Federated Heavy Hitters revisited

Graham Cormode

June 2021

1 Introduction

The problem of finding the most frequent items from a collection is a core analytics task that supports a range of objectives, from simple popularity charts, to instantiating complex language models. Due to the sensitivity of data used within these applications, it is necessary to apply strong privacy protections to the data.

The recent wok on Federated Heavy Hitters discovery [1] describes an algorithm to collect information from a set of distributed clients, who each hold a (private) item. We can treat these items as strings of characters over a fixed alphabet. The algorithm proceeds in a series of L rounds to build up a trie describing the frequent items among the client population. In each round, the server contacts a random sample of m clients, and shares the current trie with them. Each client replies if its item extends the trie, and if so the client "votes" for the prefix that its item extends, along with the next character.

The server receives these votes, and tallies them. If there are more than a threshold θ of votes for a particular extension of the trie, then this extension is added to the trie as part of the next level. We assume that the procedure stops after the trie has been built out to L levels, or if the trie cannot be extended beyond a certain level.

In the original paper, the protocol is analyzed, and it is shown that the trie of prefixes built in this fashion meets the requirements of (ε, δ)-Differential privacy. In this note, we extend this approach to heavy hitters as follows:

- We show that we can augment the trie that is output with counts of the items received from the clients and still meet (ε, δ)-DP.
- We show that these counts provide accurate frequency estimates for items and prefixes from the input
- We show that the resulting data structure can also answer quantile and range queries
- Our proofs are compact and self-contained.

2 Privacy Analysis including counts

We adopt the algorithm and notation from the TrieHH paper [1].

The algorithm proceeds over L levels. At each level, we sample m clients to report on their items that extend the current trie. We add items to the current trie based on an (absolute) threshold θ. We augment the output of TrieHH by including the observed count of prefix for each node in the trie, provided it is more than a fixed threshold θ.

Lemma 1. *The probability that the number of samples of a prefix in a round is more than θ times its expectation is at most* $\exp(-\theta \ln \theta + \theta - 1)$ *for any* $\theta > 1$.

Proof. Given a prefix that occurs k times in the input, the probability that it is picked in one sample is $p := k/n$. To simplify the analysis, we analyze the case of sampling the clients with replacement. Note that this only increases the chances that a particular prefix is picked. The expected number of sampled occurrences is then $pm = km/n$.

Let X denote the random variable that counts the number of successes (times the prefix is picked) out of the m trials, so $\mathsf{E}[X] = km/n$. Then, X is a sum of m Bernoulli random variables with parameter p. We do a case split on p:

Case: $p \le 1/m$. If $p \le 1/m$, we apply an (additive) Chernoff-Hoeffding bound to the mean of the m trials:

$$\Pr[X \ge \theta] = \Pr\left[\frac{1}{m}X - \frac{1}{m}\mathsf{E}[X] \ge (\theta p - p)\right] \le \exp\left(-D\left(\frac{\theta}{m}\middle\|\frac{1}{m}\right)m\right).$$

Here, $D(p\|q)$ denotes the K-L divergence (relative entropy) between the (Bernoulli) distributions with parameters p and q. We have

$$\begin{aligned}
-D(p\|q)m &= -\theta\ln\left(\frac{\theta}{m}\cdot\frac{m}{1}\right) - (m-\theta)\ln\left(\frac{m-\theta}{m}\cdot\frac{m}{m-1}\right) \\
&= -\theta\ln\theta - (m-\theta)\ln\left(1-\frac{\theta-1}{m-1}\right) \\
&= -\theta\ln\theta + (m-\theta)\ln\left(\frac{m-1}{m-\theta}\right) \\
&= -\theta\ln\theta + (m-\theta)\ln\left(1+\frac{\theta-1}{m-\theta}\right) \\
&\le -\theta\ln\theta + \theta - 1
\end{aligned}$$

