A model for revocation forecasting in public key infrastructures

Carlos Gañán Jose L. Munoz Oscar Esparza
Jorge Mata-Díaz Juanjo Alins
Universitat Politècnica de Catalunya (UPC)
{carlos.ganan,jose.munoz,oscar.esparza,jmata,juanjo}@entel.upc.edu

***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 (ARIMA). Our prediction model enables certification authorities to forecast the percentage of revoked certificates.**

***Index Terms*—Certification, PKI, Revocation, CRL, ARIMA.**

I. 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 to the certificate and a change of 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.

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. Each revoked certificate is identified in a CRL by its certificate serial number. Typically, CAs automatically issue new CRLs on a regular periodic basis (for example, daily, weekly or monthly). Although CRL is the most widely used mechanism to distribute certificate status information, there have been proposed other mechanisms to make revocation distribution more efficient (see Section II-A 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], [3], [4], [5], [6], [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], [9], [10] 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 the revocation process. However, they essentially just analyze the probability distribution of the revocation process. 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 weekly revocation percentage.

To build the model, we start performing a preliminary analysis using revocation codes. Revocation codes are optional parameters that can be included within revocation data. When we tried to get some insights about the revocation process using revocation codes, we found that each CA uses the parameter applying different criteria. This is mainly because the use of revocation codes is poorly specified by standardization organisms. So, we cannot use the revocation code as a parameter upon which to build our model. Thus, we build a revocation model which is exclusively based on temporal information, namely revocation dates and the life-time of certificates. These parameters are well-defined and compulsory. Using this information and the Box-Jenkins methodology [11], we are able to build a model and a predictor for revocation. More precisely, we are able to show that an ARIMA (Autoregressive, Integrated, Moving Average) model provides objective and accurate predictions of the incremental percentage of weekly 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 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 to mention that VeriSign is the only CA that makes the life-time of certificates public. Since this parameter is necessary to build the model, our ARIMA model is based only on data provided by VeriSign. Although the collected revocation data only belong to the leading certification authority[1], 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. 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 Section VII).

The rest of this paper is organized as follows. Section II briefly reviews the required background. Section III presents a preliminary analysis focused on the study of the revocation causes. In Section IV, we discuss the methodology we used to collect and analyze the revocation data. In Section V, we identify the best ARIMA model that fits the revocation events. Next, in section VI we present the predictor from the previous ARIMA model. In Section VII we discuss some possible practical applications of the revocation forecasting model. Finally, we conclude in Section VIII.

II. Background

A. Certificate Status Validation Mechanisms

In this section we briefly review the main approaches to convey Certificate Status Information (CSI). The traditional approach is to periodically publish a *certificate revocation list* (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. To minimize the network resources needed to communicate them to the end entities, CRLs are published frequently and timely when new revocations occur. However, clients tend to download CRLs according to the time established for an update. Therefore, end entities do not possess fresh revocation information on a need-to-know basis.

Two solutions were proposed to deal with the CRLs' problems [14]: 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. Delta-CRLs provide the means for constructing incremental CRLs. Whenever new revocations have taken place, the (new) CRL that end entities will have to retrieve, will contain only the new CSI.

Also to deal with the communication overhead of the CRLs, Micali [15] 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 of the CA's communication cost with the CSI repositories [16].

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) [17] 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.

The responses to CSI queries returned by OCSP are digitally signed. The authority that runs the OCSP service can either be the CA itself, or another entity designated by the CA as a CSI provider (Trusted Responder or CA Designated Responder). The CSI responses given by the authorities are always signed to let requesters verify the authenticity of the CSI. However, the signing of each OCSP response is a computational overhead and it could facilitate Denial of Service (DoS) attacks. Pre-computed responses that have a short validity interval could be a solution to this problem. These pre-signed OCSP responses are usually provided via SSL/TLS authenticated channels. 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 [18] suggested another CSI mechanism, the Certificate Revocation Tree (CRT). A CRT is based on a Merkle hash tree [19] 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 $O(1)$.

B. 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 Auto Regressive Integrated Moving Average (ARIMA) model popularized by Box and Jenkins [11]. In this paper we show that the revocation probability can be suitably forecast using an ARIMA process.

The ARIMA approach to forecasting is based on the following ideas: the forecasts are based on linear functions of

[1] VeriSign represents around half of the market according to [12].

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: 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 $ARIMA(p, d, q)$ where p describes the AR part, d describes the integrated part and q describes the MA part.

- **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 $p > 1$:

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

- **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.

  In this sense, the integrated contribution allows to capture the nonstationarity part of the moments of the stochastic process. If after differencing, a series is stationary, then the series is called integrated of order one and is denoted $I(1)$. If, however, the series is not stationary after differencing once, then we might need to take a second difference. If the series becomes stationary after the second difference then it is said to be integrated of order two and is denoted $I(2)$, and so on. That is a nonstationary 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:

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

  where:

$$c_i = \binom{d}{i} (-1)^{i+1} \;\; i \in \{1, 2, \ldots, d\}. \quad (3)$$

  For example, a process whose mean is nonstationary and the rest of high order moments are stationaries would have an integrated component of order 1. This integrated process is the so-called ”random walk”.

- **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,

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

  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 $n-q$.

III. 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 three main CAs: VeriSign, GoDaddy and Entrust.

The PKIX/X.509 certificate and CRL specification [13] defines 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 non-critical 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 revocation practice for each code. As a result, we will see that in practice, revocation reasons are poorly utilized or even ignored by most of the CAs. To illustrate this, let us consider the case of code-signing certificates. For these certificates, VeriSign, which is the overall leading CA in the marketplace, does not provide any information about the revocation reason[2]. On the other hand, certificates from the GoDaddy CA use almost always the same revocation reason. 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 Figure 1, in which we covered more than 600,000 certificates from GoDaddy.

