Discrete Distributions in the Tardos Scheme, Revisited

Thijs Laarhoven
Eindhoven University of Technology
P.O. Box 513, 5600 MB
Eindhoven, The Netherlands
t.m.m.laarhoven@tue.nl

Benne de Weger
Eindhoven University of Technology
P.O. Box 513, 5600 MB
Eindhoven, The Netherlands
b.m.m.d.weger@tue.nl

ABSTRACT

The Tardos scheme is a well-known traitor tracing scheme to protect copyrighted content against collusion attacks. The original scheme contained some suboptimal design choices, such as the score function and the distribution function used for generating the biases. Škorić et al. previously showed that a symbol-symmetric score function leads to shorter codes, while Nuida et al. obtained the optimal distribution functions for arbitrary coalition sizes. Later, Nuida et al. showed that combining these results leads to even shorter codes when the coalition size is small. We extend their analysis to the case of large coalitions and prove that these optimal distributions converge to the arcsine distribution, thus showing that the arcsine distribution is asymptotically optimal in the symmetric Tardos scheme. We also present a new, practical alternative to the optimal discrete distributions of Nuida et al. and give a comparison of the estimated codelengths for each of these distributions.

Categories and Subject Descriptors

E.4 [**Data**]: Coding and Information Theory;
G.1.4 [**Mathematics of Computation**]: Numerical Analysis—*Quadrature and Numerical Differentiation*

General Terms

Theory, Security, Design

1. INTRODUCTION

To fight against copyright infringement, distributors of copyrighted content embed hidden watermarks in the data, creating a different version of the content for each user. Then, when a user distributes his copy and the distributor finds it, the distributor extracts the watermark from this copy and traces it to the guilty user. Assuming two versions can be created for every segment of the content, it is clear that with a binary search, $\ell \approx \log_2 n$ watermarked content segments suffice to find one pirate hidden among n users.

Things become more complicated when several users *collude*, and compare their differently watermarked copies to create a new version of the content that does not exactly match any of their copies. Assuming that for each segment of the data there are two different versions, and that in each segment the colluders output one of their received versions (known in the literature as the *marking assumption*), it is impossible to trace $c \geq 2$ colluders deterministically (i.e., with no probability of error) with any fixed amount of segments. Fortunately, probabilistic schemes do exist that allow us to trace up to c colluders with at most ε probability of error, for any given $c \geq 2$ and $\varepsilon > 0$. One of the main objectives of research in this area is to construct such traitor tracing schemes, that allow us to trace colluders with as few segments ℓ as possible.

1.1 Related work

In 2003, Tardos [12] showed that the optimal order codelength of such codes (i.e., the number of segments needed) is $\ell = d_\ell c^2 \ln(n/\varepsilon_1)$ with $d_\ell = \Omega(1)$, where ε_1 is an upper bound on the probability of catching one or more innocent users.[1] In the same paper, Tardos gave a construction of a scheme with $d_\ell = 100$, known as the Tardos scheme. This shows that $d_\ell = \Theta(1)$ is optimal, and that the Tardos scheme has the optimal order codelength.

Over the last ten years, improvements to the Tardos scheme have lead to a significant decrease in the codelength parameters d_ℓ. We previously showed [5] that combining the symbol-symmetric score function of Škorić et al. [10] with the improved analysis of Blayer and Tassa [2] leads to codelength parameters of the order $d_\ell \approx 4.93$ for large c. For small coalitions, Nuida et al. [8] showed that even smaller values d_ℓ can be obtained by combining the symmetric score function with the optimized, discrete distribution functions previously obtained by Nuida et al. [9]. For large c, this lead to codelength parameters of about $d_\ell \approx 5.35$.

1.2 Contributions and outline

In this paper, we show that for large coalition sizes, the optimal discrete distributions of Nuida et al. [8, 9] converge to the arcsine distribution, thus proving that in the symmetric Tardos scheme, the arcsine distribution is asymptotically optimal. Together with results of Škorić et al. [10] and

[1] Note that ε_2, commonly used for an upper bound on the probability of not catching any pirates, does not appear in the leading term of the codelength for most practical values of ε_1 and ε_2.

us [5], this further implies that the asymptotic codelength $\ell = \frac{\pi^2}{2}c^2 \ln(n/\varepsilon_1)$ is optimal. On the practical side, we present an alternative to the optimal distributions of Nuida et al. with a simpler bias generation method, and conjecture that its performance is close to the performance of the optimal distributions of Nuida et al.

