Average Time Analysis: Searching a Signature Tree

Yangjun Chen and Yibin Chen

In [1], it is claimed that the average time of searching a signature tree is on the order of $O(n^{1-\frac{k}{m}})$, where n is the number of signatures in a signature file, m the siganture length, and k the number of bits set to 1 in a signature. In this paper, we show how this result is achieved. For this purpose, we evaluate $c_{s,n}$ given by (15) in [1] by using contour integration of complex variabled functions.

First, we define

$$\phi(x) = \sum_{h=0}^{m-1} \lambda_1\lambda_2\ldots\lambda_h \sum_{j \geq 0} 2^{j(m-k)} D_{jh}(x) \,, (x \geq 0) \qquad (1)$$

Then, we perform the following computations to evaluate $\phi(x)$:

(1) define the Mellin transformation of $\phi(x)$ ([2], p. 453):

$$\phi^*(\sigma) = \int_0^\infty \phi(x) x^{\sigma-1} dx \,. \qquad (2)$$

(2) derive an expression for $\phi^*(\sigma)$, which reveals some of its singularities.

(3) evaluate the reversal Mellin transformation

$$\phi(x) = \frac{1}{2i\pi}\int_{c-i\infty}^{c+i\infty} \phi^*(\sigma) x^{-\sigma} d\sigma \quad -1 < c < -\left(1 - \frac{k}{m}\right) \qquad (3)$$

The integral (3) is evaluated by using Cauchy's theorem as a sum of *residues* to the right of the vertical line $\{c + iy \mid y \in \Re\}$, where $\Re$ represents the set of all real numbers. This compuation method was first proposed in [3]. The following is just an extended explanation of it.

Remember that

$$D_{jh}(x) = 1 - (1 - 2^{-mj-h})^x - x2^{-mj-h}(1 - 2^{-mj-h})^{x-1}.$$

We rewrite it under the form

$$D_{jh}(x) = 1 - e^{-x\alpha_{jh}} - \beta_{jh} x e^{-x\alpha_{jh}} \qquad (4)$$

with $\alpha_{jh} = -\log(1 - 2^{-mj-h})$ and $\beta_{jh} = 2^{-mj-h}(1 - 2^{-mj-h})^{-1}$.

Now we consider the following expansion, which is valid for small values of x:

$$(-\log(1 - x))^{-\sigma} = x^{-\sigma}(1 - \frac{x\sigma}{2} + O(|\sigma|^2 x^2)). \qquad (5)$$

Let $x = 2^{-mj-h}$. Then, we have (by using the above expansion)

$$\alpha_{jh} = (-\log(1 - 2^{-mj-h}))^{-(-1)} \sim (2^{mj+h}). \qquad (6)$$

In addition, for small values 2^{-mj-h}, we also have

$$\beta_{jh} = 2^{-mj-h}(1 - 2^{-mj-h})^{-1} = O(2^{-mj}). \qquad (7)$$

Following the classical properties of Mellin transformation, we have the following proposition.

Proposition 1. Denote $D_{jh}^*(\sigma)$ the Mellin transformation of $D_{jh}(x)$. We have

$$D_{jh}^*(\sigma) = \int_0^\infty D_{jh}(x) x^{\sigma-1} dx$$

$$= -(\alpha_{jh})^{-\sigma}\Gamma(\sigma) - \beta_{jh}(\alpha_{jh})^{-\sigma-1}\sigma\Gamma(\sigma) \qquad (8)$$

provided $-1 < Re(\sigma) < 0$, where $\Gamma(\sigma)$ is the *Euler Gamma* function.

Proof. The following formulas are well-known:

$$\int_0^\infty (e^{-x} - 1) x^{\sigma-1} dx = \Gamma(\sigma)- \qquad 1 < Re(\sigma) < 0 \qquad (9)$$

$$\int_0^\infty (xe^{-x}) x^{\sigma-1} dx = \sigma\Gamma(\sigma) \qquad -1 < Re(\sigma) \qquad (10)$$

$$\int_0^\infty f(ax) x^{\sigma-1} dx = a^{-\sigma}\int_0^\infty f(x) x^{\sigma-1} dx \qquad \text{for } a > 0 \qquad (11)$$

In terms of these formulas, we have

$$D_{jh}^*(\sigma) = \int_0^\infty D_{jh}(x) x^{\sigma-1} dx \qquad (12)$$

$$= \int_0^\infty (1 - e^{-x\alpha_{jh}}) x^{\sigma-1} dx - \int_0^\infty \beta_{jh} x e^{-x\alpha_{jh}} x^{\sigma-1} dx$$

$$= -(\alpha_{jh})^{-\sigma}\Gamma(\sigma) - \beta_{jh}(\alpha_{jh})^{-\sigma-1}\sigma\Gamma(\sigma). \qquad \square$$

