Knowl Inf Syst (2015) 43:311–331
DOI 10.1007/s10115-014-0735-1

REGULAR PAPER

A model for revocation forecasting in public-key infrastructures

**Carlos Gañán · Jorge Mata-Díaz · Jose L. Muñoz ·
Oscar Esparza · Juanjo Alins**

Received: 14 March 2013 / Revised: 27 December 2013 / Accepted: 1 February 2014 /
Published online: 14 February 2014

Abstract One of the hardest tasks of a certification infrastructure is to manage revocation. This process consists in collecting and making the revocation status of certificates available to users. Research on this topic has focused on the trade-offs that different revocation mechanisms offer. Much less effort has been conducted to understand and model real-world revocation processes. For this reason, in this paper, we present a novel analysis of real-world collected revocation data and we propose a revocation prediction model. The model uses an autoregressive integrated moving average model. Our prediction model enables certification authorities to forecast the number of revoked certificates in short term.

Keywords Certification · PKI · Revocation · CRL · ARIMA

1 Introduction

In a public-key infrastructure (PKI), digital certificates are the 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 they are issued with a planned lifetime, which is defined through a validity start time and an explicit expiration date. Once issued, a certificate becomes valid when its validity start time is reached, and it is considered valid until its expiration date. However, various circumstances may cause a certificate to become invalid prior to the expiration of the validity period. Such circumstances include requiring to change the name of the subject of a certificate, the compromise or suspected compromise of the private key associated with the certificate and a change in association between subject and CA, for example, when an employee terminates employment with an organization. Thus, the PKI has to collect and distribute information about revoked certificates.

C. Gañán (✉) · J. Mata-Díaz · J. L. Muñoz · O. Esparza · J. Alins
Departament Enginyeria Telemàtica, Universitat Politècnica de Catalunya,
1-3 Jordi Girona, C3, 08034 Barcelona, Spain
e-mail: carlos.ganan@entel.upc.edu

There are several mechanisms to manage revocation information. Currently, deployed PKIs rely mostly on certificate revocation lists (CRLs) for handling certificate revocation [1]. A CRL is a list identifying revoked certificates, which is signed by a CA and made available at a public distribution point. The CRL has a validity period, and updated versions of the CRL are published before the previous CRL's validity period expires. Although CRL is the most widely used mechanism to distribute certificate status information (CSI), there have been proposed other mechanisms to make revocation distribution more efficient (see Sect. 2.1 for further details).

The work presented in this paper is motivated by the fact that despite there are many works in the literature that propose and evaluate different mechanisms for distributing revocation data like [2–7], little work has been done for analyzing the revocation process itself. For instance, many of the previous studies consider very simplistic assumptions about the revocation process like that the percentage of revoked certificates remains always constant. Only recently, we can find works like [8–10] that carry out statistical studies about the revocation process using data available from real CAs. These studies can be considered a first step toward understanding the revocation process. However, they essentially just analyze the probability distribution of the revocation process. In [11], authors analyzed empirical revocation data to measure the degree of self-similarity. However, this analysis is not useful to predict the number of revoked certificates as the fractal pattern of the revocation process has a marginal impact on future revocation events. In this paper, we go a step further and we use time series analysis to build a model and a predictor for revocation. Using this novel way of exploring the revocation data, we are able to obtain a model that predicts extremely well the daily number of revoked certificates.

We build a revocation model which is exclusively based on temporal information, namely the revocation dates. Using this information and the Box–Jenkins methodology [12], we are able to build a model and a predictor for revocation. More precisely, we are able to show that an autoregressive integrated moving average (ARIMA) model provides objective and accurate predictions of the daily number of revoked certificates. To find the number of significant coefficients and their corresponding values we used Z-transforms and a vision of the problem based on digital filters. This greatly simplified the overall process of finding the parameters of the model. The resulting model exploits autocorrelations, captures the dependencies within the revocation data and simplifies these complex relationships to linear dependencies.

Regarding the data used to build the model, we have gathered CRLs from the main CAs, i.e., Comodo, GeoTrust, GoDaddy, Thawte, Verisign and Globalsign. Using these CAs, we build a time series that captures the daily evolution of the number of revoked certificates. Although the collected revocation data only belong to the leading CAs, the model is expected to be valid for other certification providers because of the huge amount of certificates analyzed and because our model captures the general pattern of the temporal correlation that exists among the revocation events. This pattern is expected to be the same for other CAs because the theoretical reasons to revoke certificates (and therefore the temporal correlation) do not essentially differ, no matter which certificate provider you use. In this way, we validate the accuracy of the model when using other CAs such as the US Social Security Administration and the Catalan Certification Agency. As a final remark, we would like to mention that our model might be interesting for several reasons: on one hand, it provides some insights about the revocation process, like showing that the process is highly autocorrelated, and on the other hand, it allows CAs to make predictions about certificate revocation. This might be useful in some real scenarios (see Sect. 7).

The rest of this paper is organized as follows. Section 2 briefly reviews the required background. Section 3 presents a preliminary analysis of the revocation codes. In Sect. 4, we discuss the methodology we used to collect and analyze the revocation data. In Sect. 5, we identify the best ARIMA model that fits the revocation events. Next, in Sect. 6, we present the predictor from the previous ARIMA model. In Sect. 7, we discuss some possible practical applications of the revocation forecasting model. Finally, we conclude in Sect. 8.

2 Background

2.1 Certificate status validation mechanisms

In this section, we briefly review the main approaches to convey CSI. The traditional approach is to periodically publish a CRL [13], which is a list of all revoked certificates within a domain. This list is signed by the CA itself, in order to let end entities verify its authenticity. The main problem with this approach is that the list can grow for large domains, and the network load involved in all clients downloading the list can become unacceptable. Two solutions were proposed to deal with the CRLs' problems [1]: Distribution Points and Delta-CRLs. Distribution Points provide the means to partition a CRL. Delta-CRLs are there to face the problem of using up too much of the available network resources when communicating the CRL either as a whole, or even in parts through Distribution Points.

Also, to deal with the communication overhead of the CRLs, Micali [14] proposed the certificate revocation status (CRS). In CRS, a CA signs a fresh list of all not-yet-expired certificates together with selected hash chain values. The hash chain values can be used to verify whether the queried certificate is valid or not for a certain time interval. The main advantage of this mechanism is that it significantly reduces the communication costs between the CSI repository and the dependent entity, by employing a mechanism for the CSI dissemination, which contains positive statements regarding the status of a certificate. Nevertheless, the main disadvantage of this system is the increase in the CA's communication cost with the CSI repositories [15].

A different also standardized approach is to provide an on-line server and use protocols for obtaining on-line revocation information. In this case, a client issues a request for every encountered certificate instead of obtaining a full revocation list. Hence, the online certificate status protocol (OCSP) [16] allows end entities to query for CSI in a more timely fashion than CRLs. OCSP can be used to provide timely CSI, and it could be used in conjunction with CRLs. However, in case the transport protocol is not authenticated, the authority that provides the OCSP service is vulnerable against replay attacks, where someone could replay OCSP responses before their expiration date but after a certificate has been revoked.

