A Modeling of Certificate Revocation and Its Application to Synthesis of Revocation Traces

Carlos Gañán, Jorge Mata-Díaz, Jose L. Muñoz, Juan Hernández-Serrano, Oscar Esparza, and Juanjo Alins

***Abstract*—One of the hardest tasks of a public key infrastructure (PKI) is to manage revocation. New communication paradigms push the revocation system to the limit and an accurate resource assessment is necessary before implementing a particular revocation distribution system. In this context, a precise modeling of certificate revocation is necessary. In this paper, we analyze empirical data from real certification authorities (CAs) to develop an accurate and rigorous model for certificate revocation. One of the key findings of our analysis is that the certificate revocation process is statistically self-similar. The proposed model is based on an autoregressive fractionally integrated moving average (ARFIMA) process. Then, using this model, we show how to build a synthetic revocation generator that can be used in simulations for resource assessment. Finally, we also show that our model produces synthetic revocation traces that are indistinguishable for practical purposes from those corresponding to actual revocations.**

***Index Terms*—Autoregressive fractionally integrated moving average (ARFIMA), public key infrastructure (PKI), revocation, self-similarity.**

I. Introduction

DIGITAL certificates are means of accurately and reliably distributing public keys to users needing to encrypt messages or verify digital signatures. Certificates are signed by certification authorities (CAs) and managed during their life-cycle by a Public Key Infrastructure (PKI). Various circumstances may cause a certificate to become invalid prior to the expiration of its planned validity period. Thus, the PKI has to collect and distribute information about revoked certificates. Currently deployed PKIs rely mostly on Certificate Revocation Lists (CRLs) for handling certificate revocation [1]. Although CRLs are the most widely used way of distributing certificate status information, much research effort has been put on studying other revocation distribution mechanisms in a variety of scenarios [2], [3].

In the past, little work has been done for analyzing the revocation process itself. For instance, many studies in the literature compare the performance of the different revocation distribution mechanisms considering very simplistic assumptions about the revocation process such that the percentage of revoked certificates remains always constant in the system. While these assumptions might have been enough in the early deployment of PKIs, we strongly believe that they have to be refined to face the challenges of currently deployed PKIs and also of emerging scenarios such as VANETs or WSNs. Only recently, we can find works like [40], [4]–[6] that carry out statistical studies about the revocation process using data available from real CAs. These studies can be considered a first step towards understanding revocation. They essentially analyze the probability distribution of revocation and conclude that this distribution roughly follows a Poisson distribution.

Following this direction, we also analyze empirical data from real CAs and we go a step further by developing an accurate and rigorous model for certificate revocation. One of the key findings of our analysis is that the certificate revocation process is statistically self-similar. As none of the currently common formal models for revocation is able to capture the self-similar nature of real revocation data, we develop a method for modeling this behavior. The proposed model is based on an autoregressive fractionally integrated moving average (ARFIMA) process [7], which provides an accurate and parsimonious model for revocation. Once we obtain the model, we show how to use it to build a synthetic revocation generator that can be used in simulations of resource assessment. To be able to construct the revocation trace generator, we will show that we need to concatenate a zero-memory nonlinear function (ZMNL) to the ARFIMA model. The final result is that our model produces synthetic revocation traces that are indistinguishable for practical purposes from those corresponding to actual revocations.

With synthetic revocation traces, current revocation schemes can be improved to define more accurate revocation data issuance policies. Neglecting the burstiness of the revocation process leads to inefficient revocation data release strategies. We show that traditional mechanisms that aim to scale, such as delta-CRL, can benefit from our traces to improve their updating strategies.

The rest of this paper is organized as follows. In Section II, we introduce the two stochastic processes that we use to model the revocation process. In Section III, we discuss the methodology we used to collect and analyze real-world revocation data. In Section IV, we identify the best ARFIMA model that fits the revocation events. Next, in Section V we present the generator of synthetic revocation traces using the obtained ARFIMA model. In Section VI we discuss the applications of this generator and in Section VII we present the impact of our finding on the related work. Finally, we conclude in Section VIII.

Manuscript received November 04, 2011; revised June 30, 2012; accepted July 04, 2012. Date of publication July 23, 2012; date of current version November 15, 2012. This work was supported by the Spanish Ministry of Science and Education under the projects CONSOLIDER-ARES (CSD2007-00004) and TEC2011-26452 "SERVET," and by the Government of Catalonia under Grant 2009 SGR 1362. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Robert H. Deng.

The authors are with the Telematics Department, Universitat Politècnica de Catalunya, Barcelona, 08034, Spain (e-mail: carlos.ganan@entel.upc.edu; jmata@entel.upc.edu; jose.munoz@entel.upc.edu; jserrano@entel.upc.edu; oesparza@entel.upc.edu; juanjo@entel.upc.edu).

Digital Object Identifier 10.1109/TIFS.2012.2209875

II. Background

A. Self-Similar Processes

A phenomenon which is self-similar looks the same or behaves the same when viewed at different degree of magnification. Self-similarity is the property of a series of data points to retain a pattern or appearance regardless of the level of granularity used and can be the result of long-range dependence (LRD) in the data. If a self-similar process is bursty at a wide range of timescales, it may often exhibit LRD.

LRD means that all the values at any time are correlated in a positive and non-negligible way with values at all future instants. A continuous time process $Y = Y(t),\ t \geq 0$ is self-similar if it satisfies the following condition [8]:

$$Y(t) \stackrel{D}{=} a^{-H} Y(t), \quad a > 0 \ \ t \geq 0, \quad for\ 0 < H < 1,$$

where $\stackrel{D}{=}$ means equally distributed and H is the index of self-similarity, called the Hurst parameter and the equality is in the sense of finite-dimensional distributions. The values of H are in the interval $(0.5, 1)$ if there exists LRD. A value of H equal to 0.5 indicates the absence of LRD.

B. Autoregressive Fractionally Integrated Moving Average (ARFIMA) Processes

One observed property of many data series is that they appear to have long memory, either in mean or in variance. This means that the effect of shocks on the time series takes a very long time to disappear. Traditional models describing short-term memory, such as autoregressive, integrated, moving average processes (ARIMA) defined in [7], cannot precisely describe long-term memory. A set of models has been established to overcome this difficulty, and the most famous one is the autoregressive fractionally integrated moving average model (ARFIMA). ARFIMA model was established by [9] and [10]. These processes are the natural generalization of the standard ARIMA by permitting the degree of differencing (d) to take fractional values.

Each ARFIMA process has three parts: the autoregressive (or AR) part; the integrated (or I) part; and the moving average (or MA) part. The models are often written in shorthand as $ARFIMA(p, d, q)$ where p describes the AR part, d describes the integrated part and q describes the MA part. Unlike common ARIMA, in ARFIMA processes the degree of differencing d is allowed to take non-integer values [10].

- **Auto Regressive**. This part of the model describes how each observation is a function of the previous p observations. For example, if $p = 1$, then each observation is a function of only one previous observation. That is, $y(n) = a_0 + a_1 y(n-1) + w(n)$ where $y(n)$ represents the observed value at n, $y(n-1)$ represents the previous observed value at $n-1$, $w(n)$ represents some random error and a_0 and a_1 are both constants. Other observed values of the series can be included in the right-hand side of the equation if $p > 1$:

$$y(n) = a_0 + a_1 y(n-1) + \cdots + a_p y(n-p) + w(n). \quad (1)$$

- **Integrated**. This part of the model determines whether the observed values are modeled directly, or whether the differences between consecutive observations are modeled instead. If $d = 0$, the observations are modeled directly. If $d = 1$, the differences between consecutive observations are modeled. If $d = 2$, the differences of the differences are modeled. In practice, d is rarely more than 2. That is a non-stationary process is integrated of order d if we need to difference it d times to induce stationary and it is denoted $I(d)$. Although the integrated component can be considered within the AR component by its formulation, its synthesis depends on different factors. Thus, the integrated component also shows the dependence with past values of the series but its synthesis depends on the nonstationary moments of the process. The order d of the integrated component is fixed by the order of the highest nonstationary moment of the stochastic process. In general, the integrated component can be expressed:

$$s(n) = c_1 s(n-1) + \cdots + c_d s(n-d) + w(n), \quad (2)$$

where:

$$c_i = \frac{\Gamma(i+d)}{\Gamma(d)\Gamma(i+1)}, \quad (3)$$

where Γ represents the gamma function, and

$$d < \frac{1}{2},\ d \neq 0, -1, -2, \ldots, \text{and } i \in \{1, 2, \ldots, \infty\}.$$

- **Moving Average**: This part of the model describes how each observation is a function of the previous q errors. For example, if $q = 1$, then each observation is a function of only one previous error. In general,

$$x(n) = b_0 w(n) + b_1 w(n-1) + \cdots + b_q w(n-q), \quad (4)$$

where the terms b_i are constant coefficients. Here $w(n)$ represents the random error at n and $w(n-q)$ represents the previous random error at $n-q$.

