CDs Have Fingerprints Too*

Ghaith Hammouri[1], Aykutlu Dana[2], and Berk Sunar[1]

[1] CRIS Lab, Worcester Polytechnic Institute
100 Institute Road, Worcester, MA 01609-2280
{hammouri,sunar}@wpi.edu
[2] UNAM, Institute of Materials Science and Nanotechnology
Bilkent University, Ankara, Turkey
aykutlu@unam.bilkent.edu.tr

Abstract. We introduce a new technique for extracting unique fingerprints from identical CDs. The proposed technique takes advantage of manufacturing variability found in the length of the CD lands and pits. Although the variability measured is on the order of 20 nm, the technique does not require the use of microscopes or any advanced equipment. Instead, we show that the electrical signal produced by the photodetector inside the CD reader is sufficient to measure the desired variability. We investigate the new technique by analyzing data collected from 100 identical CDs and show how to extract a unique fingerprint for each CD. Furthermore, we introduce a technique for utilizing fuzzy extractors over the Lee metric without much change to the standard code offset construction. Finally, we identify specific parameters and a code construction to realize the proposed fuzzy extractor and convert the derived fingerprints into 128-bit cryptographic keys.

Keywords: Optical discs, fingerprinting, device identification, fuzzy extractor.

1 Introduction

According to the Business Software Alliance about 35% of the global software market, worth $141 Billion, is counterfeit. Most of the counterfeit software is distributed in the form of a compact disc (CD) or a digital video disc (DVD) which is easily copied and sold in street corners all around the world but mostly in developing countries. Given the severity of the problem at hand, a comprehensive solution taking into account the manufacturing process, economical implications, ease of enforcement, and the owner's rights, needs to be developed. While this is an enourmous undertaking requiring new schemes at all levels of implementation, in this work, we focus only on a small part of the problem, i.e. secure fingerprinting techniques for optical media.

To address this problem the SecuRom technology was introduced by Sony DADC. The technology links the identifiers produced to executable files which may only be accessed when the CD is placed in the reader. The main advantage of

* This material is based upon work supported by the National Science Foundation under Grant No. CNS-0831416.

C. Clavier and K. Gaj (Eds.): CHES 2009, LNCS 5747, pp. 348–362, 2009.

this technology is that it can be used with existing CD readers and writers. While the specifics of the scheme are not disclosed, in practice, the technology seems to be too fragile, i.e. slightly overused CDs become unidentifiable. Another problem is at the protocol level. The digital rights management (DRM) is enforced too harshly, therefore significantly curtailing the rights of the CD owner.

In this paper we take advantage of CD manufacturing variability in order to generate unique CD fingerprints. The approach of using manufacturing variability to fingerprint a device or to build cryptographic primitives has been applied in several contextes. A popular example is a new hardware primitives called *Physical Unclonable Functions* (PUFs). These primitives were proposed for tamper-detection at the physical level by exploiting *deep-submicron and nano-scale* physical phenomena to build low-cost tamper-evident key storage devices [7,8,6,12]. PUFs are based on the subtleties of the operating conditions as well as random variations that are imprinted into an integrated circuit during the manufacturing process. This phenomenon, i.e., manufacturing variability, creates minute differences in circuit parameters, e.g., capacitances, line delays, threshold voltages etc., in chips which otherwise were manufactured to be logically identical. Therefore, it becomes possible to use manufacturing variability to uniquely fingerprint circuits. More recently, another circuit fingerprinting technique was introduced. The technique exploits manufacturing variability in integrated chips to detect Trojan circuits inserted during the manufacturing process [5].

Another secure fingerprinting technology named RF-DNA was developed by Microsoft Research [1]. The RF-DNA technology provides unique and unclonable physical fingerprints based on the subtleties of the interaction of devices when subjected to an electromagnetic wave. The fingerprints are used to produce a cryptographic certificate of authenticity (COA) which when associated with a high value good may be used to verify the authenticity of the good and to distinguish it from counterfeit goods. Another application of manufacturing variability is fingerprinting paper objects. In [4] the authors propose Laser Surface Authentication which uses a high resolution laser microscope to capture the image texture from which the fingerprint is developed. In a more recent proposal, a cheap commodity scanner was used to identify paper documents [3]. While most of the results cited above were developed in the last decade, the idea of using physical fingerprints to obtain security primitives is not new at all. According to [1], access cards based on physical unclonable properties of media have been proposed decades ago by Bauder in a Sandia National Labs technical report [2].

Our Contribution: We introduce a method which exploits CD manufacturing variability to generate unique fingerprints from logically identical CDs. The biggest advantage of our approach is that it uses the electrical signal generated by the photodiode of a CD reader. Thus no expensive scanning or imaging equipment of the CD surface is needed. This means that regular CD readers can implement the proposed method with minimal change to their design. We investigate the new approach with a study of over 100 identical CDs. Furthermore, we introduce a new technique, called the threshold scheme, for utilizing fuzzy extractors over the Lee metric without much change to the standard code offset

construction [10]. The threshold scheme allows us to use error correcting codes working under the Hamming metric for samples which are close under the Lee metric. The threshold scheme is not restricted to CDs, and therefore can serve in any noisy fingerprinting application where the Lee metric is relevant. With the aid of the proposed fuzzy extractor we give specific parameters and a code construction to convert the derived fingerprints into 128-bit cryptographic keys.

