IEICE TRANSACTIONS

on Communications

VOL.E93-B
NO.6
JUNE 2010

A PUBLICATION OF THE COMMUNICATIONS SOCIETY

The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER Special Section on Quality of Communication Networks and Services

Forecasting of Information Security Related Incidents: Amount of Spam Messages as a Case Study

Anton ROMANOV[†a)], *Nonmember and* Eiji OKAMOTO[†], *Fellow*

SUMMARY With the increasing demand for services provided by communication networks, quality and reliability of such services as well as confidentiality of data transfer are becoming ones of the highest concerns. At the same time, because of growing hacker's activities, quality of provided content and reliability of its continuous delivery strongly depend on integrity of data transmission and availability of communication infrastructure, thus on information security of a given IT landscape. But, the amount of resources allocated to provide information security (like security staff, technical countermeasures and etc.) must be reasonable from the economic point of view. This fact, in turn, leads to the need to employ a forecasting technique in order to make planning of IT budget and short-term planning of potential bottlenecks. In this paper we present an approach to make such a forecasting for a wide class of information security related incidents (ISRI) — unambiguously detectable ISRI. This approach is based on different auto regression models which are widely used in financial time series analysis but can not be directly applied to ISRI time series due to specifics related to information security. We investigate and address this specifics by proposing rules (special conditions) of collection and storage of ISRI time series, adherence to which improves forecasting in this subject field. We present an application of our approach to one type of unambiguously detectable ISRI — amount of spam messages which, if not mitigated properly, could create additional load on communication infrastructure and consume significant amounts of network capacity. Finally we evaluate our approach by simulation and actual measurement.

key words: *information security incidents, forecasting, planning*

1. Introduction

At the current moment web-based services and other type of network based applications are becoming more and more popular. Nowadays users have become very experienced in newest technology and demand not only for new functionality and solutions but also for quality of service in terms of technology and comfort. At the same time, as more and more information is stored in the digital form (sometimes only in digital form), the loss of data, its unauthorized modification or problems with access for authorized users because of technical failures are becoming more and more critical even in terms of utility (for example lack of privacy can lead to lack of comfort of usage of some solution and thus to a negative impression). So in order to attract more customers to use new products and solutions, potential users must be provided with confidence that their data and identity are adequately protected.

At the same time as providing information security is among non-profit earning business processes, it must be organized in the cost effective manner — price of implementation and maintenance of different countermeasures must be strictly evaluated in terms of justifiability of related security investment. For example: is it really necessary to hire more IT security team members in a given enterprise? Is there any sense to buy a honey pod system for your private notebook? Or will the deployment of a fingerprint-based authentication system in a new mobile phone attract new customers for a mobile phone vendor? Thus all security related investments must be compared in terms of mitigation of amount of potentially dangerous events (for own usage) or payback (for a security feature in some product or service which you are offering).

But in order to perform such evaluation there must be a way how to estimate an amount of potential ISRI (and thus losses) related to a given technology and a given type of ISRI. Thus there must be a methodology to make an ISRI forecasting.

According to [1] there are three types of forecasts: long term (concerned with strategic decisions, looking ahead for several years), middle term (concerned with tactical decisions, looking ahead for more then a month but up to several years) and short term (concerned with operational decisions, looking ahead for the next few weeks) forecasts. As to information security, long term forecasting generally is not performed as even now this field of science is not strictly formalized and quantified (except for cryptology and side electronic radiation) and taking into account the speed of development of new technological solutions, in order to forecast severity of some threats in several years first of all it is necessary to forecast which technical solutions, protocols and etc. will be in place in several years. So security managers usually have to focus on middle-term and short-term forecasts to plan required IT budget and estimate potential problems in the nearest future accordingly.

There are many methods to perform forecasting, which are usually divided into two large groups [2]: qualitative and quantitative. Quantitative methods are again divided into two subgroups: projective methods (estimation of future events is made by looking at the pattern of past observations and trying to extend it to future) and causal methods (estimation is made by the analysis of effects of outside influence to obtain equations which are later used for forecasting of future values).

In this paper we are providing an approach to make short-term and middle-term forecasts for a wide class of

Manuscript received October 23, 2009.
Manuscript revised January 30, 2010.

[†]The authors are with University of Tsukuba, Tsukuba-shi, 305-8577 Japan.
a) E-mail: romanov@cipher.risk.tsukuba.ac.jp

DOI: 10.1587/transcom.E93.B.1411

ISRI — unambiguously detectable ISRI, applying projective methods based on different auto regression models which are already widely used for forecasting in financial time series analysis, but can not be directly applied to ISRI data due to some relevant specifics [3] which is considered in detail in Sect. 3. In our approach we address this specifics and resolve problems which it causes by introduction of two special conditions (or rules) according to which ISRI time series data collection should be organized (see Sect. 3).

To illustrate application of our approach we are taking amount of spam messages as a sample data. The spam statistics was selected as the easiest real statistics to be collected without making special efforts (like organizing special sandboxes).

The originality of this paper is summarized as follows: (1) it proposes an easy but comprehensive approach to make efficient quantitative short-term and middle-term forecasting for a wide class of information security related incidents, (2) it describes a way how to adapt already well-known auto regression models to another class of time series — ISRI time series and thus to a new subject field -information security, and (3) it points out attention of security practitioners to gaps of currently available approaches for ISRI forecasting and proposes solutions to cover these gaps.

The rest of the paper is organized as follows. In Sect. 2 we provide a survey of related research activities. In Sect. 3 we describe specifics related to information security, discuss potential problems of forecasting in this field and propose our solutions to these problems. In Sect. 4 we present the proposed approach. Section 5 is devoted to example of application of the proposed approach to real ISRI time series — amount of spam messages and evaluation of its performance characteristics. Finally we give summary in Sect. 6.

2. Related Studies

Though the problem of ISRI forecasting is a very important one, it was not covered by research efforts until the very recent time. First attempts to perform forecasting in this field were related to a rather similar problem — forecasting of yet undiscovered vulnerabilities in software and the distribution of their severity levels basing on statistical reports, previously published by vendors or security experts. This task helps to provide an indirect estimation of reliability of a given software solution. Such research work was partially performed in [4] though the authors just made a seasonal analysis of vulnerability discovery process concluding that a forecasting is intended to be their future work.