The outline of this paper is as follows. In Section 2 we describe the symmetric Tardos scheme, and different choices for the distribution function F used in this scheme. In Section 3 we state our results, and we devote Section 4 to proving the main result. In Section 5 we present what we call discrete arcsine distributions, and in Section 6 we give a heuristic comparison of the codelengths of the symmetric Tardos scheme when using these various distribution functions. Finally, in Section 7 we briefly discuss the results and mention a direction for future research.

2. THE SYMMETRIC TARDOS SCHEME

Before we describe the Tardos scheme, we introduce some more notation. The matrix $X = (X_{j,i})$, consisting of bits, is used to indicate which of the two versions of the ith content segment is assigned to user j, for each user $j \in \{1, \ldots, n\}$ and each segment $i \in \{1, \ldots, \ell\}$. We write $\vec{y} = (y_i)$ for the pirate output, consisting of ℓ bits.

The Tardos scheme roughly consists of two parts, which are outlined below. The scheme depends on appropriately chosen functions F and g, and constants ℓ and Z. The first part of the scheme is performed before the content is distributed, and focuses on generating the code matrix X. The second part is performed once the pirates have output a forged copy $\vec{y}$ and this copy has been detected by the distributor, and focuses on finding the guilty users.

(1) **Codeword generation**
- For each i, generate $p_i \sim F$.
- For each i, j, generate $X_{j,i} \sim \text{Bernoulli}(p_i)$.

(2) **Accusation algorithm**
- For each i, j, compute $S_{j,i} = g(X_{j,i}, y_i, p_i)$.
- For each j, accuse user j if $\sum_{i=1}^{\ell} S_{j,i} > Z$.

This description is very general, and covers (almost) any known version of the Tardos scheme. The choice of F and g, and the method to determine ℓ and Z, are what separates one scheme from another. In this paper we will focus on the class of *symmetric* Tardos schemes, which means choosing g as the symbol-symmetric score function of Škorić et al. [10]:

$$g(X_{j,i}, y_i, p_i) = \begin{cases} +\sqrt{(1-p_i)/p_i}, & \text{if } X_{ji} = 1, y_i = 1, \\ -\sqrt{(1-p_i)/p_i}, & \text{if } X_{ji} = 1, y_i = 0, \\ -\sqrt{p_i/(1-p_i)}, & \text{if } X_{ji} = 0, y_i = 1, \\ +\sqrt{p_i/(1-p_i)}, & \text{if } X_{ji} = 0, y_i = 0. \end{cases}$$

In this paper we will not go into detail about choosing ℓ and Z, but focus on the distribution function F.

2.1 Continuous arcsine distributions

A common choice for the distribution function F is the arcsine distribution with appropriate cutoffs. More precisely, we first compute a cutoff parameter $\delta_c > 0$, and we then use the distribution function F_c defined on $[\delta_c, 1-\delta_c]$ by:

$$F_c(p) = \frac{2\arcsin\sqrt{p} - 2\arcsin\sqrt{\delta_c}}{\pi - 4\arcsin\sqrt{\delta_c}}. \qquad (\delta_c \le p \le 1-\delta_c)$$

For small c, the parameter δ_c has to be sufficiently large for a certain proof of security to work. For large c, the cutoff δ_c tends to 0, and the distributions converge to the well-known arcsine distribution F_∞, defined on $[0, 1]$ by:

$$F_\infty(p) = \frac{2}{\pi}\arcsin\sqrt{p}. \qquad (0 \le p \le 1)$$

With these continuous arcsine distribution functions, we previously showed [5] that an asymptotic codelength of $\ell = \frac{\pi^2}{2}c^2 \ln(n/\varepsilon_1)(1 + O(c^{-1/3}))$ is optimal. For details, see [5].

