frac{1}{e}$ approximation to the best possible solution for Shannon Entropy because Shannon Entropy is submodular [51]. Additionally, real-life applications of the greedy algorithm tend to get solutions which are close to the optimal solution [9]. Inspired by this success and by the connectedness of Rényi and Shannon entropy, we use the greedy algorithm to optimize Rényi entropy on our training set.

We start by using a dummy hash which reads zero bytes of the data items. Then, we continually add new bytes to the partial key function $L$ in a way that decreases the number of collisions the most on the training data. After each new chunk of bytes, we record the entropy (either on the fixed dataset or on a validation dataset if data is not fixed) and repeat the process. We stop when $L$ has no collisions on the training data, and note that at each iteration of the algorithm we need only to look at data items which are not unique given previous bytes chosen for $L$, reducing algorithm runtime substantially (items that are not equal on a subset of bytes cannot be equal on a larger subset). At the end, we have a sequence of partial-key functions which are our solutions for $k = 1, 2, 3, \ldots$ bytes, with higher $k$ meaning more input bytes are read but also monotonically increasing the entropy of the output.

**Algorithm 1** ChooseBytes

**Input:** *train_data*: either data items or sample of past data items
**Input:** *test_data*: data to check entropy on (if not for fixed dataset)
positions = vector()
entropies = vector()
max_len = maximum length of any data item
**while** not all partial keys unique **do**
  positions.push_back(NextByte(data,max_len,positions))
  entropies.push_back(EstimateEntropy(test_data, positions))
  data = NonUnique(data, positions)
**return** positions, entropies

**Algorithm 2** NextByte

**Input:** *data*: either data items or sample of past data items
**Input:** *max_len*: maximum length item in dataset
**Input:** *past_bytes*: past bytes chosen
min_coll, min_i = ∞, −1 // track of min # collisions, most entropic byte
**for** $i = 0$ **to** max_len **do**
  count_table, num_coll = {}, 0
  **for** $j = 0$ **to** len(data) **do**
    p_key = $data[j]$ using (past_bytes, i) // form partial-key
    p_key = (len(data[j]), p_key) // length is always part of partial-key
    count_table[p_key] += 1 // increment count partial-key
    num_coll += (count_table[p_key] - 1) // add collisions (if any)
    **if** num_coll < min_coll **then**
      min_coll, min_i = num_coll, i // update best byte
**return** min_coll, min_i

Algorithms 1 and 2 give (simplified) pseudocode for this procedure. Additionally, Figure 4 shows example output from the procedure. While for simplicity Algorithm 1 is shown choosing 1 byte at a time, our implementation chooses 4 or 8 bytes at a time. This is because most modern hash functions which come after $L$ operate one word of data at a time. In addition, we limit the maximum byte being chosen for partial-key hashing so that 90% of data items are under that data size. In the end, $H'$ looks as follows:

```
if len(x) > last byte used in L:
    return H(L(x))
else
    return H(x)
```

Because we designed $L$ so that almost all keys satisfy the first if statement, this makes the full hash function have predictable branching statements. This initial if statement is also dropped if the keys are of fixed length. The result, when $L$ is tightly integrated into the hash function $H$, is that $H'$ has predictable branches and a small instruction count on average.

**Evaluating the Resulting Entropy.** To make decisions on how many bytes are needed, we need an estimate of the entropy of $L(X)$. When data is fixed, we use the training set as a ground truth value

for the entropy. When generalization to new data is needed, we use separate validation data.

To estimate the entropy of $L(X)$, we compute the empirical collision probability on the validation set $V$ by 1) computing $L(x)$ for each $x$ in $V$, 2) counting the number of collisions, and then 3) dividing this by $\frac{v^2}{2}$ where $v$ is the number of items in $V$. From Lemma 1, this gives an unbiased estimate of the collision probability. To get an estimate $\hat{H}_2$ of the entropy, we take the negative log of this number.

Given this estimator, the natural question to ask is "how many samples are needed?". The techniques of [4, 53] use the birthday paradox to answer this question; namely, if we want to say that the entropy is at least some value $H_2$ with confidence, we need $O(2^{H_2/2})$ samples. As we will show in Section 4, data structures or algorithms storing $n$ elements will generally need $H_2$ to grow at a rate of $\log_2 n$, suggesting $O(n^{1/2})$ samples is enough to say with probability approaching 1 whether or not $L(X)$ has enough entropy for a given task. Giving a concrete example, when using $v$ validation samples a 99% confidence estimator for the entropy is: $H_2 \geq \min \begin{cases} \hat{H}_2 - 2 \\ \log_2 \frac{v^2}{400^2} \end{cases}$ with probability 0.99. Thus if our data structure needs entropy $H_2 = \log_2 n$, setting $v > 400\sqrt{n}$ is enough validation samples to say with high probability whether or not $L(X)$ has the required entropy [1]. More details can be seen on this analysis in the technical report [34].

The most important takeaways are that the number of validation samples needed both varies with the data size and also grows slowly with the data size. Thus, when we want to use Entropy-Learned hashing on small data, the sample can be small because we only need to make sure it has just enough entropy. When the data is large, the number of samples needed grows but much more slowly than the data size.

# 4 CONNECTING ENTROPY TO DATA STRUCTURE PERFORMANCE

The next step in Entropy-Learned Hashing is understanding the entropy needed for a given system task, i.e., a data structure or algorithm used in a system. As Figure 1 shows, hashing is used in a range of diverse systems to implement data structures and algorithms for various complex operations. We study specifically the entropy needed by three of the mostly widely used tasks, namely:

(1) **Hash tables** which are the default way to access data by equality, and which are widely used across general purpose programs including relational systems and key-value stores.
(2) **Bloom filters** which are used to reduce accesses to a set and are used in databases to reduce the costs of joins in OLAP systems as well as point queries in key-value stores.
(3) **Partitioning** which is a core step in numerous algorithms.

Each of these tasks has multiple metrics of interest, including: CPU cost, memory footprint, throughput, false positive rate, and much more. The three hash-based operations above present a diverse set of expressions of these metrics. For example, Bloom filters have small memory footprint compared to the other components, while they all have drastically different characteristics in terms of output write patterns which affects the overall throughput.

By creating cheaper to compute hash functions we improve the computational efficiency; what is left to show is that the small increase in expected collision probability does not result in significant degradation on other metrics. For hash tables, the metric of interest for performance is the number of comparisons needed to retrieve a key. For Bloom filters, it is the false positive rate and for Partitioning the variance of the distribution of data amongst bins.

There are two takeaways from the analysis in this section. The first is that we can argue formally about the needed entropy from partial-keys for data structures to behave as desired. This allows us to design Entropy-Learned hash functions which bring end-to-end performance benefits. Second, the analysis shows that across all tasks, Hash tables, Bloom filters, and Partitioning, the needed amount of Renyi entropy in $L(X)$ is approximately $\log_2 n$ plus a constant $c$. Thus, for a fixed dataset size, hashing needs only a constant amount of randomness. If some fixed set of bytes provides this amount of randomness, then hashing only need look at these bytes and its computation becomes independent of key size in that adding more bytes to the key does not increase hash computation time. Additionally, the dependence on $n$ reaffirms our central thesis and further clarifies where Entropy-Learned Hashing is most useful: for large (hence random) objects or small datasets state-of-the-art hash functions do more work than necessary. The value of $c$ depends on how much collisions affect a data structure; for instance, hash collisions in Bloom filters produce a certain false positive and so this has a high value of $c$, whereas for hash tables a collision produces an extra comparison which is more tolerable and so $c$ is lower.

## 4.1 Hash Tables

Two prototypical designs of hash tables are separate chaining and linear probing [20]. Separate chaining stores an array of linked lists. To query for an item, separate chaining hash tables 1) perform a hash calculation to get a slot $a$ between 0 and $m-1$ and then 2) traverse the linked list at slot $a$ until either the key is found or the end of the list is reached (the key is not present). Linear probing stores an array of keys and queries the table by 1) performing a hash calculation to get an initial slot $a$, and then 2) traversing the array in sequential order until either the key is found, or until an empty slot is found (the key is not present). Separate chaining tables are easier to manage and analyze because collisions only matter for the same slot, however they have poor data locality because of many pointer traversals and require extra space for the many pointers. In contrast, linear probing offers better performance but is more difficult to analyze and manage because of complex dependencies between hash values.

### 4.1.1 *Hash Tables: Separate Chaining.*

**Fixed Data.** We first analyze separate chaining hash tables when the data is known which is an important class of indexed data. We then show this analysis translates from known data to random data.

Given $S_{|L} = (K_{|L}, z)$, when querying for an item $y$ not in $K$, the expected number of comparisons $P'$ is

$$\mathbb{E}[P'|y] = z_y + \frac{n - z_y}{m} \approx z_y + \alpha$$

[1] We note this leading constant seems higher than what is necessary in practice, suggesting possible further improvements in the analysis of this estimator.

This is because the (likely 0) $z_y$ items which have the same partial key for sure are in the same slot, and the other $n - z_y$ items have $1/m$ chance of being in the same slot. This cost of querying for a missing key is also equal to the cost of adding a new item into the hash table, and this relationship holds true for linear probing as well. This is because additions first verify the item is missing and then put the item into the first empty slot they find.

By the same logic, querying for a key $x$ in $K$ costs $1 + \frac{1}{2}(z_x - 1 + \frac{n-z_x}{m})$ comparisons on average. The leading 1 is because the query key for sure compares with itself, and the second term is $1/2$ times the expected number of items in the same slot as $x$. Summing across all data, the average cost $P$ of querying for a key satisfies:

$$\mathbb{E}[P] \leq 1 + \frac{1}{2}\alpha + \frac{1}{2}\sum_{x \in K_{|L}} \frac{z_x^2}{n}$$

**Random Data.** When generalizing partial-key hashing to unseen (random) data, the above equations can be viewed as conditional expectations where we condition on the data. By using Adam's Law, i.e. $E[X] = E[E[X|Y]]$, we can average over the possible produced datasets given by the random data. Using the union bound and Lemma 1, the expected cost of querying for a missing key and the average cost for querying for a key satisfy

$$\mathbb{E}[P'] \leq \alpha + n2^{-H_2(L(X))} \tag{1}$$

$$\mathbb{E}[P] \leq 1 + \frac{1}{2}\alpha + \frac{1}{2}(n-1)2^{-H_2(L(X))} \tag{2}$$

**Comparison with Full-Key Hashing.** For full key hashing, the corresponding costs for querying for a missing key and the average cost to query for a key are

$$\mathbb{E}[P'] = \alpha$$

$$\mathbb{E}[P] = 1 + \frac{1}{2}\frac{n-1}{m} \approx 1 + \frac{1}{2}\alpha$$

This shows the tradeoff between partial key hashing and full key hashing. The number of comparisons is lower for full-key hashing, but this advantage goes exponentially fast to 0 as the entropy of the partial key hash increases. At the same time, the partial-key hash is significantly cheaper to compute.

Looking at the required relationship between $n$ and the needed entropy of the input sub-keys further clarifies when and why partial-key hashing is useful. When $H_2(L(X)) > \log_2 n$, the number of extra comparisons needed drops below 1 and continues to drop exponentially fast with more entropy. Since hashing objects is more expensive than comparing them, this point represents near definite savings; the hash computation for the table is much faster while the work after the hash function is nearly the same.

#### 4.1.2 Hash Tables: Linear Probing.

Because of the complex dependencies between hash values and collisions, linear probing is significantly more complicated to analyze resulting in lengthier proofs. We provide a high level overview of the results while all detailed proofs can be found in the Appendices. We start with full-key hashing. We analyze the expected length of a full chain $T$ for a new item added to the hash table. The chain includes the empty position on a chain's right side but not on its left side. Figure 3 shows an example.

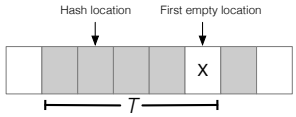



**Figure 3: Example of a linear probing chain.**

**Full-Key Hashing.** Appendix Section A provides a novel analysis of linear probing showing that the expected length of $T$ satisfies $E[T] = Q_1(m,n)$ where $x^{\underline{k}}$ is the $k$-th falling power and $Q_i(m,n) = \sum_{k \geq 0} \binom{k+i}{i} \frac{n^{\underline{k}}}{m^k}$. For a new item, each location in a probe chain is equally likely as a hash location and so the expected probe cost given $T$ is $E[P'|T] = \frac{1}{2} + \frac{1}{2}T$. Using Adam's law, it follows that

$$E[P'] = \frac{1}{2}(1 + Q_1(m,n)) \leq \frac{1}{2}(1 + \frac{1}{(1-\alpha)^2})$$

which matches the known equations given by Knuth in [38].

The average cost to query a key is then equal to the average cost to insert each key. Since the insertion cost $E[P']$ depends on $n$, we use $P'_i$ to denote the cost when there are $i$ keys in the table. The average cost to query a key is then

$$E[P] = \frac{1}{n}\sum_{i=0}^{n-1} E[P'_i] = \frac{1}{2}(1 + Q_0(m, n-1)) \leq \frac{1}{2}(1 + \frac{1}{1-\alpha})$$

**Partial-Key Hashing: Fixed Data.** When given $S_{|L} = (K_{|L}, z)$, the expected length of the probe chain $T$ depends on the number of partial key matches for the inserted key $y$, and satisfies

$$E[T] \leq Q_1(m,n) + z_y Q_0(m,n) + \sum_{x \neq y} \frac{z_x^2}{m} Q_1(m,n)$$

$$\leq \frac{1}{(1-\alpha)^2} + \frac{z_y}{1-\alpha} + \sum_{x \neq y} \frac{z_x^2}{m(1-\alpha)^2}$$

When the new key is unique, the most common scenario when $H_2(L(X))$ is high, each location in the probe chain is equally likely and so $E[P'|T] = \frac{1}{2} + \frac{1}{2}T$. However, when the new key is not unique, each position in the chain is no longer equally likely. Thus we make the worst case assumption that it is at the end of the probe chain.

$$E[P'] \leq \begin{cases} \frac{1}{2}\left(1 + \frac{1}{(1-\alpha)^2} + \sum_{x \neq y} \frac{z_x^2}{m(1-\alpha)^2}\right) & \text{if } z_y = 0 \\ \frac{z_y}{1-\alpha} + \frac{1}{(1-\alpha)^2} + \sum_{x \neq y} \frac{z_x^2}{m(1-\alpha)^2} & \text{if } z_y > 0 \end{cases} \tag{3}$$

When translating from $P'$ to $P$, we again have that $E[P] = \sum_{i=0}^{n-1} E[P'_i]$. Since the cost of inserting each key is no longer the same, there is the question of how to evaluate this expression. Here, we make use of a fact first noticed in [59], that the average cost of querying is equal for any order in which the items are inserted. Thus, in evaluating $E[P] = \sum_{i=0}^{n-1} E[P'_i]$, we may choose the insertion order of the items. Inserting all keys with non-unique partial-keys first and then inserting all keys with unique partial-keys gives the following bound for $E[P]$.

$$\begin{aligned} E[P] &\leq \frac{n-d}{2n} + \frac{1}{2}Q_0(m,n) + \frac{c}{m}Q_0(m,n) + \frac{c+d}{2n}Q_0(m,d) \\ &\approx \frac{1}{2}(1 + \frac{1}{1-\alpha}) + \frac{c}{n} + \frac{c}{m}\frac{1}{1-\alpha} \\ &\leq (\frac{1}{2} + \frac{c}{n})(1 + \frac{1}{1-\alpha}) \end{aligned} \tag{4}$$

We use $c = \sum_x z_x^2$ for the number of collisions and $d = \sum_{x:z_x \geq 2} z_x$ as the number of items that are duplicated keys. The above approximation assumes that $d/m$ is small, which is the case whenever most keys are unique. This holds true with probability near 1 if entropy is sufficiently large.