It is worth noting that ARFIMA processes are related to self-similarity. Beran showed that partial sums of an ARFIMA process have the same limiting distribution as a globally self-similar process [11]. Thus, an ARFIMA process is a self-similar process with the ability to capture both the short-range dependent (SRD) and LRD characteristics, that is, an ARFIMA process can be regarded as the increment process for a globally self-similar process. In this sense, we can relate the two parameters that define each one of these processes. The relation between the Hurst parameter (i.e., the index of self-similarity) and d (i.e., the index of fractionality) is [12]:

$$H = d + 0.5. \quad (5)$$

III. Data Collection and Preprocessing

The previous step to design our synthetic traces generator is to acquire information from real CAs about revoked certificates,

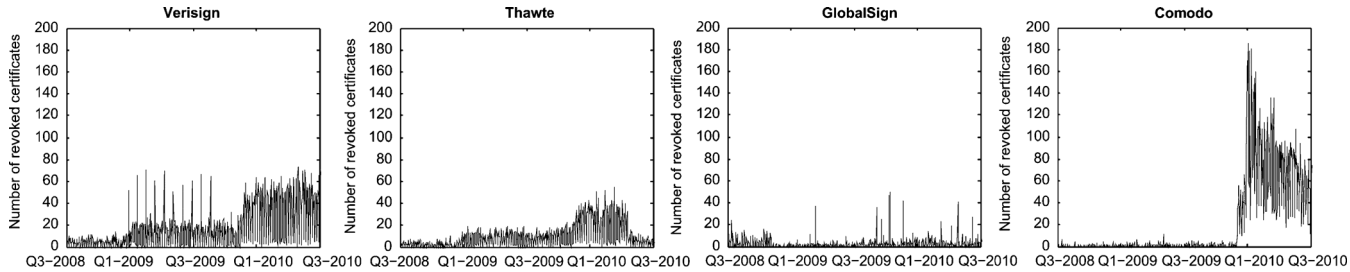

Fig. 1. Number of revoked certificates evolution for each CA.

TABLE I
DESCRIPTION OF THE COLLECTED CRLS

CRL Name	#Revoked Cert.	Issuer	Last Update	CRL Size	Cert. Type
Class3InternationalServer	16.584	VeriSign	2011/01/31	∼ 200 KB	SGC
UTN-USERFirst-Client	26.286	Comodo	2011/02/03	∼ 100 KB	SGC
PersonalSign Class 2 CA	6.695	GlobalSign	2011/02/07	∼ 300 KB	SGC
ThawteCodeSigningCA	10.418	Thawte	2011/01/31	∼ 200 KB	CSC

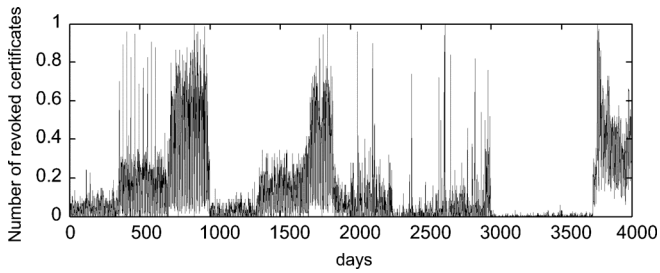

Fig. 2. Global time series.

so we can analyze the time evolution of these data. For this purpose, we collected revocation data from different certification authorities (CAs) using their available Certificate Revocation Lists (CRLs). In particular, we built some scripts to download and preprocess the CRLs from the following CAs[1]: VeriSign, Thawte, GlobalSign and Comodo.

Once downloaded the revocation data, we preprocess these data to remove duplicated information. For example, when a revoked certificate expires, it typically remains in the CRLs for one additional publication interval, so we preprocess the CRLs to remove these duplicates. On the other hand, Thawte's and GlobalSign's CRLs may contain duplicate entries for the same certificate because of their policy statements. These policy statements impose that a certificate that is revoked by several reasons must be included in the CRL as many times as the number of revocation reasons. Thus, we remove any duplicate entry from the composite dataset, and tally the number of revocations per day. Finally, we build a dataset that covers non-expired revoked certificates from 2008 to 2010. A summary about this dataset per CA is shown in Table I. Note that these CRLs cover two types of certificates: Server Gated Cryptography (SGC) and Code Signing Certificates (CSC, also known as a Software Publishing Certificate).

Then, from the dataset we could obtain the last update instant and the next update instant of the CRL, the serial number and the revocation date of each revoked certificate. With all this information, now we can analyze the time evolution of the number of revoked certificates per day. Fig. 1 shows the number of certificates revoked for the period of 2008/08/01 through 2010/08/21 for each CA. Analyzing all the collected data, we can conclude that:

- The number of revoked certificates bounces on a daily basis. Particularly, many revocations occur during weekdays, whereas few occur during weekends.
- There are some small peaks in the amount of certificates revoked per day. Moreover, there are also extraordinarily large spikes in certificate revocations at specific dates.
- Different CAs exhibit similar characteristics in terms of the revocation pattern. However, the mean number of revoked certificates depends on the CA's market share.
- Thawte and VeriSign exhibit a significant increment of the number of certificates revoked per day from 2008 to 2010. These changes might be mainly due to the changes in the total number of certificates being issued at different years. In this sense, as the number of certificates issued daily by Verisign increased approximately in 10 units from 2008 to 2009, the number of daily revoked certificates also increased in 1 unit in average. Thus, the percentage of revoked certificates remained fairly constant from one year to another. Similarly, this percentage also remains fairly constant for Thawte.
- GlobalSign market share is approximately five times lower than Verisign or Thawte. Therefore, the amount of revoked certificates that it manages is smaller. However, the time evolution of the number or revoked certificates follows a similar pattern, as there are also some small increases over the years.
- Comodo's market share has increased abruptly the last year, going from managing less than ten revoked certificates per day to more than a hundred. Despite this sudden increment, the autocorrelation function of the revocation process is still similar to the other three CAs.

As our goal is to obtain a generic model that fits the time evolution of the number of revoked certificates independently of the CA, we build a single time series by concatenating all four time series (see Fig. 2). By modeling the concatenation of all four time series, we will develop a generic model that will capture the revocation pattern independently of the CA. However, one of the difficulties in modeling revocation data is that the amount

[1]According to NetCraft's survey [13], using these CAs we cover most of the world market for SSL.

of revoked certificates that a CA manages depends on its market share. To analyze the global trend of the revocation process, we need to remove the influence of the volume of certificates that each CA manages, so that we can concentrate on the revocation pattern itself. One way to do this is to normalize the data. Therefore, we normalize each individual time series dividing by the maximum number of daily revoked certificates of each CA. This normalization will have to be undone once the model is obtained. Therefore, our synthetic revocation trace generator will need as input parameter the mean number of daily revoked certificates which is specific for each CA. Once all the data has been normalized, we create the global time series by concatenating the normalized data. In the following section we model this time series, obtaining a global model that fits well any CA.

IV. Modeling Revocation

In the previous section, we have analyzed the time evolution of the number of revoked certificates for each CA. So, we want to obtain a model from the revocation time-series to obtain a suitable model for generating synthetic revocation traces. For this purpose, one of the simplest techniques is to use a Multiple Linear Regression (MLR). Linear regression is useful for exploring the relationship of an independent variable to a dependent variable when the relationship is linear. However, MLR has drawbacks when the time-series exhibits high correlation. Correlation means that the value of the considered parameter at one time is influenced by values of the parameter at previous times. This happens when the values of the dependent variable over time are not randomly distributed. In our case, we will find that for the time-series of the number of daily revoked certificates, the error residuals are correlated with their own lagged values. This serial correlation violates the standard assumption of regression theory that disturbances are not correlated with other disturbances. The primary problems associated with serial correlation are:

- Regression analysis and basic time-series analysis are no longer efficient among the different linear estimators.
- Standard errors computed using the regression and time-series formula are not correct and are generally understated. If there are lagged dependent variables set as the regressors, regression estimates are biased and inconsistent but can be fixed using ARFIMA.

Another problem of MLR is that it fails to capture seasonal, cyclical, and countercyclical trends in time series. If there are lagged dependent variables set as the regressors, regression estimates are biased and inconsistent. Fortunately, this can be fixed using an autoregressive fractional integrated moving average Model (ARFIMA). Thus, as we have found that the simplest analysis, the MLR, is not suitable to develop our synthetic revocation trace generator, we will use ARFIMA.

The aim of this section is to show that the revocation process is a self-similar process that can be modeled as an ARFIMA process. To that end, we carry out the following steps:

1) Description of each component of the ARFIMA process in the Z-domain.
2) Formulation of the I component as a Laurent's series.
3) Characterization of the components of the ARFIMA process. This characterization consists of three steps:

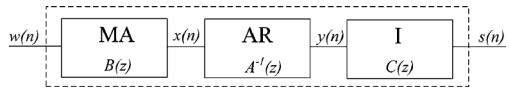

Fig. 3. Components of an ARFIMA process.

a) *Model identification*:
In this step, first we test for stationarity of the time series. Then, we determine the likely values of the order of the model, i.e., the p, d and q parameters of the ARFIMA model. The order to determine the likely values of the parameters is:
   a) d, the level of differencing. In our case, d is fractional, so we need to measure the intensity of self-similarity (H) and then we calculate d.
   b) p, the autoregression.
   c) q, the moving average.
   d) *Estimation of the ARMA components*:
   Once the order of the ARFIMA model has been determined, the values of the autoregressive component parameters and of the moving average component parameters are estimated to fit the global revocation time series.
   e) *Model Diagnostic checking*:
   Once we have identified and estimated the ARFIMA model, we assess the adequacy of the models to the revocation data. This model diagnostic checking step involves both parameter and residual analysis.

A. ARFIMA Processes in the Z-Domain

In Section II-B we have described ARFIMA processes in the time-domain. However, as any linear system, an ARFIMA process can be expressed by a difference equation involving the input series and the output series. If we Z-transform the difference equation and reorganize it, we can compute what is called the transfer function of the system.

For this purpose, we use the delay operator z^{-1} [14] to Z-transform the time-domain expression of an $ARFIMA(p,d,q)$ process. Fig. 3 shows a scheme of the ARFIMA model. Note that, as it is shown in the figure, we can express the transfer function of the $ARFIMA(p,d,q)$ process as a cascade of all three components.

The first step is to Z-transform the moving average component. A $MA(q)$ stochastic process is one that is generated using the difference equation expressed in (4). Applying the Z-transform to (4), we can express the MA process in the z-domain as:

$$B(z) = b_0 + b_1 z^{-1} + b_2 z^{-2} \cdots + b_q z^{-q}. \tag{6}$$

Note that in the previous expression, we only use previous samples of the input signal. The main features of the associated generating system are that it is Linear time-invariant (LTI), causal and stable. The MA system is Finite Impulse Response (FIR) and, therefore, an all-zero system. Fig. 4 represents the $MA(q)$ as a FIR filter whose transfer function is $B(z)$.

