# Modeling Privacy Insurance Contracts and Their Utilization in Risk Management for ICT Firms

Athanassios N. Yannacopoulos[1], Costas Lambrinoudakis[2], Stefanos Gritzalis[2],
Stylianos Z. Xanthopoulos[3], and Sokratis N. Katsikas[4]

[1] Athens University of Economics and Business, Dept. of Statistics
[2] University of the Aegean, Dept. of Information and Communication Systems Engineering
[3] University of the Aegean, Dept. of Statistics and Actuarial-Financial Mathematics
[4] University of Piraeus, Dept. of Technology Education and Digital Systems
ayannaco@aueb.gr, {clam,sgritz,sxantho}@aegean.gr, ska@unipi.gr

**Abstract.** The rapid expansion of Internet based services has created opportunities for ICT firms to collect and use, in an unauthorized way, information about individuals (e.g. customers, partners, employees etc.). Therefore, privacy issues are becoming increasingly important. In this paper we model the risk that an IT firm is exposed to, as a result of potential privacy violation incidents. The proposed model is based on random utility modeling and aims at capturing the subjective nature of the question: "how important is a privacy violation incident to someone?". Furthermore, we propose a collective risk model for the economic exposure of the firm due to privacy violation. These models are useful for the design and valuation of optimal privacy related insurance contracts for the firm and are supportive to its risk management process.

**Keywords:** Privacy, Risk Modeling, Insurance, Random Utility Models.

## 1 Introduction

The immense advances in information and communication technologies have significantly raised the acceptance rate of Internet-based applications and services. Enterprises store, manage and process large amounts of personal and sensitive data about their employees, partners, and customers. Despite the fact that this information is fundamental to enable their business processes, personal data should be accessed and used according to privacy legislation and guidelines; that is only for the purposes for which they have been collected and always after the consent of the data subjects.

Nevertheless, some people are really concerned about privacy protection issues, while others are not. The diversity of the interest level of an employee, or a partner, or a customer, may result into different estimations about the consequences for the firm in case of privacy violation incidents. It is thus necessary to develop a model capable of handling this subjective impact level for the firm, in terms of the compensation that an individual may claim after a privacy breach.

S. Jajodia, and J. Lopez (Eds.): ESORICS 2008, LNCS 5283, pp. 207–222, 2008.

For instance, when an IT firm uses personal data without the consent of its clients, it is subjective whether a client will feel upset about it and press charges or not. In fact, only a few clients may decide to press charges and claim compensation. And given that this has happened what is the likely amount of the compensation claimed?

We will introduce a simple model that incorporates the personalized view of how individuals perceive a possible privacy violation and the loss of value that such a violation represents to them.

In Section 2 of this paper, we provide a short review of the research area. In Section 3 we model the possible compensation claim of an individual after misuse of her personal information. In section 4 we model the number of compensation claims during a time period and the total amount claimed during this period, considering a homogeneous population of clients. In Section 5 homogeneity is relaxed, while in Section 6 we discuss applications of the collective risk model to insurance and risk management issues for IT firms handling personal data. In section 7 we present and discuss a simulated example to obtain a feeling of the practical usefulness and applicability of the proposed model. Finally, section 8 summarizes and concludes the paper.

## 2 Literature Review

Privacy refers to the right of someone to be left alone [1]. Information privacy refers to the right of the individual to control personal information [2,3]. Loss of information privacy may lead to loss of privacy in the above defined context. The new technologies and the expansion of the Internet have raised public concern about information privacy [4,5,6,7,8,9]. Four identified aspects of privacy concerns about organizational information privacy practices refer to (i) collection and storage of large amount of personal information data, (ii) unauthorized secondary use of personal data, (iii) errors in collected data and (iv)improper access to personal data due to managerial negligence [10], with the first two being the most important [11,12,13]. Regarding online privacy preferences, individuals are classified in three basic classes [14,15]:(i) the Privacy Fundamentalists who almost always refuse to provide personal information, (ii) the Pragmatic Majority, who exhibit privacy concerns but not as strongly as the Privacy Fundamentalists and (iii) the Marginally Concerned who are almost always willing to reveal their personal data.