**Random Data.** Using equations (3), (4), and Lemma 1 as well as Adam's Law, we have

$$E[P'] \leq \frac{1}{2}(1 + \frac{1}{(1-\alpha)^2}) + n2^{-H_2(L(X))}\frac{3}{2(1-\alpha)^2} \quad (5)$$

$$E[P] \leq \frac{1}{2}(1 + \frac{1}{1-\alpha}) + n2^{-H_2(L(X))}(1 + \frac{1}{1-\alpha}) \quad (6)$$

**Comparison With Full-Key Hashing.** The tradeoffs between partial-key hashing and full-key hashing are similar to separate chaining. Again, we have a slight increase in comparisons as a trade-off for significantly faster hash function evaluation. The expected number of comparisons again drops exponentially fast with the source entropy and $H_2(L(X))$ needs only to be in the same order of magnitude as $\log_2(n)$ for the extra needed comparisons to be small. Thus, as before, partial-key hashing makes the work of computing hash functions significantly cheaper while the work after the hash function is near identical, producing a net performance benefit.

## 4.2 Bloom Filters

For Bloom filters, the central trade-off is between the speed of the filter and the false positive rate (FPR) of the filter. As the number of bytes given as input to the hash becomes smaller, hashing becomes faster but there is a greater possibility of a partial-key collision, creating a certain false positive.

More formally, let $FPR(m, n, H)$ denote the false positive rate of a Bloom Filter using $m$ bits, storing $n$ items and using a hash function $H$. For a Bloom Filter using partial-key hash $H' = H \circ L$, its number of set bits is a function of the number of distinct items fed to $H$. If no keys collide on $L$, then it becomes a traditional Bloom Filter storing $n$ items and using $H$. If there are $n' < n$ distinct items after $L$, then the resulting filter structure has the same number of set bits as one containing $n'$ items. So for query key $y \notin K_L$, it has a false positive rate of $FPR(m, n', H)$, whereas if $y \in K_L$ it has a false positive rate of 1. It follows that our Bloom Filter using $h'$ has exactly the following false positive rate:

$$FPR(m, n, H') = \mathbb{P}(Y_{|L} \in K_L) + FPR(m, n', H) \quad (7)$$

The second term is less than $FPR(m, n, H)$ as Bloom Filters' false positive rates increase with the number of items stored. If keys and non-keys are very different conditioned on the set of bytes $L$, then it is possible to make the FPR less than that of a standard Bloom filter by having $n' << n$ and $\mathbb{P}(Y_{|L} \in K_L) \approx 0$. However, we will generally ignore this case and focus on the case where keys and non-keys have the same distribution conditioned on $L$. In this case, a convenient bound for (7) is

$$FPR(m, n, H') \leq \mathbb{P}(Y_{|L} \in K_L) + FPR(m, n, H) \quad (8)$$

which is the FPR of a standard Bloom filter plus the probability that the query key matches some item in the key set on the bytes $L$.

Using the union bound, equation (8) translates to:

$$FPR(m, n, h') \leq n2^{-H_2(L(X))} + FPR(m, n, h) \quad (9)$$

**Comparison With Full-Key Hashing.** The above analysis reaffirms the central takeaway of our analysis of hash tables; the entropy of the dataset needs to be on the order of $\log_2 n$. For Bloom filters, a reasonable additional goal is that the increase in FPR be no more than some chosen $\varepsilon$. In this case, we need $H_2(L(X)) > \log_2 n + \log_2(1/\varepsilon)$. So an additional entropy term is needed to say that collisions are very rare for new partial-keys. As we show in our experiments, datasets often have this surplus entropy and so the Bloom Filter becomes significantly faster without suffering any false positive rate increase.

## 4.3 Partitioning & Load Balancing

With Partitioning the goal is to distribute $n$ items, e.g., tuples or computational tasks, to a set of $m$ bins. Here, we characterize how even this allocation is by analyzing the variance of the number of items assigned to each bin when each input key is unique. At lower variances, each bin is distributed closely around the average number of items $n/m$ whereas higher variance suggests the bins are highly uneven. One important challenge comes when keys are skewed and heavy hitters exist. While challenging, the unevenness comes from the existence of heavy hitters rather than the quality of the hash function, and so we focus on the hash quality by considering the partitioning of all unique items.

**Full-Key Hashing.** With full-key hashing, the variance of each bin is the variance of a binomial with $n$ balls each with probability $1/m$. Thus for a specific bin, its number of assigned objects $Y$ has $Var(Y) = \frac{n}{m} - \frac{n}{m^2}$.

**Partial-Key Hashing: Fixed Data.** The probability of each key in $K_L$ being assigned to a specific bin is distributed as an independent Bernoulli trial with probability $\frac{1}{m}$. Letting $\mathbf{1}_{H(x)=i}$ be the event that $x$ was hashed to bin $i$, the variance of the number of objects $Y$ assigned to bin $i$ is

$$Var(Y|K_{|L}) = Var(\sum_{x \in K_{|L}} z_x \mathbf{1}_{H(x)=i}) = (n + \sum_{x \in K_{|L}} z_x^2)(\frac{1}{m} - \frac{1}{m^2})$$

**Partial-Key Hashing: Random Data.** For random data, we use the same conditioning arguments as before. Using Eve's Law, i.e. $Var(Y) = \mathbb{E}[Var(Y|K_L)] + Var(\mathbb{E}[Y|K_L])$, we can calculate the variance on random data. First, we note that for any set $K_L$, the value of $E[Y|K_L]$ is $n/m$ by the randomness of the hash function (each bin is equally likely to contain any item). Thus $Var(\mathbb{E}[Y|K_L]) = 0$ and again using Lemma 1, we have

$$Var(Y) \leq (1 + n2^{-H_2(L(X))})(\frac{n}{m} - \frac{n}{m^2}) \quad (10)$$

**Comparison With Full-Key Hashing.** As before, $H_2 > \log_2 n$ is enough for partial-key hashing to have similar variance to full-key hashing in terms of absolute terms. Thus, as in prior cases, once $H_2 > \log_2 n$ we have faster computation in terms of partitioning without sacrificing on the quality of our partitioning.

An important secondary argument for load balancing is whether we care about the absolute deviation from the mean or the percentage deviation away from the mean. While the absolute variance grows with $n$, the relative standard deviation, i.e. the standard deviation over the mean, of the bins decreases with $n$ so that it becomes less and less likely that some bin has $x\%$ more than its expectation.

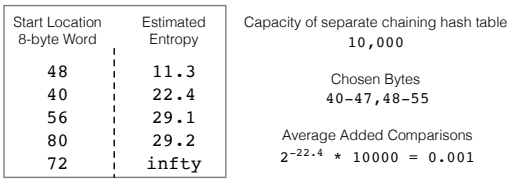



**Figure 4: The amount of bytes needed is based on the data and the current data structure capacity.**

In particular, the relative standard deviation is less than

$$\sqrt{\frac{m}{n}}\sqrt{1+n2^{-H_2(L(X))}} \approx \sqrt{m2^{-H_2(L(X))}} \tag{11}$$

Since the expected distance from the mean for a binomial is dominated by its standard deviation [12], the above statement actually says that a bin's expected proportional deviation away from its mean is less than (11). So for instance, if we want a partition to be within 5% of its mean on average, we can achieve this by having $H_2 \geq 2\log_2 \frac{1}{0.05} + \log_2 m$.

Thus partitioning and load balancing have two regimes with regards to Entropy-Learned Hashing. When small absolute variance is required, we need $H_2(L(X)) > \log_2 n$; however, when $n$ is large and we are simply interested in bins being relatively similar sizes, we can let $H_2(L(X))$ be greater than $\log_2 m$ plus a small constant, where the constant controls how much deviation is allowed.

## 5 RUNTIME INFRASTRUCTURE

Section 3 showed how to estimate the entropy of datasets when conditioned on partial-keys and Section 4 showed how much entropy is needed for important hashing-based tasks. This section brings everything together by explaining how to utilize Entropy-Learned Hashing at run time: namely, given a hash-based task and analysis of a dataset, choose the Entropy-Learned Hash function to have just enough randomness. Additionally, this section covers runtime infrastructure related to robustness so that Entropy-Learned Hashing retains the trustworthiness of traditional hash data structures.

**Creating Hash Tables.** Hash tables have a maximum capacity beyond which they need to rehash the stored items into a new larger table. This keeps the load factor low and therefore query times low. For Entropy-Learned Hashing, we use this maximum capacity before rehashing to decide $L$. In particular, for separate chaining hash tables, we choose $L$ such that $H_2(L(X)) > \log_2 n + 1$, where $n$ is the maximum number of items the current table will hold before rehashing. For linear probing hash tables, we choose $L$ so that $H_2(L(X)) > \log_2 n + \log_2 5$. Both values are chosen based on the equations governing the number of comparisons, i.e. equations (1), (2), (3), and (4), and make sure the number of comparisons executed using partial-key hashing and full-key hashing are similar. An example of how the current capacity is used to choose $L$ is shown in Figure 4, where an initial table with capacity 1000 uses just the 8-byte word at location 48 to hash keys.

As the capacity of a hash table changes (as new items are inserted), a rehash is triggered causing each item to be reinserted. Entropy-Learned Hash tables uses this opportunity to change the hash function; for instance, when key 1001 is inserted into the hash table from Figure 4, a rehash is triggered causing the table to grow. If the new capacity is above $2^{11.3} = 2521$, the partial-key function adds another word to increase entropy to the required amount. As a result, the hash table maintains just the right amount of entropy needed throughout its life cycle, using as cheap a hash function as possible without adding substantial extra collisions.

**Bloom Filters.** Bloom Filters need an estimate on the number of items they will hold before their creation. This is because, without access to their base items, they have no access to grow the number of bits being used. While there are techniques around this [5], these come with space and computation tradeoffs and it remains true that standard Bloom filters need an up-front estimate of the number of data items. For Entropy-Learned Hashing, this makes it simple to choose the hash function. Given a maximum number of items $n$ and an allowable added FPR of $\varepsilon$, we set the partial-key hash function to have entropy $H_2(L(X)) > \log_2 n + \log_2(1/\varepsilon)$.

**Partitioning.** For partitioning we require an estimate on the maximum number of items to be partitioned. We also need user input on how even they want partitions to be. If absolute variance is of primary importance (so that partitions are unlikely to vary by more than some # of tuples regardless of partition size), then setting $H_2(L(X)) > \log_2 n + c$ assures that variance is no more than $(1 + 2^{-c})$ times its usual amount. The default value of $c$ which we use is 3. When relative variance is more important, and users need partitions to be roughly even (i.e. within 100c% of each other's size), we set $H_2(L(X)) > \log_2 m - 2\log_2 c$ as dictated by equation (11). We use $c = 0.05$ by default so that partitions are expected to be within 5% of their expected size.

**Robustness.** While Entropy-Learned Hashing makes only weak assumptions, namely that data which are somewhat random remain somewhat random, it recovers good performance quickly when assumptions are violated. Entropy-Learned Hashing is the most robust for hash tables. This is for multiple reasons, namely: 1) if collisions are as expected on items in the dataset, queries for both keys in the data and not in the data return quickly (Section 4), 2) hash tables can monitor collisions during insertions with little overhead, and 3) rehashing is an acceptable operation in hash tables by default (it occurs in all standard hash table libraries). This third point is the most key, and Entropy-Learned Hashing can rehash hash tables if collisions ever deviate from what is expected, falling back to full-key hashing if needed. For Bloom filters, their # of set bits concentrates sharply around their expected value [14], and this fact is used during construction of Entropy-Learned Bloom filters to validate that the data items fit the expected level of randomness. However, if they do not, or if queries are substantially different than the inserted items, the filter must be rebuilt. For partitioning, the cost of overloaded bins depends on the context, but for many contexts, such as in-memory radix partitioning, this can be solved by dividing the one or two overloaded bins into multiple bins. Appendix Section B covers robustness in more detail.

## 6 EXPERIMENTAL EVALUATION

We now demonstrate that, by identifying and utilizing surplus randomness in data, Entropy-Learned Hashing brings critical performance benefits against the top hash functions used at scale today by Google and Meta and across a diverse set of hash-based core components of modern systems.

Our experimental evaluation consists of three parts. The first part, which contains the bulk of our experiments, shows that Entropy-Learned Hashing produces sizable benefits of up to 3.7×, 4.0×, and 14× for common medium-sized key types such as URLs and text data. The second part of our experimental section covers benefits from Entropy-Learned Hashing on large keys such as those that would appear in deduplicating file blocks, with speedups of several orders of magnitude. Finally, we cover training time for Entropy-Learned Hashing and present the run times for applying the greedy algorithm to select bytes to hash.

## 6.1 Setup and Methodology

**Data Structures and Operations.** We use a diverse set of data structures and operations to apply Entropy-Learned Hashing: we test with Hash tables, Bloom filters, and Partitioning.

For **hash tables**, we compare against Google's hardware-efficient linear-probing hash table implementation, SwissTable [32, 42]. This is the default hash table used in C++ throughout all Google operations, and has been heavily optimized as a result of the large computational footprint of hash tables at Google. A particular implementation note for SwissTable is that it first does linear probing into an array of tag bits (8 bits per key) to see if chosen bits from hash values match, and only if they do, compares the full items. This means probing for keys not in the table is cheaper than probing for keys stored in the table. We also compared against F14, the default hash table used at Facebook [16]. The results are extremely similar and so we include only results with SwissTable.

For **Bloom Filters**, we implemented register blocked Bloom filters from [43]. To cut down hashing time, and thus to be conservative with respect to our benefits, we used a variant of double hashing wherein we compute one 64 bit hash function, split it into two 32-bit hash values, and then use these as the inputs to double hashing [37]. We also utilize the techniques for fast modulo reduction by multiplication from [68].

For **partitioning**, many of the techniques devised by database research such as software write buffers [74] and non-temporal stores [10] do not apply to large data types or variable length data types. Thus our partitioning is a simple for loop that computes hash values and writes out data directly to a partition.

**Base Hash Functions.** We use three state-of-the-art hash functions. For hash tables, we use wyhash, which is one of the two default options used in SwissTable. We use both the version contained in SwissTable as well as the most recent optimized version of wyhash given directly by the author [73]. For Bloom filters we use xxh3, which is used widely at Facebook and is the default for the Bloom filters in RocksDB [19]. For partitioning we use the implementation of CRC32 from the OLAP database Clickhouse [77].