The remainder of the paper is organized as follows. In Section 2, we discuss the physical aspects of CD storage, the sources of manufacturing variability and the statistical model capturing the CD variability. Section 3 presents experimental data to verify our statistical model. In Section 4 we discuss the fingerprint extraction technique and determine the parameters necessary for key generation. We discuss the robustness of the fingerprint in Section 5 and finally conclude in Section 6.

2 Pits and Lands

On a typical CD data is stored as a series of lands and pits formed on the surface of the CD. The pits are bumps separated by the lands to form a spiral track on the surface of the CD. The spiral track starts from the center of the CD and spirals outward. It has a width of about 0.5 μm and a 1.6 μm separation. The length of the land or pit determines the stored data. The encoding length can assume only one of nine lengths with minimum value in the range 833 to 972 nm up to a maximum of 3054 to 3563 nm with increments ranging from 278 to 324 nm. Note that the range is dependent on the speed used while writing the CD. To read the data on the CD the reader shines a laser on the surface of the CD and collects the reflected beam. When the laser hits the pits it will reflect in a diffused fashion thus appearing relatively dark compared to the lands. Upon the collection of the reflected beam, the reader can deduce the location and length of the lands and pits which results in reading the data on the CD.

CDs are written in two ways, pressing and burning. In pressed CDs a master template is formed with lands and pits corresponding to the data. The master template is then pressed into blank CDs in order to form a large number of copies. In burned CDs, the writing laser heats the dye layer on the CD-R to a

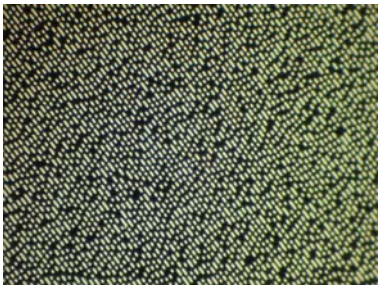

Fig. 1. Lands and pits image using an optical microscope

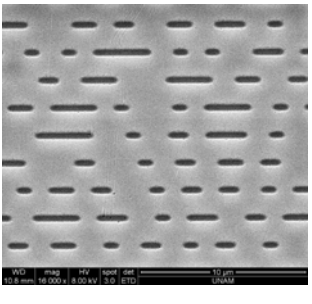

Fig. 2. Lands and pits image using a scanning electron microscope

point where it turns dark, thus reflecting the reading laser in a manner consistent with physical lands. Note that burned CDs will not have physical lands and pits but will act as if they had these features. Figures 1 and 2 show the lands and pits of a pressed CD. We captured Figure 1 using an optical microscope and Figure 2 using a scanning electron microscope.

2.1 Source of Variation

Similar to any physical process, during the writing process CDs will undergo manufacturing variation which will directly affect the length of the lands and pits. For burned CDs this variability will be a direct result of the CD velocity while writing takes place. This velocity is assumed to be at a fixed rate between 1.2 and 1.4 m/s where the velocity variation during writing should be within $\pm 0.01 m/s$ [11]. Pressed CDs are manufactured by molding thermoplastics from a micro or nanostructured master prepared by lithographic methods. The molding process itself is optimized for replication fidelity and speed with typical replication variations on the order of tens of nanometers [17]. The molding process involves contacting the thermoplastic with the master slightly above the glass transition temperature of the material, with a preset pressure for a brief amount of time, cooling the master and the thermoplastic to below the glass transition temperature and demoulding. Local variations of polymer material's mechanical and thermal properties, local variations of the temperature and pressure all potentially lead to variations in the imprinted structures. The thermal stresses induced during cooling and demoulding also potentially lead to variations. In this paper we aim at using the small variation in the length of lands and pits in order to form a unique fingerprint for each CD. In the next section we characterize the length features of lands and pits.

2.2 Single Location Characterization

Together lands and pits form the full spiral track. Therefore, it makes sense to fingerprint only lands or pits. The length of both lands and pits will follow similar distributions which is why we will simply use the term *location* to refer to either of them. We label the lengths of n consecutive locations by starting from a reference point on the track, as $L_1, L_2, \ldots, L_n$. In the ideal setting $L_i = c_i \cdot L$ for a small constant integer $c_i \in [3, 4, \ldots, 11]$ and $L \approx 300$ nm. However, due to the subtle variations we discussed in the previous section we expect $L_i = c_i \cdot L + \ell_i$. The variable ℓ_i is expected to be quite small compared to L_i, and therefore difficult to measure precisely. Still our measurements should be centered around the ideal length. Hence, quite naturally across all identical CDs we model L_i as a random variable drawn from a Gaussian distribution $\mathcal{H}_i = N(M_i, \Sigma)$ where $M_i = c_i \cdot L$ and Σ denotes the mean and the standard deviation respectively[1].

[1] $N(\mu, \sigma)$ is a normal distribution with mean μ and standard deviation σ.

Here we are assuming that regardless of the location, the standard deviation Σ will be the same. This is a quite a realistic assumption since Σ essentially captures the manufacturing variability which should affect all locations similarly. The more precise the manufacturing process is, the less of a standard deviation we would expect $\mathcal{H}_i$ to have. A perfect manufacturing process would yield $\Sigma = 0$ and would therefore give all CDs the same exact length of a specific location across all identical CDs. On the other hand, for better identification of CDs we would like $\mathcal{H}_i$ to have a relatively large Σ.

