Regular Paper

An Approach to Perform Quantitative Information Security Risk Assessment in IT Landscapes

Anton Romanov,[†1] Hiroe Tsubaki[†2] and Eiji Okamoto[†1]

The purpose of this paper is to propose a quantitative approach for the effective and efficient assessment of risks related to information security. Though there are already several other approaches proposed to measure information security (IS) related risk, they are either inapplicable to real enterprises' IT landscapes or are of a qualitative nature, i.e. based on subjective decisions of the implementation team and thus could suffer from a significant degree of speculation. In contrast, our approach is based on objective statistical data, provides quantitative results and can be easily applied to any enterprise of any industry or any non-profit organization. An example of the application of the proposed approach to a real enterprise is also provided. The only prerequisite for the proposed methodology is a sufficient amount of incidents statistics collected under conditions described later in this paper. The reason for such research is that performing of IS related risk assessment is one of the procedures required to manage information security. And the process of IS management has recently become one of the highest concerns for most organizations and enterprises. It is caused not only by the growth of hackers' activity but also because of increasing legal requirements and compliance issues.

1. Introduction

In this paper we propose a methodology to perform quantitative IS risk assessment in IT landscapes. The approach is based on objective statistical data and can be applied to any organization or enterprise despite of their specifics.

The development of such a methodology has recently become a very critical task as it is a key part of the IS management process, which uses outcomes of risk assessment as a metrics to monitor changes in its state and thus to make managerial decisions [1]. That is, in order to manage the IS of a given IT landscape, there must be some metric which allows the comparison of levels of IS in different states of an IT landscape, for example before and after the implementation of countermeasures or changes in the architecture of this IT landscape. This metric is usually assigned to each risk, associated with each information system in an IT landscape, and reflects the exposure to monetary loss presented to an organization by this risk [2]. Thus, according to the results of risk assessment, a quantitative value, assigned to each risk, can be used to rank all risks defining their criticality levels making possible a prioritization of mitigation strategies, the justification of security investments and the preparation of related contingency funds. To sum up, performing regular risk assessment is essentially necessary for the process of providing information security.

It is necessary to point out that recent attention to the IS management process in different organizations and enterprises is caused not only by frequent hackers' activity, or periodically occurring technical faults but mostly by rapidly increasing legal requirements (so-called compliance issues) which are making information security a task of the highest concern.

But as the main goal for all organizations is focused on the efficiency of its operational business processes (core activities to earn income) it is clear that an approach used for supporting business processes (to which providing information security actually belongs) will be a rather specific one - organizations' management is interested in spending for such purposes as little money as possible, but nevertheless the company must still be compliant with all relevant laws or industry regulations. So business users, like corporate auditors or governance, risk and control (GRC) officers demand accurate quantitative methodologies which could measure IS related risks exactly in the same manner as already available approaches measure financial or credit risks. But unfortunately currently available methodologies to measure IS related risks are either of a qualitative nature, (i.e. instead of providing a monetary value of a given risk they describe its severity by means of linguistic variables, for example: Low, Medium, High, which makes possible neither the justification of security investments nor the estimation of the size of contingency funds) or are inapplicable to real IT landscapes because of

[†1] University of Tsukuba
[†2] The Institute of Statistical Mathematics

unachievable or unproved assumptions. These approaches and the reasons for their inapplicability are considered in detail in the next section.

2. Current Approaches to Measure Information Security Related Risks

It is necessary to mention that originally the assessment of IS related risks (as a significant part of technology risks) mostly interested insurance and audit companies. But as such organizations are not interested in the disclosure of their methodologies, most of this knowledge became available to the public through de-facto industry standards, developed by IT and security engineers as result of multiple attempts to prove the reliability of IT infrastructure or to pass an audit. And this resulted in the fact that though some of the approaches are considered to be best-practice in the industry, it is impossible to find out where they were originally proposed and who the authors are.

As mentioned above, all of the approaches to measure IS related risks could be divided into two large classes: qualitative and quantitative.

Qualitative approaches to measure IS related risks originally appeared to meet the needs of IT managers and security administrators who needed to define IS related problems and prioritize their activities to solve these problems. That is why it was sufficient to have just a qualitative description of the criticality level of a given risk. Most typical examples of such an approach are presented in Refs. 1) and 3). And though these approaches are already widely used in the industry, they have a big problem: an outcome of assessment is always subjective. That is, different people may assign different criticality levels to the same risk, which, in turn, may lead to quite doubtful results. So the result of an assessment can be a possible subject of speculation and manipulation[4)] (for example, line managers in banks are often interested in underestimation of risks as this allows them to spend more money on operational needs instead of keeping them in a contingency fund).

As to quantitative approaches, though they can be very helpful for IT staff as well, the need for such approaches is mostly caused by the recent demand from business users who are interested in the incorporation of IS related risks in the total risk management process in order to be compliant with recent legal requirements (like ISO27001 or JSOX) and in order to achieve a higher level of control over all possible expenditures (this is necessary, for example, to stay competitive in the market or not to exceed the budget). So the deliverables of such approaches must be presented in a way familiar to business users - using one of the financial metrics to evaluate risk. One of the first remarkable research results in this field is presented in Ref. 5) where the author introduces a comprehensive methodology which is based on the financial metric described in Section 3.1. But unfortunately as the author focused on the theoretical aspects of the methodology, trying to adjust financial metrics to IS related risks, some of auxiliary variables, proposed to calculate the amount of risk, are immeasurable in a real IT landscape. The same problem also exists in approaches proposed in Refs. 6), 7), 8) and 9). Another significant attempt is presented in Ref. 10) where the authors proposed the direct application of another financial metric, described in Section 3.2. But the direct application of this metric to IS related risks is impossible because of underlying assumptions which are presented in Section 3.2. So to conclude, this class of approaches is still a subject to research.

We suppose that the main reason for the inapplicability of previously introduced quantitative approaches to real IT landscapes is that they were designed either from the view point solely of financial analysts or only of IT and security engineers. In order to clarify the difference in these concepts in Ref. 11) we proposed another classification of approaches to assess IS related risks. Here we extend the classification. This classification also consists of two trivial groups: micro and macro approaches **Fig. 1**. The first group is usually preferred by engineers, while the second one by financial analysts.