**Implementation.** We modify each of the three base hash functions. We maintain their basic interface (input is an array of bytes plus a key length), and tightly integrate Entropy-Learned Hashing. Thus there is Entropy-Learned xxh3, Entropy-Learned wyhash, and Entropy-Learned CRC32. The bytes chosen to hash are selected at hash function construction and stored in a const array. The functions read from $data[locations[i]]$ instead of $data[i]$, and we use templates to generate efficient code for partial-key hash functions using 1,2,3,4,.. words. These templates modify the initial function to reduce branching statements because of the known length of the partial-key. All implementation is in C++. All experiments for hash-based tasks are in-memory since hash-based tasks typically run in-memory. For example, a hash table should always fit in memory to get good performance while a Bloom filter will also typically reside in memory to protect from expensive disk access. Thus such structures are both created and utilized in memory. When disk is involved, the CPU cost of hashing is typically not highly visible in terms of operational latency unless very fast disk devices such as SSDs are used (although CPU usage is still reduced).

| Processor | Intel Xeon E7-4820 v2 |
|---|---|
| #sockets | 4 |
| #cores per socket | 8 |
| Hyper-threading | 2-way |
| Turbo-boost | Off |
| Clock speed | 2.00GHz |
| L1I / L1D (per core) | 32KB / 32KB |
| L2 (per core) | 256KB |
| L3 (shared) | 16MB |
| Memory | 1TB |

**Table 2: Server Parameters**

| Dataset name | Avg. key length | # keys |
|---|---|---|
| UUID | 36 | 100K |
| Wikipedia | 129 | 22K |
| Wiki | 22 | 99K |
| HN URLs | 75 | 247K |
| Google URLs | 81 | 1.2M |

**Table 3: Real-world data.**

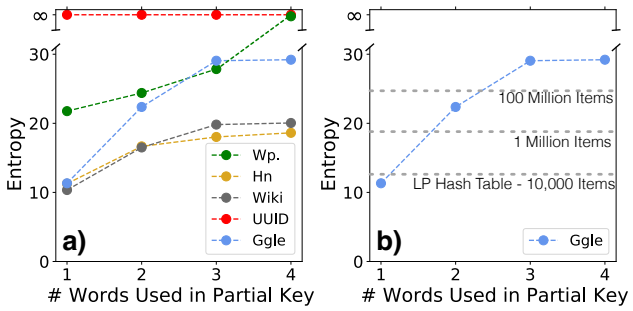



**Figure 5: The entropy of a dataset grows quickly with the amount of words being hashed. By 4 words, most datasets support data structures with millions of elements.**

**Datasets.** We use five real-world datasets for experimentation. Two datasets consist of URLs, with one containing the URLs of stored Google Landmarks and the other all URLs posted to Hacker News during 2015 [2, 52]. The other three, UUID, Wikipedia, and Wiki, are database columns taken from a recent research study [13]. They contain universally unique identifiers, sampled text from Wikipedia, and Wikipedia entry titles respectively. Table 3 presents the number of items and average key length for each real-world dataset. In addition, we use synthetic data to have finer control over key size and data size. Section 6.3 uses 80 byte keys with bytes 32-39 drawn randomly from the alphabet (26 possible values), and all other bytes constant. Section 6.6 uses 8KB keys with each byte ideally random.

**Experimental Setup.** We use an Intel Ivy Bridge server. Table 2 summarizes the server parameters. We use Debian GNU/Linux 10 operating system. Data structures are queried for a warmup phase before timing and input keys for queries are in cache. We pin the thread to a particular core and locally allocate memory. We use Intel VTune's uarch-exploration [36] for performing hardware-level time breakdown and Linux perf [45] for performing memory-level parallelism tests and software-level time breakdown.

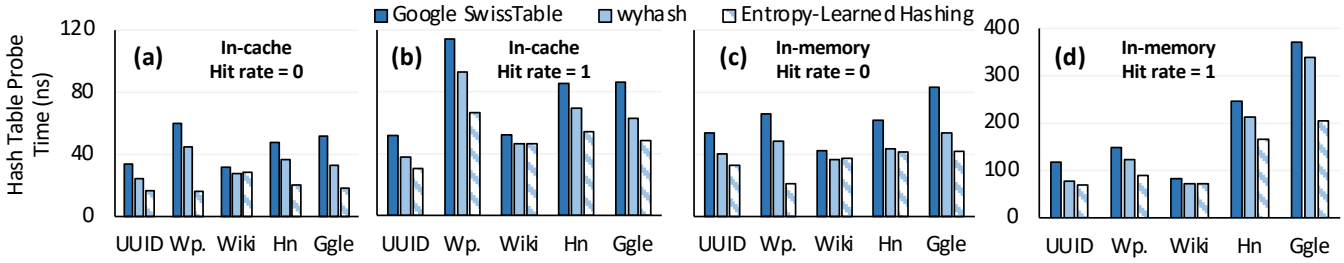



**Figure 6: Entropy-Learned Hashing reduces probe times for hash tables across datasets, data sizes, and hit rates.**

## 6.2 Number of Words vs. Entropy

Before demonstrating performance results, we first make the idea of surplus randomness more concrete with examples from real data. We show that for many datasets with medium-sized keys, good hashing properties can be achieved for data structures with millions of elements while hashing only parts of the keys. We divide each dataset in Table 3 in half. We use the first part to choose which bytes to hash in a greedy manner as described in Section 3. This produces an ordered list of bytes (or words) to choose. Choosing more bytes from the list produces a partial-key function providing more entropy. We use the second half of the dataset to get an unbiased estimate of the entropy for each combination of bytes as described in Section 3. Figure 5a shows that the entropy of the result of the partial-key function increases for all datasets with the number of words included. We see that by 3 words being included all datasets have an entropy of at least 18, and 3 of the 5 have entropies above 25. For Wikipedia and UUID, infinite entropy is estimated because no collisions are observed with the partial-key function. Figure 5b shows how this entropy translates into data structures, where we see that the Google URLs dataset is capable of using partial-key hashing with hundreds of millions of elements while hashing just a couple words. Similar results can be seen by transposing the other four datasets onto Figure 5b, with most datasets supporting hash data structures larger than the actual number of elements found in the dataset.

## 6.3 Hash Table Probe Time

After showing that datasets have enough entropy for partial-key hashing to be used,we now turn to showing the performance benefits which can be gained by using Entropy-Learned hash functions for data structures and algorithms. We first focus on hash tables. We examine the probe time per hash table lookup, where we perform the lookups one after the other without any blocking, e.g., similar to the probe-phase of the hash join algorithm.

**Entropy-Learned Hashing Reduces Hash Table Probe Time.** We first test hash table probe times on real-world datasets for small (L1-resident) and large data (L3/DRAM-resident) with 0% (hit rate = 0) and 100% (hit rate = 1) hit rates. We test with Google's SwissTable using three hash functions: (i) the default hash function provided by SwissTable (GST), (ii) the most recent version of wyhash (FK), and (iii) the Entropy-Learned wyhash hash function (ELH). The small data contains one thousand keys, and the large data contains half of the number of keys of the dataset (we use the other half to generate probes for missing keys). Figure 6 shows the results, wherein Entropy-Learned Hashing provides speedups across all data sizes, datasets, and hit rates over full-key hashing. Across the

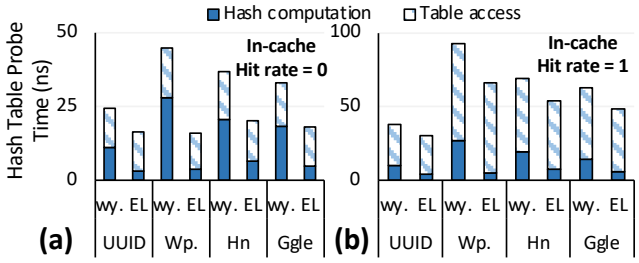



**Figure 7: Entropy-Learned Hashing significantly reduces computation time bringing speedup as high as 2.9× for cache-resident hash tables with (a) low and (b) high hit rates.**

20 experiments, the average speedup using ELH over wyhash and SwissTable's default hash function is 1.40×, with these speedups being as high 3.7× over the default hash function of SwissTable and as high as 2.9× over wyhash, both of which are well engineered functions and implementations.

**Entropy-Learned Hashing Scales with Entropy, not Key Size.** To understand the reasons behind the speed up observed in Figure 6, we first need to return to Table 3 and Figure 5. For full-key hashing, it needs to hash each byte of the dataset, and so the number of bytes processed is on average the key length given in Table 3. For Entropy-Learned Hashing, the number of bytes it hashes is when the entropy of the dataset (seen in Figure 5a) crosses the entropy needed by the data structure (seen in Figure 5b). When there is a large gap between these two numbers, Entropy-Learned Hashing produces large speedups. For instance, the large gap between the number of bytes hashed is why ELH achieves 2.9× speedup over wyhash and 3.9× speedup over default SwissTable in Figure 6a. Similarly, it is why ELH is 1.67× faster than wyhash and 1.81× faster than default SwissTable on the Google dataset in Figure 6d.

While faster hashing computation uniformly brings speedups to hash table probes, the amount of this speedup depends on other factors of hash table queries, namely the hit rate and hash table size. We now explain how the combination of these factors with Entropy-Learned Hashing affects performance.

**Computation Dominates for Cache-Resident Hash Tables.** For cache-resident hash tables, memory requests return quickly and so computation dominates the overall cost of probes. In this case, the savings created by Entropy-Learned Hashing depend on how much work there is beyond the hash function evaluation. Figure 7 shows how the work beyond hashing differs for queries for non-existing keys and for existing keys. When queries are for non-existing keys, computation usually consists of the hash function plus small amounts of computation using the tag bits. As Figure 7a shows, in this case the hash function evaluation is most of the

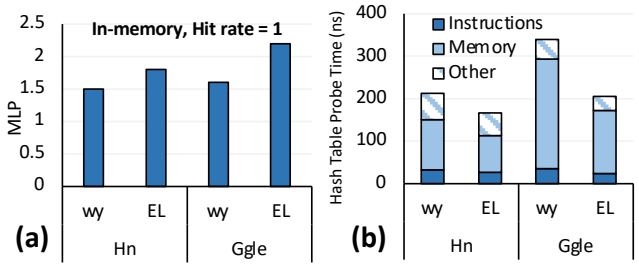



**Figure 8: (a) MLP is significantly higher for ELH than it is for full-key hashing. (b) As a result, ELH reduces both the number of instructions executed and memory waiting time.**

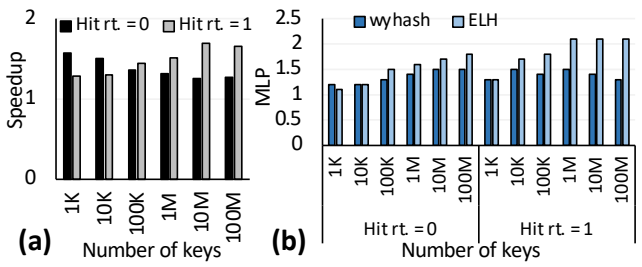



**Figure 9: (a) Entropy-Learned Hashing provides larger benefits for missing keys at small data sizes and larger benefits for existing keys at large data sizes. (b) Entropy-Learned Hashing improves memory-level parallelism.**

cost and Entropy-Learned Hashing brings significant benefits. This explains the 1.5×, 2.9×, 1.8×, and 1.8× speedup over wyhash seen in Figure 6a for the UUID, Wikipedia, Hacker News, and Google datasets, respectively. When queries are for keys in the dataset, Figure 7b shows the comparison after the hash function evaluation takes significant time. As a result, Entropy-Learned Hashing still provides benefits but not quite as large as before, with the savings being 1.23×, 1.41× 1.28×, and 1.28× for the UUID, Wikipedia, Hacker news, and Google datasets, respectively. Thus in cache, Entropy-Learned Hashing provides up to a 40% speedup for queries on existing keys and up to a 3× improvement on non-existent keys.

**Memory Parallelism Dominates for Large Hash Tables.** At large data sizes, the increase in computational performance from faster hashing leads to more efficient use of the memory hierarchy. This is due to the effects of CPU pipelining. Namely, when hash table lookups are done one after the other without blocking, then the CPU typically pipelines multiple hash table lookups which are then executed in parallel [40]. Entropy-Learned Hashing reduces the amount of computation required, and as a result, the CPU fits a larger number of hash table lookups into its pipeline. The effect of this increased pipelining is what creates the speedups seen at large data sizes in Figure 6c and 6d across datasets, with Entropy-Learned Hashing being as much as 1.67× faster than the nearest competitor.

The amount of this savings depends on the costs of memory accesses, with more expensive memory accesses leading to larger improvements. For instance, in Figure 6d we see that the larger datasets Google and Hacker News produce greater savings than the smaller datasets Wikipedia, UUID, and Wiki. Similarly, comparing Figure 6d to 6c, querying for existing keys produces greater savings because we view both tag bits and full-keys in comparison to just the tag bits most often for missing keys.

Figures 8a and 8b refine this analysis. Figure 8a shows the memory-level parallelism (MLP), which is defined as the number of L1 data cache misses per CPU cycle, for the Hacker News and Google datasets using hit rate = 1. The higher MLP in 8a indicates that a large number of data cache misses are being executed in parallel by Entropy-Learned Hashing than by full-key hashing. Figure 8b shows how this affects the overall runtime of hash table probes under the same setup, with Entropy-Learned Hashing reducing both the number of instructions executed and memory waiting time. This analysis corroborates the results seen in Figure 6c and 6d, where Entropy-Learned Hashing provides a 1.31× speedup on average over full-key hashing.

**Entropy-Learned Hashing Scales with Data.** We now turn to experiments with synthetic data so that we can more finely control the data size and experiment with larger data sizes. Figure 9a shows the main result, which is that Entropy-Learned Hashing provides benefits for hash tables across small and large data sizes. At small data sizes of 1K tuples, Entropy-Learned Hashing provides 2.33× speedups on queries for non-existing keys and 1.30× speedups for existing keys. For large data sizes of 100M tuples, this speedup is 1.3× for missing keys and 1.7× for existing keys. Figure 9b shows that the reason for these speedups is as discussed before for the real-world datasets. Namely, at small data sizes the savings in computation directly produce speedups for Entropy-Learned Hashing, whereas for large data sizes the more efficient hash computation leads to better MLP which produces faster probe times.

## 6.4 Bloom Filter Lookup Time & FPR

In this section, we evaluate Entropy-Learned Hashing for Bloom filters. We examine the lookup time and false positive rate (FPR) metrics. As input parameters, we let the FPR of the filter be 3% and allow the Entropy-Learned Hashing filter to deviate in FPR by 1%. The filter uses 3 hash functions, but computes only 1 due to double hashing. All parameters are tunable; this experimental setup is meant to reflect high-throughput filters such as those in filter push-down before joins [43]. For the small data size we use 1K keys and for the large data size we again use half the number of keys in the data.

**Entropy-Learned Hashing Reduces Filter Lookup Time.** Figures 10a and 10b present results for Bloom filters lookup time and FPR using xxHash and Entropy-Learned Hashing. Figure 10a shows that Entropy-Learned Hashing improves performance on high entropy datasets such as Google, Hacker News, UUID, and Wikipedia. The speedup is consistently between 1.85× and 4.51×. For Wiki, which has both small key size and low entropy, the speedup is small. Across all datasets, the average speedup is 2.10×, so that Entropy-Learned Hashing consistently provides drasticaly faster throughput on Bloom filter queries.

**Entropy-Learned Hashing has Tunable Added FPR.** Figure 10b presents the FPR of Bloom filters using Entropy-Learned Hashing and full-key hashing. Most importantly, as can be seen in Figure 10b, the FPR is within 1% as our tuning parameter suggests so that our analytical bounds hold. Additionally, Figure 10b shows that the increase in FPR is usually much less than this tuning parameter, in this case being only 0.1%. Thus, for most datasets the difference in

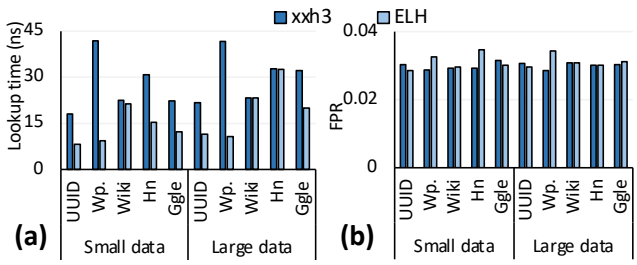



**Figure 10: Improving Bloom filter lookup time (a) and false positive rates (b) for small and large data sizes.**

FPR is negligible. Additionally, this FPR increase can be adjusted down or up as needed. Reducing the allowed increase in FPR increases the entropy needed and so requires more hash computation, and so this represents a tunable FPR vs. speed tradeoff.