In a typical CD reader, the reading laser is reflected from the CD surface back into a photodiode which generates an electrical signal that depends on the intensity of the reflected laser. Therefore, the electrical signal is expected to depict the shape of the CD surface. If these electrical signals are used to measure the length of any given location, we expect these measurements to have a certain level of noise following a Gaussian distribution. So for location i on CD_j we denote this distribution by $\mathcal{D}_{ij} = N(\mu_{ij}, \sigma)$. The noise in the length measurements is captured through the standard deviation σ. Since this quantity mainly depends on the readers noise, we assume that its the same for all CDs and all CD locations. Contrary to Σ, to identify different CDs using the length information of CD locations we would like to see a relatively small σ.

3 Experimental Validation

To validate the statistical model outlined in the previous section, we conducted extensive experiments on a number of CDs. We directly probed into the electrical signal coming out of the photodiode constellation inside the CD reader. The intensity of this signal will reflect the CD surface geometry, and therefore can be used to study the length of the CD locations. To sample the waveform we used a 20 GHz oscilloscope. Each CD was read a number of times in order to get an idea of the actual $\mathcal{D}$ distribution. Similarly, we read from the same locations of about 100 identical CDs in order to generate the $\mathcal{H}$ distribution. Each collected trace required about 100 MBytes of storage space. Moreover, synchronizing the different traces to make sure that the data was captured from the same location of the CD was quite a challenge. We had to assign a master trace which represented the locations we were interested in studying and then ran the other traces through multiple correlation stages with the master to finally extract synchronized signals from the same locations on different CDs. Automating the process in order to accurately capture this massive amount of data was a time consuming challenge. However, we note that all this work would be almost trivially eliminated if we had access to the internal synchronization signals of the CD reader chip. The captured signals were then further processed using Matlab to extract the location lengths and obtain the distributions. After processing, we extracted the length of 500 locations (lands) on the CDs. We used commercially pressed CDs for all the experiments reported in this paper.[2]

[2] We have verified a similar behavior for burned CDs. Not surprisingly, data coming from burned CDs had a much larger variation and was easier to analyze.

Figure 3 shows the histogram of lengths extracted from 550 reads for a randomly chosen location on one CD. The mean length of the histogram is about $\mu_{ij} = 958$ nm. This histogram captures the $\mathcal{D}$ distribution. The other locations observe similar distributions with different mean lengths which will depend on the encoded information. When considering data coming from different locations and different CDs we obtain $\sigma = 20$ nm (with an average standard deviation of 2 nm on σ). This will be a good estimate for the noise observed during probing of the electrical signals. These results verify the assumption that the noise in the electrical signal can be approximated as Gaussian noise. Note that with Gaussian noise simple averaging can be used to substantially reduce the noise level. As we are interested in studying the behavior of the location lengths across

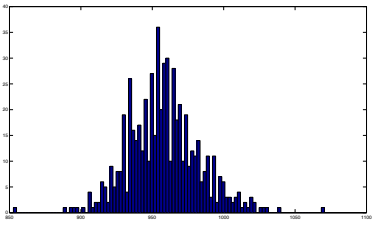

Fig. 3. Histogram of reads coming from the same location on the same CD

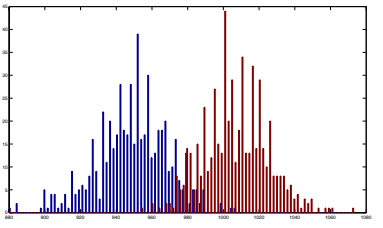

Fig. 4. Histograms of reads coming from the same location on two identical CDs

different CDs, we next shift our attention to two CDs before we look at a larger batch of CDs. Figure 4 captures a histogram for the length of the same location on two identical CDs. What is important here is the distance between the two Gaussians. The larger this distance becomes the easier it is to identify CDs. Our basic thesis for fingerprinting CDs is that the length of a single location will vary across multiple identical CDs. As pointed out earlier, this behavior can be modeled with the Gaussian distribution $\mathcal{H}_i$. The histogram in Figure 4 captures this for two CDs. To generalize these results and estimate the $\mathcal{H}_i$ distribution we need a larger sample space. The major problem here is that each data point needs to come from a different CD. Therefore, to obtain a histogram which clearly depicts a Gaussian we would need to test on the order of 500 CDs. This was not possible as each CD required substantial time, computing power and storage space in order to produce final data points. However, we were able to carry out this experiment for about 100 CDs. Each CD was read about 16 times to reduce the noise. Finally, we extracted the lengths of 500 locations for each of the CDs. Figure 5 depicts the histogram over 100 CDs for a randomly chosen location out of the 500 extracted locations. The histogram in Figure 5 has a mean of about 940 nm. Overall locations, Σ had a mean of 21 nm (with an average standard deviation of 1.8 nm on Σ). The histogram in Figure 5 looks similar to a Gaussian distribution generated from 100 data points. However, it would be interesting to get a confirmation that with more data points this histogram would actually yield a Gaussian. To do so, we normalized the lengths