In the micro-approaches group, the key focus is made on the estimation of the contribution of each risk individually and/or a group of similar risks, starting with a threat or vulnerability assessment. This is performed in order to create a low-level classification of all potentially risky events and ways to exploit available weaknesses. Afterwards it is considered what kind of loss the founded risks could lead to, and each risk's criticality is calculated. Calculated risk metric then is somehow converted to a monetary scale. The main problem of these approaches is that their application can be sensible for business users only in case if all parts of a given system have been deeply and precisely examined, which is often

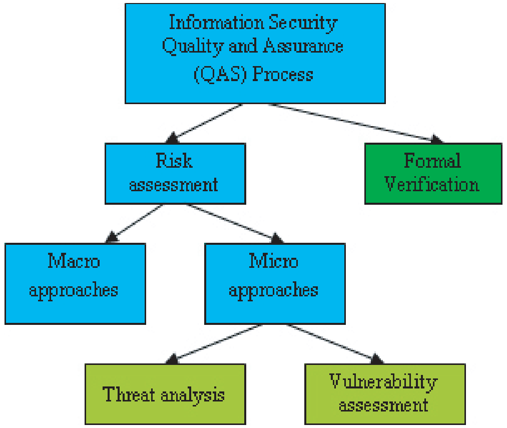


Fig. 1 Ways to perform information security QAS.

too expensive and time consuming. Thus very often it is impossible to make a deep and comprehensive threat and vulnerability assessment because of the high complexity of the underlying technology, extensively changing external conditions or lack of time.

Macro-approaches, instead, could be called money-driven, as they are focused on assessing the monetary value of risks basing on the historical data of already occurred total loss. So already occurred loss is spread between several groups of events, leading to a list of typical risks. Each of these risks afterwards could be further analyzed separately in order to define a mitigation strategy. Accordingly, as these approaches are performed by financial analysts, outcomes of such approaches may suffer from a lack of historical data of losses. And this fact may lead to the neglect of possible risks, which for various reasons has not yet occurred. Or it may lead to too high-level classification of groups of risk with the category "others" covering a significant percent of losses.

Though it may seem obvious to resolve this problem by the simultaneous application of both types of approaches, enterprises usually prefer to deploy either a macro or a micro approach, as the application of both types of approaches at the same time can lead to contradictory results which will not match each other. And such a situation may result in problems with external auditors who will consider it as a very critical event potentially leading to faulty financial statements. And as the result of the audit must be transferred to shareholders, any additional comments from auditors are very undesirable by the board of directors.

Unlike previously available approaches, our approach is an example of a combination of both micro and macro approaches. That is, we are considering all security incidents (events which could potentially lead to a monetary loss) as representatives of some class of security incidents defined with respect to threat and vulnerability assessment. Then we fix the reasons which are making possible the occurrence of such events by the deployment of configuration profiles (see Section 4.1) and afterwards we calculate risk value related to this class by the extension of a methodology which we previously proposed in Ref. 12).

On account of the potential problems described above, we are focusing on a way of providing such an output which could be easily converted to financial metric familiar to a total risk manager. This results in making it possible to determine an amount to be kept in the contingency fund, to perform cost-benefit analysis of security strategies and return on security investments in the way shown in Ref. 13).

3. Financial Metrics

In this section we analyze applicability of typical risk metrics, used by financial analysts, for the measurement of IS related risks. We also provide a rationale for selection of Value at Risk (VaR) metric as a basis for our approach.

3.1 ALE/SLE

This metric is considered as an example of the best-practice in financial risk management. In this metric a risk is defined by the amount of loss that is expected from it annually, this value is called Annual Loss Expectancy (ALE). It is defined by the multiplication of the Annual Rate of Occurrence (ARO) - the number of times a given threat to occur with a given asset and the Single Loss Expectancy (SLE), which is the monetary value expected from the single occurrence of a threat entailing this risk.

$$ALE = SLE \times ARO \quad (1)$$

SLE is defined as the multiplication of Asset Value (AV) by the Exposure Factor (EF), which represents the percentage of asset lost in an incident.

$$SLE = AV \times EF \tag{2}$$

Thus finally risk value is defined as follows:

$$ALE = AV \times EF \times ARO \tag{3}$$

Although this metric is absolutely correct and is widely used in the insurance industry to assess the consequences of, for example a flood or an earthquake, and though it seems to be very convenient to use it as an outcome of IS related risks assessment, unfortunately up to now there is no way of calculating required input parameters for IS related risks because of their specifics which can be found, for example in Ref. 14).

3.2 Value at Risk

According to Ref. 15) this metric was originally proposed as an enhancement to the previous one for a situation with high uncertainties where there is no way to calculate the precise amount of Annual Loss Expectancy. Here it is replaced with the probabilistic value which defines a risk by means of a certain amount of loss that will not be exceeded with a selected confidence level.

In other words, as defined in Ref. 16), VaR summarizes the worst loss over a target horizon with a given level of confidence (α), so it is an α-quantile of the projected cumulative distribution function (CDF) of losses over the target time horizon (see Eq. (4)). Usually in financial applications it is calculated over a 1 year horizon with the confidence level of 95% or 99%.

$$VaR^{Loss}(\alpha) = inf\left\{l \middle| CDF^{Loss}(l) \geq \frac{\alpha}{100}\right\} \tag{4}$$

In terms of this metric we defined the IS related related risk by two values: *Daily Risk* and *Annual Risk*, see Eqs. (5) and (6).

$$Daily\ Risk = VaR^{Daily\ Loss}(\alpha) = CDF^{Daily\ Loss}\left(\frac{\alpha}{100}\right) \tag{5}$$

$$Annual\ Risk = VaR^{Anual\ Loss}(\alpha) = CDF^{Annual\ Loss}\left(\frac{\alpha}{100}\right) \tag{6}$$

It is necessary to point out that in financial applications according to Markowitz theory [16] the distribution of losses caused by most risks can be represented with a normal or log-normal distribution, but it is an unproved assumption for IS related risks [4]. So to use this metric for IS related risks we must estimate the CDF of *DailyLoss* and *Annual Loss*. We describe this process in detail in Sections 4.7 and 4.8.

To conclude, though this metric seems to be more complicated than the previous one, it can be applied to measure IS related risks.

4. Proposed Approach

In this section we describe our approach to assess IS related risks.

Our approach consists of the following stages: two preliminary stages and four main stages. Preliminary Stages 1 and 2 are used to perform preparations which must be done before the execution of the main stages. Preliminary stage 0 is presented here for completeness purposes in order to define the scope of the proposed approach in the total risk management process.

- Preliminary Stage 0: Definition of risks to be assessed
- Preliminary Stage 1: Definition and deployment of configuration profiles
- Preliminary Stage 2: Collection of statistical data related to different security incidents
- Stage 1: Conversion from incidents to losses
- Stage 2: Transformation of data and its analysis
- Stage 3: Estimation of distribution function of Daily Loss
- Stage 4: Estimation of Daily Risk and Annual Risk

As input parameters our approach takes a list of risks to be assessed and *Amount of Incidents* time series, which is the daily amount of incidents related to a given type of security incident. This value must be collected according to preliminary Stages 1 and 2. We describe these procedures in Sections 4.1 and 4.2.

As an output it provides two values (for each type of security risk): *Daily Risk* and *Annual Risk*. The process of calculation of these values is described in Section 4.8.

These stages are covered in detail below, but before that we introduce some theoretical considerations which are used for the justification of our approach.

4.1 Theoretical Considerations

The contents of this subsection are based on our previous research results, provided in Refs. 11), 12) and 17).