To extend the results, we analyze the revocation causes available for 20,000 revoked certificates issued by Entrust. Figure 2 shows the probability of each revocation cause. In this case, the main reason for a certificate to be revoked is `superseding` (cause (4)) followed by `cessationOfOperation` (cause (5)). Notice that considering these two causes, we cover around the 92% of the revoked certificates.

[2] Actually, VeriSign does not provide revocation causes for any of its issued certificates.

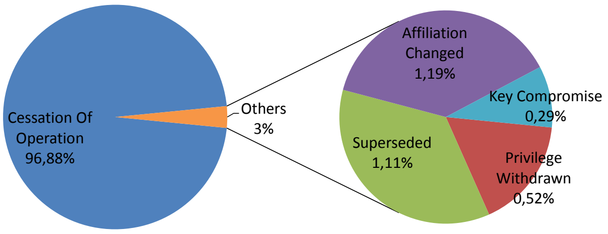


Fig. 1: Revocation causes of code-signing certificates issued by GoDaddy.

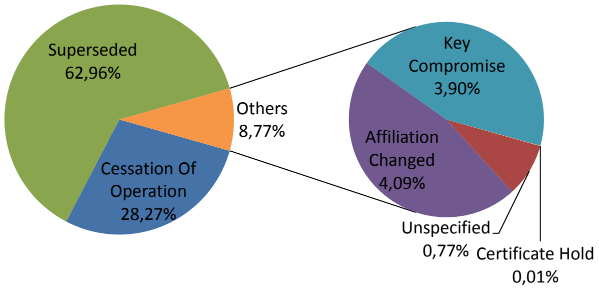


Fig. 2: Revocation causes of code-signing certificates issued by Entrust.

The rough analysis performed and the discussion presented evidences that 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. Versign, the certification market leader, does not provide revocation codes within their CRLs (this might be related with privacy issues), while other CAs like GoDaddy and Entrust use revocation causes in different ways and with a different distribution.

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 and the life-time of certificates. These parameters are well-defined, compulsory and they will allow us to build a model that will show that certificate revocations are closely related with time.

IV. 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. Those certificates were collected using VeriSign's Certificate Revocation Lists [20]. We choose VeriSign because it is the only certification authority that provides information about the lifetime of the revoked certificates. This information is crucial to develop our prediction model about the percentage of revoked certificates. VeriSign's competitors (e.g. Godaddy or Entrust) do not publicize information about the lifetime of the revoked certificates. Moreover, VeriSign is the trusted provider of Internet infrastructure services for the networked world that leads the global SSL marketplace with a 47.52% share [12]. Therefore, though the data collected belongs exclusively to a single certification authority, the model is expected to be useful for any other CA as the data covers almost half of the global marketplace.

For the purpose of this paper, we consider the two main types of certificates (see Table I for details): certificates for SSL servers and code-signing certificates. Accordingly, we also use two different types of CRL files from VeriSign website and from these CRLs we obtain the following parameters:

- Last Update instant of the CRL.
- Next Update instant of the CRL.
- Serial Number of each revoked certificate.
- Revocation Date of each revoked certificate.

It is worth to mention that typically, a revoked and expired certificate remains in the CRL for one additional CRL publication interval. However, due to VeriSign certification practice statement, certificates from VeriSign's CRLs are removed after they expire. Therefore, the collected dataset covers only non-expired revoked certificates. On the other hand, our goal is to perform a time series analysis of the percentage of weekly revoked certificates. That is to say, for each week, we want to compute the ratio between the number of certificates that are going to be revoked during the week over the number of certificates that are valid at the end of the week. Valid certificates are those certificates which are non-expired and non-revoked. Since the amount of valid certificates at each moment is not publicly made available by VeriSign, we cannot directly calculate this ratio. However, we can use an indirect strategy to calculate a proportional ratio following the same procedure as in [8], [9], [10]. The idea is to obtain a subset of valid certificates for considered weeks proportional to the full set of valid certificates. To do so, for each certificate that is in our data set (the set of collected CRLs) we obtain its serial number, its revocation date and its issuance date (this last parameter is not present in CRLs but we can obtain it through the VeriSign web interface using the serial numbers of certificates). Notice that with this information from our dataset, we are able to compute a subset of valid certificates for the considered week, that is, certificates that were issued before the considered week and that will be revoked beyond the considered week. We can name this subset of valid certificates for each week as to-be-revoked valid certificates. Figure 4 shows the different sets of certificates according to their status for any given week.