One of the first attempts to make ISRI forecasting was made in [5] where the authors proposed application of Bayesian inference for forecasting a trend of attack signatures, defined by probabilities of increase or decrease of event counts related to signatures of a given type of attack. The differences between our approach and approach from [5] are as follows: (1) difference in input data and outcomes. That is in [5] the authors consider not the actual number of concrete attacks, but number of attack signatures related to a given type of attack and they actually forecast a number of attack signatures. Moreover as they use data aggregation, it could result in information loss problem (see Sect. 3.2). In our approach we consider and forecast a number of information security related incidents (which includes number of all attack efforts) and do not use aggregation of data. (2) Difference in statistical considerations. That is as several different attacks could have same signatures, indirect forecasting of number of attacks by counting a number of attack signatures (instead of number of attacks) as proposed in [5], leads to problem of non-homogeneity of data in input time series (as number of attack signatures time series can contain data related to many different attacks). Thus such forecasting method can lead to potentially ambiguous situations like the following one: suppose there are two attacks: A and B which can be detected by presence of the same attack signature X, but they differ in the frequency of occurrence of X (for example, in a given fixed time period, X occurs one time for attack A, but two times for attack B). Thus if there is a forecast that X is going to appear two times in the next time period, there is no way to distinguish if it means two occurrences of attack A or one occurrence of attack B. In our approach we pay especial attention to homogeneity of original time series and propose a solution to this problem in Sect. 3. (3) Difference in type of forecasting. That is in [5] the authors are forecasting the trend (increase or decrease) of event counts only for the next time period. We forecast the precise number of security incidents for a set of sequential time periods. So our approach can forecast not only the trend but also the time periods with and without ISRI. (4) Difference in initial assumptions. That is in [5] the authors assume that there could be only two possible types of fluctuations in event counts of attack signatures: cyclic change and rapid increase/rapid decrease. In our approach we do not make any preliminary assumptions regarding behavior of ISRI time series, we assume only that these ISRI are unambiguously detectable (see Sect. 3).

Another interesting approach is proposed in [6] where the authors implemented Maximal Overlap Discrete Wavelet Transform to predict amount of network traffic several hours ahead in the Incident Forecast Engine. The differences between our approach and approach from [6] are as follows: (1) difference in input data and outcomes. Again in [6] the authors consider not the actual number of concrete attacks, but number of attack signatures (in this case some designated IP packets) and again they use aggregation. In our approach we consider direct number of ISRI and do not use aggregation. (2) Difference in statistical considerations. That is in [6] the authors decided considering a number of designated IP packets as an indirect method of detection of a potential ISRI, though again without assurance of explicit interrelation of these two values such assumption can lead to non-homogeneity and potentially ambiguous situations considered above (as it is clear that there can be many different attacks characterized by sending packets to a given TCP port). Thus again direct application of approach from

[6] (forecasting number of attacks by forecasting number of attack signatures) is not always possible. As was already mentioned above, in our approach we pay especial attention to the problem of homogeneity and propose a solution in Sect. 3. (3) Difference in usability. The approach presented in [6] requires non-standard software, though our approach can be executed on a wide spread set of software used for financial time series analysis.

Another remarkable attempt was made in [7] where the authors tried to apply an ARIMA (autoregressive integrated moving average) model to make middle term forecasting for several types of security incidents. Unlike our model, which is mainly intended to create short — term forecasts estimating most probable time periods where upcoming incidents can occur (of course, our model can also be used to make middle term forecasts), in [7] the authors are trying to forecast not the precise amount of upcoming incidents, but an average amount of such incidents in an upcoming week. Moreover in order to do it they use aggregated data. But, as the way the authors in [7] were collecting, storing and analyzing the data does not fit special conditions, originally introduced in our model, the applicability and the efficiency of forecasts of their approach are lower than of our approach. Another difference is that in addition, we proposed to test ARCH/GARCH models in order to avoid problems caused by heteroskedasticity (as in general, it is not possible just to assume the homoskedasticity of data related to ISRI and to apply just ARIMA model).

To sum up, we suppose that that there should be some other efforts to solve the task of ISRI forecasting, but most likely the results of these efforts were negative like in [8] because of the problems discussed in the next section.

3. Problems of Information Security Related Incidents Forecasting

In this section, we consider problems of ISRI forecasting process and propose solutions to these problems introducing rules of collection, count and storage of ISRI time series called special conditions. We also discuss the role of this rules and their contribution to forecasting process.

3.1 Problem of Information Security Related Incidents Detection

In order to perform an analysis of ISRI time series for a given type of incidents we must somehow obtain initial data. Thus there must be a detection criterion which for any fixed moment of time and a given type of ISRI can unambiguously determine if there is/was such an incident at this moment of time or not. Thus it is possible to divide all ISRI into two different groups: those which can be unambiguously detected by some criterion either right at the moment of occurrence or some time later (we call such ISRI as unambiguously detectable) and those which can be detected with less than 100% warranty or can not be detected at all (we call such ISRI as unambiguously undetectable). Thus the key features of unambiguously detectable ISRI are that an amount and a time of occurrence of such incidents can be measured precisely. Examples of unambiguously detectable ISRI are: an unscheduled shutdown of a server, a fraudulent transaction in a bank account, an attempt of robbery or a theft of a good in a shop, a user's account locked due to multiple incorrect password input, a spam message in a mailbox etc. As to unambiguously undetectable ISRI, one example is a presence of a new computer virus in a given workstation. The reason is that it is not always possible to detect a new virus if its signature is not yet familiar to any antivirus software and it has not yet made any noticeable activity (for example, it is designed to make such harmful activity at a fixed future date or it just makes this workstation a part of a botnet). Another example is an occurrence of a special network attack. The reason is that sometimes it is difficult to distinguish an attempt to perform a network attack in a network traffic flow unless it leads to any significant damage.

But even though some given class of ISRI must be considered as unambiguously undetectable at the current time, there could be a solution for unambiguous detection in future.

Further in this paper we assume that our approach is applied to any unambiguously detectable ISRI, though, of course, it is possible to try to apply it to unambiguously undetectable ISRI as well. But there is no way to confirm effectiveness of such forecasts because if we can not detect something every time when it occurs, we can not count all of its occurrences it and thus make forecasts. But if there is a criterion which makes unambiguous detection possible (like approach proposed in [6] for some network attacks) forecasting of such ISRI becomes effective as well.