As we already mentioned, one of the required parts of our approach is the estimation of the distribution function of *Daily Losses* caused by IS related risks. And as distribution of losses is defined by distribution of incidents (see Section 4.2), it is necessary to have a deeper look at the origin of security incidents and their distribution function. According to the statistical theory [18], regardless of the origin of a data sample, its distribution function can be estimated only in case this sample is homogenous (all samples were taken from one population) and stationary. So in order to estimate the distribution function of incidents we must assure that the *Amount of incidents* and *Daily Loss* time series are homogeneous and stationary.

According to Ref. 3) a security incident occurs if there is vulnerability and a threat which exploits this vulnerability. Thus the distribution of security incidents is defined by two distributions: distribution of threats and distribution of vulnerabilities (whose combination results in the occurrence of a given same security incident).

According to Ref. 19) the distribution of threats subsistent to a given organization can be considered constant for at least a certain period of time. But it is not possible to state the same about vulnerabilities. The reason for that is that an IT landscape is a set of heterogeneous IT systems and network equipment and each of these elements consists of hardware and software layers and thus can be represented as a set of software and hardware entities. But as any software or hardware entity is defined by a set of trivial components with some given properties (a set of CPU instructions for software and a set of primitive electronic elements hardware), and a vulnerability is a property subsistent to a trivial component itself or to a fixed set of such components, we can conclude that any change in the state of a given IT landscape leads to potentially different distribution of vulnerabilities subsistent to this state.

And as we showed in Ref. 11) a typical IT landscape is usually subject to superfluous changes in its states (like, installation or replacement of software, changes in system configuration or IT landscape architecture and so on). And as these changes in states lead to potential changes in distribution of vulnerabilities subsistent to these states, it results in non-homogeneity of data samples representing amount of security incidents.

To solve this problem in Ref. 17) we proposed the deployment of configuration profiles. In a wide sense a configuration profile is defined as a set of hardware and/or software which has a fixed distribution of vulnerabilities. In practice the easiest way to have two entities with the same distribution of vulnerabilities is to acquire uniform hardware and use software replication (software vendors offer this option for corporate subscribers). An example of a configuration profile is a set of unified software with fixed add-ons and version numbers. This configuration profile is deployed on all worksites in an organization by remote installation from the same distributive. The consistency of this configuration profile is achieved by a set of technical controls which enforce users not to install any additional software or any add-ons individually. We also showed that the application of such a framework leads not only to the reduction of the maintenance cost but also to a higher level of security assurance as a critical update can be deployed for all worksites in the network at once.

In terms of mathematical statistics, deployment of such configuration profiles results in the situation that if a security incident happened for a single worksite, it would happen for all others under the same external conditions. Hence observation of security incidents occurring in a large number of such typified worksites can be considered as a multiple realization of a random variable which has the same distribution of vulnerabilities subsistent to this fixed configuration profile and the same distribution of threats subsistent to the organization involved. So it means that statistics collected from different samples of the same configuration profile would represent the result of multiple experiments with same probabilistic parameters. And thus it is possible to conclude that gathered data will be homogenous and can be used to approximate the distribution function of security incidents and consequently losses.

4.2 Definition of Risks to be Assessed

As we already mentioned above, a detailed description of this stage is beyond the scope of this paper. If in short, at this stage IT and security engineers together with business users must define a list of the most critical IS related risks from a business point of view and IT and security engineers must math these risks with relevant technical systems.

4.3 Definition and Deployment of Configuration Profiles

At this stage IT and security engineers must define and deploy configuration profiles for all risks selected at the previous stage and start the collection of statistical data. In case there is an urgent need to change the configuration or implement an additional countermeasure it is recommended (if applicable) to distinguish which vulnerabilities such a replacement could fix and reflect related changes in incidents statistics.

4.4 Collection of Statistical Data Related to Different Security Incidents

At this stage, the amount of security incidents for all selected types of incidents is collected, day by day filling the *Amount of Incidents* time series for each type of incident.

4.5 Conversion from Incidents to Losses

After some amount of incidents data is collected it must be converted to losses. This conversion must be performed separately for each given type of security incident. Precise methodology for this step should be developed in advance by business users as they possess information about the cost of all assets.

We recommend taking into account the following factors (see Eq. (7)), though, of course, when applied to a concrete organization, they must be checked and in case of any trouble must be adjusted. Costs of tangible and intangible assets involved may also include all losses of all third parties involved or all losses which must be compensated for according to claims of third parties.

$$\begin{aligned} \mathit{Loss\ per\ Incident} &= \mathit{total\ man\ hours\ per\ incident} \times \mathit{average\ salary} \\ &+ \mathit{business\ cost\ of\ downtime} \\ &+ \mathit{cost\ of\ tangible\ assets\ damaged} \\ &+ \mathit{cost\ of\ intangible\ assets\ involved} \end{aligned} \tag{7}$$

Daily Loss time series then should be calculated by multiplication of *Loss per Incident* by *Amount of Incidents* for each day (see Eq. (8)).

$$\mathit{Daily\ Loss} = \mathit{Loss\ per\ Incident} * \mathit{Amount\ of\ Incidents} \tag{8}$$

4.6 Transformation of Data and its Analysis

As mentioned in Section 4.1, stationarity is a second required condition for estimation of the distribution function. So before estimation of CDF of *Daily Loss*, this time series must be checked for stationarity. We propose using a combination of tests: Augmented Dickey-Fuller (ADF) test and Phillips-Perron (PP) test (description of these tests is provided in Ref. 20)). This process is shown in detail on **Fig. 2**. Confidence levels of results of ADF and PP tests must be checked taking into account Durbin-Watson statistics 21).

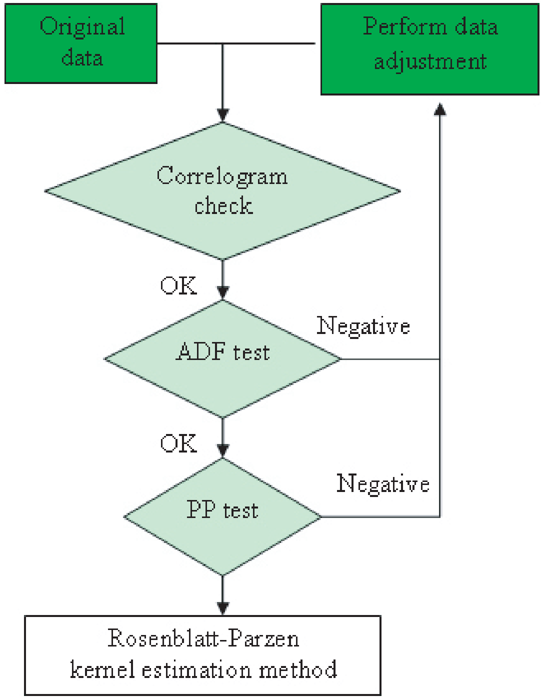


Fig. 2 Process of transformation of data to stationary state.