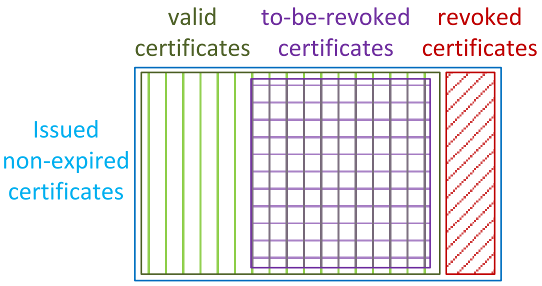
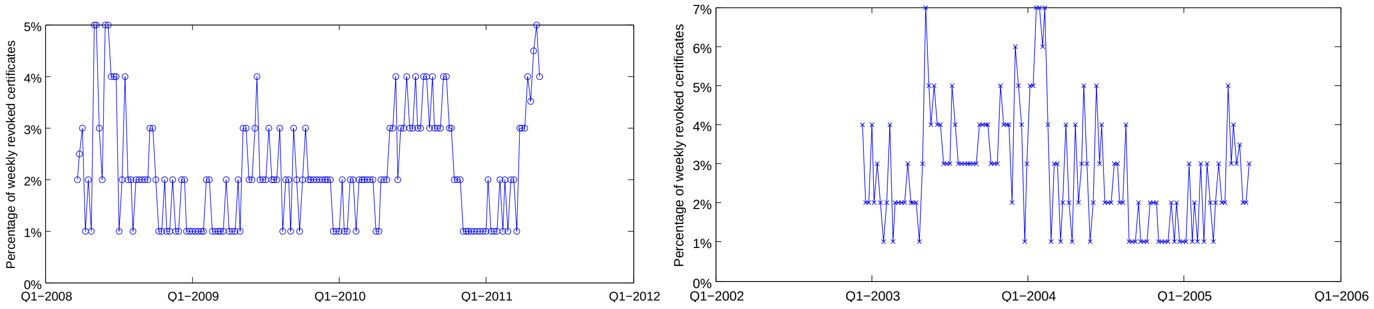


Fig. 4: Sets of certificates according to their status.

For the rest of the paper, we will simply refer to the "weekly percentage of revoked certificates" as the ratio between the

File Name	# Certificates	CA Name	Description
VeriSign International Server CA Class 3	19.345	VeriSign International Server CA Class 3	Issues Global Server SGC certificates.
Class3Code-Signing2001	1.996	VeriSign Class 3 Code Signing 2001 CA	Legacy issuer of Code Signing certificates. Issues code signing and object signing certificates for use with Netscape browsers, Microsoft Internet Explorer browsers, Microsoft Office, Sun Java Signing, Macromedia, and Marimba.

TABLE I: Description of the collected CRLs.



(a) SSL (SGC) certificates

(b) Code-signing certificates

Fig. 3: Weekly percentage of revoked certificates evolution

set of certificates that are going to be revoked during the considered week over the set of to-be-revoked valid certificates for this week. Notice that this method does not compromise the validity of our model since we are interested in modeling the revocation pattern and its statistical properties. To get absolute values, just a change of scale is needed. Figure 3 shows a plot of the weekly revocation percentages for SSL and code-signing certificates that we computed with the available data from VeriSign. Now, we build a single time series by concatenating the time-series of Figure 3 to obtain general statistical properties (the result of this concatenation is plotted in Figure 5). Analyzing the statistical properties of the concatenated time-series we observe that:

- On average, we observe that the ratio of the percentage of revoked certificates is approximately 2%.
- As we have approximated the valid certificates by the number of valid certificates to be revoked, this means that less than 2% of the issued certificates are revoked weekly.
- The percentage of weekly revoked certificates exhibits a highly variable behavior.
- The standard deviation of the percentage of revoked certificates is almost 75% of the mean ($\mu = 0.0245$, $\sigma = 0.0183$).

Following the probability analysis performed in some previous works ([8] and [9]), we also try to fit the probability density function (pdf) of the concatenated time-series of Figure 5 to an exponential distribution. As in the previous works, we also find that the empirical data, while not exactly, roughly follow an exponential distribution (see Figure 6). Fairly similar results were obtained in [8], [9]. This is remarkable since the mentioned previous works use datasets with different temporal ranges. This gives us the intuition that the revocation process follows some type of temporal pattern no matter which is the dataset used. Following this intuition, in the next section we follow a novel approach for modeling the revocation process which has not been explored by any other previous work. Our approach consists in analyzing the concatenated time-series of the weekly percentage of revoked certificates by means of an ARIMA process. Following this technique, we obtain a single model that predicts extremely well the weekly revocation percentage of both SSL and code-signing certificates. Finally, the outstanding behavior of our predictions will be asserted by applying different tests over the developed model.

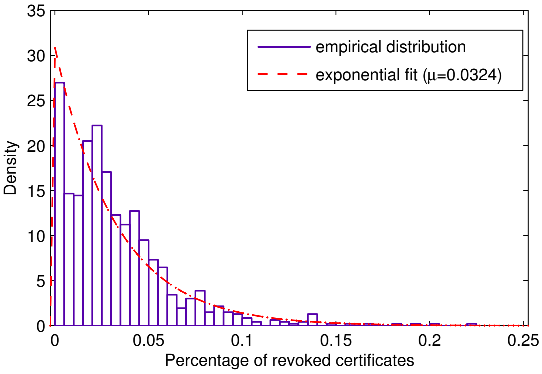


Fig. 6: Empirical data vs. fitted exponential PDF.

V. Modeling the Revocation

In this section we develop a new model for the percentage of revoked certificates. Our model might be interesting for several reasons: on one hand, it provides some insights about the revocation process. On the other hand, it allows CAs to

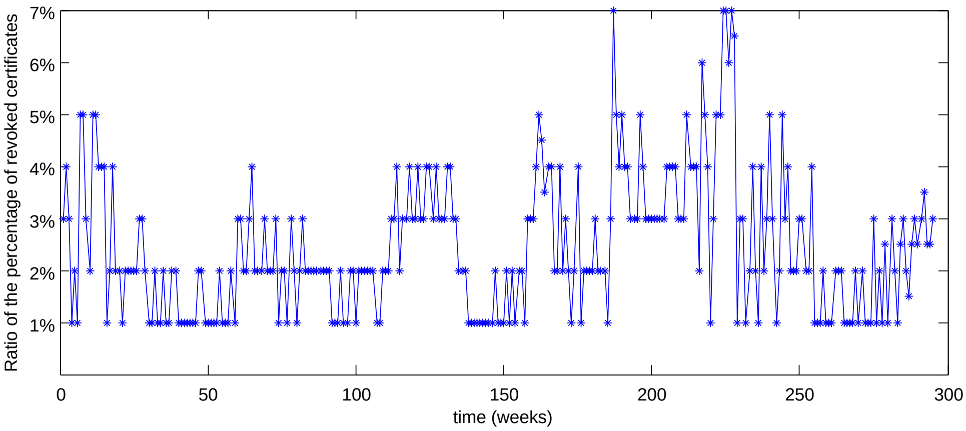


Fig. 5: Concatenated time-series.

make predictions about certificate revocation which might be useful in some scenarios (see Section VII).