3.2 Problem of Non-homogeneity

In order to make an effective forecasting, initial data in time series must come from same probabilistic distribution [9] (otherwise samples from different distributions will suppress each other). This problem is called the problem of homogeneity of original data. In general it is addressed by means of logical considerations but in ISRI forecasting it must be treated in a special way because of specifics related to this subject area. This specifics results in two additional subtasks which must be solved to build an effective forecasting methodology:

- Accurate measurement of amount of ISRI occurrences. A factor which typically leads to inaccurate measurement is an intention to use indirect estimation of amount of ISRI by measurement of related incidents' signatures. This could happen either because of complete infeasibility to measure an amount of incidents (like for some of unambiguously undetectable ISRI) or because by some reasons it is easier to measure an amount of signatures but not an amount incidents. And as we already mentioned above, as such signatures could be correlated with many different ISRI, such measurement potentially results in non-

homogeneity of collected data sample. In our approach we avoid this situation by considering only direct measurement of unambiguously detectable ISRI.

- Loss of homogeneity due to unnoticeable changes of distribution parameters. As we showed in [3], this situation happens because of general properties of all modern IT landscapes: huge heterogeneity (a typical landscape incorporates dozens of solutions from different vendors) and extremely rapid changes in its structure and each of its elements. Thus, as both of the components causing/mitigating failures or ISRI are permanently changing, it makes difficult to directly apply traditional statistical methods to evaluate distribution function of some failures or incidents. A possible solution for this task is temporary stabilization of the state of an IT landscape. In [10] we presented a way how to achieve such stabilization at least for some time by proposal to perform continuous collection of ISRI statistics only for those elements which are not changing their state in terms of being vulnerable to some specific malicious actions — threats leading to ISRI (but if the state has changed, previously collected statistics must be discarded). But as generally state must be changed rather often (as it is impossible just to sit and wait collecting statistics while your systems are being attacked or are expiring continuous technical failures), collection of appropriate amount of statistical data for further analysis is not a trivial task. And if there is a lack of statistical data, there is a lack of reasonable forecasts. In this paper we improve previous proposal by introduction of a rule of collection of ISRI time series data. We call it special condition 1 (SC-1): we collect data related not only to occurred incidents, but also to those which were successfully mitigated by all implemented countermeasures related to selected type of incidents. It means that for a given module or system (for which the statistics is collected) and for a selected type of ISRI we must understand all mechanisms which protect this module from this selected type of ISRI and we must count all attempts (both occurred and prevented) to damage it. An amount of all attempts is called an amount of potentially dangerous events. In terms of mathematical statistics it means that even if the module/system is changed, we can continue collection of data as we are measuring an amount of potentially dangerous events (in other words, original distribution of a threat related to this ISRI [11]) and it does not depend on a current configuration or set of countermeasures. Afterwards we can use this data for forecasting of both an amount of potentially dangerous events and amount of ISRI. In order to forecast an amount of ISRI we make a correction subtracting from a forecasted amount of potentially dangerous events amount of events, prevented by a newly introduced countermeasures. Thus SC-1 helps us to save significant amount of collected data which, if SC-1 is not met, must be discarded.

3.3 Problem of Information Loss

As for unambiguously detectable ISRI an amount of such incidents is a measurable variable, in terms of signal processing theory it is possible to consider it as an output generated by some random process with some properties. At the same time currently it is very popular to store and to process not the full amount of collected statistical data, but some aggregation of it. For example instead of amounts of incidents for a given hour, minute and second, there will be the whole amount of incidents per hour or per day or even per week. Thus instead of considering continuous-time process, giving mostly zeros but sometimes non-zero values, we are making it discrete by sampling its values in time domain and afterwards additionally aggregate these discrete values taking into account only non-zero values. From the theoretical point of view both of these steps lead to potential loss of information [12]. According to the sampling theorem [13], in order not to loose the information, the sampling frequency must be at least 2 times higher then the Nyquist frequency of the signal we are observing. So only having the Nyquist frequency of the generating process we can select the proper time interval when to measure amount of incidents. Of course most of the measured values will be zeros, but the problem is with the right amount of zeros between informative non-zero values. This fact, in turn, means that we must have some prior information about the Nyquist frequency of the process we are observing. Unfortunately this is usually impossible for processes generating ISRI.

In order to overcome this problem we introduce a rule of count and storage of ISRI time series data. We call it special condition 2 (SC-2): incident statistics should be collected and stored at the highest possible echelon of accuracy over the time domain. It should not be aggregated. This rule helps avoiding unnecessary aggregation in order to be able to distinguish time intervals between different incidents. Such a rule helps us to collect all possible information about a time of occurrence given technical limits of precision of a device used to measure time. Anyhow the amount of data can later be reduced if forecasting works good on the aggregated samples as well.

3.4 Role of Special Conditions in Forecasting of ISRI

Special conditions SC-1 and SC-2 introduced above should be understood as rules which must be used by an implementation team to organize a collection and processing of ISRI time series data.

The contribution of special conditions to ISRI forecasting process is as follows:

- Application of SC-1 together with direct measurement of unambiguously detectable ISRI allows collection of larger amounts of homogenous data which is required for forecasting. Moreover, as this data reflects the distribution of a threat related to selected incident, obtained

forecasts can be adjusted for simulation purposes—prediction of an amount of occurred ISRI with a specified configuration of countermeasures.
- Application of SC-2 protects from information loss caused by discrete sampling and aggregation of collected data.

Special conditions do not imply strict limitations and can be easily applied to any unambiguously detectable ISRI as according to Sect. 3.1 for any of such incidents we can precisely detect its occurrence (thus covering SC-1), measure time of occurrence and amount of such incidents (thus covering SC-2). Of course in order to perform such measurement there must be a detection criterion (which exists for any unambiguously detectable ISRI). In general, detection criteria for occurrence of ISRI are observations of some unclaimed actions. For prevented potentially dangerous events such criteria are usually built directly in countermeasures. Thus selection of a detection criterion is not a difficult task for a security specialist. For example:

- Incident: arrival of a spam message. Countermeasure: antispam system. Amount of prevented ISRI: number of filtered spam messages. Amount of occurred ISRI: number of non-filtered spam messages in mailboxes (can be counted by users).
- Incident: data loss occurred due to malware. Countermeasure: backup procedure. Amount of prevented ISRI: number of times when the data was recovered from a backup medium. Amount of occurred ISRI: number of times when the data was not recovered.
- Incident: fraudulent transaction for a credit card. Countermeasure: policy of a payment system. Amount of prevented ISRI: number of rejected transactions by policy criteria (like incorrect PIN2 code). Amount of occurred ISRI: number of disputed transactions.