2.2 Discrete Gauss-Legendre distributions

Nuida et al. [8, 9] showed that if the pirates aim to minimize their expected total score, the optimal distributions are in fact discrete distributions, and are related to Gauss-Legendre quadratures in numerical analysis. To define these distributions, we first need to introduce *Legendre polynomials*. For $c \ge 1$, the cth Legendre polynomial is given by

$$P_c(x) = \frac{1}{2^c c!}\left(\frac{d}{dx}\right)^c (x^2-1)^c.$$

This polynomial has c simple roots on $(-1, 1)$, which we will denote by $x_{1,c} < x_{2,c} < \ldots < x_{c,c}$. Now, the optimal distribution functions, for arbitrary c, are as follows. For details, see [8, 9].

LEMMA 1. *[9, Theorem 3] The optimal distribution to fight against $2c-1$ or $2c$ colluders, is*

$$F_{2c-1}(p) = F_{2c}(p) = \frac{1}{N_c}\sum_{k=1}^{c} w_{k,c} H(p - p_{k,c}), \quad (0 \le p \le 1)$$

where N_c is a normalizing constant, H is the Heaviside step function, and the points $p_{k,c}$ and weights $w_{k,c}$ are given by

$$p_{k,c} = \frac{x_{k,c}+1}{2}, \quad w_{k,c} = \frac{2}{(1-x_{k,c}^2)^{3/2} P_c'(x_{k,c})^2}.$$

For small c, this construction gives much shorter codelengths than those obtained using the arcsine distributions with cutoffs. For large c, the codelength parameter goes up, and Nuida et al. [8] showed that their results can be extended to a construction that asymptotically achieves $d_\ell \to K \approx 5.35$. Since this is higher than the asymptotic codelength parameter obtained with the arcsine distribution function, this asymptotic result is not optimal.

3. MAIN RESULTS

We will prove that letting c tend to infinity in the discrete distributions of Nuida et al. leads exactly to the arcsine distribution. This will be done by proving the following result.

THEOREM 1. *Let the parameters $p_{k,c}$, $w_{k,c}$, and N_c as in Lemma 1. Let $\alpha > 0$, and let k satisfy $\alpha c < k < (1-\alpha)c$. Then, as $c \to \infty$,*

$$p_{k,c} = \sin^2\left(\frac{\pi k}{2c}\right) + o(1), \qquad (1)$$

$$w_{k,c} = \frac{\pi}{c} + o\left(\frac{1}{c}\right), \qquad (2)$$

$$N_c = \pi - o(1). \qquad (3)$$

Note that except for the points near 0 and 1, corresponding to $k = o(c)$ or $k = c - o(c)$, the leading terms of the weights are all equal. But since these points in the 'middle' carry $1 - o(1)$ weight (cf. the proof of (3)), the points near 0 and 1 have a negligible total weight. On the other hand, the points $p_{k,c}$ converge to the expected values of the corresponding order statistics of the arcsine distribution, i.e., the value y corresponding to $F_\infty(y) = \frac{k}{c}$ is exactly $y = F_\infty^{-1}(\frac{k}{c}) = \sin^2(\frac{\pi k}{2c}) = p_{k,c} + o(1)$. Since asymptotically all these points have the same weight, after k of the c points we also have $F_{2c}(p_k^{(c)}) = \frac{k}{c} + o(\frac{1}{c})$ or $F_{2c}^{-1}(\frac{k}{c}) = p_{k,c} + o(1)$. Since the set of points $\{p_{k,c}\}_{k=1}^c$ is dense in $(0,1)$ when c tends to infinity, these results imply that $F_{2c}(p) \to F_\infty(p)$ for each $p \in (0,1)$, proving that the arcsine distribution is asymptotically optimal in the symmetric Tardos scheme.

THEOREM 2. *In the symmetric Tardos scheme, the arcsine distribution is asymptotically optimal.*

It was shown by Škorić et al. [10, Section 6] that when using the arcsine distribution, due to the Central Limit Theorem the optimal codelength inevitably converges to $\ell \to \frac{\pi^2}{2} c^2 \ln(n/\varepsilon_1)$. So the following corollary is immediate.