An $AR(p)$ stochastic process is one that is generated using the difference equation expressed in (1). This is a quite general situation, in which it is reasonable to think that a given sample

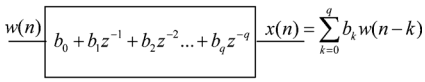

Fig. 4. MA filter.

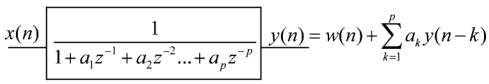

Fig. 5. AR filter.

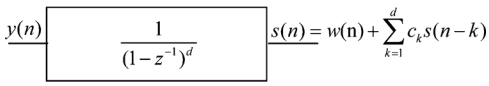

Fig. 6. Integrated filter.

of a time-series depends linearly on previous samples plus some random error. In this context, the transfer function of $AR(p)$ process in the z-domain can be expressed as:

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} \cdots + a_p z^{-p}. \tag{7}$$

The impulse response of the associated system is Infinite Impulse Response (IIR). Note that, this time, the autocorrelation is not limited and it tends to 0 when the lag tends to infinity, only if the module of all its poles is strictly smaller than 1. That means that if this condition is met, then the $AR(p)$ process is ergodic. Fig. 5 represents the $AR(q)$ as an IIR filter whose transfer function is $1/A(z)$.

Finally, we can also express the integrated component in the z-domain from (2):

$$C(z) = (1 - z^{-1})^{-d}. \tag{8}$$

In the same way as the autoregressive and moving average components of the ARFIMA process, we can represent the integrated component as a linear filter. Fig. 6 represents the $I(d)$ as a linear filter whose transfer function is $C(z)$.

Finally, the general expression of an $ARFIMA(p, d, q)$ process can be expressed by its Z-transform as:

$$S(z) = [B(z)A^{-1}(z)C(z)] \cdot W(z). \tag{9}$$

Understanding expression (9) as the relationship between the input $w(n)$ and the output $s(n)$ of a digital filter in a given instant n, the transfer function of the filter $H(z)$ could be defined as:

$$H(z) = \frac{S(z)}{W(z)} = B(z)A^{-1}(z)C(z). \tag{10}$$

It is worth noting that the factors of the transfer function follow the reverse order of the synthesis of the model. However, the order of the system in the cascade can be rearranged without affecting the characteristics of the overall combination. Hence, it is equivalent to changing the order by the commutative property of linear systems. Fig. 7 represents the ARFIMA filter with transfer function $H(z)$.

Note also that the roots of the polynomial $B(z)$ correspond to the zeros of the filter and the zeros of $A^{-1}(z)$ and $C(z)$ to

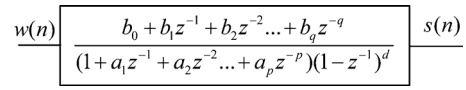

Fig. 7. ARFIMA filter.

the poles. According to the definition of the c_i values expressed in (3), the integrated order defines the multiplicity of the pole in $z = 1$. This pole generates the instability of impulsional response. The rest of obtained poles (z_k) will be found in the unit circle $(|z_k| < 1)$ of the Z plane.

B. Integrated Component as an Infinite Series

Once we have derived the expression of each one of the components of the ARFIMA process, we simplify the calculation of this integrated part to make an easier characterization of these components. This simplification will allow us building a much simpler generator as we will use sums and multiplications instead of gamma functions.

Thus, we show that we can use the Laurent's series to simplify the calculation of the integrated part expressed in (8). To that end, we show that the autocorrelation function of the coefficients c_i using gamma functions as in (3) is quite close to the autocorrelation function using (12) when limiting the number of coefficients.

A pure integrated process has the following transfer function defined in 8. According to Newton's generalized binomial theorem [15], $\left(1 - z^{-1}\right)^d$ is analytic in the open disk $\{z \in \mathbb{C} \mid |z| < 1\}$ for every $d \in \mathbb{C}$ and converges at $|z| = 1$ only for $d < 0$. Thus, the filter $C(z)$ can be expanded to the infinite series and we get the $MA(\infty)$ representation:

$$C(z) = \sum_{i=0}^{\infty} c_i z^{-i}, \tag{11}$$

where c_i is derived from the Laurent's series:

$$c_i = (-1)^i \frac{(-d)(-d-1)\ldots(-d-i-1)}{i!}, \tag{12}$$

and they can be related recursively as follows:

$$\begin{aligned} c_0 &= 1, \\ c_i &= c_{i-1} \frac{d+i-1}{i}. \end{aligned} \tag{13}$$

Moreover, these coefficients constitute the impulsional response $h(n)$ of the $ARFIMA(0, d, 0)$ filter. Fig. 8 shows the equivalence of the filters.

Let $y(n)$ be an uncorrelated process at the input of the filter, and $s(n)$ the output of the same filter. Then:

$$s(n) = y(n) * h(n). \tag{14}$$

Thus, the autocovariance functions satisfy:

$$K_{ss}(n) = K_{yy}(n) * r_{hh}(n), \tag{15}$$

where $K_{xx}(n)$ represents the autocovariance of the process $x(n)$. The input is an uncorrelated signal, so:

$$K_{yy}(n) = \delta(n), \tag{16}$$

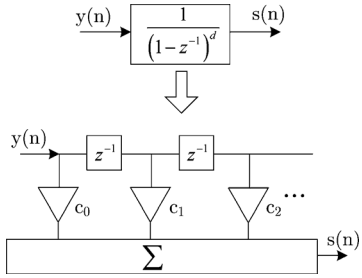

Fig. 8. Equivalent filters.

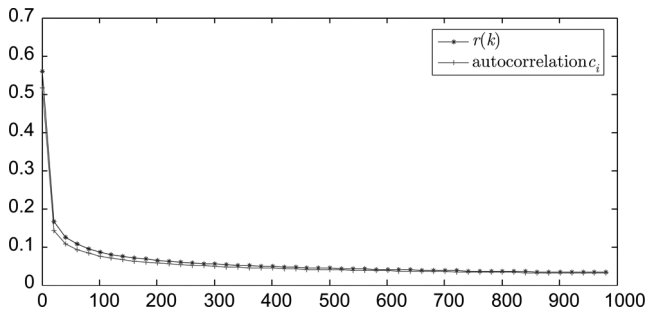

Fig. 9. Autocorrelation comparison.

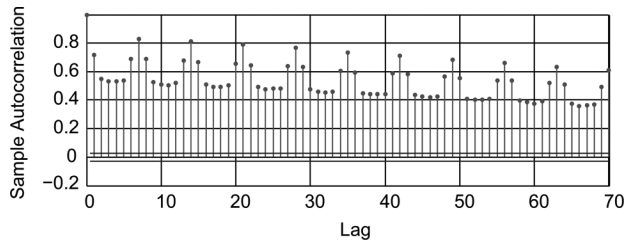

Fig. 10. ACF of the composite time series.

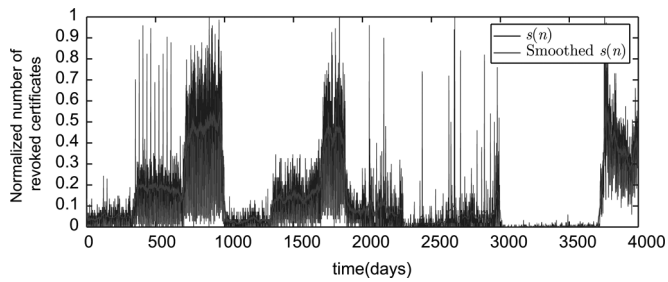

Fig. 11. Time evolution of the mean number of revoked certificates per day.

where $\delta(n)$ is the Dirac's delta function. Therefore:

$$K_{ss}(n) = K_{hh}(n). \tag{17}$$

Thus, the autocovariance of the output is the same as the autocovariance of the coefficients in expression (12). Note that as the autocorrelation is just a normalization of the autocovariance, the autocorrelation of the output is also the same as the autocorrelation of the coefficients c_i. Moreover, the output of the filter is a second order self-similar process.

Knowing that the ACF of an asymptotic second order self-similar process with self-similar parameter H is [8]:

$$r(k) = \frac{1}{2}\left[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}\right] \quad \forall k = 1, 2, 3, \ldots. \tag{18}$$

Therefore, the autocorrelation of the coefficients c_i must fit also the expression given in (18). In order to validate this study, the autocorrelation of 50000 coefficients is compared with the expression (18) with $H = 0.8$ ($d = 0.3$) in Fig. 9. As expected, the autocorrelation of the coefficients using the Laurent's series is quite close to the autocorrelation of the coefficients using the exact expression as in (3). One of the main goals of Section V will be to obtain a valid bound for the minimum number of coefficients needed to obtain the desired LRD at the output of the ARFIMA filter.

C. ARFIMA Model Identification

In this section, we identify all the necessary components and parameters of the ARFIMA model. To that end, we need to follow three steps.

1) Testing for Stationarity: Broadly speaking, a time-series is said to be stationary if there does not exist systematic change in the mean (no trend), if there does not exist systematic change in the variance, and if strict periodic variations have been removed. To test stationarity, we analyzed the Autocorrelation Function (ACF) of the global revocation time series of our dataset. The ACF plot is a plot of the partial correlation coefficients between the series and lags of itself.