Kocher [17] suggested another CSI mechanism, the Certificate Revocation Tree (CRT). A CRT is based on a Merkle hash tree [18] containing certificate serial number ranges as the tree leaves. The root of the hash tree is signed by the CA. Now, the certificate status proof for a certificate with serial number s consists of the path node siblings from the root to the appropriate leaf (having s in its range), in addition to the signature on the root of the tree. Thus, if n certificates are currently revoked, the length of the proof is $O(\log n)$. In contrast, the length of the validity proof in OCSP is $O(1)$.

2.2 Autoregressive Integrated Moving Average (ARIMA) processes

Prediction of scalar time series refers to the task of finding estimate of next future sample $\hat{s}(n)$ based on the knowledge of the history of time series, i.e., samples $s(n-1)$, $s(n-2)$, etc. Many time series can be suitably forecast using linear techniques as the ARIMA model popularized by Box and Jenkins [12]. The ARIMA approach to forecasting is based on the following ideas: the forecasts are based on linear functions of the sample observations, and the aim is to find the simplest models that provide an adequate description of the observed data. Each ARIMA process has three parts:

- *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 $p > 1$:
$$y(n) = a_0 + a_1 y(n-1) + \cdots + a_p y(n-p) + w(n). \tag{1}$$
- *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. 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:
$$s(n) = c_1 s(n-1) + \cdots + c_d s(n-d) + w(n), \tag{2}$$
where:
$$c_i = \binom{d}{i}(-1)^{i+1} \quad i \in \{1, 2, \ldots, d\}. \tag{3}$$
- *Moving Average*: This part of the model describes how each observation is a function of the previous errors $w(n)$. For example, if $q = 1$, then each observation is a function of only one previous error. In general, if we consider up to q errors, then:
$$x(n) = b_0 w(n) + b_1 w(n-1) + \cdots + b_q w(n-q), \tag{4}$$
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 $n-q$.

3 Preliminary analysis

Our preliminary analysis is focused on the study of the revocation codes. Our goal is to find whether it is possible, using this parameter, to get some insights about how the revocation process works in the real world. The real-world data that we consider are from the main CAs: Comodo, GeoTrust, GoDaddy, Thawte, Verisign and Globalsign.

The PKIX/X.509 certificate and CRL specification [13] define nine reason codes for revocation of a public-key certificate: (1) `keyCompromise`, (2) `cACompromise`, (3) `affiliationChanged`, (4) `superseded`, (5) `cessationOfOperation`, (6) `certificateHold`, (7) `removeFromCRL`, (8) `privilegeWithdrawn`, (9) `aACompromise`.

Reason codes can be included as noncritical extensions within the CSI, for instance, in an extension of a CRL. As mentioned, the standard [13] defines the possible revocation reasons and how to include them within the status data but it does not define which should be the

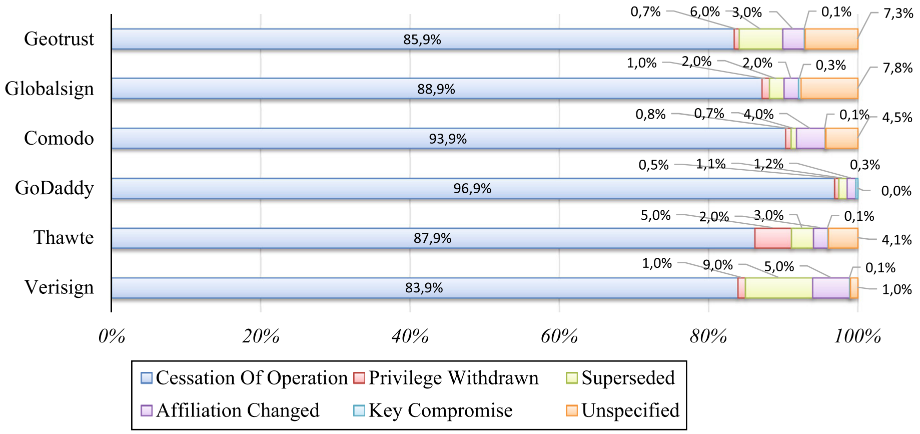

Fig. 1 Revocation causes of SSL certificates

revocation practice for each code. To illustrate this, let us consider the case of SSL certificates. We analyze these codes analyzing different CRLs issued by each CA. The main conclusion is that these CAs almost always use the same code. In particular, they use the reason number (5) `cessationOfOperation` like a kind of "default" reason for most of the revoked certificates. This is shown in Fig. 1. No matter the CA the vast majority of certificates are revoked due to `cessationOfOperation`.

Thus, the revocation causes available from the real world do not provide information which we can use to build a rigorous revocation model to predict when or how many revocations are prone to happen. CAs from the real world do not follow any clear guideline about how to use this optional parameter. Taking these facts into consideration, in the rest of the paper we build a revocation model which is exclusively based on temporal information, namely revocation dates of the certificates. These parameters are well defined and compulsory, and they will allow us to build a model that will show that certificate revocations are closely related to time.

4 Data collection and preprocessing

To analyze the time evolution of the certificate revocation process, we need information about the revoked certificates. To obtain as much information as possible, we gathered a large sample of revoked certificates. According to W^{3}Tech's survey [19], Comodo, GeoTrust, GoDaddy, Thawte, Verisign and Globalsign cover most of the world market for SSL.[1] Therefore, though the data collected belongs exclusively to the main CAs, the model is expected to be useful for any other CA as the data covers most of the global marketplace.

We analyzed all the CRLs issued from 2008 to 2012 for 6 CAs regarding SSL certificates. Even though each CA has different CRL release policies, all of them issue a new CRL every day. This CRL over-issuing policy entails that some consecutive CRL are almost identical. Thus, a total of 10,956 CRLs were downloaded. Hence, we have all the CRLs issued during the period 2008–2012, and we tail all revoked certificates during this period. Thus, we are

[1] Note that this survey mainly operates a crawler finding, and reporting on the SSL certificates it locates in the wild. Hence, we only use these data to corroborate that we are covering most of the SSL market.

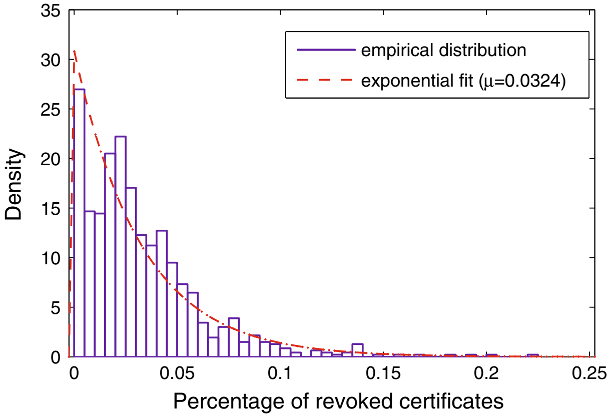

Fig. 2 Empirical data versus fitted exponential PDF (Verisign)

Table 1 Number of collected revoked certificates per CA

CA	Comodo	Geotrust	GoDaddy	Thawte	Verisign	Globalsign
#revoked certificates	14,721	17,577	1,076,479	35,976	13,779	9,480

able to capture all the revocation events that occurred during this period including those certificates that expired after being revoked.