The above studies do not claim that individuals will actually behave according to their stated preferences. In fact a dichotomy exists between stated information privacy preferences and actual behavior when individuals need to make a privacy related decision [15,16,17,18,19]. However, there is strong evidence that people are willing to exchange personal information for economic benefits or personalized services [16,20,21], giving thus ground to proposals for regulation of privacy through National Information Markets [22]. Finally, the impact of a company's privacy incidents on its stock market value is explored and analyzed in [23].

# 3 Modeling the Possible Claim of an Individual $j$ for Revealing Private Data $D_m$

We consider the case that a private data $D_m$ disclosure has occured and we wish to answer the question "'How much would an individual $j$ claim as compensation for the above mentioned privacy breach?"'

## 3.1 A Random Utility Model

Our basic working framework is the random utility model (RUM) that has been used in the past in the modeling of personalized decisions and non market valuation. We assume that the individual j may be in two different states. State 0 refers to the state where no personal data is disclosed. State 1 refers to the state where personal data has been disclosed. For simplicity we assume that there is only one sort of data that may be disclosed.

The level of satisfaction of individual j in state 1 is given by the random utility function $u_{1,j}(y_j, z_j) + \epsilon_{1,j}$ where $y_j$ is the income (wealth) of the individual and $z_j$ is a vector related to the characteristics of the individual, e.g. age, occupation, whether she is technology averse or not etc. The term $\epsilon_{1,j}$ is a term that will be considered as a random variable and models the personalized features of the individual $j$. This term takes into account effects such that, for instance one time the same individual may consider a privacy violation as annoying whereas another time she may not bother about it at all. This term gives the random features to the model and is essential for the subjective nature of it. Similarly, the level of satisfaction of the same individual j in state 0 is given by the random utility function $u_{0,j}(y_j, z_j) + \epsilon_{0,j}$ where the various terms have similar meaning.

State 1, the state of privacy loss, will be disturbing to individual j as long as

$$u_{1,j}(y_j, z_j) + \epsilon_{1,j} < u_{0,j}(y_j, z_j) + \epsilon_{0,j}$$

and that may happen with probability

$$P(\epsilon_{1,j} - \epsilon_{0,j} < u_{0,j}(y_j, z_j) - u_{1,j}(y_j, z_j))$$

This is the probability that an individual will be bothered by a privacy violation and may be calculated as long as we know the distribution of the error term. This will also depend on the general characteristics of the individual through $z_j$ as well as on her income $y_j$. The particular dependence can be deduced through statistical tests which will be sketched briefly.

Given that an individual $j$ is bothered by a privacy violation, how much would she value this privacy violation, so how much would she like to be compensated for that? If the compensation is $C_j$ then this would satisfy the random equation

$$u_{1,j}(y_j + C_j, z_j) + \epsilon_{1,j} = u_{0,j}(y_j, z_j) + \epsilon_{0,j}$$

the solution of which will yield a random variable $C_j$. This is the (random) compensation that an individual may ask for a privacy violation. The distribution of the compensation will depend on the distribution of the error terms $\epsilon_{ij}$ as well as on the functional form of the deterministic part of the utility function.

The following two cases are quite common:

1. $u_{i,j}(y_j, z_j) = a_i y_j + b_i z_j$, linear utility function. Assuming $a_0 = a_1 = 1$, the random compensation is given by $C_j = Bz_j + \epsilon_j$, where $B = b_0 - b_1$ and $\epsilon_j = \epsilon_{0,j} - \epsilon_{1,j}$. Then (since B is a deterministic vector) the distribution of $C_j$ is the distribution of the random variable $\epsilon_j$. A common assumption is that the $\epsilon_j$ are normally distributed. This leads to a normally distributed compensation, and forms the basis of the well known class of econometric models called probit models. Another common assumption is that the random variable $\epsilon_j$ is distributed by a logistic distribution. This forms the basis of the well known class of econometric models called logit models. Note that the linearity of the utility function in the income makes the compensation independent of the income.
2. $u_{ij}(y_j, z_j) = a_i ln(y_j) + b_i z_j$ i.e. the utility function is log linear in income. Again, assuming $a_0 = a_1 = 1$, the random compensation is given by $C_j = -y_j + y_j exp(Bz_j + \epsilon_j)$, where $B = b_0 - b_1$ is constant and $\epsilon_j = \epsilon_{0,j} - \epsilon_{1,j}$. The distribution of the compensation is determined by the distribution of the error term $\epsilon_j$. Normally distributed errors will lead to a probit model whereas errors distributed with a logistic distribution will lead to a logit model.