To sum up, introduction of SC-1 and SC-2 helps to mitigate problems of non-homogeneity and information loss and to achieve the highest possible accuracy of forecasting provided by selected underlying auto regression mathematical models.

The novelty of these two conditions is that they: (1) help to understand and to correct gaps introduced by currently popular assumptions in the field of ISRI forecasting, thus improving the accuracy of such forecasting, and (2) allow a practical deployment of forecasting methodology without a need to deeply learn underlying statistical theoretical considerations. This makes a deployment of our approach easier as security experts may have no deep knowledge in mathematical statistics.

4. Proposed Approach

In this section, we provide the description of the proposed approach. Its block scheme is presented on Fig. 1.

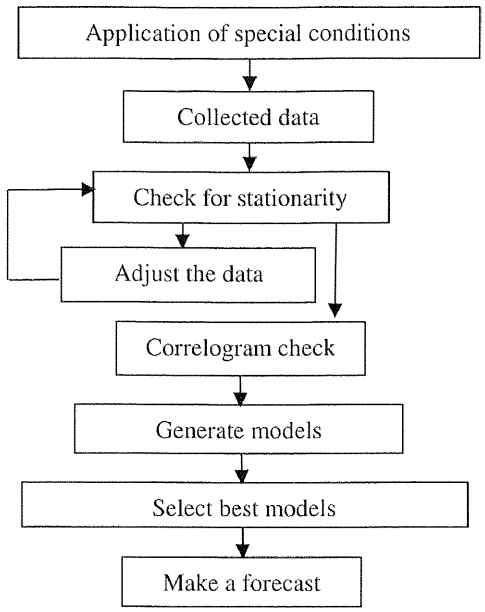


Fig. 1 Proposed approach scheme.

4.1 Collection of Data

At this stage ISRI time series data must be collected with adherence to special conditions SC-1 and SC-2, so

- Data must be collected with the highest possible accuracy over the time domain and it should not be aggregated.
- Data must be collected for all potentially dangerous events which could lead to ISRI if there is no any countermeasure in place. Afterwards the forecast must be corrected according to the average contribution of all available countermeasures for a given type of ISRI.

4.2 Stationarity Check and Data Adjustment

During these steps we must perform any of possible tests to check the stationarity of data. This step is an essential one as in order to check the significance of the coefficients in the auto regression model to be used for forecasting (introduced in detail below) the variance of the observed data must be estimated and it must be finite [14]. But as it is finite for stationary processes only, the time series to which we are trying to apply these models must be stationary. The most popular stationarity tests are Augmented Dickey Fuller test or Philip Peron tests [14], [15]. If data is stationary, we can go to the next step. If not, we must perform data adjustment taking the difference of order d ($d = 1, 2, 3 \ldots$) and perform the check again for adjusted time series till we find d when the data becomes stationary [16] (see Eq. (1)). Here y_t is initial ISRI time series at time t, $y_t^{(d)}$ is an integrated time series order d. We will use this new adjusted time series for all next steps.

$$y_t^{(1)} = y_t - y_{t-1}$$

$$\begin{aligned} y_t^{(2)} &= y_t^{(1)} - y_{t-1}^{(1)} = y_t - y_{t-1} - y_{t-1} - y_{t-2} \\ &= y_t - 2y_{t-1} + y_{t-2} \\ &\cdots \\ y_t^{(d)} &= y_t^{(d-1)} - y_{t-1}^{(d-1)} \end{aligned} \tag{1}$$

4.3 Correlogram Check

In order to generate auto regression models at the next step, we must find the maximum significant orders of autocorrelation (AC) and partial autocorrelation (PAC) functions. The theoretical background of this technique is described in [15]. In short, we must select the highest value of an AC coefficient which is significant in terms of selected confidence level for the maximum order of autoregressive (AR) part and the highest level of PAC coefficient which is significant in terms of selected confidence level for the maximum order of moving average (MA) part.

4.4 Generate Models

There are several models which should be generated. The detailed description of all these models is provided in [16].

First of all it is necessary to try a simple autoregressive (AR) model, defined in Eq. (2). Here c is the expectation of y, p — order of the autoregressive process, α_i are estimated coefficients and ε_t is an error term. In this model error term is assumed to be distributed as $N(0, \sigma^2)$.

$$y_t = c + \sum_{i=1}^{p} \alpha_i y_{t-i} + \varepsilon_t \tag{2}$$

Then autoregressive models with moving average (ARMA) of order (p, q), defined in Eq. (3), should be generated. Here p is the order of the autoregressive process, q — order of the moving average, α_i are estimated coefficients for autoregressive part, β_i are coefficients for moving average part.

$$y_t = c + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} + \varepsilon_t \tag{3}$$

If original ISRI time series, y_t was non-stationary and ARMA (p, q) model was applied to stationary adjusted time series, $y_t^{(d)}$, such a model is considered as a generalization of ARMA model to autoregressive integrated moving average (ARIMA) model of order (p, d, q) where d is order of adjusted time series [16]. Thus if ARIMA (p, d, q) model is applied to y_t it turns into ARIMA $(p, 0, q)$ which by definition is ARMA (p, q) [16].

As it was pointed out above, ARMA and thus ARIMA model relies on the assumption that an error term has a constant variance. But as there is no way to prove this property for processes generating ISRI, we must consider more general situation. Such cases can be modeled by the following models which should be applied right after ARIMA models.

An autoregressive conditional heteroskedasticity (ARCH) model considers the variance σ^2 of the current error term ε_t to be a function of the previous time periods' error terms (see Eq. (4)). Here α_i are estimated coefficients and w is the order of ARCH term. It is commonly used in modeling time series that exhibit time-varying volatility clustering, i.e. periods of swings followed by periods of relative calm.

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{w} \alpha_i \varepsilon_{t-i}^2 \tag{4}$$

ARCH can be generalized to make the conditional variance a function of past conditional variances in which case the model is called generalized autoregressive conditional heteroskedasticity (GARCH). GARCH (u, w) is defined in Eq. (5), where u is the order of GARCH term and w is the order of ARCH term.