This ACF function is shown in Fig. 10. As observed, the ACF of the time series decreases slowly, which is an indication that the mean is not stationary. Notice also that there exists certain seasonality in the ACF. This seasonality is mainly due to the fact that the number of revoked certificates decreases during the weekends. In addition, the temporal series of the number of revoked certificates (see Fig. 2) presents great variations. Thus, the first visual analysis suggests that the time series of the number of daily revoked certificates is non-stationary. To confirm this, we perform a unit root test. Those tests are based on the idea that a linear stochastic process has a unit root if 1 is a root of the process's characteristic equation. In such a case, a process is non-stationary. If the other roots of the characteristic equation lie inside the unit circle—that is, have a modulus (absolute value) less than one—then the first difference of the process will be stationary. We choose the KPSS test [16] at the 99% confidence level which rejects the null hypothesis of stationarity.

2) Finding the Integrated Component: Once we have confirmed that the time series is non-stationary, we have to find the integrated component, i.e., find d. The long range dependence complicates the characterization of the model because the temporal series shows an apparent non-stationary mean. This non-stationary mean can be observed in Fig. 11. To show the evolution of the mean number of revoked certificates, we have used a moving average filter with a 30-days span to smooth all of the data at once (by linear index). The long-range dependence produces that the mean varies. In order to characterize the ARFIMA model, it is necessary to capture this long range effect. Using the relationship between the index of self-similarity and the degree of differencing (see (5)), we can capture the LRD by means of the Hurst parameter.

TABLE II
HURST PARAMETER ESTIMATION FOR EACH CA USING DIFFERENT ESTIMATION METHOD

CA	Hurst Value			95% Confidence Interval		
	Whittle	R/S	Agg. Var.	Whittle	R/S	Agg. Var.
VeriSign	0.83	0.81	0.80	[0,79; 0,88]	[0,73; 0,84]	[0,72; 0,85]
Thawte	0.88	0.79	0.83	[0,83; 0,92]	[0,77; 0,81]	[0,79; 0,90]
GlobalSign	0.61	0.71	0.69	[0,56; 0,65]	[0,68; 0,75]	[0,67; 0,71]
Comodo	0.89	0.82	0.78	[0,80; 0,98]	[0,76; 0,88]	[0,73; 0,81]

However, while the Hurst parameter is perfectly well defined mathematically, its estimation is problematic. The accuracy and robustness of the H estimators can be influenced by processes such as periodicity, trend, length of the time series and short-range correlations which have different effects on different estimators [17]–[19] and can lead to erroneous estimation of the LRD intensity or even to reporting LRD on non-LRD series. This causes important problems when finding the most appropriate "H estimation" for the current time series. At present several methods to identify self-similar processes are known [20]. The most popular approaches are the following: analysis of R/S (rescaled adjusted range) statistics, analysis of the variance-time plot, analysis based on specific properties of $\mathrm{S}(\omega)$, Whittle estimation and analysis based on aggregated variance (see [41] for a comparison analysis). No single LRD estimator has been proved to produce more accurate estimates than the rest. We choose to use the Whittle estimation because it is based in the calculation of the FFT that has a numerical complexity of order $O[nlog_2(n)]$, so that it produces very fast algorithms for computing parameter estimations.

The Whittle method proposed in [21] calculates the Hurst index of self-similarity using the asymptotic properties of the spectral density. This method is based on a frequency domain maximum likelihood estimation of a fractionally integrated process for determining d. We estimate the Hurst parameter for all four CAs. The results are shown in Table II. It is worth noting that there exist notable differences in the Hurst value depending on the estimation method. As recently studied in [41], in general, different aspects must be taken into account before choosing the algorithm. For short inputs, only the R/S algorithm is suitable. As it is prone to noise, this algorithm must be used with care. For large input the Whittle method and wavelet-based algorithms will show better overall performance. As our dataset is quite large, we chose the Whittle method because its accuracy and simplicity. These results show that the Hurst parameter shows small variations depending on the CA. Despite these small variations, it is always above 0.5 (in the 0.61–0.89 range) which indicates the existence of the long-time correlation in the number of revoked certificates.

To build our general model, we will set the H parameter equal to the mean value of the estimated parameter $\hat{H}$ for each CA, i.e., $\hat{H} = 0.8$. The Hurst parameter is used to describe the degree of LRD and the burstiness of the traffic. Therefore, accurate characterization of LRD is very important in order to predict performance of the revocation service and to allocate network resources to provide a secure and reliable revocation mechanism. Our measured data show dramatically different statistical properties than those predicted by the stochastic models currently considered in the literature like Poisson or MMP. Almost

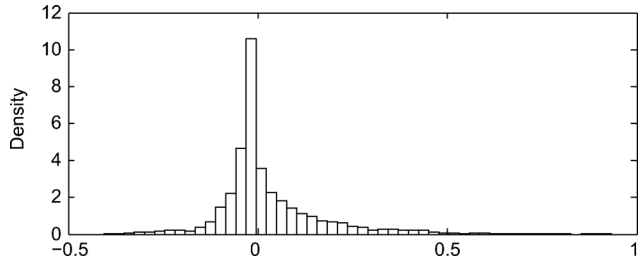

Fig. 12. PDF of the extracted integrated component time series.

all these models are characterized by an exponentially decaying autocorrelation function. As a result, they give rise to a Hurst parameter estimate of $H = .50$, producing variance-time curves, R/S plots, and frequency domain behavior strongly disagreeing with the self-similar behavior of actual revocation data. It is worth mentioning that the model derived will be rather insensitive to H (for the range of values that takes the analyzed CAs). So, the analysis of the Hurst parameter allows us to conclude that the integrated component of the model should be of order $\hat{d} = \hat{H} - 0.5 = 0.3$. Therefore, we can express its associated transfer function as $C(z) = (1 - z^{-1})^{-0.3}$.

3) Selecting the Parameters of the ARMA Process: Once we have obtained the value of d, we have to calculate the order of the autoregressive p and the order of the moving average q. To do so, it is necessary to extract the integrated component of the actual process $s(n)$. According to the scheme presented in Fig. 3, the residual ARMA series $y(n)$ and the real series $s(n)$ are related as follows:

$$Y(z) = C(z)^{-1} S(z) = \left(1 - z^{-1}\right)^{d} S(z). \tag{19}$$

Hence, the temporal series $y(n)$ can be obtained by means of a deconvolution filter. Deconvolution is, therefore, the reverse process in which an unknown input $Y(z)$ is calculated from the measured output $S(z)$ and a known transfer function $C(z)$ (approximated by the Laurent's series in expression (12) with 1,000 coefficients) . Fig. 12 shows the probability density function (PDF) of the obtained time-series $y(n)$. Note that this PDF does not fit exactly a normal distribution as there is a high peak around zero.

At this point, we have to check whether the time series without the integrated component is white noise or not. We check (at a confidence level of 99% and 70 lags) that $y(n)$ is not white noise by means of Ljung-Box Q-test [7]. The visual analysis of the autocorrelation of $y(n)$ confirms the result of this test (see Fig. 13). The fact that $y(n)$ is not white noise allows us to model it by means of an ARMA process.

Finally, we need to determine the best ARMA model that fits $y(n)$. We use the Rissanen's Minimum Description Length (MDL) criterion [22] for model selection among the different set of ARMA models with different numbers of parameters. It must be noted that information criteria penalize models with additional parameters. Therefore, the MDL model order selection criteria are based on parsimony, i.e., we adopt the simplest model that capture the self-similar pattern in accordance with the rule of Ockham's razor.

Varying the order of the AR component, we show in Fig. 14 the goodness of fit of each AR model. Note that each bar in the

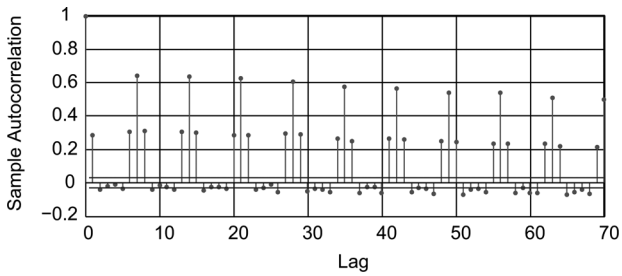

Fig. 13. ACF of the extracted integrated component time series.

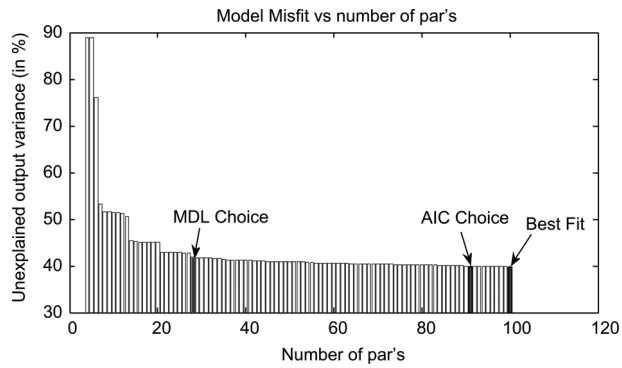

Fig. 14. AR parameter order estimation.

figure corresponds to an AR model with certain orders and delays. The x-axis shows the number of parameters in the respective models. The y-axis shows the part of the output variance, which is not explained by the model. That is, the ratio between the prediction error variance and the output variance in percent. Only that model that has the best fit for the given number of parameters is displayed. Therefore, using the MDL criteria, the order of the AR process found is $p = 29$ (see Fig. 14). In the next section, we calculate the value of the 29 coefficients of the AR component as well as the order and values of the MA coefficients.