To check the correctness of the process, we triangulate the revocation data issued by Verisign with the results obtained in [8]. Note that we only triangulate these data because it is the only CA analyzed in previous works. Following the same methodology that in [8], we use an exponential probability density function to model the distribution of the percentage of revoked certificates, i.e., $f(x) = \frac{1}{\mu}e^{-\frac{x}{\mu}}$. The dataset was discovered to follow the exponential distribution with the parameter μ being 0.0324 at a 95 % confidence interval (see Fig. 2). Walleck et al. obtained a value of $\mu = 0.0479$, which corroborates the correctness of the data collection process. Note that these results also agree with the analysis carried out by Ma et al. [9] and Hu et al. [10] who concluded that the empirical PDF followed an exponential distribution.

Once we had downloaded the revocation data, we had to preprocess these data to remove duplicates. Duplicates appear not only in different CRLs but also in the same CRL. Firstly, we parsed all the CRLs issued by the same CA to obtain a dataset containing the serial number of the revoked certificates and the date when they were revoked. Only an entry per serial number was kept. It is worth noting that Thawte's and GlobalSign's CRLs 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 had to remove any duplicate entry from the composite dataset and tally the number of revocations per day. Finally, we build a dataset that covers revoked certificates from January 2008 to December 2012. Note that in order to include expired revoked certificates, we collected any issued CRL in this period of time. A summary about this dataset per CA is shown in Table 1.

Figure 3 shows the number of certificates revoked for the period of 2008/01/01 through 2012/12/31 for each CA. Analyzing all the collected data, we can conclude that:

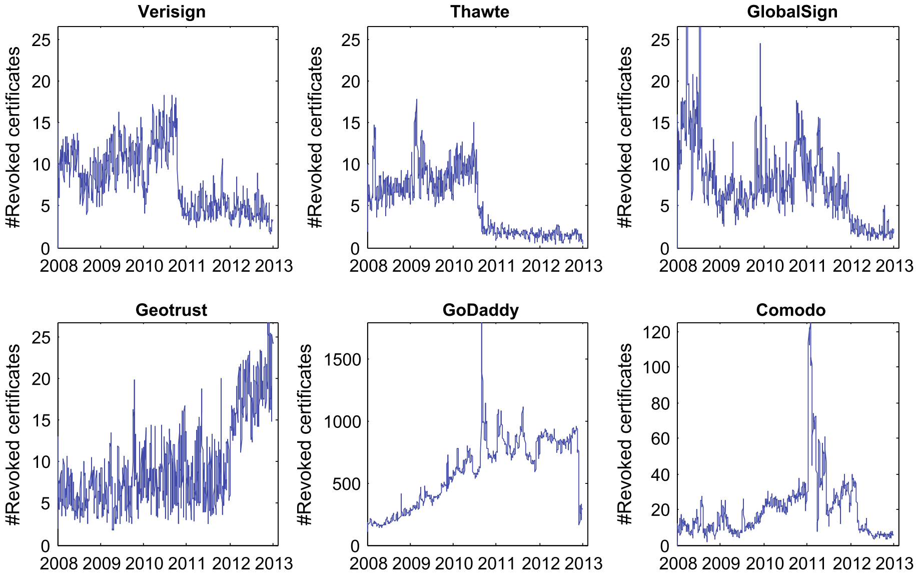

Fig. 3 Number of revoked certificates evolution for each CA

- 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 variable-length bursts. However, the mean number of daily revoked certificates highly varies depending on the CA.
- GoDaddy exhibits a significant increment in the number of certificates revoked per day from 2008 to 2011. This increase might be mainly due to the changes in the total number of certificates being issued at different years.
- Thawte suffered from a sudden decrease in the number of daily revoked certificate at the end of 2010. This decrease could be explained by Symantec's acquisition of Thawte's security business in 2010.
- Note that all the CAs that belong to the Symantec Group (i.e., Thawte, Verisign and GeoTrust) handle no more than 20 revoked certificates per day, while Comodo or GoDaddy manage more than 10 and 100 times more revoked certificates per day, respectively.

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 six time series (see Fig. 4). Time series concatenation is a methodology that has been proven to be effective [20–22]. Thus, we reduce model complexity by combining similar types of series to form a single input covariate series. By modeling the concatenation of all six 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 of revoked certificates that a CA manages highly varies. To analyze the global trend of the revocation process, we need to remove the influence of the volume of revoked 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.

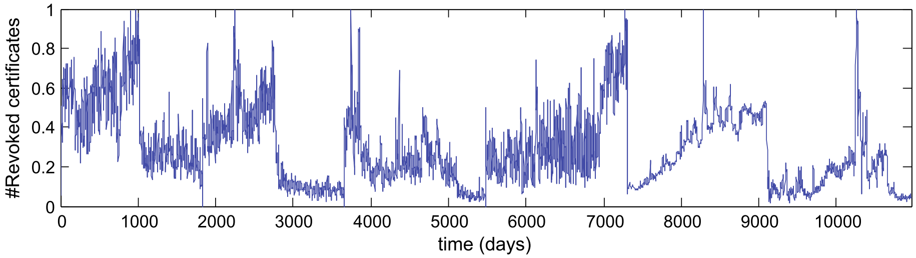

Fig. 4 Global time series

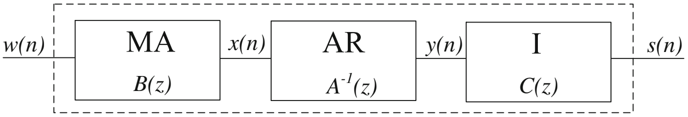

Fig. 5 Components of an ARIMA process

5 Modeling the revocation

In this section, we develop a new model for the number of revoked certificates. To do so, we use an ARIMA model in the Z-domain. We characterize the ARIMA model that best fits the revocation process. That is, we calculate the coefficients of the different ARIMA components to fit the collected revocation data. As we will show, these coefficients will allow to build a suitable predictor to forecast revocations.

5.1 ARIMA processes in the Z-domain

In Sect. 2.2, we have described ARIMA processes in the time domain. However, as any linear system, an ARIMA 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 the equation, we can compute what is called the transfer function of the system.

We use the delay operator z^{-1} [23] to Z-transform the time-domain expression of an $ARIMA(p, d, q)$ process. To that end, we Z-transform the autoregressive component, moving average component and the integrated component. In Fig. 5, a scheme of the ARIMA model is shown. Note that we can express the transfer function of the $ARIMA(p, d, q)$ process as a cascade of all three components.

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

$$B(z) = b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_q z^{-q}. \tag{5}$$

Note that it only uses previous samples of the input signal. The associated generating system is linear time invariant, causal and stable. The MA system is finite impulse response (FIR) and, therefore, an all-zero system. Figure 6 represents $MA(q)$ as a FIR filter with transfer function $B(z)$.