In case original *Daily Loss* time series is found to be non-stationary, it must be transformed to stationary state either by taking the first order difference (or difference of a higher order) or by presenting original data as a combination of intercept, trend and stationary component as described, for example in Ref. 22). For simplification of equations, later in this section we assume that the *Daily Loss* time series is stationary. If not, the steps below must be applied to a stationary transformation. Finally *Daily Risk* and *Annual Risk* values must be calculated with respect to the transformation performed. An example of such calculation for the worst case, when non-stationary *Daily Loss* time series is represented as a sum of stationary and non-stationary time series is provided in Section 5.

4.7 Estimation of Distribution Functions of Daily Loss

To obtain an estimation of the cumulative distribution functions (CDF) of *Daily Loss* we propose the application of a kernel density estimation methodology using the Rosenblatt-Parzen approach with the Epanechnikov kernel[18] considered as the best kernel for small amounts of sample data[23].

4.8 Estimation of Daily Risk and Annual Risk

In order to estimate a value of a *Daily Risk* we apply VaR methodology with the confidence level α, which is α-quantile of CDF of a daily loss. In financial applications α is usually equal to 95% or 99%.

Thus *Daily Risk* is calculated as shown in Eq. (9), where $VaR_{\alpha}^{Daily\ Loss}$ is Value at Risk for *Daily Loss* at confidence level α and $q_{\alpha}^{Daily\ Loss}$ is α-quantile of *Daily Loss* distribution.

$$\begin{aligned} Daily\ Risk &= VaR^{Daily\ Loss}(\alpha) \\ &= CDF^{Daily\ Loss}\left(\frac{\alpha}{100}\right) = q_{\alpha}^{Daily\ Loss} \end{aligned} \tag{9}$$

Afterwards we propose calculation of $q_{\alpha}^{Daily\ Loss}$ by application of Rankit-Cleveland quantile estimation method as described in Ref. 24).

As to the *Annual Risk*, as arises from Section 4.1, usually the collection of statistics for a given type of security incident during several years isn't possible. So it isn't possible to estimate Annual Risk in the same way as *Daily Risk*. To solve this problem we propose using approximation of α-quantile of *Annual Loss* distribution as follows: according to Central Limit Theorem[24] if x_i is any random variable, $\sum_{i=1}^{n} x_i \sim \mathrm{N}(\mu, \sigma^2)$ for large n, where $\mathrm{N}(\mu, \sigma^2)$ is Normal distribution with mean μ and variance σ^2.

Thus according to Ref. 15) $q_{\alpha}^{\sum x_i} = \sigma * \Phi(\frac{\alpha}{100})$, where Φ(x) is a CDF of Standard Normal distribution. Hence because of stationarity of the sample provided in Stage 5, $\sum_{i=1}^{n} Daily\ Loss \sim \mathrm{N}(\mu, \sigma^2)$ for large n - number of days. Thus

$$\begin{aligned} Annual\ Risk &= VaR^{Anual\ Loss}(\alpha) = CDF^{Annual\ loss}\left(\frac{\alpha}{100}\right) = q_{\alpha}^{Annual\ Loss} \\ &= \Phi\left(\frac{\alpha}{100}\right) * \sqrt{Variance\left(\sum_{i=1}^{n}(Daily\ Loss)\right)} \\ &= \Phi\left(\frac{\alpha}{100}\right) * \sqrt{n\sigma^2 + 2(n-1)\rho_1\sigma^2 + \ldots + 2(n-(n-1))\rho_{n-1}\sigma^2}, \end{aligned} \tag{10}$$

where n = 365, and ρ_i is i-th autocorrelation coefficient of *Daily Loss* which can be obtained form the correlogram of the data sample[23]. Φ(x) is a CDF for Standard Normal distribution. $\Phi(0.95) = 1.65$ and $\Phi(0.99) = 2.33$.

We can improve the estimation for *Annual Risk* in a special case: according to Ref. 22) Eq. (10) achieves its minimum value if all *Daily Loss*$_i$ are completely independant random variables which means there is no serial correlation in *Daily Loss* time series (or all $\rho_i = 0$) then *Annual Risk* $= \Phi(\frac{\alpha}{100})\sigma\sqrt{n}$ and Eq. (10) achieves its maximum value if all *Daily Loss*$_i$ are completely dependent variables which means there is perfect serial correlation in *Daily Loss* time series (or all $\rho_i = 1$) then *Annual Risk* $= \Phi(\frac{\alpha}{100})\sigma$n. So it is an upper limit for *Annual Risk*. Or

$$\Phi\left(\frac{\alpha}{100}\right)\sigma\sqrt{n} \le Annual\ Risk \le \Phi\left(\frac{\alpha}{100}\right)\sigma n \tag{11}$$

It is important to note here, that we can estimate *Annual Risk* even in case we have m < n = 365 observations. In this case Variance in Eq. (10) should be calculated taking all ρ_j, j = $\overline{m, 365}$ as shown in Eq. (12):

$$\begin{aligned} \text{Variance}\left(\sum_{i=1}^{n}(Daily\ Loss)\right) &= n\sigma^2 + 2(n-1)\rho_1\sigma^2 + \ldots \\ &\quad + 2(n-(m-1))\rho_{m-1}\sigma^2 \\ &\quad + 2(n-m)\sigma^2 + \ldots + (n-(n-1))\sigma^2 \end{aligned} \tag{12}$$

5. Example of Application

In this section we provide the application of the proposed approach to a real IT landscape. The data was collected in one small business organization during a time period of 199 days (August, 03 2009 - February, 20 2009) with the permission of the chief executive officer in charge. The amount of incidents was calculated on a daily basis including weekends and national holidays.

To be able to demonstrate the application of the proposed approach without making additional efforts, we selected a case where the precise calculation of the total amount of incidents is a very trivial task (it is possible to state that all of the occurred incidents were detected).

An incident investigated in this example is the event of the arrival of a single spam message to the company mail server. A spam message is denoted as a message which is not addressed to any employee of this organization or organization itself or such a message addressed to the employees or organization, which are not related to the personal or business needs of employees or the business needs of the organization. So, for example, if person A has provided his corporate email address to company B while making a personal inquiry, emails from company B addressed to person A will not be considered as spam during the processing time of his inquiry. But all emails from this company sent to him later or emails from other companies to which he has never given his corporate email address will be considered as spam emails.

Employees of the organization were asked to calculate and report the daily amount of spam messages they get, which were summed up by IT staff, thus calculating the *Amount of Incidents*. The original purpose of this investigation was to test a new antispam solution (which is beyond the scope of this research paper), which however was functioning only in a test mode without making any changes to the original emails flow (just making independent internal calculus of the total amount of messages and messages to be considered as spam) so only IT staff were able to compare results reported by employees and obtained from this system. There were no changes in the configuration of this system during the whole experiment time.