Now we try to evaluate the following two sums:

$$\omega_h(\sigma) = \sum_{j \geq 0} 2^{j(m-k)}(\alpha_{jh})^{-\sigma} \,, \qquad (13)$$

The author are with the Department of Applied Computer Science, the University of Winnipeg, winnipeg, Manitoba, Canada R3B 2E9.
E-mail: ychen2@uwinnipeg.ca, yibinchen@hotmail.com

$$\upsilon_h(\sigma) = \sum_{j \geq 0} 2^{j(m-k)} \beta_{jh} (\alpha_{jh})^{-\sigma - 1} .$$

From (6) and (7), we can see that the two sums given by (13) are uniformly and absolutely convergent when σ is in the following stripe:

$$\textit{Stripe}: -1 < \mathrm{Re}(\sigma) < -(1 - \frac{k}{m}). \tag{14}$$

Furthermore, in terms of (6) and (7), both $\omega_h(\sigma)$ and $\upsilon_h(\sigma)$ can be approximated by the following sum:

$$\hat{\omega}_h(\sigma) = \sum_{j \geq 0} 2^{j(m-k)} (2^{mj+h})^{\sigma} \tag{15}$$

When $\mathrm{Re}(\sigma) < \sigma_0 = -(1 - \frac{k}{m})$, this series can be summed exactly:

$$\hat{\omega}_h(\sigma) = 2^{h\sigma} \frac{1}{1 - 2^{m-b+m\sigma}} . \tag{16}$$

Thus, $\phi^*(\sigma)$ is defined in *Stripe* and can be computed as follows

$$\phi^*(\sigma) = \int_0^{\infty} \phi(x) x^{\sigma - 1} dx \tag{17}$$

$$= \int_0^{\infty} \left(\sum_{h=0}^{m-1} \lambda_1 \lambda_2 \ldots \lambda_h \sum_{j \geq 0} 2^{j(m-k)} D_{jh}(x) \right) x^{\sigma - 1} dx$$

$$= - \sum_{h=0}^{m-1} \lambda_1 \lambda_2 \ldots \lambda_h (\omega_h(\sigma) + \sigma \upsilon_h(\sigma)) \Gamma(\sigma)$$

$$= -\Gamma(\sigma)(1 + \sigma) \sum_{h=0}^{m-1} \lambda_1 \lambda_2 \ldots \lambda_h 2^{h\sigma} \frac{1}{1 - 2^{m-b+m\sigma}} .$$

From this, we can observe all the sigularities (poles), i.e., $\sigma = 0$, at which $\Gamma(\sigma)$ is not defined; and all those values of σ, at which $(1 - 2^{m(\sigma - \sigma_0)})$ becomes 0:

$$\sigma_j = \sigma_0 + \frac{2ij\pi}{m \log 2}, \qquad (j = 0, \pm 1, \pm 2, \ldots) \tag{18}$$

To compute the integral in (21), we consider the following integral

$$\phi_N(x) = \frac{1}{2i\pi} \int_{L_N} \phi^*(\sigma) x^{-\sigma} d\sigma, \tag{19}$$

where L_N is a rectangular contour oriented clockwise as shown in Fig. 1.

$$L_N = {}^1_N + L_N^2 + L_N^3 + L_N^4, \tag{20}$$

$$L_N^1 = \left\{ c + iu \middle| |u| \leq \frac{(2N+1)\pi}{m \log 2} \right\},$$

$$L_N^2 = \left\{ v + i\frac{(2N+1)\pi}{m \log 2} \middle| c \leq v \leq \frac{b}{3m} \right\},$$

$$L_N^3 = \left\{ \frac{b}{3m} + iu \middle| |u| \leq \frac{(2N+1)\pi}{m \log 2} \right\},$$

$$L_N^4 = \left\{ v - i\frac{(2N+1)\pi}{m \log 2} \middle| c \leq v \leq \frac{b}{3m} \right\},$$

where N is an integer. This contour is of a similar type used in ([4], p. 132).