D. Estimation of the ARMA Coefficients

Once we have identified the orders p and d of the ARFIMA model, we use the least square estimation to calculate the coefficients of the AR. After this calculation, we obtain the following coefficients:

$$\begin{aligned} A(z) = 1 &- 0.6467z^{-1} + 0.02693z^{-2} + 0.09085z^{-3} \\ &+ 0.09753z^{-4} + 0.1218z^{-5} + 0.1991z^{-6} \\ &- 0.804z^{-7} + 0.6906z^{-8} + 0.03223z^{-9} \\ &- 0.04807z^{-10} - 0.007471z^{-11} - 0.0759z^{-12} \\ &- 0.08934z^{-13} - 0.07605z^{-14} - 0.006487z^{-15} \\ &- 0.02565z^{-16} - 0.01994z^{-17} - 0.04003z^{-18} \\ &- 0.05007z^{-19} - 0.01331z^{-20} - 0.07361z^{-21} \\ &- 0.001947z^{-22} - 0.02836z^{-23} - 0.01824z^{-24} \\ &- 0.03693z^{-25} + 0.007019z^{-26} - 0.07691z^{-27} \\ &- 0.01872z^{-28} - 0.03821z^{-29}. \end{aligned} \tag{20}$$

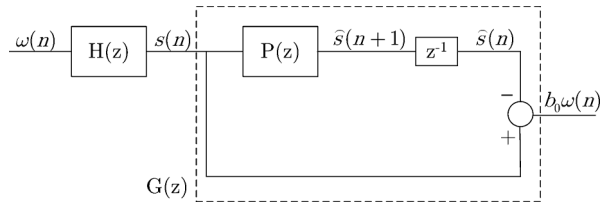

Fig. 15. Block diagram of the residual series estimation.

To obtain the MA component, the AR component of $y(n)$ is withdrawn. As shown in Fig. 3, the $x(n)$ series can be derived using the relation between the $y(n)$ and $x(n)$ series:

$$X(z) = A^{-1}(z)Y(z). \tag{21}$$

Again, we use a deconvolution to obtain $x(n)$ from the known inputs $y(n)$ and $a(n)$. To estimate the parameters of the MA process, the least square estimation is applied to fit the partial autocovariance function of $x(n)$ [7]. We use the same criteria as in the AR process for selecting the parameters of the MA process and the best adjustment is obtained with a MA process with order $q = 8$. Then, we calculate the MA coefficients, which yield the following result:

$$\begin{aligned} B(z) = 1 &- 0.6454z^{-1} + 0.005554z^{-2} \\ &+ 0.1113z^{-3} + 0.1317z^{-4} \\ &+ 0.1032z^{-5} + 0.2802z^{-6} \\ &- 0.6652z^{-7} + 0.6688z^{-8}. \end{aligned} \tag{22}$$

At this point, we have completely characterized the global revocation process as an $ARFIMA(29, 0.3, 8)$ process.

E. ARFIMA Model Diagnostic Checking

Now, we must check that the $ARFIMA(29, 0.3, 8)$ model fits the global revocation process. For that purpose, we analyze the residual series $w(n)$. The residual series can be calculated from the autoregressive component, the moving average component and the integrated component. In the following, we derive an expression of this residual series in the z-domain and analyze its statistical characteristics.

First, we express together the integrated and the moving average components in the following way:

$$B'(z) = \frac{B(z)}{C(z)} = B(z)\left(1 - z^{-1}\right)^d. \tag{23}$$

From [23] we know that:

$$s(n) - \hat{s}(n) = b_0 w(n). \tag{24}$$

Using (24), we define the transfer function of the diagnostic checker $G(z)$:

$$G(z) \triangleq b_0 \frac{W(z)}{S(z)}. \tag{25}$$

The relationship between the ARFIMA components and the residual series are shown in the scheme in Fig. 15.

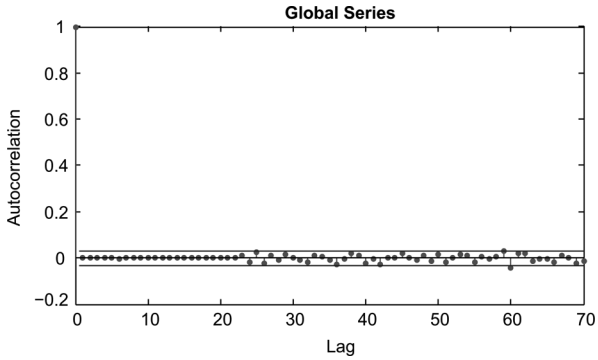

Fig. 16. ACF of the residual series for the global revocation process.

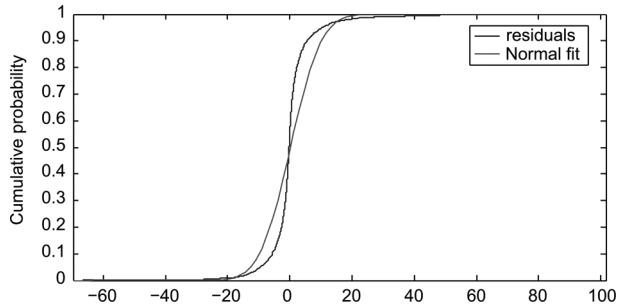

Fig. 17. CDF of the residual series versus Normal distribution.

From Fig. 15, we can express the residual series in the z-domain as:

$$W(z) = \frac{S(z)G(z)}{b_0}, \tag{26}$$

where $G(z)$ is

$$G(z) = 1 - P(z)z^{-1}, \tag{27}$$

and where $P(z)$ is the transfer function of the predictor.

Moreover, as it is shown in Fig. 15, the input of the system is $w(n)$ and the output is $b_0 w(n)$. Therefore:

$$H(z)G(z) = b_0. \tag{28}$$

Then, replacing (23) in (10), we can rewrite the transfer function of the ARFIMA filter as:

$$H(z) = \frac{B'(z)}{A(z)}. \tag{29}$$

Using (28), (29) and (27), we can express $P(z)$ as:

$$P(z) = \frac{B'(z) - b_0 A(z)}{z^{-1}B'(z)}. \tag{30}$$

Finally, replacing (30) and (27) in (26), we can express the residual series as a function of the ARFIMA components and the revocation series:

$$W(z) = \frac{S(z)(\frac{B'(z)}{b_0} - A(z))}{z^{-1}B'(z)}. \tag{31}$$

Once we have obtained the relationship between the residual series $w(n)$ and the ARFIMA components, we analyze the autocorrelation of the residuals. Fig. 16 presents the residuals' autocorrelation and the 99% confidence intervals. The residual diagnostic determines that the residuals are highly uncorrelated. Fig. 17 presents the CDF of the residuals' autocorrelation and the CDF of standard Gaussian distribution with its 99% confidence intervals. Notice that the residuals differ from the expected Gaussian distribution. The fact that the marginal distribution differs from the normal distribution will be taken into account during the design of the synthetic trace generator in the next section.

So far we have seen that the model fits quite accurately the global revocation process, next we check its suitability for each CA. For this purpose, we analyze the residuals for the individual revocation processes using the ARFIMA model obtained from the global series. Fig. 18 presents the residuals autocorrelation and the 99% confidence intervals for each CA. Although some residuals exceed the confidence intervals, the ARFIMA model still remains valid for each CA. Therefore, we can conclude that the proposed model is quite insensitive to Hurst variations, as it is able to fit well different revocation processes with different Hurst parameters.

Finally, in order to check that the model is not unspecified, we run again the Ljung-Box Q-test at a confidence level of 99% and 70 lags. The test results in the acceptance of the null hypothesis that the model fit is adequate (no serial correlation at the corresponding element of lags).

V. Revocation Traces Generation

Remember that the main goal of this work is to be able to generate synthetic revocation traces that mimic the self-similar behavior of real revocation data. To that end, we need to set a minimum bound for the number of coefficients required in the integrated component in order to achieve the desired LRD. For this purpose, at the input of the ARFIMA model, a Gaussian noise with zero mean and variance equal to 1, is applied. The number of coefficients in the integrated component is increased from 30 to 50,000 coefficients. The chosen value for the Hurst parameter is 0.8 ($d = 0.3$). The LRD level obtained in each case was measured using the Whittle method mentioned above. Table III shows the results of these tests. With 1,000 coefficients, the obtained value for $\hat{H}$ is 0.815, which is quite close to the expected value. Therefore, with only 1,000 coefficients of the equivalent filter, the generated synthetic series achieves a degree of self-similarity quite close to the expected one. Again, note that the model is rather insensitive to the H value.

Finally, to generate revocation traces the marginal distribution of our model must be adjusted to the probability distribution of the number of revoked certificates. In this step we find a problem. It is known that if the input of an ARFIMA model is Gaussian, the output will be also Gaussian. Furthermore, if the input is not Gaussian, then the output will not be Gaussian. As it was seen in the previous section, the marginal distribution of the number of revoked certificates is far from behaving as a normal distribution, rather it can be approximated as an exponential distribution (see [4]–[6] and Fig. 19). Thus, we will need to transform the normal probability distribution of our synthetic

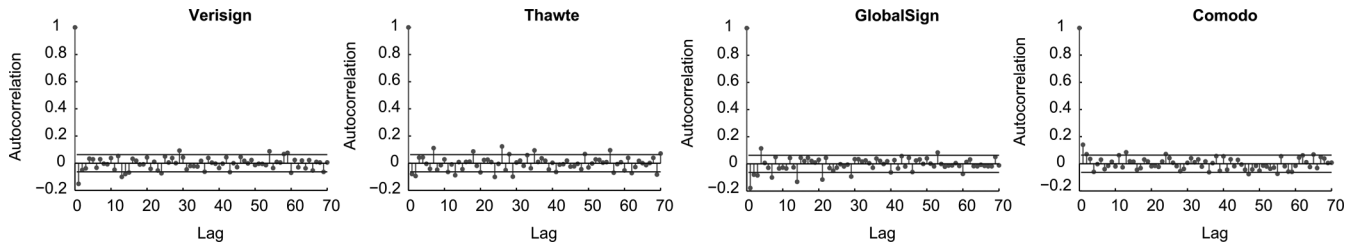

Fig. 18. ACF of the residual series for each CA.