**Bloom Filters require more entropy than Hash Tables.** For a dataset size of $n$ and added FPR of $\varepsilon$, ELH requires $\log_2 n + \log_2(1/\varepsilon)$ entropy, which is approximately $\log_2(1/\varepsilon)$ more entropy than hash tables. For certain datasets such as Wiki or Hacker News, this goes beyond the entropy they can provide using small partial-keys and so they revert to using full-key hashing at large data sizes as can be seen in Figures 10a and b. For Google URLs, Wikipedia, and UUID, they have more than enough entropy and each can support at least 100× more data or a 100× lower added FPR. Thus, these datasets maintain consistent speedups at no cost to FPR for very large data sizes as seen in Figure 10b.

## 6.5 Partitioning Time & Variance

Partitioning is used in many contexts. For instance, tuples may be sent across the network in settings such as map-reduce or simply partitioned in memory as in radix-partitioning before hash joins. Because of this, the cost of partitioning depends very heavily on the application it is used in. To help guide users in terms of whether Entropy-Learned Hashing can be useful for their application, we provide three micro-benchmarks. These benchmarks show the increased computational efficiency of Entropy-Learned Hashing on partitioning and put this computational efficiency in context. In the first micro-benchmark, we only compute the partition assigned to each input key. In the second, we keep a list of positional identifiers for each partition and write out the position of each key assigned to each partition. In the third, we write out the actual keys assigned to each partition. As we progress through the microbenchmarks, we move from a computationally heavy task with few writes to a memory bandwidth intensive task which is mostly memory bound. Depending on the setup, the benefit in performance from using Entropy-Learned Hashing may be between 14× and 18%. Thus, the benefit of Entropy-Learned Hashing for partitioning depends on whether the saved computational cycles are of use, either directly through speedups on the task at hand, or indirectly, by allowing other computation to take place while network or memory I/O is being performed. Like Bloom Filters, partitioning has a tunable parameter which allows the variance (equivalently standard deviation) to increase in exchange for faster hashing. We set this parameters so that each partition is expected to be within 5% of its mean.

**Entropy-Learned Hashing Reduces Partitioning Time.** Table 4 presents the speedups of Entropy-Learned Hashing for the three configurations we examine. Entropy-Learned Hashing dramatically improves the hashing computation as can be seen by the left side of Table 4, with increases in speed of above 3× for 4 of the 5 datasets and speedups of up to 14.1×. Partitioning by writing out positional identifiers, seen in the middle column of Table 4, is similar, with increases in speed of greater than 2× for 4 of the 5 datasets and speedups of up to 6.2×. Thus, the results show that the computational cost of partitioning is significantly cheaper using Entropy-Learned Hashing. At the same time, writing out large amounts of data can limit the benefits of using ELH for partitioning, as seen in the right side of Table 4. By writing out long-key strings at each iteration of the partitioning, limitations on write bandwidth limit gains from Entropy-Learned Hashing. Still, even in this case the speedups can be as much as 20%, and additionally CPU usage is reduced which frees up the CPU for other tasks.

| | Pure hashing | | Pos. id. | | Data | |
|---|---|---|---|---|---|---|
| # Par. | 64 | 1024 | 64 | 1024 | 64 | 1024 |
| UUID | 3.15 | 3.15 | 2.05 | 1.38 | 1.01 | 1.00 |
| Wp. | 14.10 | 14.09 | 6.18 | 2.66 | 1.23 | 1.18 |
| Wiki | 1.25 | 1.09 | 1.37 | 1.10 | 1.01 | 1.01 |
| Hn | 4.29 | 1.00 | 2.72 | 1.00 | 1.17 | 1.03 |
| Ggle | 7.83 | 7.82 | 2.51 | 1.42 | 1.01 | 1.00 |

**Table 4: Speeding up when partitioning.**

| | Pure hashing | | Pos. id. | | Data | |
|---|---|---|---|---|---|---|
| # Par. | 64 | 1024 | 64 | 1024 | 64 | 1024 |
| UUID | 1.44 | 0.95 | 1.44 | 0.95 | 1.44 | 0.95 |
| Wp. | 0.92 | 1.02 | 0.92 | 1.02 | 0.93 | 1.02 |
| Wiki | 1.35 | 1.01 | 1.35 | 1.01 | 1.35 | 1.01 |
| Hn | 2.06 | 1.00 | 2.06 | 1.00 | 2.05 | 1.00 |
| Ggle | 1.09 | 1.08 | 1.09 | 1.08 | 1.09 | 1.08 |

**Table 5: The relative standard deviations of Entropy-Learned Hashing and full-key hashing are similar.**

**Partitioning quality is maintained using Entropy-Learned Hashing.** Table 5 presents normalized relative standard deviation for partitioning, where relative standard deviation is obtained by dividing the standard deviation by the average. We calculate relative standard deviation for both full-key and Entropy-Learned Hashing and normalize the Entropy-Learned Hashing to the full-key hashing. As Table 5 shows, the normalized relative standard deviations concentrate around one, which shows that the partitions produced by the full-key hashing and the partitions produced by the entropy-learned hashing are similar. In the case they are not, such as for Hacker News with 64 partitions, the relative standard deviation of Entropy-Learned Hashing is less than 3% so that partitions are within 3% of their expected number of items on average.

## 6.6 Large Key Experiments

A key benefit of Entropy-Learned Hashing is that it creates hash functions whose runtime is independent of key size. While this provides computational benefits for data with medium sized keys such as URLs and text, we now show that the speedup is much larger for large keys such as file blocks. To demonstrate this effect, we repeat our experiments for hash tables, bloom filters, and partitioning but with synthetic random keys of 8192 bytes each. Figure 11 shows the results. For hash tables with all successful lookups, the benefits of

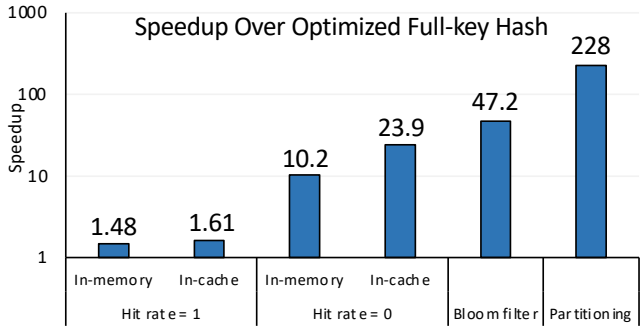



**Figure 11: Entropy-Learned Hashing provides orders of magnitude speedups with large key sizes.**

Entropy-Learned Hashing are naturally bounded because the need to compare full keys limits the throughput of this task. However, for hash table lookups that are misses, Bloom filter probes, and partitioning, Entropy-Learned Hashing brings a large speedup that is unbounded and can be one to two orders of magnitude.

## 6.7 Training Time

We now demonstrate that the training time needed for Entropy-Learned Hashing is not a bottleneck. We use the full Google dataset. Table 6 shows the results, displaying algorithm run times for a naive implementation which keeps all data points at each iteration, and for our optimized implementation which discards unique keys after each iteration. There are three main takeaways. First, the training time is reasonable for all sizes of contiguous bytes chosen, with runtimes between several minutes and several seconds. Second, pruning items which are unique from the dataset after each iteration produces substantial runtime benefits (if an item is unique on some subset of bytes, adding new bytes cannot create a collision for that item). Third, as the size of the contiguous byte locations we choose increases, the runtime decreases significantly because there are fewer options at each iteration and because after fewer iterations the number of data items that are non-unique is low (making each step, i.e. Algorithm 2, fast).

| # bytes | 1 | 4 | 8 |
|---|---|---|---|
| Optimized | 214 s | 11.6 s | 6.4 s |
| Naive | 29 min. | 13 min | 5 min |

**Table 6: Training runtime**

## 6.8 Additional Experiments

The paper focuses on a curated set of experiments which best showcase the properties of Entropy-Learned Hashing. Due to space constraints, this leaves out several experiments which cover other key metrics. Briefly, this includes experiments on 1) the efficiency of creating Entropy-Learned Hash data structures, 2) probing separate chaining hash tables, 3) experiments with dependent accesses (i.e. hash table lookups and Bloom filter lookups which must run one after the other instead of in parallel), 4) additional experiments on Bloom filters showing a range of desired false positive rates, and 5) experiments showing robustness properties. We include all of these results in the technical report [34].

# 7 RELATED WORK

**Entropy & Hashing.** Chung, Mitzenmacher, and Vadhan's work [18, 48] explains why current hash functions perform well, hypothesizing that data randomness is the reason this occurs. Our work makes a step forward to change the practice of hashing by recognizing this randomness, choosing how much and which parts of the data we need to hash, and making hash functions cheaper.

**Non-Cryptographic Hash Functions.** New hash functions are continually designed and fitted to modern processors [71]. This includes works with some form of data-independent randomness guarantees such as multiply-shift [27], CLHash [44], and tabulation hashing[58, 79]. These works are complementary to Entropy-Learned Hashing as they can be modified to work over subsets of bytes to achieve even better speeds.

**Data-Dependent hash structures.** Hash functions which depend on the data have been considered before. For point lookups, this includes perfect hashing [31] and learned hash indexes [41]. Both these methods introduce computational overhead while trying to reduce the number of collisions. Entropy-Learned Hashing is complementary to such works and it can be used in conjunction with these techniques to get both better computation and a lower number of collisions. Additional work which is related in terms of learning from data to create better hash data structures is work on using data to improve filter structures [21, 26, 47, 72]. These all improve the false positive rate of filters, and use hash computation as part of their design. Thus, Entropy-Learned Hashing is again complementary in that it can be used to speed up the processing time of these filters.

**Cryptographic Hash Functions.** Cryptographic hash functions such as MD5 [67], SHA1 [29] and newer variants have more stringent measures on the ability to invert hash function outputs, but can and have been used for hash-based data structures. A cryptographic hash function specifically designed for hash-based data structures is SipHash [8]. While development of newer cryptographic hash functions has made cryptographic hashing faster, it remains an order of magnitude slower than non-cryptographic hashing [19].

# 8 CONCLUSION & FUTURE WORK

This paper introduces Entropy-Learned Hashing, a way to produce hash functions with reduced computational cost tailored to a given context. This happens by learning the amount of randomness in input data and producing hash functions that give just enough randomness to their output. We demonstrate that Entropy-Learned Hashing leads to substantial computational benefits on hash tables, Bloom filters, and load balancing. Future work in the path of context specific hash functions includes investigating the relationship between the distribution of source data and the necessary operations inside the hash function such that we can customize functions not only with respect to the data bytes they utilize but also with respect to the exact computation that needs to be performed.

# 9 ACKNOWLEDGEMENTS

This work is partially funded by the USA Department of Energy project DE-SC0020200 and by the Swiss National Science Foundation Early Postdoc Mobility scholarship P2ELP2_199749.

# REFERENCES


[1] [n.d.]. gcc libstdc++ hash. https://github.com/gcc-mirror/gcc/blob/master/libstdc%2B%2B-v3/libsupc%2B%2B/hash_bytes.cc. Accessed: 2021-05-23.

[2] 2015. Hacker News Posts. https://www.kaggle.com/hacker-news/hacker-news-posts. Accessed: 2021-05-23.

[3] 2019. Linker Throughput Improvement in Visual Studio 2019. https://devblogs.microsoft.com/cppblog/linker-throughput-improvement-in-visual-studio-2019/.

[4] Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. 2017. Estimating Renyi Entropy of Discrete Distributions. *IEEE Transactions on Information Theory* 63, 1 (2017), 38–56. https://doi.org/10.1109/TIT.2016.2620435

[5] Paulo Sérgio Almeida, Carlos Baquero, Nuno Preguiça, and David Hutchison. 2007. Scalable Bloom Filters. *Inf. Process. Lett.* 101, 6 (March 2007), 255–261.

[6] Austin Appleby. [n.d.]. murmurhash3. https://github.com/aappleby/smhasher/wiki/MurmurHash3. Accessed: 2021-05-23.

[7] Austin Appleby. [n.d.]. smhasher suite. https://github.com/aappleby/smhasher. Accessed: 2021-05-23.

[8] Jean-Philippe Aumasson and Daniel J. Bernstein. 2012. SipHash: A Fast Short-Input PRF. In *Progress in Cryptology - INDOCRYPT 2012*, Steven Galbraith and Mridul Nandi (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 489–508.

[9] Eric Balkanski, Sharon Qian, and Yaron Singer. 2021. Instance specific approximations for submodular maximization. In *International Conference on Machine Learning*. PMLR, 609–618.

[10] Cagri Balkesen, Gustavo Alonso, Jens Teubner, and M. Tamer Özsu. 2013. Multi-Core, Main-Memory Joins: Sort vs. Hash Revisited. *Proc. VLDB Endow.* 7, 1 (Sept. 2013), 85–96. https://doi.org/10.14778/2732219.2732227

[11] Burton H Bloom. 1970. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM* 13, 7 (1970), 422–426.

[12] Colin R. Blyth. 1980. Expected Absolute Error of the Usual Estimator of the Binomial Parameter. *The American Statistician* 34, 3 (1980), 155–157. http://www.jstor.org/stable/2683873

[13] Peter Boncz, Thomas Neumann, and Viktor Leis. 2020. FSST: Fast Random Access String Compression. 13, 12 (2020), 2649–2661.

[14] Andrei Broder, Michael Mitzenmacher, and Andrei Broder I Michael Mitzenmacher. 2002. Network Applications of Bloom Filters: A Survey. In *Internet Mathematics*. 636–646.

[15] Andrei Z. Broder. 1997. On the resemblance and containment of documents.. In *SEQUENCES*, Bruno Carpentieri, Alfredo De Santis, Ugo Vaccaro, and James A. Storer (Eds.). IEEE, 21–29. http://dblp.uni-trier.de/db/conf/sequences/sequences1997.html#Broder97

[16] Nathan Bronson and Xiao Shi. [n.d.]. Open-sourcing F14 for faster, more memory-efficient hash tables. https://engineering.fb.com/2019/04/25/developer-tools/f14/.

[17] J. Lawrence Carter and Mark N. Wegman. 1977. Universal Classes of Hash Functions (Extended Abstract). In *Proceedings of the Ninth Annual ACM Symposium on Theory of Computing* (Boulder, Colorado, USA) *(STOC '77)*. Association for Computing Machinery, New York, NY, USA, 106–112.

[18] Kai-Min Chung, Michael Mitzenmacher, and Salil Vadhan. 2013. Why Simple Hash Functions Work: Exploiting the Entropy in a Data Stream. *Theory of Computing* 9, 30 (2013), 897–945. https://doi.org/10.4086/toc.2013.v009a030

[19] Yann Collet. [n.d.]. xxHash. https://cyan4973.github.io/xxHash/. Accessed: 2021-05-23.

[20] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. *Introduction to Algorithms, Third Edition* (3rd ed.). The MIT Press.

[21] Zhenwei Dai and Anshumali Shrivastava. 2020. Adaptive Learned Bloom Filter (Ada-BF): Efficient Utilization of the Classifier with Application to Real-Time Information Filtering on the Web. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 11700–11710. https://proceedings.neurips.cc/paper/2020/file/86b94dae7c6517ec1ac767fd2c136580-Paper.pdf

[22] Niv Dayan, Manos Athanassoulis, and Stratos Idreos. 2017. Monkey: Optimal Navigable Key-Value Store. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 79–94.