Hence,

$$\Pr[X \ge \theta] \le \exp(-\theta\ln\theta + \theta - 1) \tag{1}$$

Case: $p > 1/m$. If $p > 1/m$, we apply a (multiplicative) Chernoff bound:

$$\Pr[X \ge \theta\mathsf{E}[X]] \le \exp(-\theta^2\mathsf{E}[X]/3) = \exp(-\theta^2 mp/3) \le \exp(-\theta^2/3)$$

Since $\theta\ln\theta - \theta + 1 < \theta^2$ for $\theta \ge 1$, we can use $\exp(\theta\ln\theta - \theta + 1)$ as our bound on the probability of both cases. □

To achieve a target error bound δ, we rearrange to obtain $\frac{\theta}{e}\ln\frac{\theta}{e} = \frac{1}{e}\ln(1/e\delta)$, and apply Lambert's W function. This gives $\frac{\theta}{e} = W(\frac{1}{e}\ln(1/e\delta))$, i.e., $\theta = eW(\frac{1}{e}\ln\frac{1}{e\delta})$.

We next give a bound on the ratio of probabilities of seeing the same output on neighboring inputs. This follows the same proof outline as in the original paper, but now considers nodes with a count τ attached.

Lemma 2. *Given two neighboring inputs D, D', such that D differs in one item from D', the ratio of probabilities of extending the trie with a node with label τ is bounded by $\frac{k+1}{k+1-\tau}$, where $k+1$ is the number of copies of the extra item in input D and $k \ge \tau \ge \theta$.*

Proof. The case to focus on is when input D has one extra copy of a particular prefix compared to D', at some intermediate stage of the algorithm. For notation, we will write $S_k(n,m,\tau)$ to denote the number of ways to succeed in collecting exactly τ instances of the target prefix while picking m items out of n, when there are k total instances of the prefix. We can observe that there is a simple combinatorial expression for this quantity: we count the number of combinations where we pick a particular subset of size τ from the k instances, and a particular subset of size $m-\tau$ from the remaining $n-k$ examples.

$$S_k(n,m,\tau) = \binom{k}{\tau}\binom{n-k}{m-\tau} \tag{2}$$

Our goal is to bound the ratio of probabilities of seeing this part of the output between D, who has $k+1$ copies of the prefix, and D' who holds k copies. Observe that the probability for D is $S_{k+1}(n,m,\tau)/\binom{n}{m}$, and for D' it is $S_k(n,m,\tau)$. Then this ratio is given by

$$\frac{S_{k+1}(n,m,\tau)}{S_{k+1}(n,m,\tau)} = \frac{\binom{k+1}{\tau}\binom{n-k-1}{m-\tau}}{\binom{k}{\tau}\binom{n-k}{m-\tau}} = \frac{(k+1)(n-k-m-\tau)}{(n-k)(k+1-\tau)} = \left(1-\frac{m+\tau}{n-k}\right)\left(\frac{k+1}{k+1-\tau}\right) \le \frac{k+1}{k+1-\tau}$$

□

Lemma 3. *The protocol satisfies* (ε, δ)*-DP.*

Proof. The protocol proceeds in rounds. We consider two inputs D and D', such that D has one more copy of a particular string than D', and argue that the DP guarantee holds on the output of the protocol. Denote this string as t. We aim to show that the probability that the mechanism, M, applied to input D, behaves similarly to the same mechanism applied to D'. Specifically, we want to show that

$$\Pr[M(D) \in R] \leq \exp(\varepsilon) \Pr[M(D') \in R] + \delta$$

for any set of feasible outputs R.

First, we assume that the number of samples of the prefix t in any round is never more than θ times its expectation in the execution of $M(D)$. Call this event E. By Lemma 1, event E holds except with probability $p_\theta = L\exp(-\theta \ln \theta + \theta - 1)$. We condition on E holding, and just account for this probability in our final reckoning.