So, we want to obtain a model from the revocation time-series to obtain a suitable model for revocation forecasting. 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 revocation percentage, 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. If there are lagged dependent variables set as the regressors, regression estimates are biased and inconsistent. Fortunately, this can be fixed using an Autoregressive Integrated Moving Average Model (ARIMA). Thus, as we have found that the simplest analysis, the MLR, is not suitable to develop our forecasting model, we will use the more complete ARIMA.

To do so, in the rest of the section we first we describe each of the component of the ARIMA process in the Z-domain. Then, 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.

A. ARIMA processes in the Z-domain

In section II-B 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. This function completely defines the behavior of a the linear system and makes it easier its management.

In this sense we use the delay operator z^{-1} [21] 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 Figure 7 a scheme of the ARIMA model is shown. Note that, as it is shown in the figure, we can express the transfer function of the $ARIMA(p,d,q)$ process as a cascade of all three components.

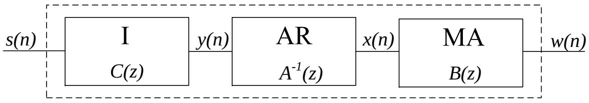


Fig. 7: Components of an ARIMA process.

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). Applying the Z-transform to equation (4) we can express the MA process in the z-domain as:

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

Note that it only uses 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. In this sense, Figure 8 represents the $MA(q)$ as a FIR filter whose transfer function is $B(z)$:

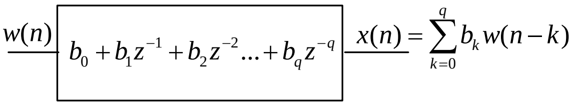

w(n)
$b_0 + b_1 z^{-1} + b_2 z^{-2} ... + b_q z^{-q}$
$x(n) = \sum_{k=0}^{q} b_k w(n-k)$

Fig. 8: MA filter.

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

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

The impulse response of the associated system is Infinite Impulse Response (IIR) and its transfer function is of the kind all-pole. 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. Figure 9 represents the $AR(q)$ as an IIR filter whose transfer function is $A(z)$:

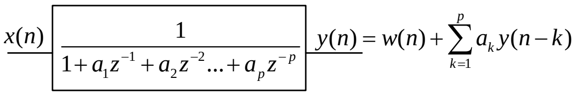

$x(n)$
$\frac{1}{1+a_1z^{-1}+a_2z^{-2}...+a_pz^{-p}}$
$y(n) = w(n) + \sum_{k=1}^{p} a_k y(n-k)$

Fig. 9: AR filter.

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

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

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 10 represents the $I(d)$ as a linear filter whose transfer function is $C(z)$:

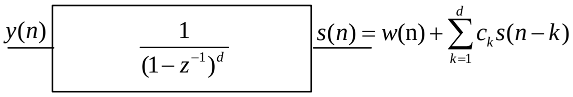


Fig. 10: Integrated filter.

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

$$S(z) = [B(z)A(z)C(z)] \cdot W(z). \quad (8)$$

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:

$$H(z) = \frac{S(z)}{W(z)} = B(z)A(z)C(z). \quad (9)$$

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 11 represents the ARIMA filter with transfer function $H(z)$:

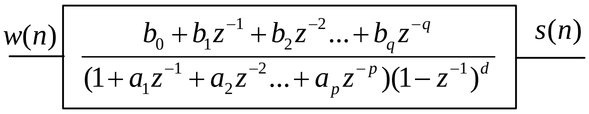

$w(n)$
$\frac{b_0+b_1z^{-1}+b_2z^{-2}...+b_qz^{-q}}{(1+a_1z^{-1}+a_2z^{-2}...+a_pz^{-p})(1-z^{-1})^d}$
$s(n)$

Fig. 11: ARIMA filter.

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 $C^{-1}(z)$ 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 impulsional response. The rest of obtained poles (z_k) will be found in the unit circle ($|z_k| < 1$) of the Z plane.

B. Characterization of the revocations as an ARIMA process

In this section, we show that the actual behavior of the revocation process can be modeled as an ARIMA process. To that end, a new ARIMA model for the revocations has been developed to find the predictor. 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 by differencing:
   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 from the stationary series.
   b) Extracting the ARMA parameters.
4) Model fit validation using residual diagnostics

The result of carrying out these steps will be an ARIMA model defined by:

- an order d that represents the number of nonseasonal differences for the I component of the process.
- q coefficients that characterize the MA component of the process.
- p coefficients that characterize the AR component of the process.

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

1) Model Identification:

The first step in developing the ARIMA model is to determine if the series is stationary and if 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 percentage of revoked certificates (see Figure 12). We can observe that the actual time series follows certain trend, so the time series is nonstationary. The temporal series of the percentage of revoked certificates presents a slow variation of the mean. To confirm this visual evaluation, we perform a KPSS test [22] at the 99% confidence level which rejects the null hypothesis of stationarity.

