Maximum Likelihood Estimation of Network Size by N Parallel k-lookups in Kademlia DHT

Abstract

..todo

1. Intro

The lookup (often informally referred to as "k-lookup") is the primary and most crucial function in the Kademlia peer-to-peer distributed hash table (DHT) protocol [1]. It is a recursive, iterative algorithm designed to locate the k closest nodes to a specific target ID within the network. The final set of k closest nodes found has XOR distances that are minimized within the network, and the magnitude of these distances is related to the density of nodes, which itself depends on network size n. It's clear that performing several lookups to random N targets nodes we get more information about anticipated network size. A method of estimation n by processing N parallel k-lookups is proposed in [2]. It's based on the fact that distances in each ordered set of k-lookup is the order statistics with specific beta distributions which depends on n. Then according to the method of moments (MoM) [3] the estimator for n is obtained by matching expected value of beta distribution with empirical result by averaging distances from lookup data. The suggested estimator is great from a practical point, however some aspects of provided solution are heuristic, e.g. the ways of averaging sample data. It's also hard to make assumptions on efficiency of the MoM estimator, because there is no way of gettins theoretical minimum variance for any unbiased estimator.

Using the same model as in [2], the current work considers Maximum Likelihood Estimation (MLE) [4]. The MLE is generally preferred over MoM because it produces more efficient, consistent, and often unbiased estimators with stronger theoretical properties. MLE requires no heuristic assumption and can be derived directly from join Probability Distribution Function (PDF) of distances from N parallel k-lookups. The join PDF also allows to get the Craner-Rao bound [5] for minimal variance of the efficient estimator.

The main results of works are:

- the MLE estimator (3) which uses average of logarithms of XOR distances from N random k-lookups;
- the Craner-Rao bound (4) for minimal variance of the efficient estimator and its approximations (6-7);
- illustration of efficiency of the MLE estimator by results of computer simulation.

2. Model

Kademlia distance estimation works by querying k “closest” nodes to a random target ID.

Let $D_1, \ldots, D_n$ be XOR distances from a probe node to all n network nodes, and $X_i = D_i/(2^L - 1)$ are normalized distances for L-bits addresses. $X_1, \ldots, X_n$ are modeled as independent random variables uniformly distributed on $(0, 1)$.

The result of a single lookup is the ordered set of k smallest distances $\{U_{(1)}, \ldots, U_{(k)}\}$ selected from $X_1, \ldots, X_n$. Here $U_{(i)}$ is the distance to the i-th closest node. So, we observe the first k order statistics [6] from a sample of normalized distances $X_1, \ldots, X_n$.

When performing several k-lookups to N random probes, we make an assumption that all N samples of distances, (so, the observed order statistics) are independent.

4. The likelihood function

For a single lookup the likelihood function (LF) is the joint PDF of the first k order statistics $U_{(1)}, \ldots, U_{(k)}$ from a sample of size n:

$$f_{U_{(1)},\ldots,U_{(k)}}(u_1, \ldots, u_k \mid n) = \frac{n!}{(n-k)!}(1-u_k)^{n-k}, \quad \text{for } 0 < u_1 < \cdots < u_k < 1 \quad (1)$$

We derive this later, but notice an important fact now: The optimal ML estimate of n can be obtained by taking into account only k-th order statistics. The statistics up to k are irrelevant to the estimate. In other words, **to estimate the network size only the last k-th distance in ordered lookup results is sufficient**.

This conclusion can be generalized for the optimal Bayesian estimator, because the LF as a part of posterior distribution is the only place which captures the observation.

Note, that heuristic solutions proposed in [2] takes in account all distances from the lookup result.

Using simplified notation, true in the area where LF differs from zero

$$f_1(u \mid n) = \frac{n!}{(n-k)!}(1-u)^{n-k}$$

and taking into account the assumption on independence of the samples, we get the LF for several k-lookups to N random nodes:

$$L(n) = \prod_{i=1}^{N} f_1(u_i \mid n) = \left[\frac{n!}{(n-k)!}\right]^N \prod_{i=1}^{N} (1-u_i)^{n-k}, \tag{2}$$

where the observation u_i is the k-th smallest normalized distance from i-th lookup.