It now suffices to consider a particular output trie $T \in R$, and argue that $\Pr[M(D) = T|E] \leq \exp(\varepsilon) \Pr[M(D') = T|E]$ (conditioned on E). We proceed round by round. Let t_i denote the string t truncated to level i. First, observe that for any string $t' \neq t$, if $t'_i \neq t_i$, then at level i, t'_i has the same number of copies in D and D', and so is treated identically in both cases. Hence, we can focus only on the treatment of t_i in the two cases. Suppose at round i, the count of t_i for D is less than m/n. Then, by our assumption on θ, D will not sample θ copies of t at this level, and so both D and D' would output the same tree T_i that does not include an extension t. Hence, in this case, $\Pr[M(D) = T_i|E] = \Pr[M(D') = T_i|E]$.

Re-derive from the first assumption: Figured it out

Otherwise, the count of t_i is at least m/n, and by our assumption D samples at most $\tau \leq \theta m(k+1)/n$ copies of t_i. Then we can state

$$\frac{\Pr[M(D) = T_i|E]}{\Pr[M(D') = T_i|E]} \leq \frac{k+1}{k+1-\tau} \leq \frac{k+1}{k+1-\theta(k+1)m/n} = \frac{n}{n-\theta m} \tag{3}$$

We will assume that $m \leq \frac{n}{10\theta}$. The effect is to ensure that the sample size m is a small fraction of n. Substituting this assumption in (3), we conclude

$$\frac{\Pr[M(D) = T_i|E]}{\Pr[M(D') = T_i|E]} \leq 1 + \frac{m\theta}{n - n/10} = 1 + \frac{10m\theta}{9n} \tag{4}$$

Unclear where this is coming from

Combining these two cases, we have

$$\frac{\Pr[M(D) = T|E]}{\Pr[M(D') = T|E]} = \prod_{i=1}^{L} \frac{\Pr[M(D) = T_i]}{\Pr[M(D') = T_i]} \leq \left(1 + \frac{10m\theta}{9n}\right)^L := \exp(\varepsilon)$$

Then we have $\varepsilon \leq L\ln(1 + \frac{10m\theta}{9n})$. Since $\frac{10m\theta}{9n} \leq 1$, we can bound $\varepsilon \leq \frac{10Lm\theta}{9n}$.

Finally, we can write

$$\begin{aligned}
\Pr[M(D) \in R] &\leq \Pr[E] \Pr[M(D) \in R|E] + (1 - \Pr[E]) \\
&\leq \Pr[E] \sum_{T \in R} \Pr[M(D) = T|E] + \Pr[\sim E] \\
&\leq \Pr[E] \exp(\varepsilon) \sum_{T \in R]} \Pr[M(D') = T|E] + \Pr[\sim E] \\
&= \exp(\varepsilon) \Pr[E] \Pr[M(D') \in R|E] + \Pr[\sim E] \\
&\leq \exp(\varepsilon) \Pr[M(D') \in R] + \Pr[\sim \mathsf{E}]
\end{aligned}$$

Therefore, we have a guarantee of (ε, δ) differential privacy, with $\varepsilon \leq \frac{10Lm\theta}{9n}$ and $\delta \leq L\exp(\theta \ln \theta + \theta - 1)$. □

[If desired, the expression for δ can be tightened to $\delta \leq \exp(\theta \ln \theta + \theta - 1) + L\exp(-\theta^2/3)$]

3 Accuracy Bounds

The trieHH algorithm is ultimately based on sampling and pruning, so for prefixes whose frequency is sufficiently above the pruning threshold, then its frequency within the trie is an (almost) unbiased estimate of its true frequency.

There is a small gap, since even for a prefix with high frequency, there is a small chance that it is not sampled often enough, and so its estimate will fall below the threshold θ (otherwise, we do not materialize the node).

We first consider the probability that a frequent item is not reported by the algorithm. For a prefix with (absolute) frequency W out of the n input items, it is reported at level i (conditioned on its length $i-1$ prefix being reported at level $i-1$) if the number of sampled occurrences exceeds θ. Similar to the analysis above, we can apply a Chernoff-Hoeffding bound to the random variable X that counts the number of occurrences of the prefix. Now, the probability of each sample picking the prefix is W/n, and the expected number in the sample is $Wm/n > \theta$. For convenience, we will write $w = Wm/n$ for this expectation.