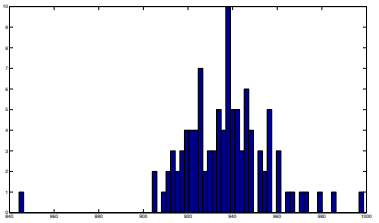

Fig. 5. Histograms of reads coming from the same location on 100 identical CDs

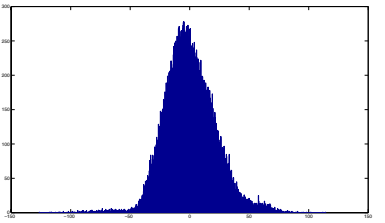

Fig. 6. Histograms of reads coming from 500 locations on 100 identical CDs

of each location by subtracting the average length for that particular location. Since the distribution for each location had roughly the same Σ the normalization process effectively made all these distributions identical with a mean of 0 and a standard deviation of Σ. We then collected all these data points (on the order of 50,000 points) and plotted the corresponding histogram. This is shown in Figure 6. The histogram of Figure 6 strongly supports our thesis of normally distributed location lengths across different CDs. One might observe a slight imbalance on the positive side of the Gaussian. This behavior seems to be a result of the DC offset observed while reading some of the CDs. Fortunately, this will not pose a problem for our fingerprinting technique as we will be normalizing each batch of data to have a mean of zero, thus removing any DC components. We finish this section by showing the histogram in Figure 7. The main purpose of this histogram is to confirm that what we are studying is in fact the length of data locations written on the CD. We elaborated earlier that on a CD data is stored in discrete lengths ranging from about 900 nm to about 3300 nm taking 9 steps in increments of about 300 nm. We build the histogram in Figure 7 using the data collected from 500 locations over the 100 CDs without normalizing each location's length to zero. In Figure 8 we show a similar histogram with data extracted by processing images coming from a scanning electron microscope.

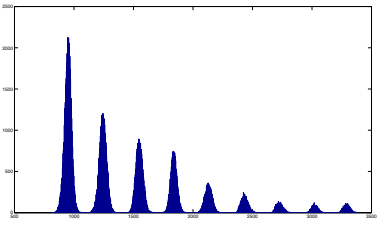

Fig. 7. Histogram of location lengths using the electrical signal

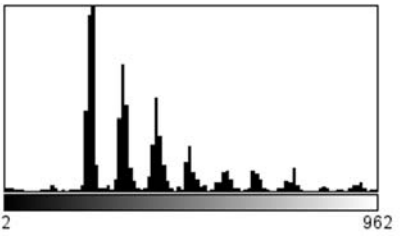

Fig. 8. Histogram of location areas using electron microscope images

4 CD Fingerprinting

There are many challenges in deriving a robust and secure fingerprint. One important issue is the reading noise. Similar to a human fingerprint, we saw in the previous section that the readings used to extract the CD fingerprint are inherently noisy. The extraction of a deterministic and secure fingerprint from noisy data has been previously studied in the literature [15,14,10]. Most relevant to our work is the fuzzy extractor technique proposed by Dodis et al. in [10]. For the remainder of this section we will present a quick review of the fuzzy extractor technique and then discuss how this technique can be modified and applied to the CD setting. Moreover, we will discuss the experimental results and present various bounds needed to achieve high levels of security.

4.1 Fuzzy Extractors

Loosely speaking a fuzzy extractor is a technique to extract an almost uniform random string from a given input such that it is possible to reproduce the same output string from a noisy version of the input. In [10] the authors show how a fuzzy extractor can be built using an error correcting code along with a universal hashing function. Their construction requires that the output of the fingerprint (the biometric data in their language) be represented as an element of $\mathcal{F}^n$ for some field $\mathcal{F}$ and an integer n which represents the size of the fingerprint. Moreover, it is naturally assumed that the noise experienced by the fingerprint is upper bounded by a constant distance from the original fingerprint in order to guarantee identical reproduction of the extracted key. We start by quoting the following theorem introduced in [10], and then give the specific construction which the theorem describes.

Theorem 1. ([10]) *Given any $[n, k, 2t+1]_{\mathcal{F}}$ code $\mathcal{C}$ and any m, ϵ, there exists an average-case $(M, m, \ell, t, \epsilon)$-fuzzy extractor, where $\ell = m + kf - nf - 2\log(\frac{1}{\epsilon}) + 2$. The generation algorithm* GEN *and the recovery algorithm* REP *are efficient if $\mathcal{C}$ has efficient encoding and decoding.*

We explain the parameters in the theorem by outlining an actual construction. This construction is proposed in [10] and further explained in [12]. As stated in the theorem, $\mathcal{C}$ is an error correcting code over the field $\mathcal{F}$, where $f = \log(|\mathcal{F}|)$.[3] For the construction we will also need a family of universal hashing functions $\mathbf{H}$.[4] The generation algorithm GEN takes the fingerprint $x \in \mathcal{F}^n$ as input and outputs the triplet (k, w, v). Here, x is drawn from some distribution X over $\mathcal{F}^n$ which has min-entropy m. Note that in our context the parameter m captures the entropy provided by the CD variability. GEN starts by computing $w = x + c$ for a randomly chosen code word $c \in \mathcal{C}$ and then computes the key $k = h_v(x) \in \{0,1\}^\ell$ for some string v chosen uniformly at random such that $h_v \in \mathbf{H}$. The recovery algorithm REP takes in the *helper data* (w, v) along with x', a noisy version of the

[3] Note that all logarithms in this paper are with respect to base 2.