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{w} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{u} \beta_j \sigma_{t-j}^2 \tag{5}$$

The way all these models are generated is as follows [16], [17]: according to the maximum order of AR and MA parts obtained at the correlogram check stage, we start generating models with all possible combinations of AR and MA parts. The total number of ARIMA models is defined in Eq. (6) as multiplication of sum of the corresponding binomial coefficients.

$$Number_{ARIMA} = \sum_{i=1}^{p} \binom{p}{i} x \sum_{i=1}^{q} \binom{q}{i} \tag{6}$$

Afterwards, we perform the next step for all generated models. And for the best of them in terms of Sect. 4.5 we perform generation of ARCH and GARCH models. The methodology of selection of orders of ARCH and GARCH terms is described in detail in [18]. In short, we must perform Ljung-Box test for randomness applied to the autocorrelation values of ε^2. The ARCH/GARCH order will be given by the number of lags in the Ljung-Box test at which the null hypothesis needs to be rejected. As it is a probabilistic test, it is recommended to take a 5% significance level. Here it is necessary to note, that most of modern statistical software allows automatic fitting of ARCH/GARCH models, so in order to use these models, there is no need to know all the mathematical background of its operation.

The best models are again selected according to the procedure in Sect. 4.5.

4.5 Selection of Best Models

During this step it is necessary to select models with minimum values of Akaike and Schwartz statistics. Related theoretical considerations can be found in [16]. If in short, the coefficients in ARIMA models are estimated according to the maximum likelihood criteria. But these criteria increase together with number of dependant variables. So in order to correct this fact, Akaike and Schwartz criteria were introduced. As in general each of these criteria can select a

Table 1 Possible situations.

Forecasting statement	Real result	
	Incident occurred	Incident did not occur
Incident will happen (c-1)	Correct forecast Situation A	Error type II (false negative or β) Situation D
Incident will not happen (c-2)	Error type I (false positive or α) Situation B	Correct forecast Situation C

different model as the best one, we recommend taking the model with the smallest criteria value.

4.6 Making a Forecast

Having obtained the best model from the previous stages, we apply it to make a forecast (a forecasting statement for each given time period). To do this, we should just put previously observed data in the obtained equation and start generating values of the forecasting function. As all the involved models are standard models for time series analysis, the variance of forecasted values can be estimated by equations related to these models, which can be found, for example in [16] or [19]. We are not providing these equations here as they do not have a direct link with the proposed methodology.

As a result, for each time period t (defined by the period of sampling used in SC-2) we obtain a value of forecasting function f_t and its variance σ_t^2. Afterwards we make a forecasting statement according to the following procedure, based on approach presented in [20]:

- We define two classes: (c-1) ISRI will happen at time period t, and (c-2) ISRI will not happen at time period t. To make a forecasting statement at a time period t we need to make a selection of a class for this time period.
- We define a classification rule to make a class selection by setting a threshold value T_t for values of forecasting function (see Eq. (7)). In our approach we decided to select a constant threshold value for any t. Here x is a parameter which together with threshold value should be estimated as described in [21].

$$\begin{aligned} &\text{If } f_t \pm x * \sigma_t \geq T_t, \text{ select } c\text{-}1 \\ &\text{If } f_t \pm x * \sigma_t < T_t, \text{ select } c\text{-}2 \end{aligned} \tag{7}$$

4.7 Accuracy of Forecasting

While making a forecasting statement we can make either a correct statement (which results in correct forecast) or an incorrect statement (which results in forecasting error). According to [22] there could be two different types of errors (see Table 1). Figure 2 shows graphical representation of possible situations.

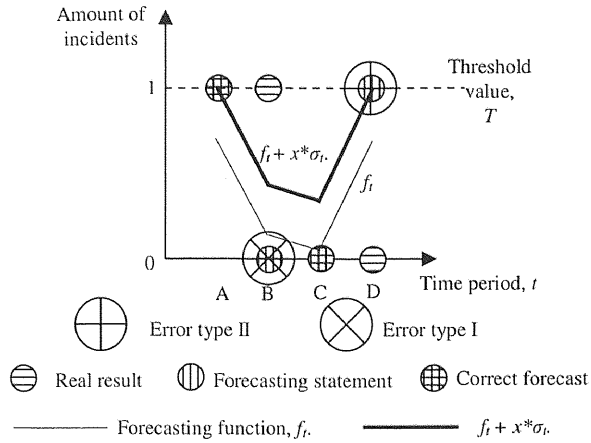


Fig. 2 Graphical representation of forecasting accuracy evaluation.

For ISRI forecasting we should understand consequences of these errors as follows:

- Error type I (false positive): we decided not to prepare for an incident at a given time period t, but it happened and we suffered from loss caused by incident.
- Error type II (false negative): we prepared for an incident at a given time period t, but it did not happen and we suffered from loss caused by erroneous preparation.

An ideal, but usually unachievable case is when both α and β are zeros [23]. In mathematical statistics a task to make a classification rule sounds like "with a given level of false positive (or false negative) develop a selection rule which provides minimum possible value of alternative error type." In information security selection of a classification rule must be made by security manager as it can be different for different types of ISRI. For example, for access control a rule usually is: with a given error type II (when we refuse to legal users), minimize error type I (when we let a fraudster go inside). But for spam detection system a rule usually is: with a given error type I (when we let a spam message go inside a mailbox), minimize error type II (when we drop a proper message).

5. Case Study and Evaluation

In this section, we provide an application of the proposed approach to forecasting of spam messages. We also evaluate its performance characteristics and demonstrate contribution of special conditions to forecasting process.

For making forecasts we used spam messages statistics collected during 68 days (07/2009–09/2009) in a small enterprise which is located in Europe and is working in marketing business. This data was collected with adherence to special conditions (introduced above) and was stored on the hourly basis. For the current example this was sufficient as the rate of spam messages was lower. If there were several messages in one single hour, we would have to switch to a shorter period up to the shortest, supported by measuring equipment.

Explanation of all metrics and abbreviations presented in tables below can be found in [24] and [25] and [16].

5.1 Case Study

5.1.1 Contribution of Special Conditions

As it is possible to count precise number of occurred incidents (amount of spam messages) and precise time of incident's occurrence (time of arrival), an arrival of a spam messages belongs to unambiguously detectable ISRI. Thus application of special conditions to this type of ISRI is feasible.

In order to illustrate contribution of special conditions, we made adjustment of originally collected data by simulation that SC-1 and SC-2 were not used.