We have that[1]

$$\Pr[X \leq \theta] = \Pr\left[X \leq \frac{\theta}{w} w\right] = \Pr\left[X \leq \left(1 - \frac{w-\theta}{w}\right) \mathsf{E}[X]\right] = \exp\left(-\frac{(w-\theta)^2}{2w}\right) \tag{5}$$

When w is sufficiently bigger than θ, this gives a very strong probability. For example, consider the case $n = 10^6$, and we set $\theta = 15$ to obtain a low δ of 10^{-12}. The sample size $m = 12,000$, and for an item that occurs 1% of the time in the input, we expect to sample it $w = 120$ times. This gives a bound of $\exp(-45) = 10^{-20}$ that such an item is not detected at any level. Over $L = 10$ levels, a union bound ensures that the chance of missing it remains at most 10^{-19}.

More generally, we can use the frequency of any prefix in the trie as an estimate for its true occurrence rate. Applying the same Chernoff-Hoeffding bound as above, we have that

$$\Pr[X \leq (1-\gamma)\mu] = \exp(-\gamma^2 \mu/2) = \beta$$

Rearranging, we obtain $\mu = \frac{2}{\gamma^2} \ln 1/\beta$. Suppose we aim to find all items whose frequency is at least ϕ, and estimate their frequency with relative error at most γ. Then we have $\mu = \phi m \geq \phi \frac{\varepsilon n}{L\theta} = \frac{2}{\gamma^2} \ln(1/\beta)$.

We can substitute values into this expression to explore the space. For example, if we set $\varepsilon = 2$, $L = 10$, $\ln 1/\beta = 10$, $\theta = 10$ and $\gamma = 1/\sqrt{10}$, then we obtain $\phi = 10^4/n$ — in other words, provided $n > 10^6$, we can accurately find estimates of frequencies that occur 1% of the time (except with vanishingly small probability).

4 Quantiles via TrieHH

We can observe that the Trie built during the TrieHH protocol can also be used to answer quantile queries, with the same privacy (and similar accuracy) guarantees.

Now each client has an input value in the range $[0,1]$ (say), and we can interpret these as prefixes, corresponding to subranges. For example, if we set the "alphabet size", α to 4, then the input value $\frac{1}{3}$ falls in the range $[0.25, 0.5]$ for a prefix of length 1; and in the range $[\frac{5}{16}, \frac{6}{16}]$ for a prefix of length 2. With this mapping of values to prefixes, the algorithm proceeds as before, and outputs the (DP) trie with weights on nodes.

To answer a range query $[0,r]$, we decompose the range greedily into chunks that can be answered by the trie. For example, if $\alpha = 4$, and we want the range $[0, 0.7]$, we find the chunks $[0, \frac{1}{4}], [\frac{1}{4}, \frac{2}{4}]$ at level 1; $[\frac{8}{16}, \frac{9}{16}], [\frac{9}{16}, \frac{10}{16}], [\frac{10}{16}, \frac{11}{16}]$ at level 2; and so on.

Due to the pruning, we will not have information on any ranges whose sampled weight is less than θ, corresponding to a θ/m fraction of mass. This will give an error bound of $(\alpha-1)\theta/m$ per level, and so $L(\alpha-1)\theta/m$ over all levels. Based on our setting of m proportional to $\varepsilon n/(L\theta)$, we obtain a total error of $(\alpha-1)(L\theta)^2/\varepsilon n$.

Picking similar test values as above shows that this can give reasonable accuracy for n large enough. For $\theta = 10$, $L = 10$, $\alpha = 4$, $\varepsilon = 2$, the error bound yields $\frac{3}{2}10^4/n$. So for $n > 10^6$, we obtain rank queries (and quantiles) in this space with error around 0.015.

References

[1] Wennan Zhu, Peter Kairouz, Haicheng Sun, Brendan McMahan, and Wei Li. Federated heavy hitters discovery with differential privacy. *CoRR*, abs/1902.08534, 2019.

[1] Applying this bound assumes that we are sampling with replacement. Sampling without replacement is possible also, but will require a slightly different analysis.