As employees were detecting spam messages manually without relying on any automated criteria and there were no any other antispam solutions filtering the emails stream going into corporate network from the outside world (internet service provider was especially asked to turn off any external antispam solutions), all employees were receiving all messages sent to their corporate email address and thus it is possible to conclude that the requirements of preliminary Stages 1 and 2 to collected data are met (as it is possible to consider that a given email client without any kind of antispam protection is a fixed software package defined in terms of Section 4.1, which is absolutely vulnerable to all realizations of a spam threat applied to the assets of this organization - corporate email addresses). And as stated above, configuration of an independent antispam solution was not changing as well.

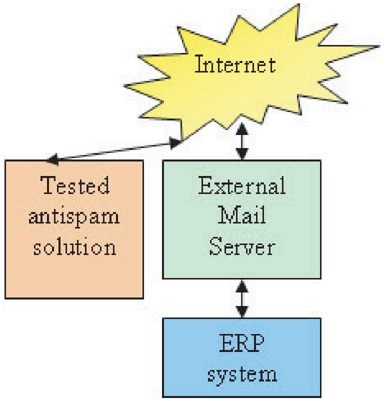


Fig. 3 Part of IT landscape related to email processing.

5.1 Landscape Description

The part of the IT landscape related to email processing and spam detection is presented on **Fig. 3**. The landscape consists of the external mail server which is connected with the internal enterprise resource planning (ERP) system which has an incorporated email client where employees usually check their emails (according to current business process). These emails are collected from two sources: the ERP system itself (which incorporates a kind of trivial mail exchange system for internal messages between employees) and the external mail server. According to the business processes of this organization most of the correspondence is usually sent between employees (thus through the ERP system) and some small amount is expected to be obtained from outside (using an external mail server).

5.2 Application Results

The collected daily incidents statistics is presented on **Fig. 4**.

Stage 1: Conversion from incidents to losses

As the arrival of a spam message does not lead to downtime or damage of tangible or intangible assets, the occurred loss depends only on man-hours spent to this incident. Here it will be the amount of time spent by employees reading this message and defining whether it is spam or not. So it is approximately possible to count it as 10 seconds per incident. The average salary of employees in the organization is $3000 per month. So as they are supposed to work about 22 days per month, 8 hours per day, the total loss because of a single spam message (*Loss per Incident*) is approximately $0.05 (as employees do not produce any useful value while they are reading a message in order to define if it is a spam one or not). Thus in this concrete case it is possible to convert data from the amount of incidents per day to the amount of loss simply by the multiplication of it by 1 incident price ($0.05). The *Daily Loss* time series obtained as a result of Eq. (8) and relevant correlogram[⋆1] (15 lags) are presented on **Fig. 5** and **Fig. 6** accordingly. In this concrete case the *Daily Loss* time series is extremely similar to daily the *Amount of Incidents* time series statistics, but in general these time series can differ significantly.

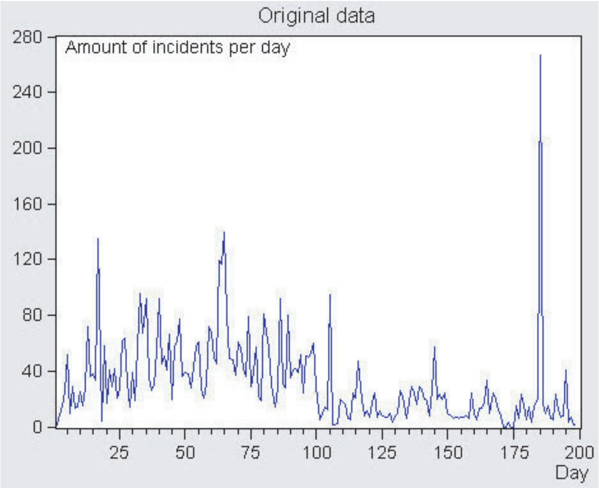


Fig. 4 Amount of Incidents graph.

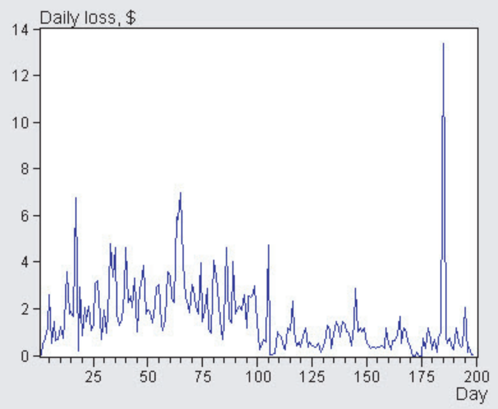
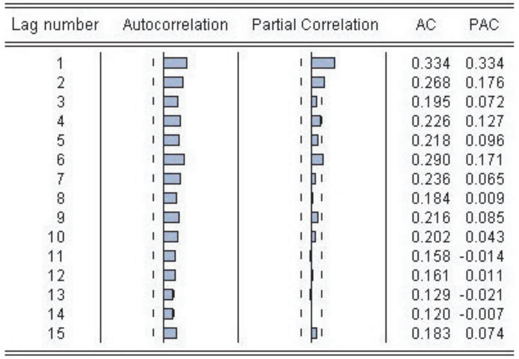


Fig. 5 Daily Loss graph.

Stage 2: Transformation of data and its analysis

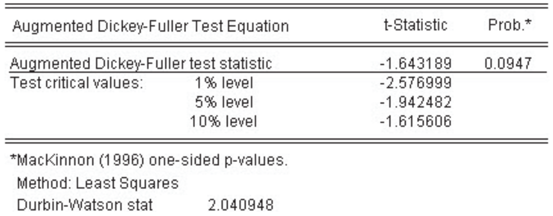
As seen from the correlogram all lags exceed the border and there is no trend in the decreasing of these values. So it is possible to make a preliminary conclusion

[⋆1] A correlogram is a plot of values of the sample autocorrelation function (AC), r_k and partial (auto)correlation function (PAC), r_k^* versus time lag, k. These values are calculated as follows: $r_k = \frac{\sum_{t=k+1}^{T} (Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{T} (Y_t - \overline{Y})^2}$, where $\overline{Y}$ is a sample mean of time series Y, T is number of observations in time series Y. And $r_k^* = \frac{\sum_{t=k+1}^{T} Y_t^* Y_{t-k}^*}{\sum_{t=k+1}^{T} (Y_{t-k}^*)^2}$, where Y_t^* and Y_{t-k}^* are residuals from the regression of Y_t and Y_{t-k} on $(1, Y_{t-1}, Y_{t-2}, ..., Y_{t-k+1})$ and T is number of observations in time series Y. The dotted lines in the plots are the approximate two standard error bounds computed as $\pm\frac{2}{\sqrt{T}}$. If the autocorrelation or partial autocorrelation is within these bounds, it is not significantly different from zero at (approximately) the 5% significance level [25].
The shape of AC and PAC plots can be used for preliminary judgment of stationarity (as a check of necessary conditions) and, in case original time series is not stationary, it can suggest a way to transform it to a stationary state or combination of stationary and non-stationary components. These techniques are described in detail in Refs. 22), 26).