TABLE III
HURST PARAMETER VERSUS NUMBER OF COEFFICIENTS

Number of coefficients	30	50	100	2,000	50,000
Hurst Parameter Value	0,899	0.852	0.827	0.809	0.806

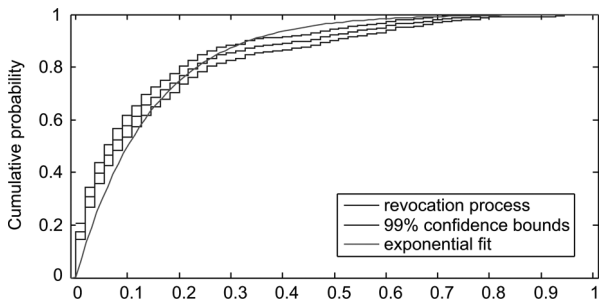

Fig. 19. CDF of the revocation process versus an exponential fit.

traces generator into an exponential distribution. For this purpose, we will use a zero-memory nonlinear (ZMNL) function as in [24], [25].

In more detail, we are going to use a monotonic ZMNL function $g(\cdot)$ relating the Gaussian distribution to the desired realization. This ZMNL is based on the cumulative distribution function of the revocation process and the known CDF of a Gaussian distribution. The established function is used to transform the Gaussian distribution into a realization of the exponential distribution. Since the transformation preserves the times of the zero-crossings and peaks of the original Gaussian distribution, and does not introduce any substantial discontinuities, the power spectral density is not substantially changed [26]. As the PDF of the revocation process can be approximated by an exponential distribution, the resulting CDF can be used with the known distribution of a Gaussian distribution to establish a ZMNL transformation function. This function relates a random variable with a Gaussian distribution to a random variable with exponential distribution.

Fig. 20 shows the block diagram of the synthetic revocation generator, where the ZMNL function is placed at the output of the ARFIMA filter. The values of the white noise sequence $w(n)$ at the input of the ARFIMA filter are chosen such that $Var(w(n)) = 1$ and $E[w(n)] = 0$. In turn, the output of the ARFIMA filter $s(n)$ becomes the input of the ZMNL function. In this way, the ARFIMA model transforms the colored $N(0,1)$ sequence in a colored $N(0,\sigma_s^2)$ sequence. Then, the ZMNL function transforms the colored $N(0,\sigma_s^2)$ sequence in

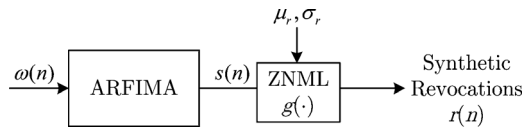

Fig. 20. Synthetic Revocation trace generator.

an $Exponential(\mu_r)$ sequence, where μ_r is the measured average of daily revoked certificates.

Hence, the ARFIMA filter output $s(n)$ is followed by a ZMNL $g(\cdot)$, which is chosen so that:

1) All the random samples $g(s(n))$ are positive.
2) The expected value $E[g(s(n))]$ matches well with the average of daily revoked certificates μ_r.
3) The influence on the autocovariance function of the synthetic sequence is low.

The conditions i) and ii) are obviously necessary in order to convert the Gaussian sequence (output of the ARFIMA filter) to a sequence which represents revocations. Fulfilling condition iii) allows to fit the measured autocovariance function directly to the ARFIMA filter. Thus, with ARFIMA and ZMNL together, it becomes possible to recreate the self-similar nature of the revocation process.

In order to fulfill the aforementioned conditions, we use a series of algebraic manipulations to obtain the desired ZMNL function that transforms $s(n)$ to an exponential distribution with specific parameters. First, we use the probability integral transformation that allows to convert the Gaussian distribution $(s(n))$ to a uniform distribution $(u(n))$:

$$u(n) \triangleq \int_{-\infty}^{s(n)} e^{-x/2\,\sigma_s^2} dx. \tag{32}$$

Once we have a uniform distribution, we convert it to an exponential distribution $(e(n))$ applying a logarithmic transformation (see Inverse Transformation Techniques in [27]).

$$e(n) \triangleq -\mu_r ln(u(n)). \tag{33}$$

Then, we modify the mean and variance of the exponential distribution in order to fit the desired parameters of the synthetic revocation process. Finally, we truncate the output allowing only nonnegative integers. Thus, the ZMNL function can be expressed as:

$$g(s(n)) = \max\left(0, \left\lceil a - b \cdot ln\left(\frac{1}{\sqrt{2\pi\,\sigma_s^2}} \int_{-\infty}^{s(n)} e^{-x/2\,\sigma_s^2} dx\right)\right\rceil\right). \tag{34}$$

TABLE IV
MEAN AND STANDARD DEVIATION OF THE NUMBER OF REVOKED CERTIFICATES FOR EACH CA

Certification Authority	μ_r	σ_r
VeriSign	15.7077	18.1256
Thawte	7.9970	10.1251
GlobalSign	2.7658	5.4362
Comodo	47.8001	42.6027

where the parameters a and b have to be chosen so that the measured mean and variance match well with that of the revocation process. In this manner,

$$a = \mu_r, \quad b = \mu_r \frac{\sigma_r}{\sigma_s},$$

where μ_r and σ_r are the mean and standard deviation of the synthetic revocation process to generate respectively.

Note that the variance of the $s(n)$ only depends on the value of the coefficients of the ARFIMA filter. In short, the amplitude of the autocovariance at zero displacement provides a measure of the signal magnitude [14]. This variance can be expressed as:

$$\sigma_s^2 = K_{hh}(0).$$

The expression of the $g(\cdot)$ in (34) can be simplified using the Q-function [28] to facilitate the implementation of the ZMNL function:

$$g(s(n)) = \max\left(0, \left\lceil a - b \cdot ln\left(1 - Q\left(\frac{s(n)}{\sigma_s}\right)\right)\right\rceil\right). \quad (35)$$

Using the Chernoff bound [29] of the Q-function, we can express the $g(\cdot)$ as:

$$g(s(n)) \simeq \max\left(0, \left\lceil a - b\left(ln(2) + \frac{s(n)^2}{2\sigma_s^2}\right)\right\rceil\right). \quad (36)$$

The ZMNL function allows to amplify the output of the fractional ARIMA model to tailor the standard deviation of the synthetic revocation process. In the same way, the mean of the revocation process is increased allowing to model different revocation rates. Note that the mean number of revoked certificates per day (μ_r) varies depending on the CA. Therefore this parameter must be set according to Table IV.

It is worth noting that the mean number of revoked certificates per day (μ_r) follows a specific pattern. Mainly, this parameter depends on the market share of the certification authority. Using the market share of each CA [30] as exogenous variable, we can perform a simple lineal regression analysis to estimate the endogenous variable μ_r. The result of the regression analysis using least squares is:

$$\mu_r = -10.2 + 3.22\ Market\ Share(\%).$$

The p-value of the regression analysis $(p = 0.099)$ indicates that the relationship between μ_r and $MarketShare(\%)$ is statistically significant at an α-level of 0.1. This is also shown by the p-value for the estimated coefficient of $MarketShare(\%)$, which is 0.102. Analysis the coefficient of determination R^2 value, it shows that the $MarketShare(\%)$ explains 59,8% of the variance in μ_r, indicating that the model fits the data well. Because the model is significant and explains a large part of the variance in μ_r, the generator could have as input the $MarketShare(\%)$ of the desired CA instead of μ_r.

Summing up, using as ZMNL function (36), and concatenating it to the ARFIMA filter as in Fig. 20, we are able to generate synthetic revocation traces that mimic the actual behavior of the revocation process.

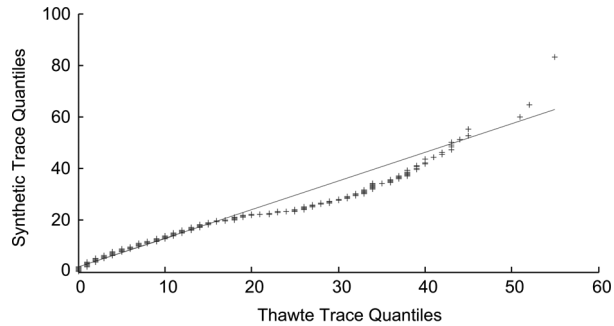

Fig. 21. Quantile-to-quantile plot of the generated synthetic trace and Thawte's revocation trace.

A. Quality of the Traces

In the following, we present a summary of a comprehensive study that evaluates the quality of the synthetic revocation traces. The goal of this study is to prove that the method really synthesizes revocations corresponding to an $ARFIMA(29, 0.3, 8)$ process. We show that the proposed generator produces synthetic revocations that are indistinguishable for practical purposes from those corresponding to actual revocations. Therefore, its marginal distribution is exponential and its autocorrelation function adequately fits the actual ACF, the estimation of H is close to the true value, and its spectral density is consistent with ARFIMA.

Exponentiality. The synthesized revocation traces follow an exponential distribution for practical purposes. In order to analyze the marginal distribution we have carried out the χ^2, Kolmogorov-Smirnov, and Anderson-Darling goodness-of-fit tests with satisfactory results. Moreover, visual tests such as the QQ-plot show that the traces follow quite well an exponential distribution (see Fig. 21).

Correlation Structure. To see how the autocorrelation function (ACF) of the revocation traces fits the correlations of an actual revocation trace, we generate a sequence based on Thawte's revocation statistics. Fig. 22 compares the ACF of the generated trace and the ACF of Thawte's revocation trace. The slope of both functions for large lags, which is related with the LRD level, is very similar. This allows to confirm the goodness of fit.

Spectral Density. Our synthesizing method easily passes the strict Beran goodness-of-fit test for the spectral density [11]. In this sense, the percentage of rejections is always lower than or equal to the level of significance.