The above mentioned models may in principle lead to unbounded claims, though with diminishing probability. As an attempt to remedy this situation we may resort to bounded logit or probit models. Such models have been used in the literature for valuation of environmental and natural resources with great success. An example of such a model may be the model

$$C_j = \frac{y_j}{1 + exp(-z_j \gamma - \epsilon)}$$

where the error may be taken as either logistic or normal.

In general the RUMs may lead to a wide variety of individual claim distributions, depending on the choice of the utility function and the distribution of the random terms. Therefore, one may obtain heavy tailed distributions, characteristic of large claims, or distributions with thin tails, characteristic of the small claims that insurance companies may deal with in everyday practice. The distribution of claims depends heavily on the fall off of the inverse of the utility function in the range of large values of its argument; this is evident since

$$C_j = u_{1,j}^{-1}(u_{0,j}(y_j, z_j) + \epsilon_{0,j} - \epsilon_{1,j}) \; - \; y_j$$

almost surely, where $u_{1,j}^{-1}$ is the inverse of the function $u_{1,j}(y_j, z_j)$ with respect to the first argument, keeping $z_j$ fixed. The above formula shows that slow decay of the inverse utility function may lead to heavy tails for the distribution of the claims, thus leading to typical large claim distributions such as the Pareto distribution.

### 3.2 Estimation of such Models

The estimation of such models may be made using appropriately chosen questionnaires in order to obtain enough data for the proposed claims so that a logit or probit distribution may be fitted into them. An appropriate form of the questionnaire could be for instance: Would you be ready to accept a sum of $t$ euros in order to reveal this data (e.g. telephone number, credit card number etc). The test will be made for a vector of $t$'s and the answer will be in the form of yes (1) and no's (0). The answers to the test will provide estimates for the probability that $P(C_j > t)$ and these results will then be fitted into a logit or probit model using standard statistical procedures which are now well implemented in commercial packages. A possible procedure for the model estimation could be for instance a maximum likelihood method, where the likelihood of the observed answers to the survey is computed as a function of the parameters of the model obtained by the RUM and then the parameter values are chosen to be such that the likelihood is maximized. For the RUMs described above, i.e. the logit and probit model, there exist analytic formulae for the likelihood, thus facilitating the maximization.

After estimating the model we have a good approximation of the probability distribution of the compensation claim of an individual $j$ with characteristics $z_j$ and income $y_j$ for revealing some private data $D_m$.

## 4 The Temporal Structure of the Risk Model

In the previous section we established a personalized model for the compensation that an individual may claim from a firm that caused a privacy incident. In this section we use the methods of non-life insurance mathematics [24], in order to model the total compensation that may be claimed during some time period from the firm by a class of individuals who were affected by the privacy incident.

### 4.1 Modeling the Number of Claims

We now assume that a series of claims $C_j$ may arrive at certain random times $N_j$. Each of these claims may be distributed as determined by the RUM. Of paramount importance to the construction of a satisfactory model for the liabilities of a firm handling privacy related data is to model the distribution of random times when claims concerning privacy breaches may occur.

The distribution of random times may be modeled as a Poisson distribution $Pois(\lambda)$ or as a geometric distribution.

Another possible model for the distribution of the arrival times may be a renewal process. This may be seen as a generalization of the homogeneous Poisson process, allowing for the modeling of large gaps between the arrival of claims. A renewal process may be constructed as a random walk $T_0 = 0$, $T_n = W_0 + W_1 + \cdots + W_n$ where $W_i$ is an i.i.d. sequence of almost surely positive random variables. The special case where $W_i \sim Exp(\lambda)$ generates the homogeneous Poisson process. However, the use of interarrival time distributions such as the lognormal or the Pareto distribution may model long interarrival times, which may be better fitted for the description of claims connected to privacy related incidents.