Lag number	Autocorrelation	Partial Correlation	AC	PAC
1			0.334	0.334
2			0.268	0.176
3			0.195	0.072
4			0.226	0.127
5			0.218	0.096
6			0.290	0.171
7			0.236	0.065
8			0.184	0.009
9			0.216	0.085
10			0.202	0.043
11			0.158	-0.014
12			0.161	0.011
13			0.129	-0.021
14			0.120	-0.007
15			0.183	0.074

Fig. 6 Daily Loss correlogram.



Augmented Dickey-Fuller Test Equation		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.643189	0.0947
Test critical values:	1% level	-2.576999	
	5% level	-1.942482	
	10% level	-1.615606	

*MacKinnon (1996) one-sided p-values.
Method: Least Squares
Durbin-Watson stat 2.040948

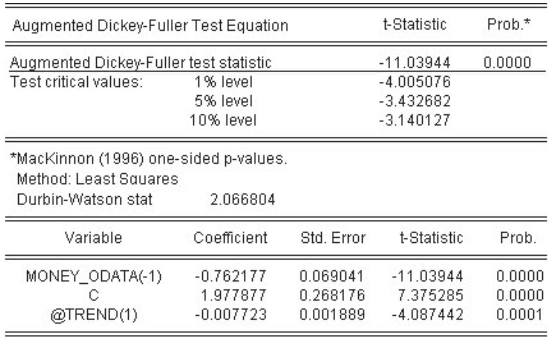
Fig. 7 ADF test for losses.

that most likely *Daily Loss* time series is not stationary. But at first it is better to check it with ADF test **Fig. 7**.

As ADF test statistics is higher then t-Statistics for 1% and 5% confidence level, the original data is non stationary. Let's try to check if *Daily Loss* can be presented as a sum of intercept and trend and make ADF test again. **Figure 8** shows that in that case ADF test statistics is much lower then t-Statistics for 1% level and as Durbin-Watson statistics is close to 2, the result of the test is trustworthy. Estimated coefficients for trend and constant are shown on Fig. 8.

The result obtained means that the *Daily Loss* time series can be represented as the sum of new adjusted stationary time series - *New Stationary Data*, linear trend and constant as shown in Eq. (13).

$$New\ Stationary\ Data = Daily\ Loss - ((-0.007723) * trend + 1.977877) \quad (13)$$

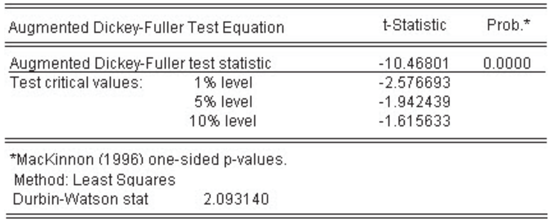


Augmented Dickey-Fuller Test Equation		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-11.03944	0.0000
Test critical values:	1% level	-4.005076	
	5% level	-3.432682	
	10% level	-3.140127	

*MacKinnon (1996) one-sided p-values.
Method: Least Squares
Durbin-Watson stat 2.066804

Variable	Coefficient	Std. Error	t-Statistic	Prob.
MONEY_ODATA(-1)	-0.762177	0.069041	-11.03944	0.0000
C	1.977877	0.268176	7.375285	0.0000
@TREND(1)	-0.007723	0.001889	-4.087442	0.0001

Fig. 8 ADF with trend and intercept assumption.



Augmented Dickey-Fuller Test Equation		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-10.46801	0.0000
Test critical values:	1% level	-2.576693	
	5% level	-1.942439	
	10% level	-1.615633	

*MacKinnon (1996) one-sided p-values.
Method: Least Squares
Durbin-Watson stat 2.093140

Fig. 9 ADF test for new data sample.

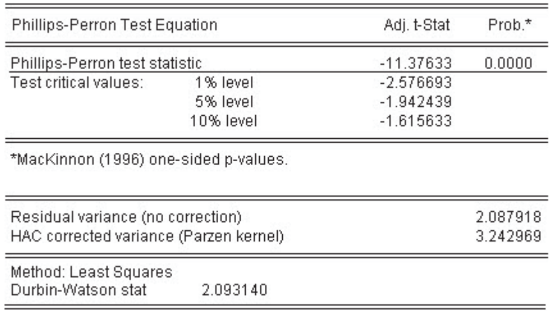
It means that values from the new adjusted stationary data will fluctuate around the trend line shifted by a constant.

The results of ADF and PP tests applied to new data and its correlogram are presented on **Fig. 9**, **Fig. 10** and **Fig. 11** accordingly. As ADF statistics is less than t-Statistics for 1% value, the data sample is stationary.

As Durbin-Watson statistics is close to 2, results of tests are trustworthy.

Stage 3: Estimation of distribution function of Daily Loss

Having done the adjustment of data to a stationary state it is possible to perform kernel density estimation as described in Section 4.7. An outcome of this stage is presented in **Fig. 12**. Thus the original *Daily Loss* distribution function for a day number d can be approximated by a function with the same density function and mean value shifted each day by (−0.007723)*trend(d) + 1.977877.



Phillips-Perron Test Equation		Adj. t-Stat	Prob.*
Phillips-Perron test statistic		-11.37633	0.0000
Test critical values:	1% level	-2.576693	
	5% level	-1.942439	
	10% level	-1.615633	

*MacKinnon (1996) one-sided p-values.

Residual variance (no correction)	2.087918
HAC corrected variance (Parzen kernel)	3.242969

Method: Least Squares
Durbin-Watson stat 2.093140

Fig. 10 PP test for new data sample.

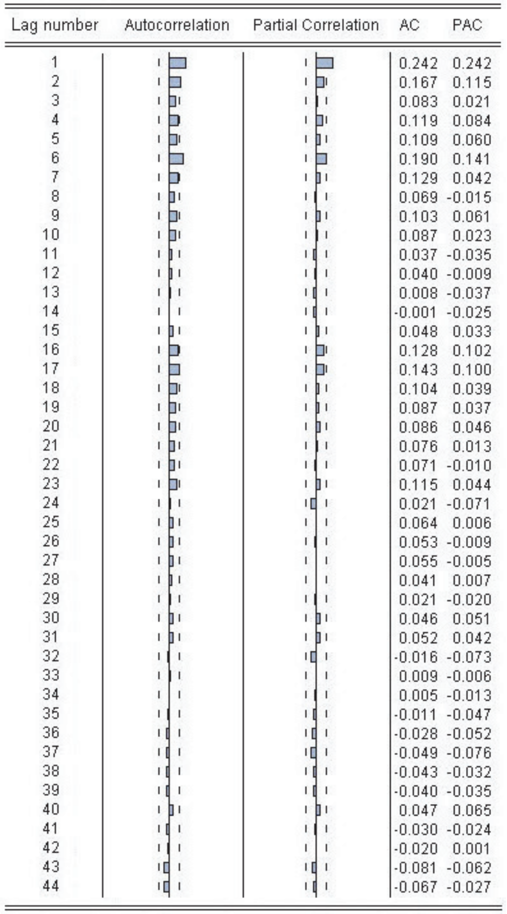
Stage 4: Estimation of Daily Risk and Annual Risk