5. Maximum Likelihood Estimate

The MLE estimator $\hat{n}$ maximizes (2). It means that $\hat{n}$ is an integer at which the likelihood sequence $L(n)$ stops increasing and starts decreasing. To find it, consider the ratio function $G(n) = L(n)/L(n-1)$ and conditions the $\hat{n}$ conforms to:

1. The likelihood is increasing up to $\hat{n}$: $G(\hat{n}) \geq 1$.
2. The likelihood starts decreasing after $\hat{n}$: $G(\hat{n}+1) < 1$.

As follows from (2), the ratio function is:

$$G(n) = \left(\frac{n}{n-k}\right)^N \prod_{i=1}^{N} (1-u_i)$$

and within rounding error the MLE is the root of equation $G(n) = 1$:

$$\hat{n} = \frac{k}{1 - \left[\prod_{i=1}^{N} (1-u_i)\right]^{1/N}}$$

From practical point the last expression is easier to implement in the equivalent logarithmic form:

$$\hat{n} = \frac{k}{1 - \exp(\overline{L}_u)}, \quad \text{where} \quad \overline{L}_u = \frac{1}{N} \sum_{i=1}^{N} ln(1-u_i) \tag{3}$$

Algorithmically, the estimator (3) performs following steps:

1. For each lookup, consisted of k smallest distances to random targets:
   - select the maximal distance d_i;
   - calculate logarithmic metric $ln(1-u_i)$ for the normalized distance $u_i = d_i/(2^L - 1)$.
2. Calculate average $\overline{L}_u$ on the metrics.
3. Estimate the network size.

As expected, the estimator (3) always produces result greater than k, because $\overline{L}_u$ is negative. The two edge cases when u_is are nearing to 0 ($\hat{n} \to \infty$) ot to 1 ($\hat{n} \to k$) also intuitively are understandable.

6. Efficiency of the estimate

MLEs are known for their strong theoretical properties. For large sample sizes, they are consistent (converge to the true parameter), asymptotically normal, and asymptotically efficient (achieve the lowest possible variance).

Treating n as a continuous parameter, the variance of the efficient estimator [5] is given by Cramér–Rao Low Bound (CRLB):

$$\mathrm{Var}(\hat{n}) \geq \frac{1}{I(n)}$$

where the Fisher Information is obtained by averaging of second derivative of the log-likelyhood (2):

$$I(n) = -E\left[\frac{\partial^2}{\partial n^2} \ln L(n)\right]$$

The log-likelihood for N observation is:

$$\ln L(n) = N \cdot \ln \frac{\Gamma(n+1)}{\Gamma(n-k+1)} + (n-k)\sum_{i=1}^{N} \ln(1-u_i),$$

where $\Gamma(x)$ is gamma function. Differentiating this two times, obtain:

$$I(n) = -N\left[\psi'(n+1) - \psi'(n-k+1)\right]$$

where $\psi'(x) = \frac{\partial^2}{\partial x^2} \ln \Gamma(x)$ is trigamma function [7].

So, the result for minimal variance is:

$$\mathrm{Var}(\hat{n})_{min} = \frac{1}{I(n)} = \frac{1}{N\left[\psi'(n-k+1) - \psi'(n+1)\right]} \tag{4}$$

Next, the difference of trigamma functions in (4) we approximate with:

$$\psi'(n-k+1) - \psi'(n+1) \approx \frac{1}{n-k} - \frac{1}{n} \tag{5}$$

The approximation (5) is quite good, especially when n is large and k is small relative to n. Indead, we get (5) using a known series representation [7] and substituting sum by an integral:

$$\psi'(n-k+1) - \psi'(n+1) = \sum_{j=n-k+1}^{n} \frac{1}{j^2} \approx \int_{n-k}^{n} \frac{1}{x^2}\, dx$$