[4] For details on universal hashing the reader is referred to [9].

fingerprint x, and returns the key k. REP starts by computing $c' = w - x'$ which is a noisy version of c. If the Hamming distance between x and x' is less than t then so will the Hamming distance between c and c'. Therefore, using the error correcting code $\mathcal{C}$, REP can reproduce c from c'. Next, REP computes $x = w - c$ and consequently compute $k = h_v(x)$ which will conclude the recovery algorithm. All that remains to be defined is the parameter ϵ which captures the security of the fuzzy extractor. Specifically, if the conditional min-entropy[5] $H_\infty(X|I)$ (meaning X conditioned on I)[6] is larger than m then $\mathbf{SD}((k,(w,v),I),(U_\ell,(w,v),I) \leq \epsilon)$ where $\mathbf{SD}(A,B) = \frac{1}{2}\sum_v |\Pr(A = v) - \Pr(B = v)|$ is the statistical distance between two probability distributions A and B. Finally, U_ℓ is the uniform distribution over $\{0,1\}^\ell$ and I is any auxiliary random variable.

With this construction we will have a clear way to build a fuzzy extractor. However, the key size ℓ and the security parameter ϵ will both depend on m and the code used. Moreover, the code will depend on the noise rate in the fingerprint. We finish this section by relating the min-entropy and the error rate of the fingerprint. Recall, that x is required to have a min-entropy of m and at the same time using the above construction x will have n symbols from $\mathcal{F}$. To merge these two requirements we define the average min-entropy in every symbol $\delta = m/n$. We also define ν to be the noise rate in the fingerprint x and $F = |\mathcal{F}|$. With these definitions we can now prove the following simple bound relating the noise rate and the min-entropy rate δ/f.

Proposition 1. *For the fuzzy extractor construction of Theorem 1, and for any meaningful security parameters of $\epsilon < 1$ and $\ell > 2$ we have $H_F(\nu) < \frac{\delta}{f}$. Where H_F is the F-ary entropy function.*

Proof. From Theorem 1 we now that $\ell = m + kf - nf - 2\log(\frac{1}{\epsilon}) + 2$. Let $A = \ell + 2\log(\frac{1}{\epsilon}) - 2 = m + kf - nf$. From the conditions above we now that $A > 0$ and therefore $m + kf - nf > 0$. Let $R = k/n$ which yields $(\delta + Rf - f)n > 0$ and therefore $R > 1 - \delta/f$. Using the sphere packing bound where $R \leq 1 - H_F(\nu)$ we immediately get $H_F(\nu) < \frac{\delta}{f}$.

As it is quite difficult to calculate the min-entropy for a physical source we will estimate this quantity over the symbols of x. The bound given above will give us an idea whether the min-entropy in the symbols of x will be sufficient to handle the measured noise rate. Next we shift our attention to the fingerprint extraction technique. Note here that we still did not address how the data extracted from the CDs will be transformed into the fingerprint x.

4.2 Fingerprint Extraction

In Section 3 we described how the empirical data suggests that every CD has unique location lengths. These location lengths as can be seen from Figure 7

[5] The definition of min entropy is $H_\infty(A) = -\log(\max_a \Pr[A = a])$.

[6] Typically we use the | operator to mean concatenation. This will be the only part of the paper where it will have a different meaning.

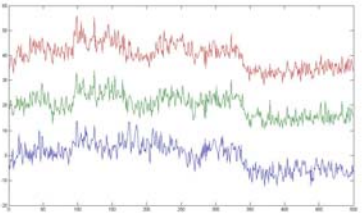

Fig. 9. Length variation over 500 locations from CD1 with the bottom trace taken 3 months after the top two traces

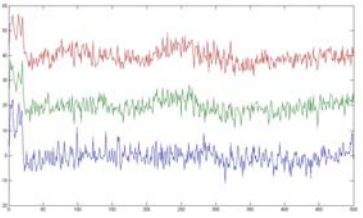

Fig. 10. Length variation over 500 locations from CD2 with the bottom trace taken 3 months after the top two traces

will have different values depending on the encoded information. Moreover, we discussed earlier that the raw data measured from the electrical signal will sometimes have different DC offsets. Therefore, it is important to process the data before the different locations can be combined together in order to produce the final fingerprint x. The first step in processing the data coming from every location on every CD is to remove the signal noise. To achieve this, the length of every location on a CD is averaged over a number of readings. Since we are assuming Gaussian noise, the noise level σ will scale to $\sigma/\sqrt{a}$ where a is the number of readings used for averaging. Next, we normalize the data using the ideal average of each location. As the ideal location lengths are discretized it becomes easy to find the ideal length for every location and subtract it from the measured lengths. This will guarantee that all location lengths have similar distributions as we saw in Figure 6. Finally, to remove the DC component we need a second normalizing step. We subtract the mean of the reading coming from different locations of the same CD. Figures 9 and 10 show the variation in the length of 500 locations for two identical CDs after being averaged and normalized. Each figure contains three traces with an added horizontal shift to set the traces apart. The top two traces in each figure are obtained from readings taken at different times using one CD reader. The bottom trace in each figure was obtained three months after the first two traces using a second CD reader with a different brand and model. The vertical axis represents the variation in nanometers from the ideal length of that location. These figures clearly support the idea of identical CDs having different fingerprints which are reproducible from different readers. We still need to outline a technique to extract a final fingerprint. Even after the previous averaging and normalization steps we will still have errors in the length readings. Although we will be using a fuzzy extractor to correct the errors, the biggest challenge towards achieving an efficient extraction technique will be the nature of these errors. The noise is Gaussian over the real values of the lengths. This means that even when the data is discretized the error will manifest itself more as a shift error from the ideal length rather than a bit flip error. Unfortunately, the Hamming metric does not naturally accommodate for this kind of error. Moreover, if we assume that every location length of the CD will be a symbol in the extracted fingerprint, then the error rate would be very high as it is very difficult to get the same exact