To show contribution of SC-1 we simulated deployment of the following countermeasure: antispam system with the policy to filter all messages from free public mail servers (yahoo, hotmail etc.) and from each address from other mail servers which were previously used to deliver spam (to simulate a kind of adaptive learning). We simulated that this new countermeasure was added after 1 day of initial observations. And as almost every day there was a spam message from non free mail server, the list of filtered addresses was frequently changing. Thus when we applied this countermeasure to original data it resulted in dropping of more than 95% of it. Thus if SC-1 is not met and we count only amount of occurred ISRI, we will loose more than 95% of collected data.

In order to show contribution of SC-2 and the loss of information occurred as the result of aggregation, the original ISRI time series data was aggregated to obtain daily values. The Fourier spectrum obtained by calculation of Discrete Fourier Transformation for both time series is displayed on Fig. 3 (presented as a line plot, covering value for each frequency). It is clear that spectrums are significantly different from each other. The aggregated data spectrum is much smoother and also does not have some of original harmonics.

Results of Augmented Dickey Fuller test for stationarity of original and aggregated data are presented in Table 2 and Table 3 accordingly. In both cases the data is stationary. So we can proceed with the correlogram analysis.

Correlograms for aggregated and original data are pre-

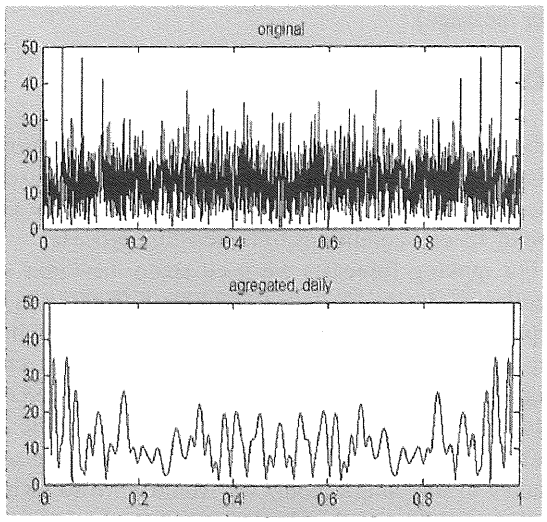


Fig. 3 Fourier transformation for original and aggregated data.

Table 2 Augmented Dickey Fuller test for original data.

Augmented Dickey-Fuller Test Equation			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-19.19870	0.0000
Test critical values:	1% level		-2.566382	
	5% level		-1.941018	
	10% level		-1.616569	
Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**Prob.**
INCIDENT(-1)	-0.735758	0.038323	-19.19870	0.0000
D(INCIDENT(-1))	-0.164461	0.033041	-4.977441	0.0000
D(INCIDENT(-2))	-0.090522	0.024698	-3.665224	0.0003

Table 3 Augmented Dickey Fuller test for aggregated data.

Augmented Dickey-Fuller Test Equation			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-7.419528	0.0000
Test critical values:	1% level		-3.531592	
	5% level		-2.905519	
	10% level		-2.590262	
Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**Prob.**
INCIDENT(-1)	-0.927221	0.124970	-7.419528	0.0000
C	2.863610	0.432371	6.623039	0.0000

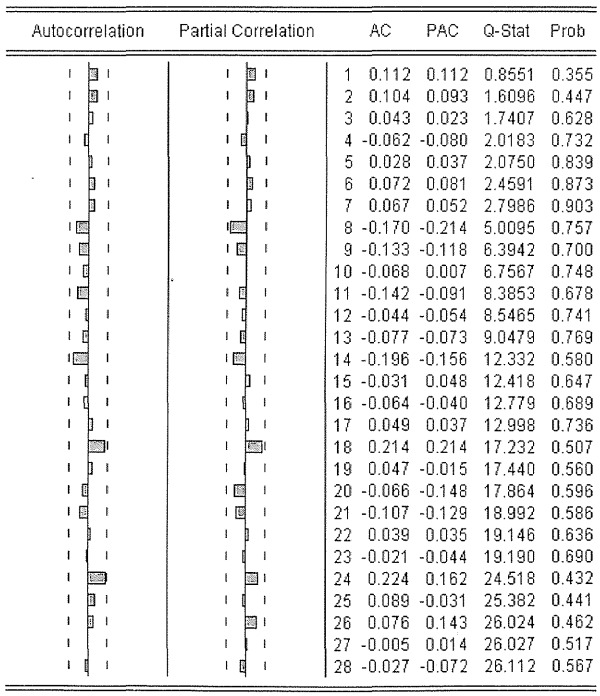


Fig. 4 Correlogram of aggregated data.

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.006	0.006	0.0539	0.816
		2	-0.018	-0.018	0.5778	0.749
		3	-0.004	-0.004	0.6019	0.896
		4	-0.027	-0.027	1.7846	0.775
		5	-0.032	-0.032	3.4275	0.634
		6	-0.018	-0.018	3.9356	0.685
		7	0.006	0.005	3.9923	0.781
		8	-0.046	-0.048	7.4709	0.487
		9	-0.037	-0.038	9.6816	0.377
		10	0.006	0.002	9.7340	0.464
		11	-0.017	-0.020	10.231	0.510
		12	-0.013	-0.016	10.499	0.572
		13	-0.018	-0.023	11.005	0.610
		14	0.020	0.016	11.670	0.633
		15	-0.013	-0.016	11.948	0.683
		16	0.039	0.036	14.428	0.567
		17	-0.008	-0.016	14.545	0.628
		18	-0.023	-0.023	15.392	0.635
		19	-0.013	-0.015	15.684	0.678
		20	0.024	0.023	16.663	0.675
		21	-0.004	-0.007	16.690	0.730
		22	0.015	0.015	17.050	0.761
		23	0.072	0.069	25.598	0.320
		24	0.091	0.093	39.206	0.026
		25	0.006	0.012	39.260	0.035
		26	0.020	0.023	39.908	0.040
		27	-0.018	-0.013	40.447	0.047
		28	0.001	0.015	40.450	0.060
		29	0.002	0.013	40.456	0.077
		30	-0.017	-0.014	40.939	0.088
		31	-0.008	0.002	41.038	0.107
		32	-0.017	-0.005	41.531	0.121
		33	0.003	0.013	41.544	0.146
		34	0.007	0.012	41.636	0.173
		35	-0.030	-0.025	43.174	0.162
		36	0.026	0.027	44.313	0.161
		37	-0.002	0.001	44.318	0.190
		38	-0.025	-0.028	45.386	0.191
		39	-0.020	-0.025	46.058	0.203
		40	0.027	0.018	47.280	0.200
		41	-0.025	-0.021	48.319	0.201
		42	-0.020	-0.017	49.008	0.213
		43	0.013	0.006	49.276	0.237
		44	0.022	0.017	50.088	0.245
		45	-0.030	-0.032	51.592	0.232

Fig. 5 Correlogram of original data.

sented on Fig. 4 and Fig. 5 accordingly. As we can see form correlogram for aggregated data, there are no neither AC nor PAC coefficients which are higher then significance level. This means that obtained data has properties of white noise and there is no way to make any other forecast except to state that the data will fluctuate around its mean value. As to the non-aggregated original data, there are rather many significant AC and PAC coefficients.