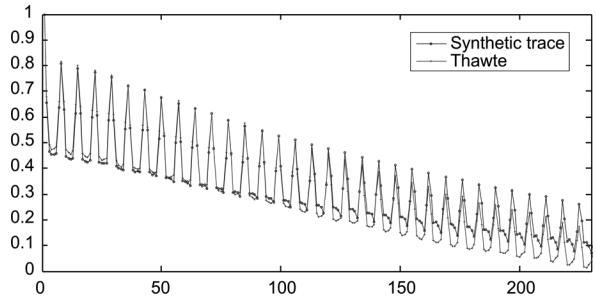

Fig. 22. ACF of a generated trace versus ACF of the revocation time series from Thawte.

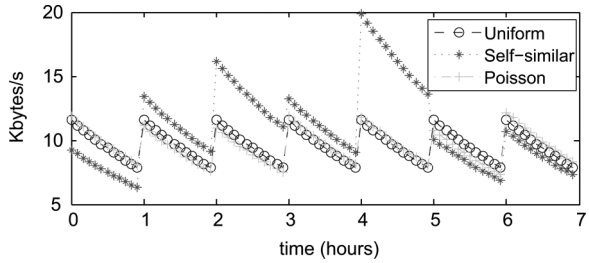

Fig. 23. BW consumption depending on the revocation process distribution.

VI. Synthetic Revocation Traces

In this section we describe the potential applications of synthetic revocation traces, because they are desirable for many reasons. Firstly, thorough revocation data are difficult to obtain, mostly because of security or privacy concerns. Only a few CAs (e.g., Verisign) allow to obtain additional information about the revoked certificates such as the issued time, the country or the issuer. However, revocation traces are necessary to evaluate the performance of certificate status validation mechanisms. To compare the performance of different mechanisms, researchers run simulations based on theoretical assumptions. For example, Naor and Nissim calculate the communication cost by assuming a fixed length CRL [31]. Cooper [32] and Arnes [33] model the distribution of revocation information by assuming an exponential inter-arrival probability for the requests for CRLs. Their theoretical assumptions turn their results into qualitative information, as they do not use neither real nor synthetic revocation traces to develop their models.

Secondly, there do not exist CRLs large enough to test the revocation mechanism proposed for new environments (e.g., VANETs) where these lists are expected to contain millions of revoked certificates [34]. In this sense, revocation mechanisms proposals for VANETs (e.g., [35]–[37]) that are tested without taking into account the self-similarity pattern of the revocation process will not be completely accurate.

Finally, synthetic trace generation is appealing because is simple, fast, and controllable. The key issue is the fidelity of a synthetic revocation trace, i.e., how well it mimics the relevant statistical properties of real revocation data. The synthetic revocation traces must fit not only the probability distribution but also the temporal correlation of the revocation process.

VII. Related Work

Most of previous works are not based on any empirical analysis of real-world data; instead, they focus on theoretical aspects of certificate revocation including the cause of revocation [38], the model of revocation [32] and communication cost of revocation [31]. However, these theoretical models are not able to capture the actual behavior of the revocation data. Most recently, authors have studied the statistical characteristics of real revocation data [4]–[6]. However, the bursty pattern of the revocation process has been neglected. In the previous sections we have shown that revocation data is statistically self-similar.

Regarding the traditional way of issuing CRLs, X.509 defines one method to release CRLs. This method involves each CA periodically issuing CRLs. Using this method the issued CRLs will contain a number of revoked certificates that will differ significantly from one CRL to another CRL. Thus, each CRL will have a different size, and the transmission of these lists will be bandwidth inefficient. Ma *et al.* in [5] already pointed out the inefficiencies of the traditional method, and proposed releasing CRLs based on a series of economic and liability costs. However, when they defined the function that calculates the number of new certificate revocations between two days, they assumed a Poisson process neglecting the burst pattern. Therefore, the CRL release policies they obtained could be highly improved by taking into account the self-similarity of the revocation process. Specifically, the liability costs could be reduced by issuing CRLs in order to palliate the burst pattern of the revocation data. To give an idea of the impact of using synthetic revocation traces, we analyze the work of Cooper in [32] and in [39]. In these works, Cooper analyzed the best way to issue CRLs, segmented CRLs and delta-CRLs in order to decrease the request peak bandwidth. The author assumed that an average of 1,000 certificates are revoked each day and that the CRLs have a fixed validity time. By doing these assumptions, the self-similar behavior of the revocation process is neglected and the results need to be adapted to the reality. From [39], we can calculate the bandwidth for a delta-CRL system can be computed as:

$$B = \frac{Nve^{-vt}((51+4.5rL_c)e^{-(w+l/O-l)v} + (51+9rw))}{(O-1)1 - e^{vl/O} + 1},$$

where N is the number of valid certificates, v is the validation rate, l is the amount of time that a delta-CRL is valid, L_c is the certificate lifetime, r is the number of certificates revoked per day, w is the window size of the delta-CRL and O is the number of delta-CRLs that are valid at any given time.

Using the bandwidth as comparison metric, we can evaluate the impact of the self-similarity. Fig. 23 shows the bandwidth necessary to download the revocation data using a sliding window delta-CRL scheme. We have assumed that there are 300,000 relying parties each validating an average of 10 certificates per day; delta-CRLs are issued once an hour, are valid for 4 hours, and have a window size of 9 hours. We have also assumed that an average of 10 certificates are revoked each day and that certificates are valid for 365 days. Note that depending on the distribution of the revocation process, the required bandwidth presents significant variations. Uniform and poisson distributions present a similar behavior. On the opposite, a self-similar process makes the delta-CRL's size to vary. Thus, the optimal window to issue delta-CRLs should be calculated taking into account the bursty pattern of the self-similar process. If this pattern is neglected, the peak bandwidth will vary with each delta-CRL issuance making the revocation service bandwidth-inefficient. When with a poisson or uniform

process the maximum peak bandwidth is of $\sim 12\ \mathrm{Kb/s}$, a burst of revocation events causes that during the 15th delta-CRL issuance there are required $\sim\ 20\ \mathrm{Kb/s}$. Therefore, ignoring the self-similar pattern of the revocation process will led to inaccurate network planning.

Authors in [5], [6] suggest a functional form for the probability density function of certificate revocation requests. They choose a exponential distribution function because it adequately approximates the data they collected from a single CA. Based on this assumption, they provide an economic model based on which a CA can choose what they state to be the optimal CRL release interval. However, they do not take into account the self-similar behavior of the revocation data. Intuitively, the critical characteristic of this self-similar pattern is that there is no natural length of a "burst" of revoked certificates: at every time scale ranging from a few days to weeks and months, similar-looking revocation bursts are evident. This bursty pattern could be taken into account by CAs to derive better strategies to release CRLs.

Walleck *et al.* in [4] carried out a deeper empirical analysis of the revocation data not only taking into account the number of revoked certificates, but also other variables such as geographical factors. They also conclude that their collected CRLs exhibit exponential distribution patterns. Though they acknowledge the existence of revocation burst, they do not capture this fractal behavior.

This self-similar or apparently fractal-like behavior of revocation data is very different both from currently considered formal models (e.g., pure Poisson or Poisson-related models). These differences require a new look at modeling the data and performance of the revocation service. For example, our analysis of the revocation data shows that the generally accepted argument for the "Poisson-like" nature of revocation events, namely, that the number of revoked certificates becomes smoother (less bursty) as the number of certificate sources increases, has very little to do with reality. In fact, the burstiness (degree of self-similarity) is expected to be intensified as the number of active certificate sources increases, contrary to commonly held views. Thus, in novel environments such as VANETs, where digital certificates will be used to provide anonymity, increasing the number of valid certificates, the self-similar behavior of the revocation data cannot be neglected. Thus, self-similarity is both ubiquitous in our data and unavoidable in future, and more in highly populated networks. However, none of the currently common formal models for revocation data is able to capture it.

VIII. Conclusions

In this paper, we have analyzed real empirical data collected from the leading CAs. The main findings of our analysis about the revocation process are that (I) this process is statistically self-similar (irrespective of when data were collected during the 3-year period 2008–2011 or from which CA), (II) the degree of self-similarity, which can be measured in terms of the Hurst parameter H, is a function of the overall utilization of the revocation service and can be used for measuring the "burstiness" of the revocation process (i.e., the more bursts in the revocation process the higher H), and (III) the leading CAs share similar Hurst parameters even though they operate in different market segments.

Moreover, as none of the currently common formal models for revocation is able to capture the self-similar nature of real revocation data, we have developed a method for modeling this behavior. The proposed model is based on an ARFIMA process, that provides an accurate and parsimonious model. In this context, this research represents a step towards linking empirical observations to mathematical models in description of the complex process of certificate revocation. We believe that this is going to be necessary in traditional scenarios with a high number of users as well as in incipient certification scenarios such as vehicular communications, in which CAs will have to deal potentially billions of issued certificates.

Finally, for practical purposes, we have shown how the developed model can be easily used as a synthetic revocation generator. We have also shown that our model produces synthetic revocations that are indistinguishable for practical purposes from those corresponding to actual revocations.

References

[1] R. Housley, W. Polk, W. Ford, and D. Solo, Internet X.509 Public key Infrastructure Certificate and Certificate Revocation List (CRL) Profile, Internet Engineering Task Force RFC 3280 Apr. 2002 [Online]. Available: http://www.rfc-editor.org/rfc/rfc3280.txt

[2] T. P. Hormann, K. Wrona, and S. Holtmanns, "Evaluation of certificate validation mechanisms," *Comput. Commun.* vol. 29, pp. 291–305, Feb. 2006 [Online]. Available: http://portal.acm.org/citation.cfm?id=1646655.1646878

[3] M. Lippert, V. Karatsiolis, A. Wiesmaier, and J. Buchmann, "Lifecycle management of x.509 certificates based on ldap directories," *J. Comput. Secur.* vol. 14, pp. 419–439, Sep. 2006 [Online]. Available: http://portal.acm.org/citation.cfm?id=1239313.1239316

[4] D. Walleck, Y. Li, and S. Xu, "Empirical analysis of certificate revocation lists," in *Proc. 22nd Ann. IFIP WG 11.3 Working Conf. Data and Applications Security*, 2008, pp. 159–174.