Table 1. Formulation of the threshold scheme for CD fingerprint extraction

Threshold Scheme: (GEN,REP) parameterized by $M, m, \ell, t, \epsilon, l, \mathcal{C}, \mathbf{H}, \tau = 2^s$
GEN: $(k, w, v) \leftarrow \text{GEN}(\text{CD}j)$
1. Obtain (a) samples for the length of each of the n locations on CDj.
2. Generate $z = z_n \dots z_1$:
a. Average the lengths over a samples,
b. Subtract the ideal mean from the averaged reads,
c. Normalize the sequence to have a zero mean and set that to z.
3. Find u such that $-2^{u-1} \leq z_i \leq 2^{u-1} - 1$ for all i, and shift z_i to $0 \leq z_i \leq 2^u - 1$.
4. Shift the binary representation of z_i left by l bits, round to an integer and set to $\hat{z}_i$.
5. Form $z_{2,i}$, the lowest $s+1$ bits of $\hat{z}_i$, and $x_i = z_{1,i}$, the remaining bits of $\hat{z}_i$.
6. Set $x = x_n \dots x_1$ to be the fingerprint template.
7. Choose a random code word $c \in \mathcal{C}$, such that $c = c_n \dots c_1$.
8. Compute $w_i = (x_i \| z_{2,i}) + (c|\tau)$ and form $w = w_n \dots w_1$.
9. Randomly choose v to compute $k = h_v(x)$ where $h_v \in \mathbf{H}$, and output (k, w, v).
REP: $k \leftarrow \text{REP}(\text{CD}j, w, v)$
1. Generate $z' = z'_n \dots z'_1$ as $\hat{z} = \hat{z}_n \dots \hat{z}_1$ was generated in Steps 1 through 4 of GEN.
2. Set c'_i to be the highest $u + l - s - 1$ bits of $w_i - z'_i$.
3. Use $\mathcal{C}$ to correct $c' = c'_n \dots c'_1$ to $c = c_n \dots c_1$.
4. Compute $x_i = w_i - c_i$.
5. Form $x = x_n \dots x_1$ and return $k = h_v(x)$.

length for the CD locations. A more natural distance metric in this situation would be the Lee metric [16]. However, this will require finding long codes that have good decoding performance under the Lee metric. To solve this problem we propose a *threshold* scheme which uses the Hamming distance while allowing a higher noise tolerance level. The threshold scheme also works naturally with the fuzzy extractor construction of Theorem 1. Table 1 shows a formulation of the threshold scheme applied to the CD setting. The threshold τ solves the error correcting problem with respect to the Lee distance. In particular, τ helps control the error rate which arises when treating the real values as symbols over some field. Without a threshold scheme ($\tau = 0$), the error rate will be very high. On the other hand, if τ grows too large then the error rate will be low. However, the Hamming distance between the extracted fingerprint originating from different CDs will decrease thus decreasing distinguishability between CDs. An important aspect about the threshold scheme is that it is very simple to compute and does not require previous knowledge of the distribution average.

4.3 Entropy Estimation and 128-Bit Security

The previous sections dealt with the theoretical aspects of extracting the CD fingerprint. In this section we take more of an experimental approach where we are interested in computing actual parameters. The most important parameters that we need to estimate are the entropy of the source (the CD variability) and the noise level. With these two parameters the rest of the parameters can be determined. The first and hardest task here will be to decide the amount of

entropy generated by the source. In [12] and [13] the authors use a universal source coding algorithm in order to estimate the secrecy rate. In particular it was proposed to use the Context-Tree Weighting Method (CTW) [19]. What is quite useful about the CTW algorithm is that in [18] it was shown that for any binary stationary and ergodic source X, the compression rate achieved by CTW is upper bounded by the min-entropy $H_\infty(X)$ as the length of the input sequence approaches infinity. This is a good indication about the entropy produced by the source provided enough bits are fed to the algorithm. To apply this algorithm to our setting we start by using the data coming from the 100 CDs. On each CD we collected data from 500 locations and processed the data with a threshold value of $\tau = 2^2$. The final data came out to be in the range $[0, 2^5 - 1]$ and we did not use any fractional bits so $l = 0$. With these parameters the size of the symbols was $f = 2$. This means that every CD produced 1000 bits. The data was fed into the CTW algorithm which resulted in a compression rate of about 0.83 bits of entropy per extracted bit. Recall here that these samples were not averaged over multiple reads. Therefore the error rate is quite high. When we averaged over 16 samples the combined entropy rate became 0.71. This is expected since the noise will add to the entropy. In order to get a more precise estimate for the min entropy we decided to average over 225 reads. With this many reads we had to restrict our sample to only 14 CDs as the amount of data quickly becomes large. With the new sample the compression rate of the CTW algorithm was about 0.675 which seemed to be a good estimate of our min-entropy. For this sample, the average error rate is $P_e = 0.08$. On the other hand the collision probability P_c, the probability of extracting similar bits between two different CDs, is about 0.46.