An $AR(p)$ stochastic process 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 of a time series depends linearly on previous samples plus some random error. The transfer function

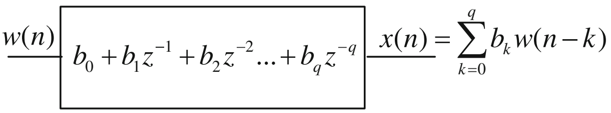
$w(n)$
$b_0 + b_1 z^{-1} + b_2 z^{-2} \ldots + b_q z^{-q}$
$x(n) = \sum_{k=0}^{q} b_k w(n-k)$

Fig. 6 MA filter

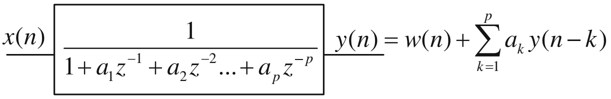
$x(n)$
$\frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} \ldots + a_p z^{-p}}$
$y(n) = w(n) + \sum_{k=1}^{p} a_k y(n-k)$

Fig. 7 AR filter

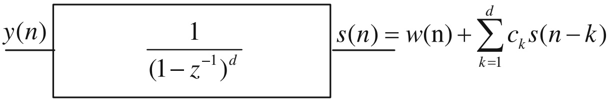
$y(n)$
$\frac{1}{(1 - z^{-1})^d}$
$s(n) = w(\mathrm{n}) + \sum_{k=1}^{d} c_k s(n-k)$

Fig. 8 Integrated filter

of $AR(p)$ process in the z-domain is:

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p}. \tag{6}$$

It is worth noting that the AR process determines the poles of a linear time invariant system. As long as all the poles lie strictly inside the unit circle, the region of convergence will include the unit circle and the system will be stable. A discrete-time LTI system is stable if and only if $H(z)$ has all its poles strictly inside the unit disk $|z| < 1$. The impulse response of the associated system is IIR, and its transfer function is of the kind all-pole. Note that the module of all its poles must be strictly smaller than 1 so that the autocorrelation is limited and tends to 0 when the lag tends to infinity. That means that if this condition is met, then the $AR(p)$ process is ergodic. Figure 7 represents the $AR(q)$ as an IIR filter whose transfer function is $\frac{1}{A(z)}$.

We can also express integrated component in the z-domain from Eq. (2):

$$C(z) = (1 - z^{-1})^{-d}. \tag{7}$$

In the same way, as the autoregressive and moving average components of the ARIMA process, we can represent the integrated component as a linear filter. Figure 8 represents the $I(d)$ as a linear filter whose transfer function is $C(z)$.

Finally, the general expression of an $ARIMA(p, d, q)$ process is:

$$S(z) = [B(z)A^{-1}(z)C(z)] \cdot W(z). \tag{8}$$

Understanding expression (8) 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:

$$H(z) = \frac{S(z)}{W(z)} = B(z)A^{-1}(z)C(z). \tag{9}$$

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 for linear systems. Figure 9 represents the ARIMA filter with transfer function $H(z)$:

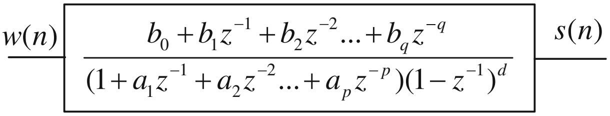
$w(n)$
$\frac{b_0 + b_1 z^{-1} + b_2 z^{-2} \ldots + b_q z^{-q}}{(1 + a_1 z^{-1} + a_2 z^{-2} \ldots + a_p z^{-p})(1 - z^{-1})^d}$
$s(n)$

Fig. 9 ARIMA filter

Note 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 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 impulse response. The rest of obtained poles (z_k) will be found in the unit circle ($|z_k| < 1$) of the Z plane.

5.2 Characterization of the revocations as an ARIMA process

The steps we follow to model the revocation process can be summarized as follows:

1. Testing for stationarity of the time series.
2. If the series is not stationary, transforming it to a stationary series:
   (a) Estimating the degree of differencing.
   (b) Differencing the time series to obtain a stationary series.
3. Identifying the ARMA components.
   (a) Estimating the order of the AR component and the MA component.
   (b) Extracting the ARMA parameters.
4. Model fit validation using residual diagnostics

Knowing the ARIMA coefficients, any CA could use the ARIMA model to predict which weeks are more prone to suffer from revocation in the near future.

5.2.1 Model identification

The first step in developing the ARIMA model is to determine whether the series is stationary and whether there is any significant seasonality that needs to be modeled. In the following, we show that the global revocation time series is nonstationary and does not present a seasonal pattern. The aim of this first step is to calculate the order of differencing d to achieve stationarity. Once we have obtained a stationary time series, we will be able to model it as an ARMA process.

First of all, we start testing for stationarity. To test stationarity, first we analyzed the autocorrelation function (ACF) of the number of revoked certificates (see Fig. 10). We can observe that the actual time series follows certain trend, so the time series is nonstationary. The temporal series of the number of revoked certificates presents a slow variation in the mean. To confirm this visual evaluation, we perform a KPSS test [24] at the 99 % confidence level, which rejects the null hypothesis of stationarity.

Once we have confirmed that the time series is nonstationary, we have to find the integrated component. The long-range dependence complicates the development of a predictor because the temporal series shows an apparent nonstationary mean. To synthesize a good predictor, it is necessary to capture this long-term effect. The long-term dependence produces a mean that varies. This variation reaches maximum and minimum levels, which are very distant.

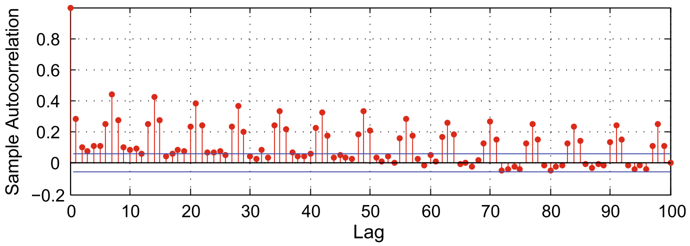

Fig. 10 ACF of the number of revoked certificates

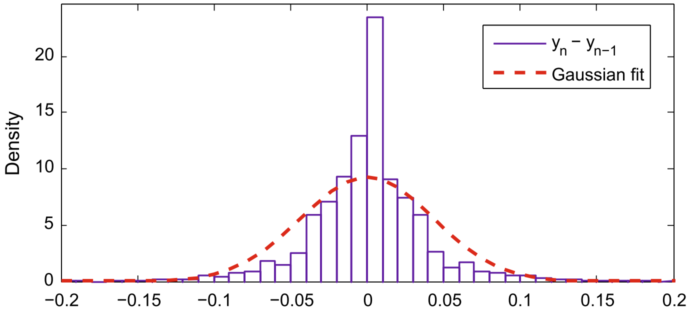

Fig. 11 PDF of the extracted integrated component time series

However, the variance remains almost constant. Hence, the integrated component of the model should be of order 1 and its associated transfer function $C(z) = (1 - z^{-1})^{-1}$.

Next, it is necessary to extract the integrated component of the actual process $s(n)$ to determine which are the values of the AR and MA components. According to the scheme presented in Fig. 5, the residual ARMA series $y(n)$ and the real series $s(n)$ are related as follows:

$$Y(z) = \frac{1}{C(z)} S(z) = \left(1 - z^{-1}\right) S(z). \tag{10}$$

The temporal series $y(n)$ will be obtained at the output of the FIR filter, whose transfer function is $(1-z^{-1})$, when it is excited with the concatenated time series. It can be checked that the temporal series $y(n)$ is a stochastic process with mean 0 and an invariant autocorrelation coefficients. The probability distribution function fits a Gaussian distribution (see Fig. 11).