COROLLARY 1. *In the symmetric Tardos scheme, the following codelength is asymptotically optimal:*

$$\ell = \left(\frac{\pi^2}{2} + o(1)\right) c^2 \ln(n/\varepsilon_1).$$

In addition to these theoretical results, we present a new class of distribution functions, which can be obtained by disregarding some of the order terms in Theorem 1. Compared to the optimal Gauss-Legendre distributions, these distributions are much simpler, but seem to achieve comparable codelengths. For details, see Sections 5 and 6.

4. PROOF OF THEOREM 1

(1): Let $\theta_{k,c} = \arccos(x_{k,c})$. From [1, Eq. (22.16.6)] we have

$$\theta_{k,c} = \left(\frac{4(c-k)+3}{4c+2}\right)\pi + o(1) = \pi - \frac{\pi k}{c} + o(1). \qquad (4)$$

Using $\cos(\pi - \phi) = 2\sin^2(\frac{\phi}{2}) - 1$ for $\phi \in \mathbb{R}$, we get

$$x_{k,c} = \cos\left(\pi - \frac{\pi k}{c} + o(1)\right) = 2\sin^2\left(\frac{\pi k}{2c}\right) - 1 + o(1).$$

Since $p_{k,c} = \frac{1}{2}(x_{k,c} + 1)$, Equation (1) follows.
(2): Combining [11, Eq. (15.3.1) and Eq. (15.3.10)], and using $2\sin(\frac{\theta_{k,c}}{2})\cos(\frac{\theta_{k,c}}{2}) = \sin(\theta_{k,c})$, we get

$$\frac{2}{(1-x_{k,c}^2)P_c'(x_{k,c})^2} = \frac{\pi}{c}\sin(\theta_{k,c}) + o\left(\frac{1}{c}\right).$$

Dividing both sides by $\sqrt{1-x_{k,c}^2} = \sin\theta_{k,c}$ leads to (2).
(3): The Gauss-Legendre quadrature rule states that, for analytic f, there exist $A_c > 0$, $\xi \in (-1,1)$, with [1, (25.4.29)]

$$\int_{-1}^{1} f(x)dx = \sum_{k=1}^{c} \frac{2f(x_{k,c})}{(1-x_{k,c}^2)P_c'(x_{k,c})^2} + A_c f^{(2c)}(\xi).$$

Let $f(x) = (1-x^2)^{-1/2}$. Then $f^{(2c)} > 0$ on $(-1,1)$, so

$$\pi = \int_{-1}^{1} \frac{dx}{\sqrt{1-x^2}} > \sum_{k=1}^{c} w_{k,c} = N_c.$$

On the other hand, from (2) and $w_{k,c} > 0$ for all k, we have

$$N_c > \sum_{k=o(c)}^{c-o(c)} w_{k,c} = (c - o(c))\left(\frac{\pi}{c} + o\left(\frac{1}{c}\right)\right) = \pi - o(1).$$

So $\pi - o(1) < N_c < \pi$, which proves (3).

5. DISCRETE ARCSINE DISTRIBUTIONS

By making a slight refinement to (1) using (4), we get that for large c and almost all values of k, the parameters of the optimal distributions satisfy

$$p_{k,c} \approx \sin^2\left(\frac{4k-1}{8c+4}\pi\right), \quad w_{k,c} \approx \frac{\pi}{c}, \quad N_c \approx \pi.$$

To get the exact values of these parameters for large c requires quite some effort, so in practice one may consider using an approximation of these distributions. An obvious approximation to the above weights and points would be

$$p'_{k,c} = \sin^2\left(\frac{4k-1}{8c+4}\pi\right), \quad w'_{k,c} = \frac{\pi}{c}, \quad N'_c = \pi.$$

Generating biases p from the associated distribution function is equivalent to drawing r uniformly at random from $\{\frac{3\pi}{8c+4}, \frac{7\pi}{8c+4}, \ldots, \frac{\pi}{2} - \frac{3\pi}{8c+4}\}$, and setting $p = \sin^2(r)$. Note that if we were to draw r uniformly at random from the complete interval $[0, \frac{\pi}{2}]$, this would correspond to the arcsine distribution, while drawing r uniformly at random from $[\arcsin(\sqrt{\delta}), \frac{\pi}{2} - \arcsin(\sqrt{\delta})]$ corresponds to the arcsine distribution with cutoff δ. So these distributions may be appropriately called *discrete arcsine distributions*, and needless to say, for large c these distributions also converge to the arcsine distribution.