Proposition 1 suggests that for a noise rate of 0.08 and $f = 2$ the entropy of the source should be at least 0.40 which translates to $\delta = 0.8 < 1.35$, and therefore we conclude that we have enough entropy in our source. However, with this level of entropy we are placing stringent conditions on R, i.e. the rate of the error correcting code.[7] To relax the restriction on the code rate we took a closer look at our source bits. Ideally the two bits would have the same entropy. However, looking at Figure 9 and 10 and multiple similar figures we clearly see that there is a degree of dependency between the adjacent locations. There is a low probability of a sharp change in the length variability from one location to its neighbor. With this observation we would suspect that the most significant bit will have less entropy as it is less likely to change across adjacent locations. To verify this observation, we applied the CTW algorithm to each of the two extracted bits separately. For the most significant bit, the entropy for the cases of no averaging, averaging over 16 reads, and averaging over 225 reads was $1, 0.9$ and 0.6-bits of entropy, respectively. When we repeated this process for the least significant bit we obtained $1, 1$ and 0.98-bits of entropy, respectively. Clearly, we have more entropy in the least significant bit. It seems reasonable to only use the least significant bit to form the fingerprint and the final key. This would

[7] Recall from the prof of Proposition 1 that $R \geq A/nf + (1 - \delta/f)$ for a security level of at least $A = \ell + 2\epsilon - 2$.

effectively increase the entropy of our source while very slightly affecting the error rate and the collision rate. For this least significant bit scheme we obtained $P_e = 0.08$ and $P_c = 0.46$.

We now have $P_e = 0.08$, $\delta = 0.98$ and $f = 1$. With these parameters we can build a fuzzy extractor which can extract secure keys from CD fingerprints. For a 128-bit key we set $\ell = 128$. Similarly, to achieve a fuzzy extractor output which reveals very little information about the fingerprint we set $\epsilon = 64$. Using the equation of Theorem 1 we require that the error correcting code in the fuzzy extractor should satisfy $k \geq 190 + 0.02n$. Note that although $P_e = 0.08$, this is the expected error rate. For a practical scheme we require the fuzzy extractor to correct around a 0.17 error rate. These parameters can now be satisfied using a binary BCH code of $[255, 45, 88]$. More specifically, we define a code word containing 7 code words of this BCH code, which will make $n = 1785$. With this construction the failure probability[8] P_{fail} will be on the order of 10^{-6}. Note here that treating the 7 code words separately to generate separate parts of the key would significantly decrease ϵ but will decrease the failure probability. Therefore, in our failure probability we treat the 7 code words as a single entity. As we noted earlier, our data suffers from higher error rates due to the external connections which we used. With an on-chip process we can expect the error rate to drop significantly.

5 Robustness of the Fingerprint

A CD fingerprint can be used to tie software licenses to individual CDs where the software is stored. Under this use scenario it becomes important to address the robustness of the fingerprint. In all our experiments the data collected came from locations in the same sector of the CD. In a real application readings would typically be collected from different sectors. Thus ensuring that a scratch or any physical damage to a specific location will not render the CD fingerprint useless.

Another important concern regarding the robustness of the fingerprint is that of aging. Although no quantitative estimate of fingerprint durability can be given within the scope of this work, mechanisms related to viscoelastic relaxation in optical disc patterns need to be discussed briefly. Optical discs are printed on polymeric substrates, which have glass transition temperatures typically above 150 °C. The viscosity of such materials are temperature dependent and governed by an Arrhenius type exponential temperature dependence, described by an activation energy defined by the glass transition temperature. In its simplest form, the Arrhenius model assumes that the rate of change is proportional to $e^{\frac{-E_a}{kT}}$ where E_a is the activation energy, k is the Boltzmann constant (an invariant physical parameter) and T is the absolute temperature (temperature in degrees Kelvin). Even at lower temperatures (natural operating and storage temperature range of the optical disc), viscosity of the polymer remains finite. During the molding process, most of the internal stresses are relieved upon cooling, resulting in fluctuations in the nanoscale structure of the bit patterns. The

[8] Here, $P_{fail} = 1 - \left(1 - \sum_{i=0}^{t=43} \binom{n}{i} P_e^i (1 - P_e)^{n-i}\right)^7$.

pressed discs have a thin metal coating, which is typically coated on to the polymer disc by evaporation or sputter coating, that results in the increase of the surface temperature by up to 50 °C. This process is also likely to be a source of local thermoelastic stress buildup which relaxes over the lifetime of the CD. In a first order approximation, the disc material can be thought of as a Kelvin-Voigt material, and creep relaxation can be approximated by a single time-constant exponential behavior. In such a case, most of the viscoelastic relaxation will occur at the early stages of disc production, and latter time scales will have less of an effect. It may be speculated that the fingerprints due to length fluctuations of 25 nm upon 300 nm characteristic bit length will persist within at least 10% of the CD lifetime, which is predicted to be 217 years at 25 °C and 40% relative humidity conditions. This gives an estimated 20 year lifetime for the fingerprint [20]. Due to the exponential dependence of the relaxation on time, by recording the signature on a slightly aged optical disc (months old), the persistance of the signature can be increased.