Applying the Rankit-Cleveland quantile estimation method, 95%-quantile of this distribution equals approximately 2.8023 per each day. Thus according to Eq. (5), *Daily Risk* for a given day number d is equal to \$(2.8023 + (−0.007723)*trend(d) + 1.977877).

$$\begin{aligned} \mathit{Daily\ Risk} &= \$(2.8023 + (-0.007723)^{*}\mathrm{trend(d)} + 1.977877) \\ &\simeq \$(4.7801 + (-0.007723)(d-1)) \qquad (14) \end{aligned}$$

Thus according to Eq. (10) modified by Eq. (12), taking into account that *Daily Loss* time series contains non-stationary deterministic components (trend and intercept), the *Annual Risk* for this organization for spam messages with 95% confidence level can be calculated as presented in Eq. (15):

$$\begin{aligned} \mathit{Annual\ Risk} &= VaR^{\mathit{Annual\ Loss}}(95\%) = CDF^{\sum_{i=1}^{365}(\mathit{Daily\ Loss})}(0.95) \\ &= CDF^{\sum_{i=1}^{365}(\mathit{New\ stationary\ Data})} \\ &\quad + \sum_{i=1}^{365}(trend(i) * (-0.007723) + 1.9778) \\ &= \Phi(0.95) * \sqrt{Variance\left(\sum_{i=1}^{365}(\mathit{New\ stationary\ Data})\right)} \\ &\quad + 208.8862 \\ &= 1.65 * \sqrt{60521.11} + 208.8862 \\ &\simeq (\mathit{rounded\ up\ to\ integer\ value})\ \$615 \qquad (15) \end{aligned}$$



Lag number	Autocorrelation	Partial Correlation	AC	PAC
1			0.242	0.242
2			0.167	0.115
3			0.083	0.021
4			0.119	0.084
5			0.109	0.060
6			0.190	0.141
7			0.129	0.042
8			0.069	-0.015
9			0.103	0.061
10			0.087	0.023
11			0.037	-0.035
12			0.040	-0.009
13			0.008	-0.037
14			-0.001	-0.025
15			0.048	0.033
16			0.128	0.102
17			0.143	0.100
18			0.104	0.039
19			0.087	0.037
20			0.086	0.046
21			0.076	0.013
22			0.071	-0.010
23			0.115	0.044
24			0.021	-0.071
25			0.064	0.006
26			0.053	-0.009
27			0.055	-0.005
28			0.041	0.007
29			0.021	-0.020
30			0.046	0.051
31			0.052	0.042
32			-0.016	-0.073
33			0.009	-0.006
34			0.005	-0.013
35			-0.011	-0.047
36			-0.028	-0.052
37			-0.049	-0.076
38			-0.043	-0.032
39			-0.040	-0.035
40			0.047	0.065
41			-0.030	-0.024
42			-0.020	0.001
43			-0.081	-0.062
44			-0.067	-0.027

Fig. 11 Correlogram for new data.

Of course as VaR methodology does not offer full warranty (we just know that the daily loss will be lower than a given value only with a given level of confidence, usually 95 or 99%) the organization should take additional measures to mitigate residual risk if it considers that the maximum expected loss is much higher than it can accept (for example loss leading to a bankruptcy). So if they keep in a contingency fund an amount which equals to VaR they know that a bigger, extremely critical loss can happen only in 1% or 5% of cases (depending on confidence level applied before). So annually it will be about $365*(\frac{100-99}{100}) \simeq 4$ times per year for 99% confidence level. Thus they can buy a year insurance policy with required coverage amount which can be executed not more then 4 times per single year.

5.3 Cost Benefit Analysis of Selection Antispam Solution

Now it is very easy to estimate the efficiency of security investments on an antispam solution. One way to do it is the calculation of Net present value (NPV), see Ref. 13).

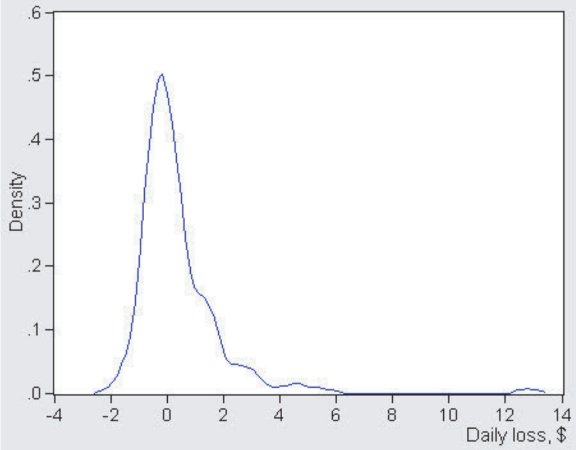


Fig. 12 Density estimation ⋆2.

$$NPV = \frac{Expected\ loss}{(1+r)} - \frac{Price\ of\ solution}{Depreciation\ period} \tag{16}$$

where r is discount (interest) rate.

6. Interpretation of Results

Having done the calculus above it is possible to conclude that the amount of risk related to spam activities for this organization is $615 per year.

Figure 13 shows the collected incidents statistics with the trend line. The trend is negative, which, means that the amount of spam has a tendency to reduce in future. At first this result could seem to be unnatural, but according to reports of several IT security consulting companies [27),28)], it is possible to see that they have got exactly the same result for the 2008 financial year. The explanation could be as follows: according to Ref. 27) on the 11th of November 2008 the US hosting provider was taken offline by its upper level network provider because of the very high spam activity of its clients. And this step has drastically decreased the amount of spam worldwide.

In the data observations (Fig. 13) November, 11 has the number 101. It is possible to see that the behavior of the function of total spam emails per day really seems to change after the specified date.

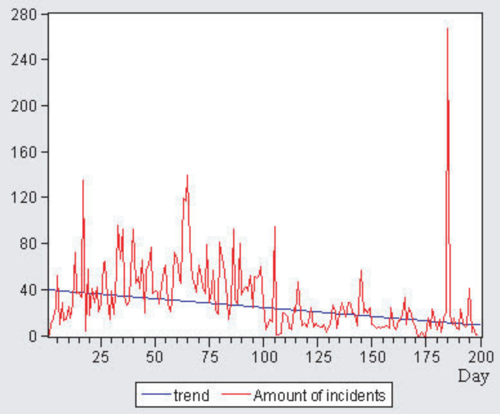


Fig. 13 Daily incident data with trend.