[23] Niv Dayan, Manos Athanassoulis, and Stratos Idreos. 2018. Optimal Bloom Filters and Adaptive Merging for LSM-Trees. *ACM Transactions Database Systems (TODS)* 43, 4 (2018), 16:1–16:48.

[24] Niv Dayan and Stratos Idreos. 2018. Dostoevsky: Better Space-Time Trade-Offs for LSM-Tree Based Key-Value Stores via Adaptive Removal of Superfluous Merging. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 505–520.

[25] Niv Dayan and Moshe Twitto. 2021. Chucky: A Succinct Cuckoo Filter for LSM-Tree. In *Proceedings of the 2021 International Conference on Management of Data* (Virtual Event, China) *(SIGMOD/PODS '21)*. Association for Computing Machinery, New York, NY, USA, 365–378. https://doi.org/10.1145/3448016.3457273

[26] Kyle Deeds, Brian Hentschel, and Stratos Idreos. 2020. Stacked Filters: Learning to Filter by Structure. *Proc. VLDB Endow.* 14, 4 (dec 2020), 600–612.

[27] Martin Dietzfelbinger, Torben Hagerup, Jyrki Katajainen, and Martti Penttonen. 1997. A Reliable Randomized Algorithm for the Closest-Pair Problem. *Journal of Algorithms* 25, 1 (1997), 19–51.

[28] Markus Dreseler, Martin Boissier, Tilmann Rabl, and Matthias Uflacker. 2020. Quantifying TPC-H Choke Points and Their Optimizations. *Proc. VLDB Endow.* 13, 8 (April 2020), 1206–1220. https://doi.org/10.14778/3389133.3389138

[29] D. Eastlake and P. Jones. 2001. RFC3174: US Secure Hash Algorithm 1 (SHA1).

[30] P. Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. 2007. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. *Discrete Mathematics & Theoretical Computer Science* (2007), 137–156.

[31] Michael L. Fredman, Michael L. Fredman, Michael L. Fredman, Michael L. Fredman, Janos Komlos, Janos Komlos, Janos Komlos, Janos Komlos, Endre Szemeredi, Endre Szemeredi, Endre Szemeredi, and Endre Szemeredi. 1982. Storing a sparse table with O(1) worst case access time. In *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*. 165–169. https://doi.org/10.1109/SFCS.1982.39

[32] Google. [n.d.]. Abseil Common Libraries. https://github.com/abseil/abseil-cpp.

[33] Jason Gregory. 2009. *Game engine architecture* (1 ed.). Taylor & Francis Ltd.

[34] Brian Hentschel, Utku Sirin, and Stratos Idreos. [n.d.]. Entropy-Learned Hashing Technical Report. http://daslab.seas.harvard.edu/EntoryLearnedHashing/EntropyLearnedHashingTechnicalReport.pdf.

[35] Stratos Idreos and Mark Callaghan. 2020. Key-Value Storage Engines. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 2667–2672.

[36] Intel. 2021. Intel VTune Amplifier XE Performance Profiler. http://software.intel.com/en-us/articles/intel-vtune-amplifier-xe/.

[37] Adam Kirsch and Michael Mitzenmacher. 2006. Less hashing, same performance: building a better bloom filter. In *European Symposium on Algorithms*. Springer, 456–467.

[38] Donald E. Knuth. 1998. *The Art of Computer Programming, Volume 3: (2nd Ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., USA.

[39] Chun-Wa Ko, Jon Lee, and Maurice Queyranne. 1995. An exact algorithm for maximum entropy sampling. *Operations Research* 43, 4 (1995), 684–691.

[40] Onur Kocberber, Babak Falsafi, and Boris Grot. 2015. Asynchronous Memory Access Chaining. *Proc. VLDB Endow.* 9, 4 (2015), 252–263.

[41] Tim Kraska, Alex Beutel, Ed H Chi, Jeffrey Dean, and Neoklis Polyzotis. 2018. The case for learned index structures. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, 489–504.

[42] Matt Kulukundis. [n.d.]. Designing a Fast, Efficient, Cache-friendly Hash Table, Step by Step. https://www.youtube.com/watch?v=ncHmEUmJZf4.

[43] Harald Lang, Thomas Neumann, Alfons Kemper, and Peter Boncz. 2019. Performance-optimal filtering: Bloom overtakes cuckoo at high throughput. *Proceedings of the VLDB Endowment* 12, 5 (2019), 502–515.

[44] Daniel Lemire and Owen Kaser. 2016. Faster 64-bit universal hashing using carry-less multiplications. *Journal of Cryptographic Engineering* 6, 3 (2016), 171–185.

[45] Linux. 2021. Perf Wiki. https://perf.wiki.kernel.org/.

[46] Zaoxing Liu, Ran Ben-Basat, Gil Einziger, Yaron Kassner, Vladimir Braverman, Roy Friedman, and Vyas Sekar. 2019. Nitrosketch: Robust and General Sketch-Based Monitoring in Software Switches. In *Proceedings of the ACM Special Interest Group on Data Communication* (Beijing, China) *(SIGCOMM '19)*. Association for Computing Machinery, New York, NY, USA, 334–350. https://doi.org/10.1145/3341302.3342076

[47] Michael Mitzenmacher. 2018. A Model for Learned Bloom Filters and Optimizing by Sandwiching. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc.

[48] Michael Mitzenmacher and Salil Vadhan. 2008. Why simple hash functions work: Exploiting the entropy in a data stream. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 746–755.

[49] Michael David Mitzenmacher and Alistair Sinclair. 1996. *The Power of Two Choices in Randomized Load Balancing*. Ph.D. Dissertation. AAI9723118.

[50] Hamid Mohamadi, Justin Chu, Benjamin P. Vandervalk, and Inanc Birol. 2016. ntHash: recursive nucleotide hashing. *Bioinformatics* 32, 22 (07 2016), 3492–3494. https://doi.org/10.1093/bioinformatics/btw397 arXiv:https://academic.oup.com/bioinformatics/article-pdf/32/22/3492/19397493/btw397_Sup.pdf

[51] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions—I. *Mathematical programming* 14, 1 (1978), 265–294.

[52] Hyeonwoo Noh, Andre Araujo, Jack Sim, and Bohyung Han. 2016. Large-Scale Image Retrieval with Attentive Deep Local Features. *International Conference on Computer Vision (ICCV)* (2016). http://arxiv.org/abs/1612.06321

[53] Maciej Obremski and Maciej Skorski. 2017. Renyi Entropy Estimation Revisited. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2017, August 16-18, 2017, Berkeley, CA, USA (LIPIcs, Vol. 81)*, Klaus Jansen, José D. P. Rolim, David Williamson, and Santosh S. Vempala (Eds.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 20:1–20:15.

[54] Patrick E. O'Neil, Edward Cheng, Dieter Gawlick, and Elizabeth J. O'Neil. 1996. The Log-Structured Merge-Tree (LSM-Tree). *Acta Inf.* 33, 4 (1996), 351–385. http://dblp.uni-trier.de/db/journals/acta/acta33.html#ONeilCGO96

[55] Anna Pagh, Rasmus Pagh, and Milan Ružić. 2011. Linear Probing with 5-Wise Independence. *SIAM Rev.* 53, 3 (Aug. 2011), 547–558. https://doi.org/10.1137/

110827831

[56] Rasmus Pagh and Flemming Friche Rodler. 2004. Cuckoo Hashing. *J. Algorithms* 51, 2 (May 2004), 122–144. https://doi.org/10.1016/j.jalgor.2003.12.002

[57] Mihai Pătraşcu and Mikkel Thorup. 2010. On the k-Independence Required by Linear Probing and Minwise Independence. In *Automata, Languages and Programming*. Springer Berlin Heidelberg, Berlin, Heidelberg, 715–726.

[58] Mihai Patrascu and Mikkel Thorup. 2011. The Power of Simple Tabulation Hashing. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing* (San Jose, California, USA) *(STOC '11)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/1993636.1993638

[59] W. W. Peterson. 1957. Addressing for Random-Access Storage. *IBM Journal of Research and Development* 1, 2 (1957), 130–146. https://doi.org/10.1147/rd.12.0130

[60] Geoff Pike and Jyrki Alakuijala. 2011. CityHash. https://github.com/google/cityhash.

[61] Geoff Pike and Jyrki Alakuijala. 2014. FarmHash. https://github.com/google/farmhash.

[62] Orestis Polychroniou and Kenneth A. Ross. 2014. A Comprehensive Study of Main-Memory Partitioning and Its Application to Large-Scale Comparison- and Radix-Sort. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data* (Snowbird, Utah, USA) *(SIGMOD '14)*. Association for Computing Machinery, New York, NY, USA, 755–766. https://doi.org/10.1145/2588555.2610522

[63] M. V. Ramakrishna. 1988. Hashing Practice: Analysis of Hashing and Universal Hashing. In *Proceedings of the 1988 ACM SIGMOD International Conference on Management of Data* (Chicago, Illinois, USA) *(SIGMOD '88)*. Association for Computing Machinery, New York, NY, USA, 191–199. https://doi.org/10.1145/50202.50223

[64] M. V. Ramakrishna. 1989. Practical Performance of Bloom Filters and Parallel Free-Text Searching. *Commun. ACM* 32, 10 (Oct. 1989), 1237–1239. https://doi.org/10.1145/67933.67941

[65] Raghu Ramakrishnan and Johannes Gehrke. 2002. *Database Management Systems* (3 ed.). McGraw-Hill, Inc., USA.

[66] Stefan Richter, Victor Alvarez, and Jens Dittrich. 2015. A Seven-Dimensional Analysis of Hashing Methods and Its Implications on Query Processing. *Proc. VLDB Endow.* 9, 3 (Nov. 2015), 96–107. https://doi.org/10.14778/2850583.2850585

[67] R. Rivest. 1992. RFC1321: The MD5 Message-Digest Algorithm.

[68] Kenneth A. Ross. 2007. Efficient Hash Probes on Modern Processors. In *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, Rada Chirkova, Asuman Dogac, M. Tamer Özsu, and Timos K. Sellis (Eds.). IEEE Computer Society, 1297–1301.

[69] Utku Sirin and Anastasia Ailamaki. 2020. Micro-Architectural Analysis of OLAP: Limitations and Opportunities. *Proc. VLDB Endow.* 13, 6 (2020), 840–853.

[70] Oracle ZFS Steve Tunstall. 2017. DeDupe 2.0. https://blogs.oracle.com/wonders-of-zfs-storage/dedupe-20-v2. Accessed: 2021-05-23.

[71] Reini Urban. [n.d.]. SMHasher - Reini Urban Fork. https://github.com/rurban/smhasher. Accessed: 2021-05-23.

[72] Kapil Vaidya, Eric Knorr, Michael Mitzenmacher, and Tim Kraska. 2021. Partitioned Learned Bloom Filters. In *International Conference on Learning Representations*. https://openreview.net/forum?id=6BRLOfrMhW

[73] Yi Wang, Diego Barrios Romero, Daniel Lemire, and Li Jin. 2020. Modern Non-Cryptographic Hash Function and Pseudorandom Generator. (2020).

[74] Jan Wassenberg and Peter Sanders. 2011. Engineering a Multi-Core Radix Sort. In *Proceedings of the 17th International Conference on Parallel Processing - Volume Part II* (Bordeaux, France) *(Euro-Par'11)*. Springer-Verlag, 160–169.

[75] Mark N. Wegman and J.Lawrence Carter. 1981. New hash functions and their use in authentication and set equality. *J. Comput. System Sci.* 22, 3 (1981), 265–279. https://doi.org/10.1016/0022-0000(81)90033-7

[76] Oracle ZFS. 2019. ZFS Deduplication. https://blogs.oracle.com/bonwick/zfs-deduplication-v2. Accessed: 2021-05-23.

[77] Tianqi Zheng, Zhibin Zhang, and Xueqi Cheng. 2020. SAHA: A String Adaptive Hash Table for Analytical Databases. *Applied Sciences* 10, 6 (2020). https://doi.org/10.3390/app10061915

[78] Zichen Zhu, Ju Hyoung Mun, Aneesh Raman, and Manos Athanassoulis. 2021. Reducing Bloom Filter CPU Overhead in LSM-Trees on Modern Storage Devices. In *Proceedings of the 17th International Workshop on Data Management on New Hardware (DaMoN 2021)* (Virtual Event, China) *(DAMON'21)*. Association for Computing Machinery, New York, NY, USA, Article 1, 10 pages. https://doi.org/10.1145/3465998.3466002

[79] A. Zobrist. 1990. A New Hashing Method with Application for Game Playing. *ICGA Journal* 13 (1990), 69–73.

# Appendix: Entropy Learned Hashing 10x Faster Hashing with Controllable Uniformity

BRIAN HENTSCHEL, Harvard University
UTKU SIRIN, Harvard University
STRATOS IDREOS, Harvard University

This document accompanies Entropy-Learned Hashing and covers aspects of the paper which were not covered in the main body of the paper. This covers proofs regarding Entropy-Learned Hashing and linear probing, discussion of robustness for Entropy-Learned Hashing data structures, and additional experiments.

## Appendix A Proofs for Linear Probing

### A.1 Notation & Problem Setup

The notation and problem setup are mostly the same as in the original paper. We give a brief recap of the notation and problem setup here for convenience.

Hash functions in Entropy-Learned hashing consists of two parts, $L$ and $H$. The first part $L$ chooses which locations of an input item $x$ to hash and maps a key $x$ to any subkey of $x$. The second part is a general hash function $H$ which gets applied to $L(x)$ so that the full partial key hash $H' = H \circ L$. It is assumed that $H$ follows the model of a perfectly random hash function, so that each unique input is given an independent uniform output between 0 and $m - 1$.

To analyze algorithms which use partial-key hashing, we need to deal with the fact that the keys entered into $H$ are no longer unique. Towards this end we introduce notation for the set of keys $K$ contained in a hash based data structure. The multi-set $S_{|L} = (K_{|L}, z)$ is made of $K_L$, the set of all partial-keys (outputs of $L$ applied to keys in $K$), and $z$, which maps each key in $K_L$ to the cardinality of its pre-image in $K$. For instance, if $L$ takes the first two characters of an input and $K = \{\text{dog, dot, cat, fan}\}$, then $K_{|L} = \{\text{"do", "ca", "fa"}\}$, $z(\text{"ca"}) = 1$, and $z(\text{"do"}) = 2$. We use as shorthand throughout $z_x$ for $z(x)$.

Finally, we assume some familiarity with the linear probing algorithm, wherein collisions are resolved by checking the next slot in a circular array until an empty one is found. In this document, we resolve collisions by going to the left, i.e. checking $i, i-1, i-2, \ldots$ rather than $i, i+1, \ldots$. This is to be consistent with Knuth's proof for linear probing given in [3]. A reminder of the rest of the notation can be seen in the table below.

### A.2 Linear Probing

As in the main document, we start our proof by going over full-key hashing, then analyze partial-key hashing with a fixed dataset, and finally cover partial-key hashing with random data. Before starting the proof, we briefly cover the results. When the multi-set $S_{|L}$ is fixed, the expectation is an expectation over the randomness of the hash function $H$. When it is not fixed, the expectation is taken over both the random multi-set $S_{|L}$ and the random hash function $H$.