Let ϕ_N^i be the integral along L_N^i ($i = 1, 2, 3, 4$). Then,

$$\phi_N(x) = \phi_N^1(x) + \phi_N^2(x) + \phi_N^3(x) + \phi_N^4(x) .$$

Furthermore, we have the following results:

$$\lim_{N \to \infty} \phi_N^1(x) = \phi(x) ,$$

$$\lim_{N \to \infty} \phi_N^2(x) = O(1) ,$$

$$\left| \phi_N^3(x) \right| \leq x^{-k/(3m)} \int_{L_\infty} \left| \phi^*(\sigma) \right| d\sigma = O(x^{-k/(3m)}), \text{ and}$$

$$\lim_{N \to \infty} \phi_N^4(x) = O(1).$$

Thus, we have

$$\lim_{N \to \infty} \phi_N(x) = \phi(x) + O(x^{-k/(3m)}) \tag{21}$$

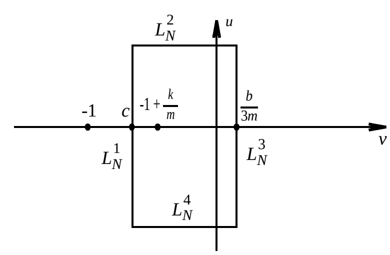

Fig. 1 The rectangular contour L_N

On the other hand, $\lim_{N \to \infty} \phi_N(x)$ can be evaluated as the sum of the residues of the integrand, i.e., $\phi^*(\sigma)x^{-\sigma}$, inside L_N. Concretely, we have

$$\lim_{N \to \infty} \phi_N(x) = - \sum_{\alpha \in \mathrm{Pole}(\phi^*(\sigma))} (\phi^*(\sigma) x^{-\sigma}, \sigma = \alpha)$$

$$= - \sum_{\alpha \in \mathrm{Pole}(\phi^*(\sigma))} \lim_{\sigma \to \alpha} (\sigma - \alpha) \phi^*(\sigma) x^{-\sigma} . \tag{22}$$

Within L_∞, $\phi^*(\sigma)$ has the following poles:

$\alpha = 0$, and

$$\alpha = \sigma_j = \sigma_0 + \frac{2ij\pi}{m \log 2} \qquad (j = 0, \pm 1, \pm 2, \ldots)$$

The contribution of the pole $\alpha = 0$ is O(1); and the contribution of $\alpha = \sigma_0$ is

$$\lim_{\sigma \to \sigma_0} (\sigma - \sigma_0) \phi^*(\sigma) x^{-\sigma}$$

$$= x^{-\sigma_0}\frac{(1+\sigma_0)\Gamma(\sigma_0)}{m\log 2}\sum_{h=0}^{m-1}\lambda_1\lambda_2\ldots\lambda_h 2^{h\sigma_0}. \quad (23)$$

Finally, the contribution of each σ_j ($j = \pm 1, \pm 2, \ldots$) is

$$\lim_{\sigma\to\sigma_j}(\sigma-\sigma_j)\phi^*(\sigma)x^{-\sigma} \quad (24)$$

$$= x^{-\sigma_0}\exp\left(-\frac{2ij\pi}{m}\log_2 x\right)(1+\sigma_j)\Gamma(\sigma_j)\sum_{h=0}^{m-1}\lambda_1\lambda_2\ldots\lambda_h 2^{h\sigma_j}$$

So we have

$$\lim_{N\to\infty}\phi_N(x) = x^{-\sigma_0}\frac{(1+\sigma_0)\Gamma(\sigma_0)}{m\log 2}\sum_{h=0}^{m-1}\lambda_1\lambda_2\ldots\lambda_h 2^{h\sigma_0} +$$

$$\sum_{j=-\infty}^{-1} x^{-\sigma_0}\exp\left(-\frac{2ij\pi}{m}\log_2 x\right)(1+\sigma_j)\Gamma(\sigma_j)\sum_{h=0}^{m-1}\lambda_1\lambda_2\ldots\lambda_h 2^{h\sigma_j}$$

$$+$$

$$\sum_{j=1}^{+\infty} x^{-\sigma_0}\exp\left(-\frac{2ij\pi}{m}\log_2 x\right)(1+\sigma_j)\Gamma(\sigma_j)\sum_{h=0}^{m-1}\lambda_1\lambda_2\ldots\lambda_h 2^{h\sigma_j}$$

$$= x^{-\sigma_0}\frac{(1+\sigma_0)\Gamma(\sigma_0)}{m\log 2}\sum_{h=0}^{m-1}\lambda_1\lambda_2\ldots\lambda_h 2^{h\sigma_0}. \quad (25)$$

From this, we know that

$$C_{s,n} = O(n^{-\sigma_0}) = O(n^{1-\frac{k}{m}}). \quad (26)$$

REFERENCES

[1] Y. Chen and Y. Chen, On the Signature Tree Construction and Analysis, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, NO. 9, September 2006.

[2] R.V. Churchill, *Operational Mathematics*, McGraw-Hill Book Company, NewYork, 1958.

[3] P. Flajolet and C. Puech, Partial match Retrieval of Multidimentional Data, *J. ACM*, Vol. 33, No. 2, April 1986, pp. 371-407.

[4] D.E. Knuth, *The Art of Computer Programming: Sorting and Searching*, Addison-Wesley Pub. London, 1973.