⋆2 Negative loss values on the graph arose due to specifics of selected kernel density estimation method and should be understood as nearly equal to zero. So the graph should be understood as the density estimation of loss function in case loss is greater than zero. And the probability that the loss is equal to zero is around 40% (computed using triangle approximation).

In this paper we use the definition of effectiveness as an ability to produce an intended or expected result and efficiency - as an ability to accomplish a given task with reasonable efforts. The effectiveness of the proposed approach is justified by underlying mathematical considerations. And its efficiency is justified by the fact that the deployment of the proposed approach does not require significant expenditures or human efforts.

7. Conclusion

In this paper we proposed an approach to perform the quantitative assessment of risks related to information security. The proposed approach is based on statistical data and not on subjective qualitative assumptions and can be easily applied in any IT landscape in any organization. As an outcome of assessment our approach provides two values for each selected risk and given confidence level α: *Daily Risk* and *Annual Risk*. And as these variables are defined through values of a well-known financial risk metric $Value\ at\ Risk$ (see Eqs. (5) and (6)), outcomes of our approach can be easily used for IS management and incorporation of IS management into Total Risk Management process.

A possible limitation of our approach is the necessity to make a preliminary adjustment of a given IT landscape, deploying configuration profiles. Though preparation of such profiles may definitely require additional efforts from security team, in some cases (like in example from Section 5) it turns into a very trivial process.

But in spite of some additional efforts, the pay off provided by the deployment of the configuration profiles will be not just a possibility to quantify risks but also the reduction of future maintenance efforts, the reduction of cost of a single worksite (as unification leads to the possibility of buying software and hardware with significant discounts) and the possibility to save on unprofitable security initiatives.

References

1) Brotby, K.: *Information Security Management Metrics*, CRC Press (2009).
2) Jaquith, A.: *Security Metrics. Replacing Fear, Uncertainty and Doubt*, Addison-Wesley (2007).
3) Tipton, H. and Henry, K.: *Official (ISC) Guide to the CISSP CBK*, Auerbach Publications (2007).
4) Landoll, D.: *The Security Risk Assessment Handbook*, Auerbach Publications (2006).
5) Dillard, K. and Pfost, J.: *The Security Risk Management Guide*, Microsoft Press (2004).
6) Zhao, D., Wang, J., Wu, J. and Ma, J.: Using Fuzzy Logic and Entropy Theory to Risk Assessment of the Information Security, *Proc. International Conference on Machine Learning and Cybernetics*, Vol.4, pp.2448–2453 (2005).
7) Lin, M., Wang, Q. and Li, J.: Methodology of Quantitative Risk Assessment for Information System Security, *Lecture Notes in Computer Science*, Vol.3802/2005, pp.526–531 (2005).
8) Ekelhart, A., Fenz, S., Klemen, M. and Weippl, E.: Security Ontologies: Improving Quantitative Risk Analysis, *40th Annual Hawaii International Conference on System Sciences*, p.156a (2007).
9) Wawrzyniak, D.: Information Security Risk Assessment Model for Risk Management, *Lecture Notes in Computer Science*, Vol.4083/2006, pp.21–30 (2006).
10) Ozcelik, Y. and Rees, J.: *A New Approach for Information Security Risk Assessment: Value at Risk* (2005). http://ssrn.com/abstract=1104264
11) Romanov, A. and Okamoto, E.: An approach for designing of enterprise IT landscapes to perform quantitave information security risk assessment, *Proc. International Conference on Security and Cryptography, SECRYPT 2009*, pp.313–318 (2009).
12) Romanov, A. and Okamoto, E.: A Quantitative Approach to Assess Information Security Related Risks, *Proc. International Conference on Risks and Security of Internet and Systems, CRISIS 2009*, pp.117–123 (2009).
13) Tipton, H. and Krause, M.: *Information Security Management Handbook*, CRC Press LLC (2005).
14) Peltier, T.: *Information Security Risk Assessment*, CRC Press LLC (2005).
15) Jorion, P.: *Value at Risk. The new benchmark for managing financial risk*, McGrow-Hill (2001).
16) Pearson, N.: *Risk budgeting: Portfolio Problem Solving with Value at Risk*, John Wiley and Sons Inc. (2002).
17) Romanov, A. and Okamoto, E.: A framework for building and managing secured ERP landscape, *Proc. 2009 International Conference on Security and Management, SAM 2009*, pp.490–495 (2009).
18) Racine, J.: *Nonparametric Econometrics*, Now Publishers Inc. (2008).
19) Lenstra, A. and Voss, T.: *Information Security Risk Assessment, Aggregation, and Mitigation*, Springer (2004).
20) Baltagi, B.: *A companion to theoretical econometrics*, Wiley-Blackwell (2003).
21) Zellner, A. and Palm, F.: *The structural econometric time series analysis approach*,

Cambridge University Press (2004).
22) Brockwell, P. and Davis, A.: *Time Series: Theory and Methods*, Springer (2009).
23) Martinez, A.: *Computational statistics handbook with MATLAB*, CRC Press LLC (2002).
24) Berthouex, P. and Brown, L.: *Statistics for environmental engineers*, CRC Press LLC (2002).
25) I Gusti Ngurah Agung: *Time series data analysis using EViews*, John Wiley and Sons (Asia) Pte Ltd. (2009).
26) Greene, W.: *Econometric analysis*, Prentice Hall (2003).
27) ENIDAN Technologies GmbH of Herrliberg: *The Spamchek Report* (2009). http://www.spamchek.com/company/press
28) Kaspersky Lab Inc.: *Kaspersky Security Bulletin: Spam Evolution 2008* (2009). http://www.viruslist.com/en/analysis?pubid=204792053#2

(Received November 30, 2009)
(Accepted June 3, 2010)
(Released September 8, 2010)

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 worlds largest software vendors. Currently he is taking a Ph.D. course at the Graduate School of Systems and Information Engineering, University of Tsukuba (Japan) and is a Visiting Researcher at the Institute of Statistical Mathematics (Japan).

Hiroe Tsubaki is the Director and Professor of Risk Analysis Research Center at the Institute of Statistical Mathematics and a Professor of applied statistics at the Graduate School of Business Sciences at University of Tsukuba. Professor Tsubaki was formerly a lecturer of statistics at Keio University from 1987 to 1997. He earned his Bachelor, Master and Doctor of Engineering from the University of Tokyo in 1979, 1982 and 1988, respectively. He is also an expert member of several Japanese government committees such as the Statistics Committee of the Cabinet Office, the Japanese Industrial Standard Committee and the Pharmaceutical Affairs and Food Sanitation Council. His primary research interests are mainly applications of statistical methods to different fields of business and industry.

Eiji Okamoto received his B.Sc., M.S. and Ph.D. degrees in electronics engineering from Tokyo Institute of Technology. Since 1978 he has been working for NEC central research laboratory. Then in 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 the International Journal of Information Security.