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 100 lags) that $y(n)$ is not white noise by means of Ljung–Box Q-test [12]. The fact that $y(n)$ is not white noise allows us to model it by means of an ARMA process.

5.2.2 Estimation of the ARMA coefficients

In this section, we determine the best ARMA model that fits $y(n)$. To that end, first we estimate the number of coefficients needed to capture the autoregressive component of the process. Then, we calculate the value of these coefficients. Once we have modeled the AR(p) component, we estimate the number of coefficients that model the moving average component of the process. Finally, we calculate the values of the MA coefficients.

It must be noted that we use the Bayesian Information Criterion (BIC) [12] for model selection among the different set of ARMA models with different numbers of parameters. That information criterion penalizes models with additional parameters. Therefore, the BIC model order selection criteria are based on parsimony. Along with the BIC criterion, we also try to minimize the correlation of the residuals. Using these criteria, the order of the AR process found is 10. Once the order the AR process has been determined, we use least squares estimation to calculate the coefficients of the AR component.

$$A(z) = 1 + 1.631z^{-1} + 0.9213z^{-2} + 0.7102z^{-3} + 0.3692z^{-4} - 0.7699z^{-5} - 0.9442z^{-6} \\ -0.5766z^{-7} - 0.3447z^{-8} + 0.01869z^{-9} + 0.1073z^{-10}. \tag{11}$$

The MA component can be analyzed when the AR component of the $y(n)$ series is withdrawn. Applying the above explained technique for the integrative component the $x(n)$ series can be derived using the relation between the $y(n)$ and $x(n)$ series:

$$X(z) = A^{-1}(z)Y(z) = (1+1.631z^{-1}+0.9213z^{-2}+0.7102z^{-3}+0.3692z^{-4} - 0.7699z^{-5} \\ -0.9442z^{-6}-0.5766z^{-7}-0.3447z^{-8}+0.01869z^{-9}+0.1073z^{-10})^{-1}Y(z). \tag{12}$$

Using a FIR filter with transfer function $A^{-1}(z)$, the series $x(n)$ can be obtained at the output of this filter when $y(n)$ is applied at the input. To estimate the parameters of the MA process, least square estimation is applied to fit the partial autocovariance function of $x(n)$ [12]. We use the same criteria as in the AR process for selecting the parameters of the MA process. The best adjustment is obtained with a MA process with order 6.

$$B(z) = 1 - 0.3549z^{-1} + 0.6133z^{-2} + 0.0976z^{-3} - 0.7463z^{-4} + 0.190z^{-5} - 0.7706z^{-6}. \tag{13}$$

As the integrated component of the obtained model has order 1, the integrated and the autoregressive components can be written together in the following way:

$$A'(z) = A(z)C^{-1}(z) = A(z)\left(1 - z^{-1}\right). \tag{14}$$

Therefore, the generated series can be expressed as:

$$s(n) = b_0 w(n) + \cdots + b_q w(n-q) + a'_1 s(n-1) + \cdots + a'_{p+1} s(n-p-1), \tag{15}$$

where the coefficients are obtained applying the inverse Z transform to $A'(z)$:

$$A'(z) = 1 + 0.631z^{-1} - 0.7097z^{-2} - 0.2111z^{-3} - 0.341z^{-4} - 1.1391z^{-5} - 0.1743z^{-6} \\ +0.3676z^{-7} + 0.2319z^{-8} + 0.3634z^{-9} + 0.08861z^{-10} - 0.1073z^{-11}. \tag{16}$$

5.2.3 Model diagnostic checking

Finally, to check that the model is not misspecified, 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).

6 Revocation forecasting

Once we have developed an ARIMA model for the time series of the daily number of revoked certificates, in this section we use this model to build an ARIMA predictor. Then, we check the accuracy of the predictions for the global revocation time series and also for each CA.

We start by obtaining an expression of the residual series to check whether it follows a Gaussian distribution or not. From (15), the $(n+1)$ sample prediction is:

$$\hat{s}(n+1) = b_0\hat{w}(n+1) + \cdots + b_q w(n-q+1) + a'_1 s(n) + \cdots + a'_{p+1} s(n-p). \quad (17)$$

Nevertheless, in the prediction context, the values of the $w(n)$ series must be figured out. The predictor will have only the previous values of the $s(n)$ series. Moreover, the $\hat{w}(n+1)$ is a future value. Using an average forecasting procedure, the forecast value of $\hat{w}(n+1)$ is calculated as the mean value of the $w(n)$ series. In this case, this mean value is 0 because the $w(n)$ series must follow a normal distribution. Thus, the $(n+1)$ sample prediction can be simplified as:

$$\hat{s}(n+1) = b_1 w(n) + \cdots + b_q w(n-q+1) + a'_1 s(n) + \cdots + a'_{p+1} s(n-p). \quad (18)$$

To determine $w(n)$ as a function of $s(n)$, we conduct some algebraic manipulations. From (15), it is also possible to write:

$$\hat{s}(n) = b_0\hat{w}(n-1) + \cdots + b_q w(n-q) + a'_1 s(n-1) + \cdots + a'_{p+1} s(n-p-1). \quad (19)$$

Subtracting (19) from (15):

$$s(n) - \hat{s}(n) = b_0\left(w(n) - \hat{w}(n)\right). \quad (20)$$

As it has been mentioned, the forecast value of $\hat{w}(n)$ will be 0, so:

$$s(n) - \hat{s}(n) = b_0 w(n). \quad (21)$$

Therefore,

$$w(n) = \frac{s(n) - \hat{s}(n)}{b_0}. \quad (22)$$

Thus, using expression (22), we can calculate the errors of our forecasts. Next, we analyze the statistical properties of these errors. As expected, these errors have the statistical properties of white noise, i.e., the residual diagnostic determines that the forecast errors are clearly uncorrelated. This means that we have predicted all the predictable, and thus, that our predictions are the most accurate possible.

Once we have seen that the residuals are uncorrelated, we obtain an expression for the "1 ahead" prediction. Replacing this expression in the Eq. (18), the $(n+1)$ sample prediction can be written as:

$$\begin{aligned}\hat{s}(n+1) = {} & \frac{1}{b_0}\left[b_1(s(n) - \hat{s}(n)) + b_2\left(s(n-1) - \hat{s}(n-1)\right)\right.\\ & \left.+ \cdots + b_q\left(s(n-q+1) - \hat{s}(n-q+1)\right)\right]\\ & + a'_1 s(n) + \cdots + a'_{p+1} s(n-p). \quad (23)\end{aligned}$$

From this expression, we can plot the derived ARIMA predictor using the coefficients obtained in the previous section (see Fig. 12; Table 2). The predictor supplies the estimated value for the $(n+1)$ sample as a function of the n previous ones. This set of samples can be used also to obtain a prediction of the samples "2 ahead", "3 ahead", etc. Running the predictor with the $(n+j)$ sample estimation, it supplies an estimated value of the $(n+j+1)$ sample.

Finally, to check that the model is not misspecified, 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).