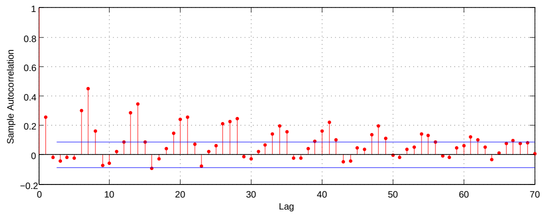


Fig. 12: ACF of the percentage of revoked certificates.

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 non stationary mean. To synthesize a good predictor it is necessary to capture this long term effect. The long term dependence produces that the mean varies. This variation reaches maximum and minimum levels which are very distant (see Figure 13). However, the variance remains almost constant. This allows to conclude that the integrated component of the model should be of order 1 and its associated transfer function $C(z) = (1 - z^{-1})^{-1}$.

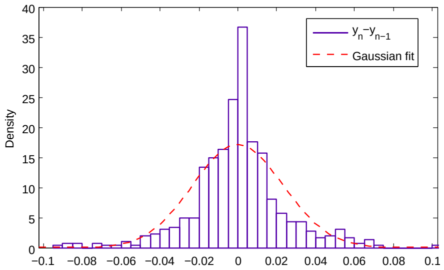


Fig. 13: PDF of the extracted integrated component time series.

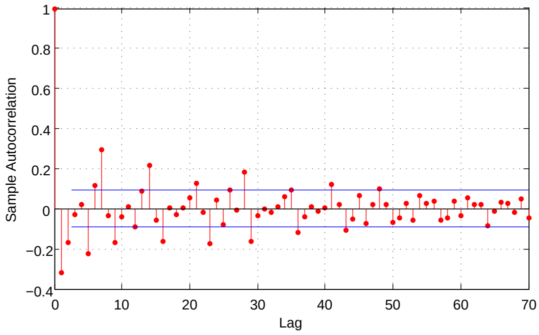


Fig. 14: ACF of the extracted integrated component time series.

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 Figure 7 the residual ARMA series $y(n)$ and the real series $s(n)$ are related as follows:

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

In this way, the temporal series $y(n)$ will be obtained at the output of the FIR (Finite Impulse Response) filter, whose transfer function is $(1 - z^{-1})$, when it is excited with the temporal series generated by the coder. It can be checked that the temporal series $y(n)$ is a stochastic process with mean 0 and an invariant autocorrelation coefficients. This statistical analysis has been carried out with the data gathered from VeriSign and with autocorrelation lags of 100 units. The probability distribution function fits a Gaussian distribution.

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 [11]. The visual analysis of the autocorrelation of $y(n)$ confirms the result of this test (see Figure 14). The fact that $y(n)$ is not white noise allows us to model it be means of an ARMA process.

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) [11] 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.

$$\begin{aligned} A(z) = 1 &+ 1.642z^{-1} + 0.9225z^{-2} + 0.7091z^{-3} + 0.3704z^{-4} \\ &- 0.7782z^{-5} - 0.9453z^{-6} - 0.5768z^{-7} - 0.3449z^{-8} \\ &+ 0.01871z^{-9} + 0.1073z^{-10}. \end{aligned} \tag{11}$$

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:

$$\begin{aligned} X(z) = A(z)Y(z) = (1 &+ 1.642z^{-1} + 0.9225z^{-2} + 0.7091z^{-3} \\ &+ 0.3704z^{-4} - 0.7782z^{-5} - 0.9453z^{-6} - 0.5768z^{-7} \\ &- 0.3449z^{-8} + 0.01871z^{-9} + 0.1073z^{-10})Y(z). \end{aligned} \tag{12}$$

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 $x(n)$ [11]. We use the same criteria as in the AR process for selecting the parameters of the MA

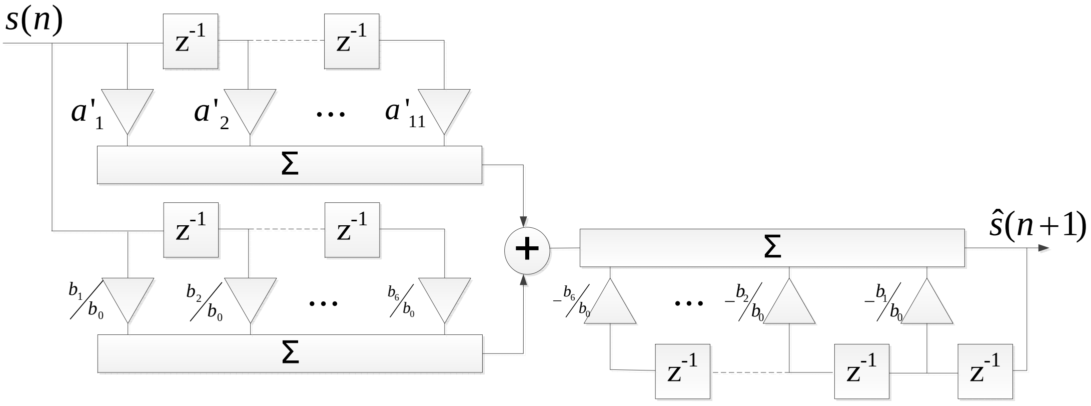


Fig. 15: ARIMA Predictor.

i	0	1	2	3	4	5	6	7	8	9	10	11
a'_i	–	0.6417	-0.7192	-0.2134	-0.3387	-1.149	-0.1671	0.3685	0.2319	0.3636	0.08859	-0.1073
b_i	1	-0.3558	0.6139	0.09758	-0.7464	0.1897	-0.7705					

TABLE II: Coefficients of the ARIMA predictor.

process. The best adjustment is obtained with a MA process with order 6.

$$B(z) = 1 - 0.3558z^{-1} + 0.6139z^{-2} + 0.09758z^{-3} - 0.7464z^{-4} + 0.1897z^{-5} - 0.7705z^{-6}. \tag{13}$$

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:

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

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 $A'(z)$, that is:

$$A'(z) = 1 + 0.6417z^{-1} - 0.7192z^{-2} - 0.2134z^{-3} - 0.3387z^{-4} - 1.149z^{-5} - 0.1671z^{-6} + 0.3685z^{-7} + 0.2319z^{-8} + 0.3636z^{-9} + 0.08859z^{-10} - 0.1073z^{-11}. \tag{16}$$

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).