5.1.2 Generation of Models and Forecasting

According to Sect. 4, now we need to generate auto regression models. In order to do it, we again use correlogram of original data (Fig. 5).

The highest significant coefficient's number is 169. So it means that we must generate all combinations with orders p and q of AR and MA up to 169. The result of the best model selection according to both Akaike and Schwartz criteria is shown in Table 4. According to z-Statistics values,

Table 4 Best model.

Variable	Coefficient	Std. Error	z-Statistic	Prob.
AR(24)	0.268798	0.046207	5.817249	0.0000
AR(72)	0.339394	0.049067	6.916916	0.0000
AR(168)	0.241110	0.044762	5.386447	0.0000
AR(169)	0.122613	0.022188	5.526029	0.0000
MA(24)	-0.227439	0.049841	-4.563296	0.0000
MA(72)	-0.319726	0.051677	-6.186990	0.0000
MA(168)	-0.273580	0.049981	-5.473688	0.0000
MA(169)	-0.121631	0.032083	-3.791145	0.0001
Variance Equation				
Constant	0.013055	0.001343	9.722474	0.0000
ARCH(-1)	-0.007839	0.001332	-5.884501	0.0000
GARCH(-1)	0.816526	0.024004	34.01561	0.0000
GARCH(-2)	0.925229	0.013612	67.97274	0.0000
GARCH(-3)	-0.839327	0.020838	-40.27797	0.0000

R-squared	0.035147	Mean dependent var	0.128503
Adjusted R-squared	0.027162	S.D. dependent var	0.360348
S.E. of regression	0.355420	Akaike info criterion	0.755476
Sum squared resid	183.1690	Schwarz criterion	0.802466
Log likelihood	-539.6304	Hannan-Quinn criter.	0.773003
Durbin-Watson stat	2.028150		

all the coefficients are significant.

It means that our process is described as a sum of AR(24), AR(72), AR(168), AR(169), MA(24), MA(72), MA(168) and MA(169). As the data was observed on the hourly basis, it means that amount of spam in a given hour depends on amounts of spam one day before (AR(24) and MA(24)), three days ago (AR(72) and MA(72)) and about a week ago (AR(168), AR (169) and MA(168), MA(169)).

This dependency in the spam generating process shows us its internal characteristics. In this case it can indicate that a spam agent which is sending us messages has a regular schedule.

As there are several significant ARCH/GARCH terms, it means that the variance of each error term is changing over the time and can be represented as a sum of ARCH and GARCH terms. According to Table 4 this process is GARCH order (3, 1). It means that the data is heteroskedastic.

Finally we can proceed with forecasting. An example of a real short-term forecasting with obtained model on a one day period ahead is shown on Fig. 6. Here forecasting statements are compared with real result values measured in addition to those initial 68 days considered for building the model. As sampling period in this case study was set to one hour, in total we have 24 time periods for one day ahead forecasting. Here we obtained constant threshold value T_t=1 and parameter x=2 (see Eq. (7)), standard error values were calculated according to procedure presented in [16].

We can also use Fig. 6 to perform rough estimation of accuracy of the proposed model: we can clearly see that the model has precisely indicated 2 of 3 hours where we should expect the next spam message (so there is only one occurrence of error type I) and all safe hours without spam were successfully detected (there are completely no occurrences of error type II).

A middle-term forecast can be obtained by making

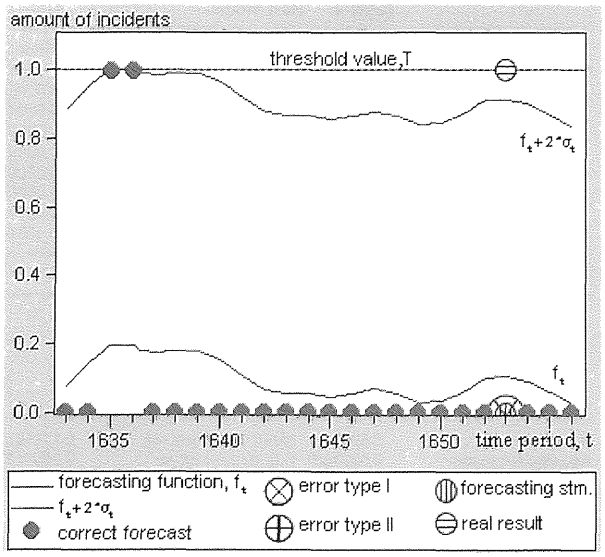


Fig. 6 Forecasting result.

Table 5 Model comparison and performance characteristics.

	ARIMA	ARIMA and ARCH	ARIMA, ARCH and GARCH
RMSE	0.3565	0.3565	0.3560
FPR	0.12	0.11	0.10
FNR	0	0	0

short-term forecast for a larger sequence of hours and then calculating mean and variance for this sequence taking into account that each value in the sequence has its own variance because of heteroskedasticity.

5.2 Evaluation

In order to evaluate the proposed model, we are using a cross validation technique presented in [26]. With this method, all observed data is divided into two parts: training data and test data. Training data is used as an initial data to start forecasting and test data is supposed to be yet unknown. After the forecasting procedure, forecasted values are compared with test data. There are no recommendations about the length of these parts (but the shorter is the training data the higher could be possible errors in forecasted data), so we selected 15% of all data as training data and 85% as test data. The quality of forecasting can be estimated by parameters introduced in [16].

Root Mean Squared Error (RMSE) allows estimating which forecasting model is the best one for a given set of data. But more critical task is to compare models between deferent sets of data. For such purpose we decided to use False Positive Rate (FPR) and False Negative Rate (FNR). Comparison of different models is presented in Table 5, which shows that consideration of ARCH and GARCH terms improved the accuracy of forecasting.