*A.2.1 Results* For full-key hashing, given any set of $n$ keys and a hash table with $m$ slots, the expected number of comparisons for querying a key not in the dataset and the average number of comparisons when querying for

Authors' addresses: Brian Hentschel, Harvard University, bhentschel@g.harvard.edu; Utku Sirin, Harvard University, wilsonqin@seas.harvard.edu; Stratos Idreos, Harvard University, stratos@seas.harvard.edu.

| Notation | Definition |
|---|---|
| $X, x$ | keys stored in the filter or hash table |
| $Y, y$ | key which is queried for |
| $H$ | hash function |
| $m$ | size of table (in slots) |
| $n$ | number of keys in table |
| $K$ | set of keys |
| $S_{\|L}$ | multi-set of keys conditioned on $L$. Equal to $(K_{\|L}, z)$ where $K_{\|L}$ consists of all partial keys, and $z$ maps each key $x \in K_{\|L}$ to the number of times it appears in $K$. $z_x$ will be used as shorthand for $z(x)$ throughout. |
| $\alpha$ | fill of hash table: $\frac{n}{m}$ |
| $P'$ | number of comparisons to find non-existing key |
| $P$ | average # of comparisons to retrieve an existing key |

Table 1. Notation used throughout the paper

an existing key are:

$$\mathbb{E}[P'] \leq \frac{1}{2}\left(1 + \frac{1}{(1-\alpha)^2}\right) \tag{1}$$

$$\mathbb{E}[P] \leq \frac{1}{2}\left(1 + \frac{1}{1-\alpha}\right) \tag{2}$$

Letting $c = \sum_x z_x^2$ be the number of collisions in $S_{|L}$ and $d = \sum_{x:z_x \geq 2} z_x$ the number of duplicated keys in $S_{|L}$, the costs for partial-key hashing are

$$\mathbb{E}[P'] \leq \begin{cases} \frac{1}{2}\left(1 + \frac{1}{(1-\alpha)^2} + \sum_{x \neq y} \frac{z_x^2}{m(1-\alpha)^2}\right) & \text{if } z_y = 0 \\ \frac{z_y}{1-\alpha} + \frac{1}{(1-\alpha)^2} + \sum_{x \neq y} \frac{z_x^2}{m(1-\alpha)^2} & \text{if } z_y > 0 \end{cases} \tag{3}$$

$$\begin{aligned} \mathbb{E}[P] &\leq \frac{n-d}{2n} + \frac{1}{2}Q_0(m,n) + \frac{c}{m}Q_0(m,n) + \frac{c+d}{2n}Q_0(m,d) \\ &\approx (\frac{1}{2} + \frac{c}{n})(1 + \frac{1}{1-\alpha}) \end{aligned} \tag{4}$$

where here $y$ is the queried for key in $P'$ and the approximation in (4) uses the assumption that $d$ is small compared to $m$. For random data from an i.i.d. source with Renyi entropy $H_2$, these translate to the following bounds:

$$\mathbb{E}[P'] \leq \frac{1}{2}(1 + \frac{1}{(1-\alpha)^2}) + n2^{-H_2(L(X))}\frac{3}{2(1-\alpha)^2}$$

$$\mathbb{E}[P] \leq \frac{1}{2}(1 + \frac{1}{1-\alpha}) + n2^{-H_2(L(X))}(1 + \frac{1}{1-\alpha})$$

**Comparing Partial-Key Hashing with Full-Key Hashing.** The above equations, when fully analyzed, make a rather intuitive point: when there are only a few duplicates in $K_L$ so that it closely approximates full-key hashing, the number of comparisons required by full-key and partial-key hashing are close. The general form of each equation for partial key hashing is that it is the same as full-key hashing plus a small penalty term. These penalty terms go to 0 as the keys in partial-key hashing become more distinct.

For all equations, we note the equations above are relatively tight bounds as long as $\alpha$ is not close to 1. This is reasonable to assume as all hash tables keep $\alpha < 0.9$ in practice because of the extremely bad behavior of linear

probing tables as $\alpha \to 1$, wherein they degrade to linear scans. The most common range for $\alpha$ is between 0.25 and 0.85, with higher alpha leading to better memory usage but slower query times.

*A.2.2 Analysis of Full Key Linear Probing* To prove (3) and (4), we start by proving (1) and (2), first proven by Donald Knuth. Our proof initially follows steps taken in The Art of Computer Programming[3], however we deviate from the proof given there because that approach does not generalize to partial-key hashing. Additionally, we believe this proof is simpler to follow than the proof in [3].

**Counting Hash Sequences.** For $n$ distinct items, we have $m^n$ possible ways to assign them to $[m] = \{0, \ldots, m-1\}$, all of which are equally likely by our hashing assumptions. Of all possible hash sequences, we first consider how many leave position 0 empty in the hash table.

Theorem 1. *The number of hash sequences such that $n$ items are hashed into $[1, m-1]$, and no overflow occurs so that position 0 is empty is*

$$\begin{cases} (1 - \frac{n}{m}) \cdot h(m, n) & \text{if } 0 \le n \le m \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

*where $h(m, n)$ is the number of ways to hash $n$ objects into $[0, m-1]$.*

This is a slight reformulation of the way it is stated in [3] so that it generalizes to partial-key hashing. To prove the theorem, note that hash sequences which hash $n$ items into $[1, m-1]$ and do not overflow into position 0 are precisely the same as those that leave position 0 empty when hashing into a hash table of size $m$ and resolving collisions via linear probing. This probability is $1 - \frac{n}{m}$ by the circular symmetry of linear probing. Thus, for full-key hashing there are $(1 - \frac{n}{m})m^n$ sequences which leave location 0 empty.

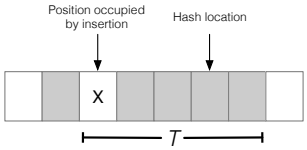
The next theorem gives the probability that a run of occupied slots begins at location 1 in the table and stops at location $k$.

Theorem 2. *Let $g(m, n, k)$ be the number of sequences when hashing into a hash table of size $m$ that leave position 0 empty, positions 1 to $k-1$ filled, and $k$ empty. This is*

$$g(m, n, k) = \binom{n}{k-1} \frac{1}{k} \frac{m-n-1}{m-k} h(k, k-1) h(m-k, n-k+1) \tag{6}$$

This theorem simply builds off of Theorem 1 as there are $\binom{n}{k-1}$ ways to choose the $k-1$ elements in the run starting at 1, and then by Theorem 1 there are $\frac{1}{k}h(k, k-1)$ and $\frac{m-n-1}{m-k}h(m-k, n-k-1)$ ways to create the two subsequences with locations $0, k$ empty for each choice of the $k-1$ elements.

**Analyzing Chain Length.** We use the two theorems above to analyze a random variable $T$ which represents the length counting from the empty position which a new item is inserted into to the final occupied position before the next empty position (going from left to right). The figure below shows an example.




If we know the expected value for $T$ for a new item to be inserted, we can get its average distance from its hash location by the uniform randomness of hashing. Namely, for any chain of length $T$, each location in the chain is equally likely as an initial hash location, so $\mathbb{E}[P'|T] = \frac{1}{T}\sum_{i=1}^{T} i = \frac{1}{2} + \frac{1}{2}T$. It follows that

$$\mathbb{E}[P'] = \mathbb{E}[\mathbb{E}[P'|T]] = \frac{1}{2} + \frac{1}{2}\mathbb{E}[T]$$

To see how the two theorem above apply to our analysis of $T$, assume that the insertion key hashes to location $a$. If the insertion key is hashed into a chain of length $t$, then for some $t \geq 1, 0 \leq k \leq t-1$, we have $a-k$ empty, $a-k+1, \ldots, a+(t-k-1)$ full, and $a+(t-k)$ empty. We note by symmetry that the same number of sequences produce a chain of this nature as which produce a chain such that 0 is empty, $1, \ldots, t-1$ is full, $t$ is empty. Since we have $t$ choices for $k$, it then follows that there are $t \cdot g(m,n,t)$ hash sequences for which the insertion item is inserted into a chain of length $t$.

The approach taken in Knuth's art of computer programming is to analyze $E[T] = \sum_{t \geq 1} t^2 g(m,n,t)$ directly by using Abel's Binomial Theorem, i.e.

$$(x+y)^n = \sum_k \binom{n}{k} x(x-kz)^{k-1}(y+kz)^{n-k}$$

While correct, this approach feels magical as the use of Abel's Binomial is unintuitive (at least to the authors). Additionally, it does not generalize to the case of partial-key hashing, and so we take a different approach.

**Analyzing Chain Length By Connecting Chain Length and the Probability of Chain Inclusion.** Our approach to analyze $\mathbb{E}[T]$ is to look at the probability that each item is in the same probe chain as the inserted item. Namely, if we let $x_1, \ldots, x_n$ be the keys inserted into the hash table, then we have that

$$\mathbb{E}[T] = 1 + \sum_{i=j}^{n} P(x_j \in T)$$

Here we slightly abuse notation to have $T$ both represent the length of the chain on the left and the set of objects in the chain on the right. We now connect this probability that $x_j \in T$, which is equal for all $j$, to the expected chain length conditioned on a single new item $x$ being part of the chain $T$.

More concretely, let $T_i$ be the chain length of an inserted item if $i$ of the $n$ items in the hash table have the same hash value as the item to be inserted. All other items are distributed uniformly at random and independently of the hash value of the to be inserted item. The following theorem holds.

Theorem 3. *If $C$ represents a set of $i$ values which are fixed to have the same hash location as the newly inserted item, then for $x \notin C$, we have*

$$P(x \in T_i) = \frac{1}{m}\mathbb{E}[T_{i+1}]$$

Proof. We first derive the formula for $\mathbb{E}[T_i]$. Assume as before that the insertion key hashes to location $a$. If the new item is hashed into a chain of length $t$, then for some $t \geq 1, 0 \leq k \leq t-1$, we have $a-k$ empty, $a-k+1, \ldots, a+(t-k-1)$ full, and $a+(t-k)$ empty. Again by symmetry, the number of sequences that produce a chain of this nature is the same as the number of sequences that produce a chain such that 0 is empty, $1, \ldots, t-1$ is full, $t$ is empty, and the $i$ items hash to location $k$. Let this number be $g(m,n,t,i,k)$. Then $f(m,n,t,i) = \sum_{k \geq 0} g(m,n,t,i,k)$ is the total number of sequences such that the new item is hashed into a chain of length $t$.

The value for $f(m,n,t,i)$ is calculable directly as this is identical to hashing $t-i-1$ items of multiplicity 1 and 1 value of multiplicity $i$ into the first $t$ slots keeping slot 0 empty, and hashing $n-(t-1)$ items into the final $m-t$ slots. Using theorem 1, there are $\frac{1}{t}h(t,t-i)$ ways to do the first and $\frac{m-n-1}{m-t}h(m-t,n+1-t)$ ways to do the second, and $\binom{n-i}{t-i-1}$ ways to which $t-i-1$ elements from $K \setminus C$ are the elements in the run starting at 1. It follows that

$$f(m,n,t,i) = \binom{n-i}{t-i-1}\frac{1}{t}\frac{m-n-1}{m-t}h(t,t-i)h(m-t,n-t+1)$$

The expected value of $T_i$ is then $m^{-(n-i)} \sum_{t \geq 1} t \cdot f(m, n, t, i)$.

We now calculate the probability that $x \notin C$ is in $T_i$. To do so, we count the number of chains which contain $x$, and then divide by the total number of possible hash sequences. Using the same logic as before, for a key hashing to location $a$, the number of chains containing $x$ is equivalent to the number of chains such that 0 is empty, $1, \ldots, t-1$ is full and contains $x$, $t$ is empty, and $i$ items hash to location $k$. For a chain of length $t$, this means we have $\binom{n-i-1}{t-i-2}$ choices for the elements in the run between 0 and $t$, and we have $\frac{1}{t}h(t, t-i)$ and $\frac{m-n-1}{m-t}h(m-t, n-t+1)$ ways to hash these items to chain 1 and chain 2 respectively such that 0 and $t$ are empty. Summing across all possible chain lengths, the number of chains which contain $x \notin C$ is

$$\begin{aligned} &= \sum_t \binom{n-i-1}{t-i-2} \frac{1}{t} \frac{m-n-1}{m-t} h(t, t-i) h(m-t, n-t+1) \\ &= \sum_t t f(m, n, t, i+1) \end{aligned}$$

where we use that $h(t, t-i-1) = t \cdot h(t, t-i)$. Dividing by the number of hash sequences, $m^{n-i} = m^{n-(i+1)} m$ gives that $P(x \in T_i) = \frac{1}{m} \mathbb{E}[T_{i+1}]$. □

**Finishing the Proof.** Using these relationship between $\mathbb{E}[T_i]$ and $P(x \in T_i)$, we have

$$\begin{aligned} E[T] &= 1 + \sum_{i=1}^{n} P(x_i \in T_0) \\ &= 1 + \frac{n}{m} \mathbb{E}[T_1] \\ &= 1 + \frac{n}{m}(2 + \frac{n-1}{m} \mathbb{E}[T_2])) \end{aligned}$$

where here we use that $\mathbb{E}[T_i] = (i+1) + \sum_{x \notin C} P(x \in T_i)$. If we continue expanding the values of $\mathbb{E}[T_i]$, we get that

$$\begin{aligned} \mathbb{E}[T] &= 1 + 2\frac{n}{m} + 3\frac{n^{\underline{2}}}{m^2} + \cdots + \frac{(n+1)n!}{m^n} \\ &= Q_1(m, n) \end{aligned}$$

where

$$Q_r(m, n) = \sum_{k \geq 0} \binom{k+r}{k} \frac{n^{\underline{k}}}{m^k}$$

It follows that

$$\mathbb{E}[P'] = \frac{1}{2} + \frac{1}{2} Q_1(m, n)$$

recovering the value proven in [3]. Noting that $\frac{n^{\underline{k}}}{m^k} \leq \alpha^k$ and using that $\sum_{k \geq 0} (k+1)\alpha^k = d/d\alpha(\sum_{k \geq 0} \alpha^k) = (1-\alpha)^{-2}$ gives the desired bound $\mathbb{E}[P'] \leq \frac{1}{2}(1 + (1-\alpha)^{-2})$

**Average Cost to Query an existing item.** To go from $\mathbb{E}[P']$, the cost for a newly inserted key, to $\mathbb{E}[P]$, the average cost to query a key in the table, we average over the cost to insert the first $n$ keys. Because the cost of an unsuccessful search in linear probing is the same as the cost to insert a key, this is the same as averaging

over unsuccessful searches over tables storing 1 through $n$ elements. So $\mathbb{E}[P] = \frac{1}{n}\sum_{i\geq 0}^{n}\mathbb{E}[P_i']$ where $\mathbb{E}[P_i']$ is the expected cost to insert a key if there are $i$ items in the table. Thus

$$\begin{aligned}\mathbb{E}[P] &= \frac{1}{2} + \frac{1}{2n}\sum_{i=0}^{n-1} Q_1(m,i) \\ &= \frac{1}{2} + \frac{1}{2n}\sum_{i=0}^{n-1}\sum_{k\geq 0}(k+1)\frac{i^{\underline{k}}}{m^k} \\ &= \frac{1}{2} + \frac{1}{2n}\sum_{k\geq 0}\sum_{i=0}^{n-1}(k+1)\frac{i^{\underline{k}}}{m^k}\end{aligned}$$

Now we use the rules of "finite calculus", which says that if $\Delta f(x) = f(x+1) - f(x) = g(x)$, then $\sum_{i=a}^{b} g(x) = f(b+1) - f(a)$. Here we have $\Delta x^{\underline{k}} = kx^{\underline{k-1}}$, so $\sum_{x=0}^{n-1} kx^{\underline{k-1}} = n^{\underline{k}}$. Plugging this in above, we have

$$\begin{aligned}\mathbb{E}[P] &= \frac{1}{2} + \frac{1}{2n}\sum_{k\geq 0}\frac{n^{\underline{k+1}}}{m^k} \\ &= \frac{1}{2} + \frac{1}{2}Q_0(m,n-1)\end{aligned}$$

Using the same bound of $\frac{n^{\underline{k}}}{m^k} \leq \alpha^k$, it follows that $E[P] \leq \frac{1}{2}(1 + \frac{1}{1-\alpha})$.