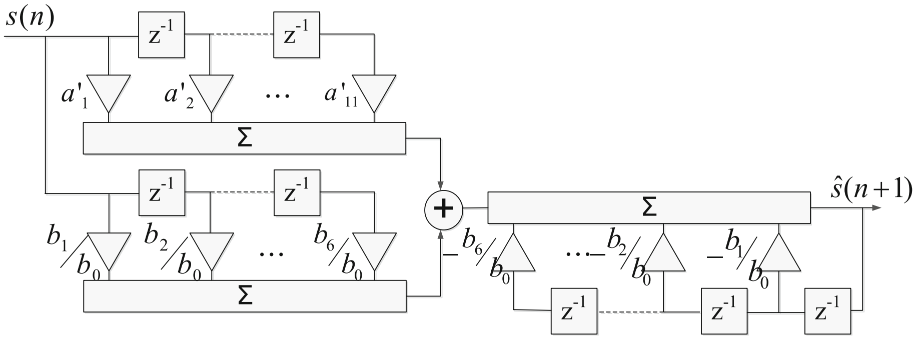

Fig. 12 ARIMA predictor

Table 2 Coefficients of the ARIMA predictor

a'_i	–	0.631	−0.7097	−0.2111	−0.341	−1.1391	−0.1743	0.3676	0.2319	0.3634	0.08861	−0.1073
b_i	1	−0.3549	0.6133	0.0976	−0.7463	0.190	−0.7706	–	–	–	–	–

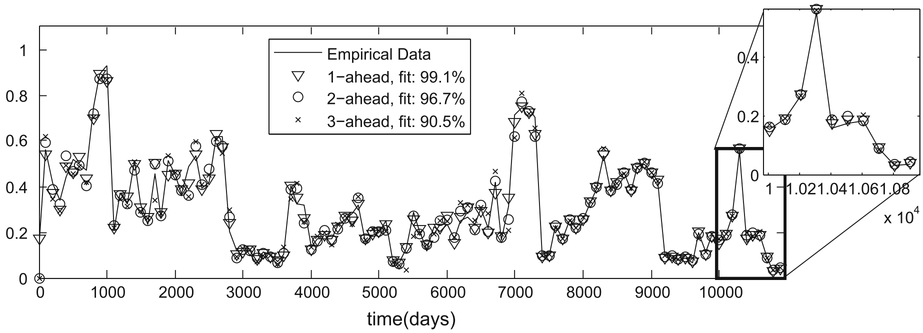

Fig. 13 Predictions for the global time series

Therefore, using the predictor diagram depicted in Fig. 12 where the value of each coefficient is stated in Table 2, we can obtain a prediction of the samples “1 ahead”, “2 ahead” and “3 ahead” for the global time series (see Fig. 13). It is evidenced from the figure that one-step ahead out-sample forecasts follow the actual revocation data more closely than *k*-step ahead out-sample forecasts. As expected, the *k*-step ahead out-sample forecasts accumulate the error terms resulting in low accuracy in forecasting performances. Note that for the one-step ahead prediction, the accuracy of the model is above 98 %. Thus, using the proposed ARIMA predictor, users should be able to forecast the number of revoked certificates for the next day with more than 98 % accuracy.

Once we have seen that the model fits quite accurately the global revocation process, we must check its suitability for each individual CA. For this purpose, we analyze the residuals for individual revocation processes using the ARIMA model obtained from the global series. We conclude that no matter the CA, there is no residual that exceeds the confidence intervals. Therefore, the ARIMA model developed from the concatenated time series captures the revocation pattern of all the selected CAs.

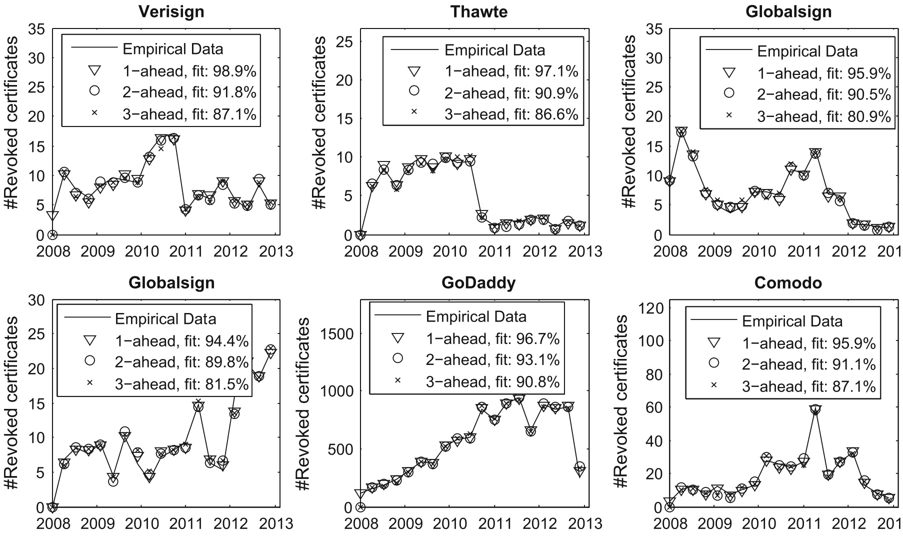

Fig. 14 Predictions for the daily number of revoked certificates per CA

Finally, in Fig. 14, the “1 ahead”, “2 ahead” and “3 ahead” predictions are presented for the number of revoked certificates per CA. Again, the predictions are valid for all the CAs, corroborating the results obtained from the residual analysis. It is worth noting that the accuracy of the predictor depends not only on the prediction horizon but also on the CA. However, no matter the CA, the proposed predictor has an accuracy for the 1-ahead prediction above the 94 %. As in the case of the global time series, as the prediction horizon increases, the accuracy decreases. This is because the k-step ahead forecasts accumulate the error terms resulting in low accuracy in forecasting performances. Makridakis and Hibon [25] already proved that the accuracy of various forecasting methods depends upon the length of the forecasting horizons involved. Meade [26] confirmed that forecasting accuracy was less accurate for longer horizons. The results from the current study as shown in Fig. 14 support Makridakis and Hibon [25] and Meade’s [26] findings.

6.1 Forecasting in other domains

Previously, we have shown the validity of the predictor for the case of certificates for SSL servers. Henceforward, we analyze the accuracy of the predictions in the case of other type of certificates. Particularly, we test three different cases:

– Digital certificates issued by the Social Security Administration Certification Authority (SSACA) [27] to enable secure communications between the Social Security Administration and customers transacting business and for identity proofing of individuals.
– CatCert’s free email certificates [28] that allow digitally signing and encrypting emails to secure and authenticate them.
– Verisign’s code signing certificates which are used by developers on all platforms to digitally sign the applications and software they distribute over the Internet.