Unfortunately, the authors in [7] provide neither their data sets nor FPR/FNR values, so the direct comparison is not possible. But if to apply approach presented in [7] to the type of ISRI we were analyzing (thus collecting data without adherence to special conditions, introduced in our approach) we will have to use aggregated daily statistics (as it is proposed in [7]) which leads us to the series with properties of white noise (see Sect. 5.1), and, in turn, to impossibility to make a meaningful short-term forecast. To conclude, application of approach presented in [4] lead to negative results while our approach helped to resolve this problem.

6. Conclusion

In this paper we have proposed an approach to make a short-term and a middle-term forecasting for a wide class of ISRI — unambiguously detectable ISRI.

In order to do that, we have considered specifics related to ISRI forecasting process and investigated its theoretical background. As a result we proposed solutions to these problems for unambiguously detectable ISRI by introduction of rules of data collection and storage (special conditions) adherence to which improves forecasting in this subject field.

We have confirmed the applicability and effectiveness of the proposed theoretical model to real data by an example of application of our approach to statistics of spam messages. Finally we have made an evaluation of performance characteristics of the proposed approach which confirmed its robustness.

The proposed approach is significantly different from previously available approaches by the way of collection of initial data, its further storage and underlying statistical considerations which makes model applicable in more situations, and by extension of underlining mathematical models to more general heteroskedastic models which help to increase accuracy of forecasting.

The proposed approach improves quality and applicability of previously available approaches and thus can be useful for practical applications.

Acknowledgments

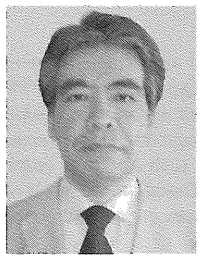
We wish to express our highest gratitude to Professor Hiroe Tsubaki, Director of Risk Analysis Research Center in the Institute of Statistical Mathematics, for his valuable comments.

References

[1] D. Waters, Operations Management, K Press, 1999.
[2] M. Armstrong, A handbook of management techniques, K Press, 2001.
[3] A. Romanov and E. Okamoto, "An approach for designing of enterprise IT landscapes to perform quantitative information security risk assessment," Proc. International Conference on Security and Cryptography (SECRYPT), pp.313–318, 2009.
[4] H.C. Joh and Y.K. Malaiya, "Seasonal variation in the vulnerability discovery process," Proc. International Conference on Software

Testing Verification and Validation, pp.191–200, 2009.
[5] C. Ishida, Y. Arakawa, I. Sasase, and K. Takemori, "Forecast techniques for predicting increase or decrease of attacks using Bayesian inference," Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), pp.450–453, 2005.
[6] D. Inoue, K. Yoshioka, M. Eto, M. Yamagata, E. Nishino, J. Takeuchi, K. Ohkouchi, and K. Nakao, "An incident analysis system NICTER and its analysis engines based on data mining techniques," Lect. Notes Comput. Sci., vol.5506/2009, pp.579–586, June 2009.
[7] E. Condon, A. He, and M. Cukier, "Analysis of computer security incident data using time series models," Proc. International Symposium on Software Reliability Engineering, pp.77–86, 2008.
[8] E. Rescorla, "Is finding security holes a good idea?," IEEE Security and Privacy, vol.3, no.1, pp.14–19, Jan.-Feb. 2005.
[9] A. Harvey, Forecasting, structural time series models and the Kalman filter, Cambridge University Press, 1991.
[10] A. Romanov and E. Okamoto, "A quantitative approach to assess information security related risks," Proc. 4th International Conference on Risks and Security of Internet and Systems (CRISIS), pp.117–123, 2009.
[11] M. Whitman and H. Mattord, Principles of Incident Response and Disaster Recovery, Thomson Course Technology, 2006.
[12] K. Morita, Applied Fourier Transform, IOS Press, 1995.
[13] C.-H. Chen, Signal Processing Handbook, Marcel Dekker, 1988.
[14] R. DeFusco, D. McLeavey, and J. Pinto, Quantitative investment analysis, John Wiley & Sons, 2007.
[15] R. Shumway and D. Stoffer, Time Series Analysis and its Applications, Springer-Verlag, 2000.
[16] C. Brooks, Introductory Econometrics for Finance, Cambridge Press, 2008.
[17] D. Hanssens, L. Parsons, and R. Schultz, Market Response Models: Econometric and Time Series Analysis, Kluwer Academic Publishers, 2001.
[18] R. Engle, ARCH: Selected readings, Oxford University Press, 2005.
[19] R. Yaffee and M. McGree, An Introduction to Time Series Analysis and Forecasting: With Applications of SAS and SPSS, Academic Press, 2000.
[20] B. Kisacanin, V. Pavlovic, V. Pavlovic, and T. Huang, Real-time vision for human-computer interaction, Springer, 2005.
[21] N. Baba, L. Jain, and R. Howlett, Knowledge-based intelligent information engineering systems and allied technologies, IOS Press, 2001.
[22] W. Mendenhall, R. Beaver, and B. Beaver, Introduction to probability and statistics, Cengage Learning, 2008.
[23] M. Lipsey, Design sensitivity: Statistical power for experimental research, Sage Publications, 1990.
[24] B. Vogelvang, Econometrics: Theory and applications with EViewsý, Pearson Press, 2005.
[25] G. Agung, Time series data analysis using EViewsý, John Wiley & Sons, 2009.
[26] R. Pickard and R. Cook, "Cross-validation of regression models," J. American Statistical Association, vol.79, no.387, pp.575–583, Sept. 1984.

Anton Romanov received his Master's degree in information security from the Moscow Engineering and Physics Institute (MEPhI). Before 2009 he was working as an information security consultant for one of TOP5 world's largest software vendors. At the current moment he is taking Ph.D. course at Graduate School of Systems and Information Engineering, University of Tsukuba, Japan.



Eiji Okamoto received his B.Sc., M.S. and Ph.D. degrees in electronics engineering from the Tokyo Institute of Technology. From 1978 he was working for NEC central research laboratory. Then from 1991 he became a professor at Japan Advanced Institute of Science and Technology. Now he is a professor at Graduate School of Systems and Information Engineering, University of Tsukuba, Japan. He is a member of IEEE and coeditor-in-chief of International Journal of Information Security.