VI. Revocation Forecasting

Once we have developed an ARIMA model for the time-series of the weekly revocation percentage, in this section we use this model to build an ARIMA predictor. Then, we check the goodness of the predictions for the concatenated revocation time-series and also for each type of certificate.

First, 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:

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

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. The forecast value of $\hat{w}(n+1)$ will be the mean value of the $w(n)$ series. In this case, the mean value is 0. Thus, the $(n+1)$ sample prediction can be simplified as:

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

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

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

Subtracting (19) to (15):

$$s(n) - \hat{s}(n) = b_0(w(n) - \hat{w}(n)). \tag{20}$$

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

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

Therefore:

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

To evaluate the behavior of the above transfer function, an analysis of the forecast errors has been done. Figure 16 presents the residuals autocorrelation and the 99% confidence intervals. The residual diagnostic determines that the forecast errors are clearly uncorrelated.

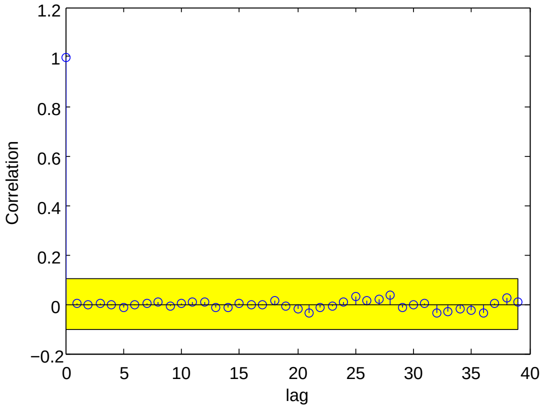


Fig. 16: Autocorrelation function of the forecast errors using the ARIMA predictor.

Figure 17 presents the CDF of the residuals autocorrelation and the CDF of standard Gaussian distribution with its 99% confidence intervals. Notice that the residuals are close to the Gaussian distribution. As it is shown below, the estimated ARIMA model is appropriate for forecasting. Therefore, the ARIMA model fits well the behavior of the revocations.

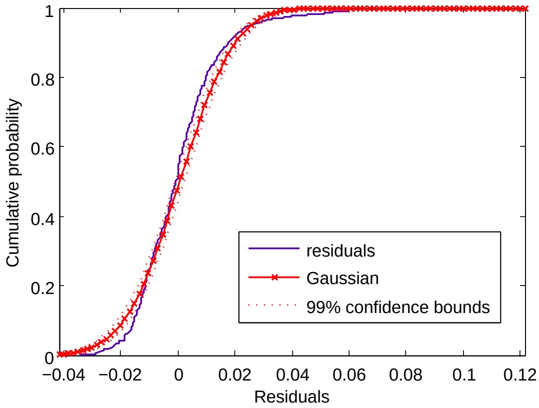


Fig. 17: CDF of the ARIMA prediction residuals.

Once we have seen that the residuals are quite close to the Gaussian distribution, we obtain an expression for the ”1 ahead“ prediction. Replacing this expression in the equation (18), the $(n+1)$ sample prediction can be written as:

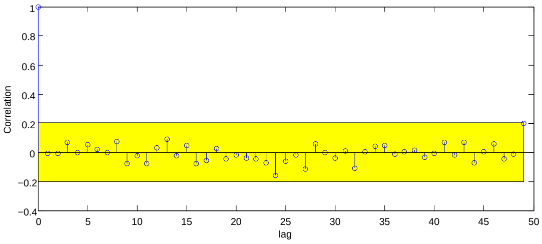


Fig. 18: ACF of the residual series for SGC certificates

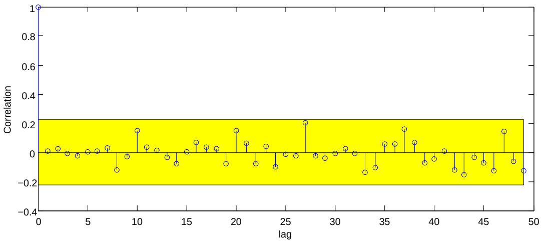


Fig. 19: ACF of the residual series for Code-signing certificates

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

From this expression, we can plot the derived ARIMA predictor using the coefficients obtained in the previous section (see Figure 15). 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).

Once we have seen that the model fits quite accurately the global revocation process, we must check its suitability for each type of certificate. For this purpose, we analyze the residuals for individual revocation processes using the ARIMA model obtained from the global series. Figures 18 and 19 present the residuals autocorrelation and the 99% confidence intervals for each type of certificates. Note that there is no residual that exceeds the confidence intervals. Therefore, we can conclude that the ARIMA model developed from the concatenated time-series captures the revocation pattern of both certificate types.