*A.2.3 Proof of Costs Using Partial-Key Hashing* To reason about the costs for partial-key hashing, we need to show how insertion and probe costs depend on $S_{|L}$, the multi-set of keys. To clear up notation, we will refer to all multisets in theorems as $C$, $C'$, or $C''$. For any multiset $C = (K, z)$, we have that $|C| = |K|$ is the number of distinct partial-key elements and $||C|| = \sum_{x\in C} z(x)$ is the total number of elements in $C$. We additionally extend our shorthand notation $z_x = z(x)$ to sets, i.e. $z_C = \sum_{x\in C} z_x = ||C||$. Additionally, we define $C_{2+} = x : z_x \geq 2$ to be the set of keys which are not unique in $C$.

**Theorem 1,2 and 3 Restatements.** Looking at theorem 1, it's proof holds for partial-key hashing. Namely, if $C$ is a multi-set with $||C|| = n \leq m$, and $C$ is hashed into locations $[1, m-1]$, then the number of hash sequences such that 0 is empty is

$$(1 - \frac{n}{m})h(m, C)$$

where $h(m, C)$ is the number of ways to hash the objects into $[0, m-1]$. For a given multiset $C$, we have $h(m, C) = m^{|C|}$.

To generalize theorem 2, we start by analyzing the form of equation (6) for $g(m, n, k)$. The equation consists of three parts: 1) ways to divide the set of $n$ objects into one set with $k-1$ items and another with $n-k+1$ items, 2) ways to hash the $k-1$ and $n-k+1$ items into their arrays of size $k$,$m-k$, and 3) the scaling factors that say only $\frac{1}{k}, \frac{m-n-1}{m-k}$ of those leave slots 0 and k empty. It follows that when hashing a multi-set of items $C$, that

$$g(m, C, k) = \sum_{\substack{C'\subseteq C \\ ||C'||=k-1}} \frac{1}{k}\frac{m-n-1}{m-k}h(k, C')h(m-k, C \setminus C')$$

is the number of hash combinations that lead to a hash table with 0 empty, positions 1 to $k-1$ filled, and $k$ empty.

Theorem 3 also holds for multisets, but we need new notation to be clearer about what items are included in the new chain. We define $T_{C'}$ to be the expected chain length of a new item if all items in the multiset $C' \subseteq C$ are guaranteed to have the same hash value as the newly inserted item.

Theorem 4. *Let $C$ be the multi-set of items being hashed into the hash table and let $C'$ be a multi-set of items such that the items in $C'$ are fixed to have the same hash value as a newly inserted item for the hash table. Let $x \notin C'$, and define $C'' = C' \cup \{(x, z_x)\}$. Then*

$$P(x \in T_{C'}) = \frac{1}{m}\,\mathbb{E}[T_{C''}]$$

Proof. We start by deriving the equation for $\mathbb{E}[T_{C''}]$. Using the same reduction as in the proof of Theorem 3, the number of sequences for which the a new item is hashed into a chain of length $t$, given that the items in $C''$ hash to the same location as the new item, is the same as the number of chains with 0 empty, $1, \ldots, t-1$ full, $t$ empty, and for which the items in $C''$ hash to some location $k$ between 0 and $t-1$. For a fixed chain $C^* \subset (C \setminus C'')$ with $||C^*|| = t - 1 - ||C''||$, the number of ways to hash $C^*$ and $C''$ into $0, \ldots, t-1$ with the elements of $C''$ fixed to have the same hash location is equivalent to the number of ways to hash a new multi-set consisting of $C^* \cup (a, z_{C''})$ for an arbitrary new element $a$. By theorem 1 for multi-sets, it follows that the number of hash combinations such that $C^* \cup C''$ are between 0 and $t-1$ and location 0 is empty is $(1 - \frac{t-1}{t})\big(h(t, C^*) \cdot t\big)$. It follows that $f(m, C, t, C'')$, the number of hash combinations that produce a chain of length $t$, satisfies

$$f(m, C, t, C'') = \sum_{\substack{C^* \subset (C \setminus C'') \\ ||C^*|| = t-1-||C''||}} \frac{m-n-1}{m-t} h(t, C^*) h(m-t, C \setminus (C^* \cup C''))$$

There are $m^{|C \setminus C''|}$ possible hash sequences and so it follows that $E[T_{C''}] = m^{-|C \setminus C''|} \sum_{t \geq 1} t f(m, C, t, C'')$.

We now derive the equation for $P(x \in T_{C'})$. The same logic again applies, with the exception that now $x$ has a degree of freedom in its hash location. So for a fixed chain $C^* \subset (C \setminus C'')$ with $||C^*|| = t - 1 - ||C''||$, the number of ways to hash $C^*$, x, and $C'$ into $0, \ldots, t-1$ with the elements of $C'$ fixed to have the same hash location is equivalent to the number of ways to hash a new multi-set consisting of $C^* \cup \{(a, z_{C'}), (x, z_x)\}$ for an arbitrary new element $a$. Again by theorem 1 for multi-sets, this implies the number of hash cominbations where $x, C'$, and $C^*$ are hashed between 0 and $t-1$ with location 0 empty is $(1 - \frac{t-1}{t})\big(h(t, C^*) \cdot t^2\big)$. It follows that

$$\begin{aligned}
P(x \in T_{C'}) &= m^{-|C \setminus C'|} \sum_{t \geq 1} \sum_{\substack{C^* \subset (C \setminus C'') \\ ||C^*|| = t-1-||C''||}} \frac{m-n-1}{m-t} t \cdot h(t, C^*) h(m-k, S_{|L} \setminus C^*) \\
&= m^{-1} m^{-|C \setminus C''|} \sum_{t \geq 1} t \cdot f(m, C, t, C'') \\
&= \frac{1}{m} E[T_{C''}]
\end{aligned}$$

□

**Bounds on $\mathbb{E}[T_{C'}]$.** Theorem 4 again gives us a bridge between the probability of an item being in the same chain as a new item and the expected chain length when that item is hashed to the same location as a new item. We now use Theorem 4 to prove the following bound by induction on the set of unallocated objects, i.e. $C \setminus C'$:

$$\mathbb{E}[T_{C'}] \leq z_{C'} Q_0(m, n - z_{C'}) + Q_1(m, n - z_{C'}) + \sum_{x \in (C \setminus C')_{2+}} \frac{z_x^2}{m} Q_1(m, n - z_{C'})$$

Base case: Our base case is all sets such that $C' = C$, i.e. all objects in $C$ are guaranteed to hash to the same location as the newly inserted item. The length of $T$ is guaranteed to be $z_{C'} + 1 \leq z_{C'} Q_0(m, 0) + Q_1(m, 0)$.
Induction Steps: Assume our induction hypothesis holds for all sets $C, C' \subset C$ with $|C \setminus C'| \leq k$. Assume now that are given sets $C, C'$ such that $|C \setminus C'| = k + 1$, and that we wish to calculate $E[T_{C'}$. As shorthand, let $C^- = C \setminus C'$. Then

$$
\begin{aligned}
\mathbb{E}[T_{C'}] &\leq (1+z_{C'}) + \sum_{x\in C^-} z_x P(x \in T_{C'}) \\
&= (1+z_{C'}) + \frac{1}{m}\sum_{x\in C^-} z_x E(T_{C'\cup\{(x,z_x)\}}) \\
&\leq (1+z_{C'}) + \frac{1}{m}\sum_{x\in C^-} z_x\Big[(z_x+z_{C'})Q_0(m,n-z_{C'}-z_x) + Q_1(m,n-z_{c'}-z_x) + \sum_{y\in C^-, y\neq x}\frac{z_y^2}{m}Q_1(m,n-z_{C'}-z_x)\Big] \\
&\leq (1+z_{C'}) + \frac{n-z_{C'}}{m}z_{C'}Q_0(m,n-z_{C'}-1) + \sum_{x\in C^-}\frac{z_x^2+z_x}{m}Q_0(m,n-z_{C'}-1) \\
&\quad + \frac{n-z_{C'}}{m}Q_1(m,n-z_{C'}-1) + \sum_{x\in C^-}\frac{z_x^2}{m}\frac{n-z_{C'}}{m}Q_1(m,n-z_{C'}-1) \\
&= z_{C'}(1+\frac{n-z_{C'}}{m}Q_0(m,n-z_{C'}-1)) && (7)\\
&\quad + 1 + \frac{n-z_{C'}}{m}Q_0(m,n-z_c-1) + \frac{n-z_{C'}}{m}Q_1(m,n-z_{C'}-1) && (8)\\
&\quad + \sum_{x\in C^-_{2+}}\frac{z_x^2}{m}(Q_0(m,n-z_c-1) + \frac{n-z_{C'}}{m}Q_1(m,n-z_{C'}-1)) && (9)
\end{aligned}
$$

Now, using the rules $\frac{n}{m}Q_1(m,n-1) = Q_1(n,m) - Q_0(n,m)$ and $\frac{n}{m}Q_0(m,n-1) = Q_0(m,n) - 1$ we can reduce equations (7),(8), and (9) respectively to $z_{C'}Q_0(m,n-z_{C'})$, $Q_1(m,n-z_{C'})$, and $\sum_{x\in(C\setminus C')_{2+}}\frac{z_x^2}{m}Q_1(m,n-z_{C'})$ respectively. This gives the required result on $\mathbb{E}[T_{C'}]$.

It follows for a query on a new item $k$ such that $L(k)=y$, that it has a bound on its expected chain length of

$$E[T] \leq z_y Q_0(m,n-z_y) + Q_1(m,n-z_y) + \sum_{x\neq y\in C}\frac{z_x^2}{m}Q_1(m,n-z_y)$$

**Connecting $\mathbb{E}[T]$ to $\mathbb{E}[P']$.** Depending on whether $z_y = 0$, the connection between chain length and probe length changes. For items which match no other partial-keys in the dataset, we can use the same uniformity assumptions as before to recover that $\mathbb{E}[P'] = \frac{1}{2} + \frac{1}{2}\mathbb{E}[T]$. For items which match at least one partial-key, this uniformity assumption does not hold, and indeed items with larger $c_y$ values are more likely to be hashed towards the beginning of their chain. Thus, we use that $P \leq T$ to say $\mathbb{E}[P] \leq \mathbb{E}[T]$, and note this is a loose bound when $z_y$ is small. The result, overall, is equation 3, restated here for convenience.

$$\mathbb{E}[P'] \leq \begin{cases} \frac{1}{2}\big(1 + \frac{1}{(1-\alpha)^2} + \sum_{x\neq y}\frac{z_x^2}{m(1-\alpha)^2}\big) & \text{if } z_y = 0 \\ \frac{z_y}{1-\alpha} + \frac{1}{(1-\alpha)^2} + \sum_{x\neq y}\frac{z_x^2}{m(1-\alpha)^2} & \text{if } z_y > 0 \end{cases}$$

**Average query cost for an existing key.** The average cost to query a key in linear probing is unaffected by the order in which keys are inserted. Thus, we may assume any insertion order we desire in order to ease the analysis of the average number of probes for existing keys.

Because our analysis in the prior section created looser bounds for $z_y > 0$, namely in that there was no uniformity assumption on where items were in a chain, we add all duplicate keys first and then all non-duplicate

keys after. Expanding this out, we have

$$\begin{aligned}
\mathbb{E}[P] &\leq \frac{1}{n}\Bigg(\sum_{x\in C_{2+}}\frac{z_x^2}{2}Q_0(m,d)+\sum_{i=0}^{d-1}\Big[Q_1(m,i)+\sum_{x\in C_{2+}}\frac{z_x^2}{m}Q_1(m,i)\Big]\Bigg)+\frac{1}{2n}\Big[\sum_{i=d}^{n-1}\big(1+Q_1(m,i)+\sum_{x\in C_{2+}}\frac{z_x^2}{m}Q_1(m,i)\big)\Big]\\
&\leq \frac{n-d}{2n}+\frac{1}{2}Q_0(m,n-1)+\frac{c+d}{2n}Q_0(m,d-1)+\frac{c}{2m}(Q_0(m,n-1)+\frac{d}{n}Q_0(m,d-1))\\
&\approx \frac{1}{2}(1+\frac{1}{1-\alpha})+\frac{c}{n}+\frac{c}{m}\frac{1}{1-\alpha}\\
&\leq (\frac{1}{2}+\frac{c}{n})(1+\frac{1}{1-\alpha})
\end{aligned}$$

Here we use that $\alpha_d = \frac{d}{m} \approx 0$ so that $\frac{1}{1-\alpha_d} \approx 1$. Because we are interested in the case that duplicate items are rare, this is a very good approximation.

**Random Data.** To go from fixed data to random data, we condition via Adam's law. For querying for a missing key, we first rewrite the expectation as

$$\begin{aligned}
\mathbb{E}[P'|y,S_{|L}] &\leq \frac{1}{2}+\mathbb{1}_{z_y\neq 0}(\frac{z_y}{1-\alpha}-\frac{1}{2})+(\frac{1}{2}+\frac{1}{2}\mathbb{1}_{z_y\neq 0})\frac{1}{(1-\alpha)^2}+(\frac{1}{2}+\frac{1}{2}\mathbb{1}_{z_y\neq 0})\sum_x\frac{z_x^2}{m}\frac{1}{(1-\alpha)^2}\\
&\leq \frac{1}{2}(1-\frac{1}{(1-\alpha)^2})+\frac{\mathbb{1}_{z_y\neq 0}z_y}{1-\alpha}+\frac{\mathbb{1}_{z_y\neq 0}}{(1-\alpha)^2}+\sum_x\frac{z_x^2}{m}\frac{1}{(1-\alpha)^2}
\end{aligned}$$

Then using that $E[\mathbb{1}_{z_y\neq 0}] \leq \mathbb{E}[z_y] = \mathbb{E}[\mathbb{1}_{z_y\neq 0}z_y]$, that $\mathbb{E}[z_y] \leq n2^{-H_2(L(X))}$ by the union bound, and that $E[\sum_{x\in S_L} z_x^2] = n^2 2^{-H_2}$, we have

$$\begin{aligned}
\mathbb{E}[P'] &\leq \frac{1}{2}(1+\frac{1}{(1-\alpha)^2})+\frac{n2^{-H_2(L(X))}}{1-\alpha}+\frac{n2^{-H_2L(X))}}{2(1-\alpha)^2}+\sum_x\frac{\alpha n2^{-H_2(L(X))}}{(1-\alpha)^2}\\
&\leq \frac{1}{2}(1+\frac{1}{(1-\alpha)^2})+n2^{-H_2(L(X))}\frac{3}{2(1-\alpha)^2}
\end{aligned}$$

More straightforwardly,

$$\mathbb{E}[P] \leq \frac{1}{2}(1+\frac{1}{1-\alpha})+n2^{-H_2(L(X))}(1+\frac{1}{1-\alpha})$$

This concludes the analysis of linear probing, proving the initial equations given in the paper.