Finally, in practice it is often covenient to normalize estimation error to the estimated value. Substituting (5) into (4) and dividing result by n^2, we obtain an expression for normalized variance, which defines the potential precision of the MLE estimator:

$$\sigma^2_{min} = \text{Var}\left(\frac{\hat{n}}{n}\right)_{\text{min}} \approx \frac{1}{N}\left(\frac{1}{k} - \frac{1}{n}\right), \quad \text{where} \quad n > k \tag{6}$$

It is worth to note that in practice when network size n is much larger than k the normalized variance of efficent MLE is mainly determined by the values N and k, rather than n. For example, having $k = 8$ and $N = 10$, the standard deviation σ_{min} increases only from 10% to 11%, as network size grows from 50 to 10^6 nodes. Without the dependence on $1/n$ in (6), the the standard deviation of MLE estimator can be calculated as

$$\lim_{n\to\infty} \sigma_{min} = \max_n \sigma_{min} = \frac{1}{\sqrt{Nk}} \tag{7}$$

7. Results of computer simulation

The simulation [8] was performed to compare precision of the MLE estimator (3) with the theoretical low bound (6). We process 10,000 estimations for each combination of parameters $k = 8, 20, N = 10, 20, 40$ and network sizes $n = 25, 10^2, 10^3, 10^4, 10^5$. Each estimation receives N statistics of k-th order from n independent random variables uniformly distributed on $(0, 1)$. The normalized errors then used to calculate the sample variance σ^2 which is the subject of comparison with the theoretical low bound σ^2_{min} (6).

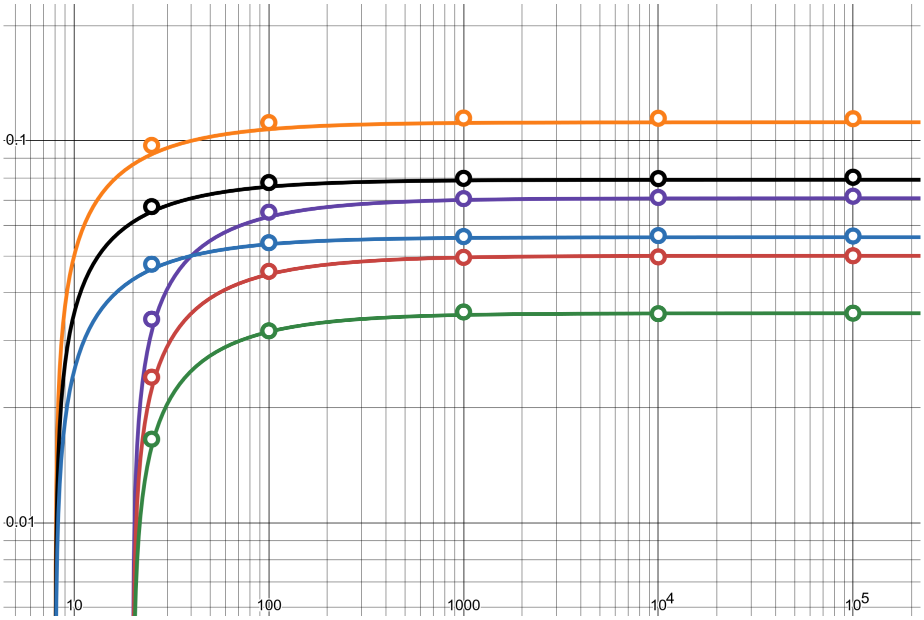

Figure 1: Standard deviation of MLE estimate in comparison with theretical low bound

The solid curves in Figure 1 show the theoretical low bounds for standard deviation σ_{min}. They are plotted according to (6) as function of network size. The simulation results, related to the sample deviation σ, are shown as points for fixed network sizes. The results prove the efficency of the MLE estimate. For the worse case $(k = 8, N = 10, n = 25)$ the MLE deviation (10%) differs in less than 1% from the low bound. At increasing k and N the estimate precision and low bound are practically indistinguishable. For combination $(k = 20, N = 40, n = 25)$ the MLE deviation (1.7%) differs in less than 0.01% from the low bound.

8. Conclusion

..todo

9. Appendix. Joint PDF of the first k order statistics from a sample of size n

..todo

10. References