Finally, in figures 20 and 21 the “1 ahead“, ”2 ahead“ and ”3 ahead“ predictions are presented for the percentage of revoked certificates. 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

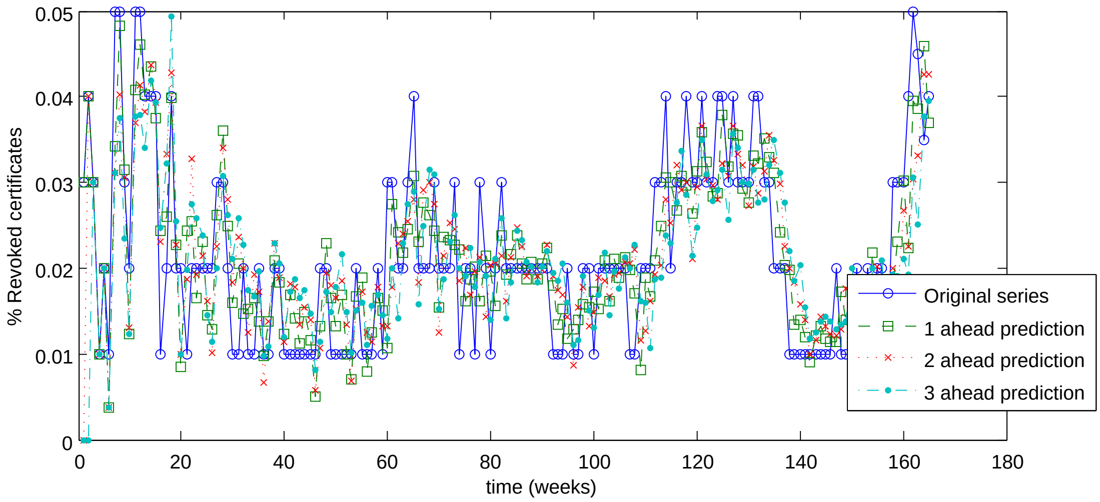


Fig. 20: Predictions for SGC certificates.

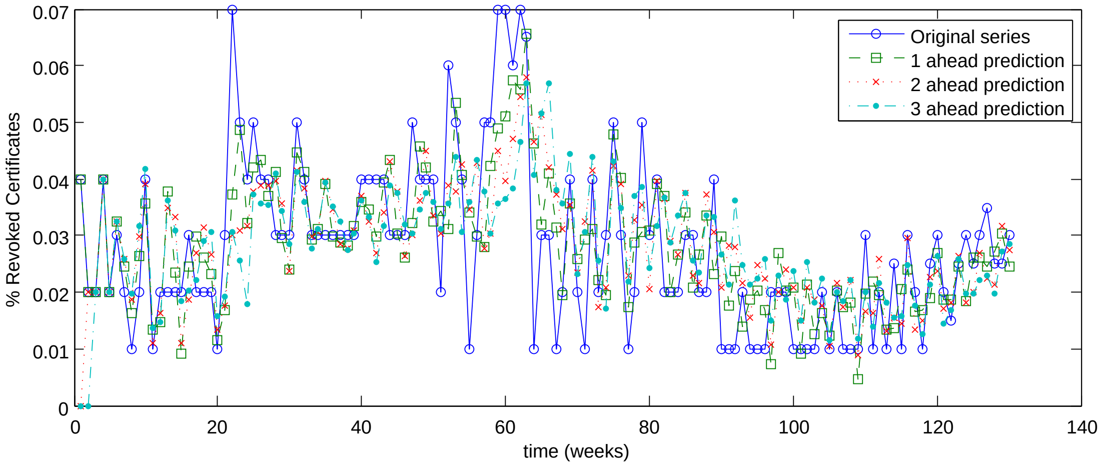


Fig. 21: Predictions for code-signing certificates.

terms resulting in low accuracy in forecasting performances. Again, the predictions are valid for both certificate type, corroborating the results obtained from the residual analysis.

VII. Some Applications of Revocation Forecasting

In this section, we briefly describe some possible scenarios in which revocation forecasting might be useful. However, a detailed study of all the possible applications of revocation forecasting is beyond the scope of this paper.

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 revocation percentage. 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 [23]. Regarding security, one of the main challenges in DTNs is how to create and distribute keys and credentials. To this respect, many authors [24], [25] and the current security draft specification for DTN [26] agree that the most promising solution is to use public-key cryptography with digital certificates.

An example of this paradigm are vehicular communications (VANET) in which vehicles might not be always connected to the infrastructure (Internet). In this case, PKI users (vehicles) might not dispose of the latest CRL available. 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 may not be considered a long time if the revocation percentage is around 10^{-5} but this interval can be considered quite long if we are going to have a 10% of revoked certificates during the next two hours. 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 [27], [28], [29]. Dynamic delegation is devised for highly distributed scenarios like Web Services [30]. 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 [31]. 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 [32]. In this scenario, a revocation forecasting model could be useful to set the validity period of the certificates and/or its associated CRLs.

VIII. Conclusions

We have analyzed real empirical data to derive a model which allows to forecast the percentage of revocations in the near future. Our research represents a step towards 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 certificate status information distribution system. The ARIMA model completely considers the dynamic process of data series and the autocorrelation of residuals to achieve precise forecasting of the percentage 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 both short (a few weeks) and long (a few months) time scales.

Although the collected revocation data only belong to the leading certification authority 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 P2PSEC (TEC2008-06663-C03-01), and by the Government of Catalonia under grant 2009 SGR 1362 to consolidated research groups.

References

[1] R. Housley, W. Polk, W. Ford, and D. Solo. Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile. RFC 3280, Internet Engineering Task Force, April 2002.

[2] Andre Arnes, Mike Just, Svein J. Knapskog, Steve Lloyd, and Henk Meijer. Selecting revocation solutions for PKI. In *NORDSEC '95*, 1995.

[3] M. Lippert, V. Karatsiolis, A. Wiesmaier, and J. Buchmann. Life-cycle management of x.509 certificates based on ldap directories. *J. Comput. Secur.*, 14:419–439, September 2006.