Remark. Interestingly, slightly different parameters,

$$p''_{k,c} = \sin^2\left(\frac{4k-2}{8c}\pi\right), \quad w''_{k,c} = \frac{\pi}{c}, \quad N''_c = \pi,$$

correspond exactly to the parameters of the so-called *Chebyshev-Gauss quadratures* [1, Eq. (25.4.38)]. These quadratures allow one to approximate integrals of the form

$$\int_{-1}^{1} \frac{g(x)}{\sqrt{1-x^2}}dx \approx \frac{1}{N''_c}\sum_{k=1}^{c} w''_{k,c} g(x''_{k,c}),$$

where $x''_{k,c} = 2p''_{k,c} - 1$. The distribution functions generated by these weights and points are very similar to the discrete arcsine distributions described above. The main difference seems to be the "cutoff", which would be about a third smaller (i.e., $\frac{2\pi}{8c}$ compared to $\frac{3\pi}{8c+4}$). Since these distributions are worse approximations of the optimal Gauss-Legendre distributions, it seems that the discrete arcsine distributions are a better alternative.

6. ESTIMATING CODELENGTHS

Let us now try to give a qualitative comparison of the several classes of discrete and continuous distribution functions, in terms of codelengths. Since the (tails of) distributions of user scores are hard to estimate, and known proof methods are not tight, we will only give a heuristic estimate of the codelengths. Getting more accurate estimates remains an open problem.

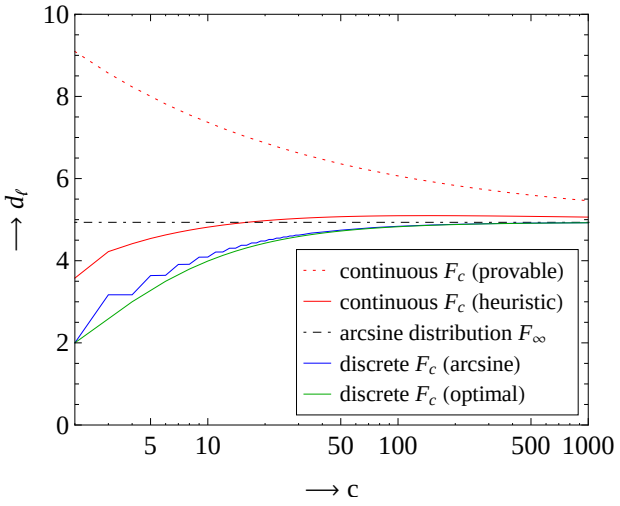

Figure 1: Estimates of the codelength parameters d_ℓ for several types of distribution functions F. The dashed line shows the asymptotic optimal value $d_\ell = \frac{\pi^2}{2}$, corresponding to the arcsine distribution F_∞.

Assuming that the scores of users are Gaussian, we can get a reasonable estimate for the optimal codelength parameter as $d_\ell \approx 2/\tilde{\mu}^2$, where $\tilde{\mu}$ is the expected average pirate score per content segment [10, Corollary 2]. In the case of the discrete distributions of Nuida et al., $\tilde{\mu}$ does not depend on the pirate strategy, so we can compute $\tilde{\mu}$ exactly. For the arcsine distributions with cutoffs and the discrete arcsine distributions, $\tilde{\mu}$ does depend on the pirate strategy, but by considering the attack that minimizes $\tilde{\mu}$ we can obtain lower bounds on $\tilde{\mu}$.

Figure 1 shows the resulting estimates of d_ℓ, as well as the provable upper bounds on d_ℓ of [5] (for constant ε_2). Note that the heuristic estimates for the continuous distributions are based on the arcsine distributions with cutoffs optimized for the proof technique of [5]. A different optimization of the cutoffs would lead to different values of d_ℓ.