6 Conclusion

In this paper we showed how to generate unique fingerprints for any CD. The proposed technique works for pressed and burned CDs, and in theory can be used for other optical storage devices. We tested the proposed technique using 100 identical CDs and characterized the variability across the studied CDs. We also gave specific parameters and showed how to extract a 128-bit cryptographic keys. This work opens a new door of research in the area of CD IP-protection.

References

1. DeJean, G., Kirovski, D.: RF-DNA: radio-frequency certificates of authenticity. In: Paillier, P., Verbauwhede, I. (eds.) CHES 2007. LNCS, vol. 4727, pp. 346–363. Springer, Heidelberg (2007)
2. Bauder, D.W.: An anti-counterfeiting concept for currency Systems. Research Report PTK-11990, Sandia National Labs, Albuquerque, NM, USA (1983)
3. Clarkson, W., Weyrich, T., Finkelstein, A., Heninger, N., Halderman, J.A., Felten, E.W.: Fingerprinting blank paper using commodity scanners. In: Proceedings of S&P 2009, Oakland, CA, May 2009. IEEE Computer Society, Los Alamitos (to appear, 2009)
4. Cowburn, R.P., Buchanan, J.D.R.: Verification of authenticity. US Patent Application 2007/0028093, July 27 (2006)
5. Agrawal, D., Baktir, S., Karakoyunlu, D., Rohatgi, P., Sunar, B.: Trojan detection using IC fingerprinting. In: Proceedings of S&P 2007, Oakland, California, USA, May 20-23, 2007, pp. 296–310. IEEE Computer Society, Los Alamitos (2007)
6. Lim, D., Lee, J.W., Gassend, B., Suh, G.E., van Dijk, M., Devadas, S.: Extracting secret keys from integrated circuits. IEEE Transactions on VLSI Systems 13(10), 1200–1205 (2005)
7. Ravikanth, P.S.: Physical One-Way Functions. PhD thesis, Department of Media Arts and Science, Massachusetts Institute of Technology, Cambridge, MA, USA (2001)

8. Tuyls, P., Schrijen, G.J., Skoric, B., van Geloven, J., Verhaegh, N., Wolters, R.: Read-proof hardware from protective coatings. In: Goubin, L., Matsui, M. (eds.) CHES 2006. LNCS, vol. 4249, pp. 369–383. Springer, Heidelberg (2006)
9. Carter, L., Wegman, M.: Universal hash functions. Journal of Computer and System Sciences 18(2), 143–154 (1979)
10. Dodis, Y., Ostrovsky, R., Reyzin, L., Smith, A.: Fuzzy extractors: how to generate strong keys from biometrics and other noisy data. SIAM Journal on Computing 38(1), 97–139 (2008)
11. European Computer Manufacturers' Association. Standard ECMA-130: Data interchange on read-only 120mm optical data disks (CD-ROM) (2nd edn.). ECMA, Geneva, Switzerland (1996)
12. Guajardo, J., Kumar, S.S., Schrijen, G.J., Tuyls, P.: FPGA intrinsic PUFs and their use for IP protection. In: Paillier, P., Verbauwhede, I. (eds.) CHES 2007. LNCS, vol. 4727, pp. 63–80. Springer, Heidelberg (2007)
13. Ignatenko, T., Schrijen, G.J., Skoric, B., Tuyls, P., Willems, F.: Estimating the secrecy-rate of physical unclonable functions with the context-tree weighting method. In: Proceedings of ISIT 2006, Seattle, Washington, USA, July 9-14, 2006, pp. 499–503. IEEE, Los Alamitos (2006)
14. Juels, A., Sudan, M.: A fuzzy vault scheme. Designs, Codes and Cryptography 38(2), 237–257 (2006)
15. Juels, A., Wattenberg, M.: A fuzzy commitment scheme. In: Proceedings of CCS 1999, pp. 28–36. ACM Press, New York (1999)
16. Lee, C.: Some properties of nonbinary error-correcting codes. IRE Transactions on Information Theory 4(2), 77–82 (1958)
17. Schift, H., David, C., Gabriel, M., Gobrecht, J., Heyderman, L.J., Kaiser, W., Köppel, S., Scandella, L.: Nanoreplication in polymers using hot embossing and injection molding. Microelectronic Engineering 53(1-4), 171–174 (2000)
18. Willems, F.M.J.: The context-tree weighting method: extensions. IEEE Transactions on Information Theory 44(2), 792–798 (1998)
19. Willems, F.M.J., Shtarkov, Y.M., Tjalkens, T.J.: The context-tree weighting method: basic properties. IEEE Transactions on Information Theory 41(3), 653–664 (1995)
20. Stinson, D., Ameli, F., Zaino, N.: Lifetime of Kodak writable CD and photo CD media. Eastman Kodak Company, Digital & Applied Imaging, NY, USA (1995)