[4] John Iliadis, Diomidis Spinellis, Dimitris Gritzalis, Bart Preneel, and Sokratis Katsikas. Evaluating certificate status information mechanisms. In *Proceedings of the 7th ACM conference on Computer and communications security*, CCS '00, pages 1–8, New York, USA, 2000. ACM.

[5] G. F. Marias, K. Papapanagiotou, and P. Georgiadis. Adopt. a distributed ocsp for trust establishment in manets. *11th European Wireless Conference 2005*, April 2005.

[6] T. Perlines Hormann, K. Wrona, and S. Holtmanns. Evaluation of certificate validation mechanisms. *Comput. Commun.*, 29:291–305, February 2006.

[7] A. Arnes. Public key certificate revocation schemes, 2000. Queen's University. Ontario, Canada. Master Thesis.

[8] Daryl Walleck, Yingjiu Li, and Shouhuai Xu. Empirical analysis of certificate revocation lists. In *Proceeedings of the 22nd annual IFIP WG 11.3 working conference on Data and Applications Security*, pages 159–174, 2008.

[9] Chengyu Ma, Nan Hu, and Yingjiu Li. On the release of crls in public key infrastructure. In *Proceedings of the 15th conference on USENIX Security Symposium - Volume 15*, Berkeley, CA, USA, 2006.

[10] Nan Hu, Giri K. Tayi, Chengyu Ma, and Yingjiu Li. Certificate revocation release policies. *J. Comput. Secur.*, 17:127–157, April 2009.

[11] George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.

[12] Netcraft. Market share of certification authorities, 2009. https://ssl.netcraft.com/ssl-sample-report/CMatch/certs Accessed on 05/2011.

[13] D. Cooper, S. Santesson, S. Farrell, S. Boeyen, R. Housley, and W. Polk. Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile. RFC 5280, Internet Engineering Task Force, May 2008.

[14] R. Housley, W. Ford, W. Polk, and D. Solo. Internet X.509 Public Key Infrastructure Certificate and CRL Profile. RFC 2459, Internet Engineering Task Force, January 1999.

[15] S. Micali. Efficient certificate revocation. Technical Report TM-542b, MIT Laboratory for Computer Science, 1996.

[16] M. Naor and K. Nissim. Certificate Revocation and Certificate Update. *IEEE Journal on Selected Areas in Communications*, 18(4):561–560, 2000.

[17] M. Myers, R. Ankney, A. Malpani, S. Galperin, and C. Adams. X.509 Internet Public Key Infrastructure Online Certificate Status Protocol - OCSP, 1999. RFC 2560.

[18] P.C. Kocher. On certificate revocation and validation. In *International Conference on Financial Cryptography (FC98)*, number 1465 in Lecture Notes in Computer Science, pages 172–177, feb 1998.

[19] R.C. Merkle. A certified digital signature. In *Advances in Cryptology (CRYPTO89)*, number 435 in Lecture Notes in Computer Science, pages 234–246, 1989.

[20] Legal repository from verisign: Crl reference table. http://www.verisign.com/repository/crl.html Accessed on 05/2011.

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

[22] Denis Kwiatkowski, Peter C.B. Phillips, and Peter Schmidt. Testing the null hypothesis of stationarity against the alternative of a unit root. Technical Report 979, Cowles Foundation for Research in Economics, Yale University, May 1991.

[23] Thrasyvoulos Spyropoulos, Thierry Turletti, and Katia Obraczka. Routing in delay-tolerant networks comprising heterogeneous node populations. *IEEE Transactions on Mobile Computing*, pages 1132–1147, 2008.

[24] N. Bhutta, G. Ansa, E. Johnson, N. Ahmad, M. Alsiyabi, and H. Cruickshank. Security analysis for delay/disruption tolerant satellite and sensor networks. In *Satellite and Space Communications, 2009. IWSSC 2009. International Workshop on*, pages 385–389, Sept. 2009.

[25] S. Farrell, S. Symington, H. Weiss, and P. Lovell. Delay-tolerant networking security overview. *IRTF, DTN research group*, March 2009. Draft version -06.

[26] S. Symington, S. Farrell, and H. Weiss. Bundle security protocol specification. *IRTF, DTN research group*, November 2009. Draft version -12.

[27] D. Chadwick. Dynamic Delegation of Authority in Web Services. In Panayiotis Periorellis, editor, *Securing Web Services: Practical Usage of Standards and Specifications*, pages 111–137. Idea Group Inc, 2007.

[28] W. She, I-L. Yen, B. Thuraisingham. Enhancing Security Modeling for Web Services Using Delegation and Pass-On. In *IEEE International Conference on Web Services (ICWS)*, pages 545–552, Sept. 2008.

[29] M. Francisca Hinarejos, Jose L. Muñoz, Jordi Forné, and Oscar Esparza. Preon: An efficient cascade revocation mechanism for delegation paths. *Computers & Security*, 29(6):697 – 711, 2010.

[30] W3C Working Group. Web Services Architecture. http://www.w3.org/TR/ws-arch/.

[31] S. Tuecke, V. Welch, D. Engert, L. Pearlman, and M. Thompson. Internet X.509 Public Key Infrastructure (PKI) Proxy Certificate Profile. RFC 3820, Internet Engineering Task Force, June 2004.

[32] J. Luna, M. Medina, O. Manso. Towards a Unified Authentication and Authorization Infrastructure for Grid Services: Implementing an Enhanced OCSP Service Provider into GT4. In *Public Key Infrastructure*, LNCS, pages 36–54, Berlin, Heidelberg, 2005. Springer-Verlag.