[5] C. Ma, N. Hu, and Y. Li, "On the release of crls in public key infrastructure," in *Proc. 15th Conf. USENIX Security Symp.*, Berkeley, CA, 2006, vol. 15.

[6] N. Hu, G. K. Tayi, C. Ma, and Y. Li, "Certificate revocation release policies," *J. Comput. Secur.* vol. 17, pp. 127–157, Apr. 2009 [Online]. Available: http://portal.acm.org/citation.cfm?id=1544133.1544134

[7] G. E. P. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. Holden-Day: Incorporated, 1990.

[8] W. Willinger, V. Paxson, and M. S. Taqqu, "Self-similarity and heavy tails: Structural modeling of network traffic," in *A Practical Guide to Heavy Tails*, R. J. Adler, R. E. Feldman, and M. S. Taqqu, Eds. Cambridge, MA: Birkhauser Boston Inc., 1998, pp. 27–53.

[9] C. W. J. Granger and R. Joyeux, "An introduction to long-memory time series models and fractional differencing," *J. Time Series Anal.*, vol. 1, no. 1, pp. 15–29, 1980.

[10] J. R. M. Hosking, "Fractional differencing," *Biometrika*, vol. 68, no. 1, pp. 165–176, Apr. 1981.

[11] J. Beran, *Statistics for long-memory processes, ser. Monographs on statistics and applied probability*. London, U.K.: Chapman & Hall, Oct. 1994, vol. 61.

[12] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Trans. Netw.*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[13] Netcraft, Market Share of Certification Authorities 2009 [Online]. Available: https://ssl.netcraft.com/ssl-sample-report/CMatch/certs

[14] J. G. Proakis, *Digital communications/John G. Proakis*. New York: McGraw-Hill, 1983.

[15] C. Liu, "The essence of the generalized newton binomial theorem," *Commun. Nonlinear Sci. Numerical Simulation*, vol. 15, no. 10, pp. 2766–2768, 2010.

[16] D. Kwiatkowski, P. C. Phillips, and P. Schmidt, Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root, Cowles Foundation for Research in Economics, Yale University, May 1991.

[17] A. Montanari, M. S. Taqqu, and V. Teverovsky, "Estimating long-range dependence in the presence of periodicity: An empirical study," *Mathemat. Comput. Modelling*, vol. 29, no. 10–12, pp. 217–228, 1999.

[18] M. J. Cannon, D. B. Percival, D. C. Caccia, G. M. Raymond, and J. B. Bassingthwaighte, "Evaluating scaled windowed variance methods for estimating the hurst coefficient of time series," *Physica A, Statist. Theoret. Phys.*, vol. 241, no. 3–4, pp. 606–626, 1997.

[19] T. Karagiannis, M. Molle, and M. Faloutsos, Understanding the Limitations of Estimation Methods for Long-Range Dependence, UC Riverside, 2006.
[20] R. G. Clegg, A Practical Guide to Measuring the Hurst Parameter, 21st UK Performance Engineering Workshop, School of Computing Science, University of Newcastle, Tech. Rep. Series, CSTR-916, 2006, pp. 43–55.
[21] P. Whittle, "Estimation and information in stationary time series," *Arkiv für Matematik*, vol. 2, pp. 423–434, 1953.
[22] J. Rissanen, *Information and Complexity in Statistical Modeling*, 1st ed. New York: Springer, 2007.
[23] L. de la Cruz, E. Pallarès, J. Alins, and J. Mata, "Self-similar traffic generation using a fractional arima model—Application to vbr video traffic," *J. Brazilian Telecommun. Soc.*, no. 14, p. 1, 1999.
[24] P.-R. Chang and J.-T. Hu, "Optimal nonlinear adaptive prediction and modeling of mpeg video in atm networks using pipelined recurrent neural networks," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 6, pp. 1087–1100, Aug. 1997.
[25] R. Gruenenfelder, "Stochastic Modelling of the Traffic and its Properties in an ATM Network," Ph.D. thesis, Lausanne, 1991.
[26] G. Wise, A. Traganitis, and J. Thomas, "The effect of a memoryless nonlinearity on the spectrum of a random process," *IEEE Trans. Inf. Theory*, vol. 23, no. 1, pp. 84–89, Jan. 1977.
[27] V. Krishnan, *Probability and Random Processes, ser. Wiley Survival Guides in Engineering and Science*. New York: Wiley-Interscience, 2006.
[28] S. Stein, The q Function and Related Integrals Applied Research Laboratory, Sylvania Electronic Systems, Res. Rep. 467, 1965.
[29] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Stat.*, vol. 23, pp. 493–507, 1952.
[30] WhichSSL, SSL Market Share, Tech. Rep. 979, 2010 [Online]. Available: http://www.whichssl.com/ssl-market-share.html
[31] M. Naor and K. Nissim, "Certificate revocation and certificate update," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 4, pp. 561–560, Apr. 2000.
[32] D. Cooper, "A model of certificate revocation," in *Proc. 15th Ann. Computer Security Applications Conf.*, 1999, pp. 256–264.
[33] A. Arnes, M. Just, S. J. Knapskog, S. Lloyd, and H. Meijer, "Selecting revocation solutions for PKI," in *Proc. NORDSEC '95*, 1995.
[34] M. Raya and J.-P. Hubaux, "The security of vehicular ad hoc networks," in *Proc. 3rd ACM Workshop on Security of ad hoc and sensor networks, ser. SASN '05*, 2005, pp. 11–21.
[35] G. F. Marias, K. Papapanagiotou, and P. Georgiadis, "ADOPT. A distributed OCSP for trust establishment in MANETs," in *Proc. 11th Eur. Wireless Conf. 2005—Next Generation Wireless and Mobile Communications and Services (European Wireless)*, Apr. 10–13, 2005, pp. 1–7 [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5755348&isnumber=5755255
[36] A. Wasef and X. Shen, "EDR: Efficient decentralized revocation protocol for vehicular Ad HOC networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 5214–5224, Nov. 2009.
[37] P. Papadimitratos, L. Buttyan, T. Holczer, E. Schoch, J. Freudiger, M. Raya, Z. Ma, F. Kargl, A. Kung, and J.-P. Hubaux, "Secure vehicular communication systems: Design and architecture," *IEEE Commun. Mag.*, vol. 46, no. 11, pp. 100–109, Nov. 2008.
[38] B. Fox and B. LaMacchia, "Certificate revocation: Mechanics and meaning," in *Proc. Int. Conf. Financial Cryptography (FC98)*, Feb. 1998, pp. 158–164.
[39] D. Cooper, "A more efficient use of delta-CRLs," in *Proc. 2000 IEEE Symposium on Security and Privacy. Computer Security Division of NIST*, 2000, pp. 190–202.
[40] M. Ofigsbø, S. Mjølsnes, P. Heegaard, and L. Nilsen, "Reducing the cost of certificate revocation: A case study," in *Proc. Public Key Infrastructures, Services and Applications*, 2010, vol. 6391, pp. 51–66, ser. Lecture Notes in Computer Science.
[41] R. Racine, "Estimating the hurst exponent," Master's thesis, ETH, Zurich, Apr. 2011.

Carlos Gañán was born in Barcelona, Spain, in 1984. He received the B.S. degree in electrical engineering and the M.S. degree in telematics from the Universitat Politècnica de Catalunya (UPC) in 2008 and 2009, respectively. In 2008, he joined the Information Security Group, with the Department of Telematics Engineering at UPC, Barcelona. He is currently pursuing the Ph.D. degree, carrying out research on security for vehicular communications. His academic interests span multimedia communications, network security, and vehicular ad-hoc networks.

Jorge Mata-Díaz received the M.S. degree in telecommunications engineering in 1991, and the Ph.D. degree in 1996, both from the Universitat Politècnica de Catalunya (UPC). The Spanish Association of Telecommunications Engineers rewarded him for his doctoral thesis with the Telefonica Award in Networks and Telecommunications Services. He is currently a research staff member of the Telematics Engineering Department. His research interests include network services, audiovisual appliance, streaming QoS, traffic modeling, and statistical performance analysis.

Jose L. Muñoz received the M.S. degree in telecommunication engineering from the Technical University of Catalonia (UPC) in 1999. In the same year, he joined the AUNA Switching Engineering Department. Since 2000, he works in the Department of Telematics Engineering of the UPC, currently as Associate Professor. In 2003, he received the Ph.D. degree in network security.

Juan Hernández-Serrano was born in Salamanca, Spain, in 1979. He received the M.S. degree in electrical engineering in 2002, and the Ph.D. degree in 2008, both from the Universitat Politècnica de Catalunya (UPC). In 2002, he joined the Information Security Group (ISG) within the Telematics Services Research Group at the Department of Telematics Engineering of the UPC. He currently works as assistant professor at the UPC. His research interests include security for large deployment of sensor networks, autonomous cognitive networks, smart grids, digital forensics, and e-voting.

Oscar Esparza received the M.S. degree in telecommunication engineering from the Technical University of Catalonia (UPC) in 1999. In the same year, he joined the AUNA Switching Engineering Department. Since 2001, he works in the Department of Telematics Engineering of the UPC, currently as Associate Professor. In 2004, he received the Ph.D. degree in mobile agent security and network security.

Juanjo Alins received the M.S. degree in telecommunications engineering in 1994, and the Ph.D. degree in 2004, both from the Polytechnic University of Catalonia, Spain. He is currently a research staff member with the Telematics Services Research Group in the Telematics Engineering Department. His research interests include network services to the home, audiovisual appliance, streaming QoS, secure multimedia transmission, traffic modeling, and statistical performance analysis. He works as a Professor in the Polytechnic University of Catalonia.