We gather the corresponding CRL from the CA’s repositories and perform the same procedure as in Sect. 4. After that, we obtain the time series corresponding to the number of revoked certificates per day. Using these time series as input of the predictor, we obtain the “1

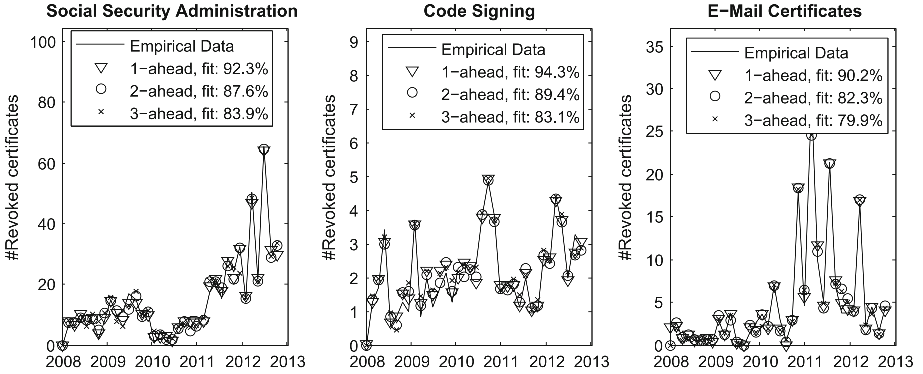

Fig. 15 Predictions for the daily number of revoked certificates

ahead", "2 ahead" and "3 ahead" predictions. Figure 15 plots the corresponding time series and the predictions. Again, the accuracy of the "1 ahead" prediction is approximately 90 % for all 3 cases. Note that this accuracy is achieved although we have not changed the ARMA coefficients, i.e., the predictor remains accurate and is able to capture the correlation among revocation events.

7 Some applications of revocation forecasting

The first quite straightforward application of revocation forecasting is to use it to set the validity period of the CRLs. More precisely, the validity period of the CRLs could be set as a function of the predicted number of revoked certificates. Currently, CAs release CRLs at a fixed time interval because consumers may not need the most current CRL. However, this is not an optimal strategy. Authors in [9,10] state that CAs must understand the revocation process for setting an optimal CRL releasing policy. On the one hand, CA should take different strategies for publishing CRLs when dealing with a new type of certificates versus a re-serving type of certificates. On the other hand, to derive an optimal releasing interval prescription, the CA must balance the trade-off between cost and risk. Our predictor could directly help in the estimation of that risk. As stated in [29], this risk is directly proportional to the number of revoked certificates. Thus, by forecasting the number of revoked certificates, CAs could estimate the risk of their CRLs release policy and optimize it. In this way, the CA could issue CRLs with short validity periods if the predictor forecasts many revocations, and viceversa. In a wired network, this might seem a subtle enhancement since there is always the possibility of issuing CRLs with small validity periods. As in wired networks, bandwidth is not scarce and connections are stable; the risk of operating with a revoked certificate can be made fairly small by frequently issuing CRLs. However, this is not the case of some new communication paradigms like *Delay* and *Disruption Tolerant Networks* (DTN). In a DTN, applications must opportunistically exploit connectivity over intermittent links [30]. Regarding security, one of the main challenges in DTNs is how to create and distribute keys and credentials. To this respect, many authors [31,32] and the current security draft specification for DTN [33] agree that the most promising solution is to use public-key cryptography with digital certificates.

An example of this paradigm are vehicular network (VANET) communications in which vehicles might not be always connected to the infrastructure (Internet). In this case, PKI users (vehicles) might not have the latest CRL available [4,34–36]. The question is whether to operate or not with a certificate that might be revoked, considering that the only information that we have is an obsolete CRL. Obviously, if the CRL is old, the risk is higher than if it is recent. The problem is to distinguish between what is old and what is recent. For instance, let us consider that a PKI user has a copy of a CRL, which was issued a couple of hours ago. Two hours without new revocation data will become critical for high revocation rates, e.g., if the revocation rate is 2 cert/min, after these two hours there will be 240 unknown revoked certificates. Here, the forecasting mechanism could be used to properly set the time-stamps (validity periods) of the CRLs so that they provide the user with an idea about how revocation process is behaving, and thus, how risky is operating with other users' certificates. Furthermore, if a more precise criterion is desired, the CA could include the parameters of the forecasting model inside the CRL in one of the so-called extension fields. In this way, PKI users might use predictions to evaluate the risk of operating with other user's certificates when connection to the infrastructure is not available.

Another type of scenario in which revocation forecasting can be applied is dynamic delegation [37–39]. Dynamic delegation is devised for highly distributed scenarios such as Web Services [40]. In these scenarios, users delegate certain credentials or attributes to perform a certain task by issuing certificates. In this context, some authors propose to use short-lived certificates avoiding the need of a revocation system.

However, if we use a short-lived certificate to perform some tasks and the validity period of the certificate is lower than the one required by the task, we will need to contact the certification authority to get a new certificate to finish the task. This is a problem mainly in long-term jobs [41]. For this reason, the GT4 group of the Globus Consortium considers that a good option for these long-term jobs is to use long-lived certificates and a revocation mechanism [42]. In this scenario, a revocation forecasting model could be useful to set the validity period of the certificates and/or its associated CRLs.

8 Conclusions

We have analyzed real empirical data to derive a model, which allows to forecast the number of daily revocations in the near future. Our research represents a step toward linking empirical observations to mathematical models in description of the complex issue of certificate revocation.

This paper has proposed an ARIMA model for short-range forecasting in a CSI distribution system. The ARIMA model completely considers the dynamic process of data series and the autocorrelation of residuals to achieve precise forecasting of the number of revoked certificates. We used the Box-Jenkins methodology as a framework for our modeling procedure, and we built the best ARIMA model possible for the available data. The model exhibits great accuracy for short time scales.

Although the collected revocation data only belong to the leading CAs, the model is expected to be valid for other certification providers because of the huge amount of certificates analyzed and because our model captures the general pattern of the temporal correlation that exists among the revocation events. This pattern is expected to be the same for other CAs because the theoretical reasons to revoke certificates (and therefore the temporal correlation) do not essentially differ, no matter which certificate provider you use.

Acknowledgments This work is funded 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.