## Appendix B Robustness of Entropy-Learned Hashing Data Structures

**Robustness.** For most algorithms and data structures, including those listed here, there is a trade-off between their robustness and their expected performance. How to choose between the expected performance and robustness is highly dependent on the application; certain applications might accept 50% better average performance for a 1% chance at 2X worse performance whereas others may not. For Entropy-Learned Hashing and other techniques which exist as a base layer of potential use to many applications, the goal is to make this trade-off minimal, so that large gains in expected performance come at little cost for robustness. For Entropy-Learned Hashing, the amount of robustness for the hashing task depends on 1) the assumptions of Entropy-Learned Hashing, and 2) the task at hand. We start by discussing the assumptions of Entropy-Learned Hashing and then discuss Entropy-Learned Hashing for hash tables, Bloom Filters, and partitioning individually.

**Assumptions of Entropy-Learned Hashing.** Entropy-Learned Hashing makes weak assumptions, namely that data which are somewhat random remain somewhat random. Thus, unlike the usual case for learning

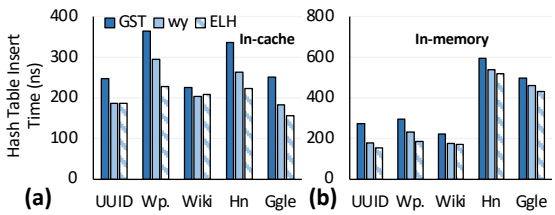



Fig. 1. Entropy-Learned Hashing reduces insert times for hash tables across datasets, and data sizes.

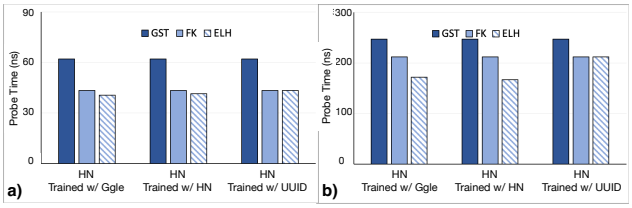



Fig. 2. Entropy-Learned Hashing keeps good performance under some data shifts and is never worse than traditional hashing.

where shifts in data distributions almost always cause degradations in performance, shifts in data distribution for Entropy-Learned Hashing cause no change in performance as long as data stays random. In particular, if a distribution $D_1$ shifts to distribution $D_2$ this only causes a performance shift if $H_2(L(D_2)) << H_2(L(D_1))$. Thus the main robustness of Entropy-Learned Hashing comes from the fact that changes in distribution are unlikely to degrade performance. Still, we must address the case where $D_2$ actually becomes predictable on the bytes we choose to hash. We now cover this for each data structure.

**Hash Tables.** Entropy-Learned Hashing for hash tables is the most robust. This is for multiple reasons. First, if collisions are as expected on keys in the dataset, queries for both keys in the data and not in the data return quickly (see equations (4) and (5) for example). Second, we can monitor collisions during insertions with little overhead by counting the displacement from the initial hash position. Thus, this together with reason 1 means we have guaranteed good performance at all times. Third, Entropy-Learned Hashing can rehash if collisions ever deviate what is expected. The simplest way to do this is to revert to a full-key hash function, although other approaches such as backup bytes to include are an interesting future direction. Thus performance for hash tables can be guaranteed to be at least as good as traditional hashing (at the expense of code complexity) while usually being substantially better.

**Bloom Filters.** Bloom filters are less robust than hash tables. Their first defense in terms of robustness is the weak assumptions, namely that different distributions work for Entropy-Learned Hashing as long as the distribution is still random. Their second defense is that we can check that the data matches the desired distribution, namely because the # of set bits concentrates sharply around their expected value [1]. We use this fact to estimate the number of unique items in the filter. If data items are not as expected, or if queries are substantially different than the inserted items leading to a larger FPR, the filter must be rebuilt.

**Partitioning.** For partitioning, the cost of overloaded bins depends on the context, but for many contexts, such as in-memory radix partitioning, this can be solved by dividing the one or two overloaded bins into multiple bins. This is simple to implement and does not affect in-memory tasks like radix-partitioning much. For more complex tasks for which partitioning happens across a more expensive medium such as a network, approaches such as using the least loaded of d-bins would be of use [2].

## Appendix C Additional Experiments

This Section covers the additional experiments described in Section 6.6. This includes experiments on 1) the efficiency of creating Entropy-Learned Hash data structures, 2) probing separate chaining hash tables, 3) experiments showing robustness properties, 4) experiments with dependent accesses (i.e. hash table lookups and Bloom filter lookups which must run one after the other instead of in parallel), and 5) additional experiments on Bloom filters showing different false positive rates.

**1. Entropy-Learned Hashing Reduces Hash Table Insert Time.** Entropy-Learned Hashing works similarly for inserts as it works for probes with hit rate = 1. Figure 1 presents hash table insert times for the five real-world

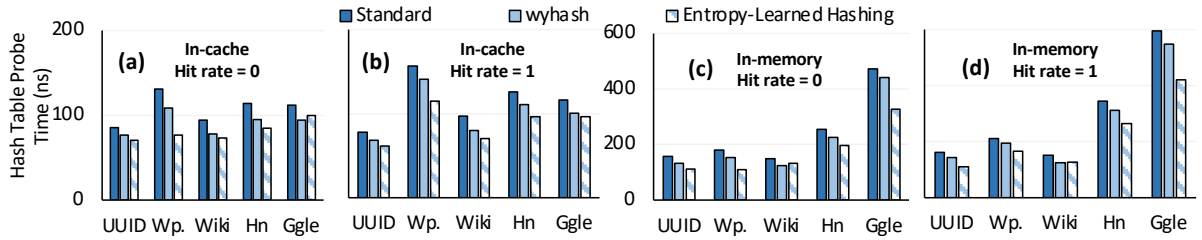



Fig. 3. Entropy-Learned Hashing reduces the probe time of standard chaining hash tables.

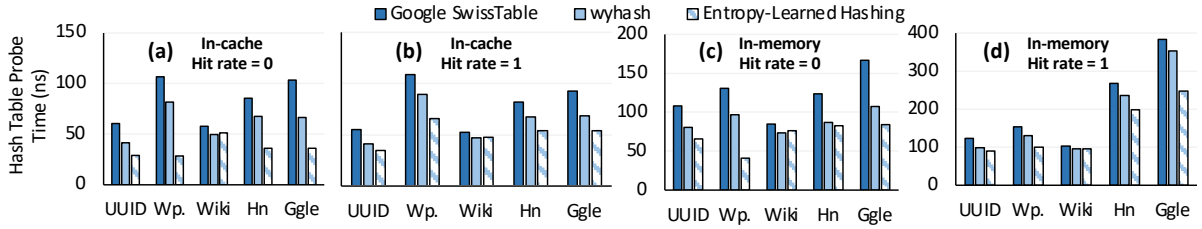



Fig. 4. Entropy-Learned Hashing reduces dependent probe times for hash tables across datasets, data sizes, and hit rates.

datasets we examine. The figure shows that Entropy-Learned hashing provides a 1.16× to 1.3× speedup over its closest competitor thanks to its reduced hash computation both for in-cache and in-memory data sizes. Thus, the speedups seen in query performance on Entropy-Learned hash tables carry over to performance on the insertion of data.

**2. Separate Chaining Experiments.** To show that Entropy-Learned Hashing works with separate chaining hash maps as well as linear probing hash maps, we integrate Entropy-Learned Hashing into std::unordered_map. The results can be seen in Figure 3. The key takeaway is that, as expected, Entropy-Learned Hashing reduces the probe times across datasets, data sizes, and hit rates, with this improvement being up to 1.72×. Thus Entropy-Learned Hashing works for separate chaining tables in addition to linear probing tables. The improvement is slightly lower than what is seen for SwissTable; this is mainly because std::unordered_map is much slower as a baseline and so portions of the hash table probe that are not the hash function take up more of the computation. A second change is that std::unordered_map is only a tiny amount faster when querying for non-existent keys than when querying for existent keys. This is again because unordered_map lacks optimizations present in SwissTable. Thus, a second takeaway of Figure 3 is that Entropy-Learned Hashing provides larger speedups for more optimized hash tables.

**3. Robustness of Entropy-Learned Hashing.** To test the robustness of Entropy-Learned hash tables, we test the performance of them when their training sets, i.e the set of sampled past queries and data items, does not match their test distribution of actual inserted and queried items. To test this, we vary the distribution given as a data sample and then insert and query items from the Hacker News dataset. Figures 2a,b show the results when using the large data size and querying for missing items (Fig. 2a) and existing items (Fig. 2b). As expected, when the Hacker News dataset is used to gather data as well as to insert and query items, we see speedups of 5% to 30%. More importantly, as Figures 2a and 2b show, when using Google URLs as a dataset to analyze ahead of time, using the Hacker News dataset to insert and query items still produces speedups of 5% to 27%. This shows the case where the test data is different in distribution but still random enough on the bytes of

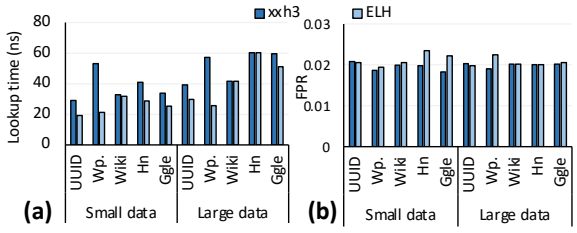



Fig. 5. Improving Bloom filter lookup time (a) and false positive rates (b) for small and large data sizes for dependent lookups.

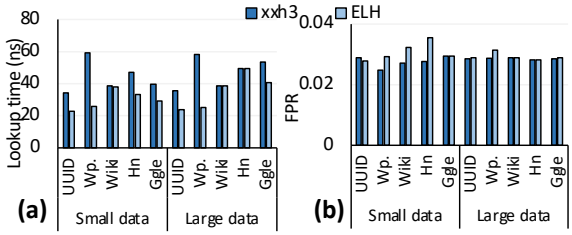



Fig. 6. Improving regular Bloom filter lookup time (a) and false positive rates (b) for small and large data sizes.

choice; in this case Entropy-Learned Hashing still provides speedups over full-key hashing. When using the UUID dataset as a training dataset, the user ids from UUID and posted urls from Hacker News are very different and Entropy-Learned Hashing defaults to using the full-key hash function as described in Section 2 of the technical report (and Section 5 of the main paper). Thus, in this case it provides no speedups but does not degrade performance.

**4. Entropy-Learned Hashing Reduces Dependent Probe Times.** Entropy-Learned Hashing reduces probe times for data structures when the probes are dependent (i.e. they cannot be batched). Figure 4 presents hash table probe times, and shows Entropy-Learned Hashing provides 1.18x to 2.9x speedups across different datasets, data sizes and hit rates. Figure 5 shows similar results for dependent Bloom filter lookups, where every lookup requires the result of the preceding lookup to start. The figure shows that Entropy-Learned Hashing significantly reduces the filter lookup times and provides a 1.16× to 2.5× speedup across datasets and data sizes.

The speedups are consistently lower for dependent lookups than they are for independent lookups. The reason is that independent lookups benefit from inter-lookup parallelism, as we discussed in Section 6.3. However, we observe a significant speedup even for dependent lookups. To illustrate, Entropy-Learned Hashing is 1.42× faster than wyhash for hash table probes on the Google dataset with large data and high hit rate. This is because there are two types of memory-level parallelism that Entropy-Learned Hashing benefits from: (i) inter-lookup parallelism, and (ii) intra-lookup parallelism. While independent hash table probes benefit from both inter- and intra-lookup parallelism, dependent hash table probes benefit from only intra-lookup parallelism. As a result, Entropy-Learned Hashing improves hash table probe time more when the probes are independent than it improves when the probes are dependent. To verify our hypothesis, we examined the MLP value for Entropy-Learned Hashing and wyhash for dependent hash table probes and found the MLP value is 2.0 for Entropy-Learned Hashing compared to 1.6 for wyhash. Thus ELH provides both parallelism benefits and computational benefits and produces consistent speedups for dependent lookups across small and large data sizes.

**5. Experiments with different FPRs for Bloom Filters.** The experiments in the main body of the text focused on throughput optimized filters, in particular using register-blocked Bloom filters from [4]. Figure 6 shows the performance of Entropy-Learned Hashing on traditional Bloom filters at a lower false positive rate of 1%. The results are extremely similar to Section 6.4, and so we don't cover them in detail. Entropy-Learned Hashing again provides benefits in lookup performance of up to 2.4× at negligible (and tunable) increases to the false positive rate.

**6. Entropy-Learned Hashing scales linearly with the number of threads.** Up to now, we have focused on single-threaded behavior for Entropy-Learned Hashing. In this section, we examine how Entropy-Learned Hashing behaves as the number of threads is increased. We increase the number of threads concurrently probing

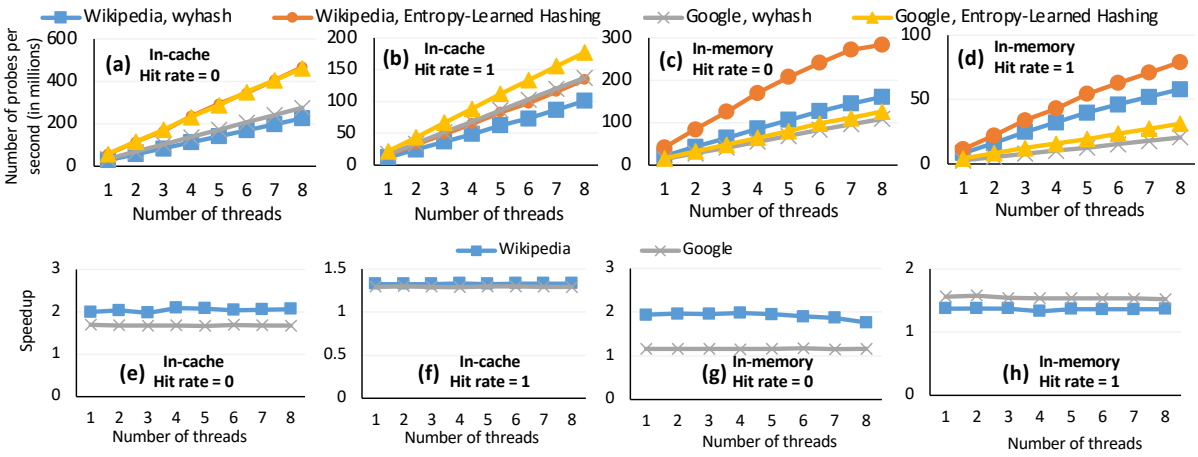



Fig. 7. Entropy-Learned Hashing scales linearly with the number of cores.

the same table and examine the number of probes per second by all the concurrently running threads. We pin each thread to a separate physical core without any hyper-thread sharing.

Figure 7 presents the results. The figure shows that both Entropy-Learned Hashing and wyhash scale linearly with the number of threads. As the number of threads is increased, the delivered throughput is increased in proportion with the number of threads. As a result, Entropy-Learned Hashing provides constant speedups over wyhash for all the number of threads.

## References

[1] A. Broder, M. Mitzenmacher, and A. B. I. M. Mitzenmacher. Network applications of bloom filters: A survey. In *Internet Mathematics*, pages 636–646, 2002.

[2] R. M. Karp, M. Luby, and F. Meyer auf der Heide. Efficient pram simulation on a distributed memory machine. In *Proceedings of the Twenty-Fourth Annual ACM Symposium on Theory of Computing*, STOC '92, page 318–326, New York, NY, USA, 1992. Association for Computing Machinery.

[3] D. E. Knuth. *The Art of Computer Programming, Volume 3: (2nd Ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., USA, 1998.

[4] H. Lang, T. Neumann, A. Kemper, and P. Boncz. Performance-optimal filtering: Bloom overtakes cuckoo at high throughput. *Proceedings of the VLDB Endowment*, 12(5):502–515, 2019.