7. CONCLUSION

We have shown that the optimal discrete distributions of Nuida et al. converge to the arcsine distribution, hence showing that the arcsine distribution is asymptotically optimal in the symmetric Tardos scheme. This connects the world of the discrete distributions to the world of the continuous distributions, as both converge to the same distribution.

In practice, the question remains which distribution function to choose. In the static Tardos scheme, choosing one of the discrete distributions seems logical, as this may drastically reduce the codelength. Recently, it was shown that the Tardos scheme can be extended to the *dynamic* traitor tracing setting, allowing efficient tracing of pirates when the colluders broadcast their forged copy in real-time [6, 7]. The construction of the universal Tardos scheme in [6] uses the fact that the continuous distributions are very similar for different values of c, so in this setting it seems that the continuous arcsine distributions are more practical.

An interesting open problem is what happens when the number of versions per content segment increases from 2 to q. Recent results show [3, 4] that with unlimited computing power, the optimal asymptotic codelength decreases linearly in q. This suggests that the optimal codelength in a q-ary Tardos scheme possibly decreases linearly in q as well. Škorić et al. [10] analyzed a natural generalization of the Tardos scheme to the q-ary setting, but did not obtain this linear decrease in q in their codelengths. The question remains whether their construction is suboptimal (and if so, whether this has to do with the choice of F or the choice of g), or if approaching the fingerprinting capacity for higher q with a single decoder traitor tracing scheme is simply impossible.

8. ACKNOWLEDGMENTS

The authors thank Jeroen Doumen, Wil Kortsmit, Jan-Jaap Oosterwijk, Georg Prokert, Berry Schoenmakers, and Boris Škorić for valuable discussions and comments.

9. REFERENCES

[1] M. Abramowitz and I.A. Stegun, editors. *Handbook of Mathematical Formulas*. Dover Publications, 1972.

[2] O. Blayer and T. Tassa. Improved versions of Tardos' fingerprinting scheme. *Designs, Codes and Cryptography*, 48(1):79–103, 2008.

[3] D. Boesten and B. Škorić. Asymptotic fingerprinting capacity for non-binary alphabets. In *Proc. 13th Conf. Information Hiding (IH)*, pages 1–13, 2011.

[4] Y.-W. Huang and P. Moulin. On fingerprinting capacity games for arbitrary alphabets and their asymptotics. In *Proc. International Symposium on Information Theory (ISIT)*, pages 2571–2575, 2012.

[5] T. Laarhoven and B. de Weger. Optimal symmetric Tardos traitor tracing schemes. *Designs, Codes and Cryptography*, 2012.

[6] T. Laarhoven, J. Doumen, P. Roelse, B. Škorić, and B. de Weger. Dynamic Tardos traitor tracing schemes. *IEEE Transactions on Information Theory*, 2013.

[7] T. Laarhoven, J.-J. Oosterwijk, and J. Doumen. Dynamic traitor tracing for arbitrary alphabets: Divide and conquer. In *Proc. 4th Workshop on Information Forensics and Security (WIFS)*, pages 240–245, 2012.

[8] K. Nuida, S. Fujitsu, M. Hagiwara, T. Kitagawa, H. Watanabe, K. Ogawa, and H. Imai. An improvement of discrete Tardos fingerprinting codes. *Designs, Codes and Cryptography*, 52(3):339–362, 2009.

[9] K. Nuida, M. Hagiwara, H. Watanabe, and H. Imai. Optimization of Tardos's fingerprinting codes in a viewpoint of memory amount. In *Proc. 9th Conf. Information Hiding (IH)*, pages 279–293, 2007.

[10] B. Škorić, S. Katzenbeisser, and M.U. Celik. Symmetric Tardos fingerprinting codes for arbitrary alphabet sizes. *Designs, Codes and Cryptography*, 46(2):137–166, 2008.

[11] G. Szegö. *Orthogonal Polynomials*. American Mathematical Society, 4th Edition, 1975.

[12] G. Tardos. Optimal probabilistic fingerprint codes. In *Proc. 35th Symposium on Theory of Computing (STOC)*, pages 116–125, 2003.