References

1. Housley R, Polk W, Ford W, Solo D (2002) Internet X.509 public key infrastructure certificate and certificate revocation list (CRL) profile. RFC 3280, Internet Engineering Task Force
2. Narasimha M, Tsudik G (2007) Privacy-preserving revocation checking with modified crls. In: Lopez J, Samarati P, Ferrer J (eds) Public key infrastructure, vol 4582, lecture notes in computer science. Springer, Berlin, pp 18–33
3. Lippert M, Karatsiolis V, Wiesmaier A, Buchmann J (2006) Life-cycle management of x.509 certificates based on ldap directories. J Comput Secur 14:419–439
4. Gañán C, Muñoz JL, Esparza O, Mata-Díaz J, Hernández-Serrano J, Alins J (2013) COACH: cOllaborative certificate stAtus CHecking mechanism for VANETs. J Netw Comput Appl 36(5):1337–1357
5. Tsang PP, Au MH, Kapadia A, Smith SW (2010) Blac: revoking repeatedly misbehaving anonymous users without relying on TTPs. ACM Trans Inf Syst Secur 13:39:1–39:33
6. Solis J, Tsudik G (2006) Simple and flexible revocation checking with privacy. In: Danezis G, Golle P (eds) Privacy enhancing technologies, vol 4258, lecture notes in computer science. Springer, Berlin, pp 351–367
7. Caubet J, Gañán C, Esparza O, Muñoz JL, Mata-Díaz J, Alins J (2014) Certificate revocation list distribution system for the KAD network. Comput J 57(2):273–280
8. Walleck D, Li Y, Xu S (2008) Empirical analysis of certificate revocation lists. In: Proceedings of the 22nd annual IFIP WG 11.3 working conference on data and applications security, pp 159–174
9. Ma C, Hu N, Li Y (2006) On the release of CRLs in public key infrastructure. In: Proceedings of the 15th conference on USENIX security symposium, vol 15. Berkeley, CA, USA
10. Hu N, Tayi GK, Ma C, Li Y (2009) Certificate revocation release policies. J Comput Secur 17:127–157
11. Gañán C, Mata-Diaz J, Munoz JL, Hernandez-Serrano J, Esparza O, Alins J (2012) A modeling of certificate revocation and its application to synthesis of revocation traces. IEEE Trans Inf Forensics Secur 7(6):1673–1686
12. Box GEP, Jenkins G (1990) Time series analysis: forecasting and control. Holden-Day, Incorporated
13. Cooper D, Santesson S, Farrell S, Boeyen S, Housley R, Polk W (2008) Internet X.509 public key infrastructure certificate and certificate revocation list (CRL) profile. RFC 5280, Internet Engineering Task Force
14. Micali S (1996) Efficient certificate revocation. Technical Report TM-542b. MIT Laboratory for Computer Science
15. Naor M, Nissim K (2000) Certificate revocation and certificate update. IEEE J Sel Areas Commun 18(4):561–560
16. Myers M, Ankney R, Malpani A, Galperin S, Adams C (1999) X.509 internet public key infrastructure online certificate status protocol—OCSP. RFC 2560
17. Kocher PC (1998) On certificate revocation and validation. In: International conference on financial cryptography (FC98), number 1465 in lecture notes in computer science, pp 172–177
18. Merkle RC (1989), A certified digital signature. In: Advances in cryptology (CRYPTO89), number 435 in lecture notes in computer science, pp 234–246
19. W^{3}Techs—Web technology surveys. Usage of SSL certificate authorities, December 2012 [Online] http://w3techs.com/technologies/overview/ssl_certificate/all
20. Jeon TJ, Park SJ (1988) Multiple time series model identification using concatenated sample cross-correlations. Commun Stat Theory Methods 17(1):1–16
21. Zhang B-L, Coggins R, Jabri MA, Dersch D, Flower B (2001) Multiresolution forecasting for futures trading using wavelet decompositions. Neural Netw IEEE Trans 12(4):765–775
22. Kang S, Lee S, Won Y, Seong B (2010) On-line prediction of nonstationary variable-bit-rate video traffic. Signal Process IEEE Trans 58(3):1219–1237
23. Proakis JG (1983) Digital communications / John G. Proakis. McGraw-Hill, New York
24. Kwiatkowski D, Phillips PCB, Schmidt P (1991) Testing the null hypothesis of stationarity against the alternative of a unit root. Technical Report 979. Cowles Foundation for Research in Economics, Yale University
25. Makridakis S, Hibon M (2000) The M3-Competition: results, conclusions and implications. Int J Forecast 16(4):451–476

26. Meade N (2000) A note on the robust trend and ararma methodologies used in the M3 competition. Int J Forecast 16(4):517–519
27. US Department of the Treasury. Treasury Public Key Infrastructure (PKI) and Shared Service Provider (SSP) Portal, December 2012. [Online] https://pki.treas.gov/crl_certs.htm
28. Catalan Certification Agency. Certificate Revocation List Repository, December 2012. [Online] http://www.catcert.cat/esl/RECURSOS/Comproveu-el-vostre-certificat/Llista-de-certificats-revocats
29. Gañán C, Munoz JL, Esparza O, Mata-Diaz J, Alins J, Silva-Cardenas C, Bartra-Gardini G (2012) RAR: risk aware revocation mechanism for vehicular networks. In: 2012 IEEE 75th vehicular technology conference (VTC Spring), vol 7. IEEE, Yokohama, pp 1–5
30. Spyropoulos T, Turletti T, Obraczka K (2008) Routing in delay-tolerant networks comprising heterogeneous node populations. IEEE Trans Mobile Comput, 1132–1147
31. Bhutta N, Ansa G, Johnson E, Ahmad N, Alsiyabi M, Cruickshank H (2009) Security analysis for delay/disruption tolerant satellite and sensor networks. In: Satellite and space communications. IWSSC 2009. International Workshop on, pp 385–389
32. Farrell S, Symington S, Weiss H, Lovell P (2009) Delay-tolerant networking security overview. IRTF, DTN research group, March 2009. Draft version-06
33. Symington S, Farrell S, Weiss H (2009) Bundle security protocol specification. IRTF, DTN research group, November 2009. Draft version-12
34. Gañán C, Muñoz JL, Esparza O, Mata-Día J, Alins J (2014) PPREM: privacy preserving REvocation mechanism for vehicular ad hoc networks. Comput Stand Interfaces 36(3):513–523
35. Gañán C, Muñoz JL, Esparza O, Loo J, Mata-Día J, Alins J (2013) BECSI: bandwidth efficient certificate status information distribution mechanism for VANETs. Mobile Inf Syst 9(4):347–370
36. Gañán C, Muñoz JL, Esparza O, Mata-Día J, Alins J (2014) EPA: an efficient and privacy-aware revocation mechanism for vehicular ad hoc networks. Pervasive Mobile Comput, ISSN 1574-1192, doi:10.1016/j.pmcj.2014.01.002
37. Chadwick D (2007) Dynamic delegation of authority in web services. In: Periorellis P (eds) Securing web services: practical usage of standards and specifications. Idea Group Inc, pp 111–137
38. She W, Yen I-L, Thuraisingham B (2008) Enhancing security modeling for web services using delegation and pass-on. In: IEEE international conference on web services (ICWS), pp 545–552
39. Hinarejos MF, Muñoz JL, Forné J, Esparza O (2010) PREON: an efficient cascade revocation mechanism for delegation paths. Comput Secur 29(6):697–711
40. W3C Working Group. Web Services Architecture. http://www.w3.org/TR/ws-arch/
41. Tuecke S, Welch V, Engert D, Pearlman L, Thompson M (2004) Internet X.509 public key infrastructure (PKI) proxy certificate profile. RFC 3820, Internet Engineering Task Force
42. Luna J, Medina M, Manso O (2005) Towards a unified authentication and authorization infrastructure for grid services: implementing an enhanced OCSP service provider into GT4. In: Public key infrastructure, LNCS. Springer, Berlin, pp 36–54

Carlos Gañán received the B.S. in Electrical Engineering and the M.S. in Telematics at the Technical University of Catalonia (UPC) in 2008 and 2009, respectively. In 2008, he joined the Information Security Group, with the Department of Telematics Engineering at UPC, Barcelona. In 2013, he received a Ph.D. in Telematics and currently is 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 Polytechnic University of Catalonia, Spain. He is currently a research staff member with the Telematics Services Research Group at 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 has published over 20 papers in these areas in international journals and conferences.

Jose L. Muñoz received the M.S. degree in Telecommunication Engineering in 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.

Oscar Esparza received his M.S. degree in Telecommunication Engineering in 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 at 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.