Yet another possibility for modeling the distribution of arrival times for the individual claims may be a mixed Poisson process. This, generally speaking, is a Poisson process, whose rate is no longer deterministic but rather a random variable, that is $N(t) = \bar{N}(\theta\, \mu(t))$ where $\bar{N}$ is a standard homogeneous Poisson process, $\mu$ is the mean value function of a Poisson process and $\theta$ a positive random variable, independent of $N$. The random variable $\theta$ is called the mixing variable. Mixture models may provide a wide variety of distributions for $N(t)$. For example, if $\mu(t) = t$ and $\theta$ is assumed to follow the gamma distribution with parameters $\gamma$ and $\beta$ then $N(t)$ is distributed by the negative binomial with parameter $(p, v) = (\frac{\beta}{t+\beta}, \gamma)$.

Such a model may be reasonable into taking account of the randomness included into whether somebody suffering a privacy incident will finally decide to act and demand compensation or not, and if yes when. It is a nice complement to the RUM, since the RUM was used to estimate the size of the claims, given that the person suffering the privacy incident had decided to act and claim compensation. The mixed Poisson case is a nice way to model the probability and the waiting time distributions of the events related to when and how often the person suffering the privacy incident will decide to act. An interesting fact concerning models using mixed Poisson processes is that the increments now may be dependent, in contrast to the situation for the Poisson process. This introduces difficulties in calculating the statistical characteristics of the total claim $L(t)$ (defined formally in the next paragraph), but it offers realistic effects to the model. For instance, the intention of somebody to act against a privacy breach, may depend on the number of previous breaches that passed without taking any action. On the same argument, if one has already taken legal action in protest to a privacy breach, she is more likely to do it again, since she has overtaken once the "barrier" of the legal and formal measures to be followed.

### 4.2 Modeling the Total Claim Amount

The total claim up to time $t$ will be given by the random sum

$$L(t) = \sum_{i=0}^{N(t)} C_i$$

This is a compound random variable and forms the basis of the model of collective risk in actuarial mathematics. The distribution of $L(t)$ depends on the distribution of $C_i$ and on the distribution of the counting process $N(t)$. In this subsection we assume that our population is homogeneous, i.e. the $C_i$'s are i.i.d.

Assuming independence between $N(t)$ and the size of the arriving claims $C_j$, we may calculate the expected total claim and its variance

$$\mathbb{E}[L(t)] = \mathbb{E}[N(t)]\, \mathbb{E}[C]$$

$$Var(L(t)) = Var(N(t))(\mathbb{E}[C])^2 + \mathbb{E}[N(t)]Var(C).$$

For instance, when $N(t) \sim Pois(\lambda t)$, straightforward calculations imply

$$\mathbb{E}[L(t)] = \lambda\, t\, \mathbb{E}[C_1]$$
$$Var(L(t)) = \lambda\, t\, \mathbb{E}[C_1^2]$$

where $\mathbb{E}[C_1]$ and $\mathbb{E}[C_1^2]$ can be estimated by the use of the RUM.

In the case where $N(t)$ is modeled with the use of a renewal process, we may have a more realistic and robust model. The price one has to pay though when abandoning the nice Poisson type structure of the model is that the statistical properties of the stochastic process $L(t)$ may no longer be as easily calculated analytically and one may have to resort to simulation studies. However, approximate limiting results are available, allowing to state general approximate but robust results, since they hold under quite general conditions. For example, an important result from renewal theory states that if $\mathbb{E}[W_1] = \lambda^{-1}$, then the counting process $N(t)$ that counts the number of claims up to time $t$ satisfies, almost surely, $\lim_{t\to\infty} \frac{N(t)}{t} = \lambda$. This suggests that for a general model utilizing a renewal process, $\mathbb{E}[N(t)]$ is of order $\lambda t = \frac{t}{\mathbb{E}[W_1]}$ for large $t$ and this can be turned into a rigorous limiting argument in the sense that $\lim_{t\to\infty} \frac{\mathbb{E}[N(t)]}{t} = \lambda$. Similar asymptotic results can be shown to hold for the variance. For instance, assuming that $\mathbb{E}[W_1^2] < \infty$,

$$\lim_{t\to\infty} \frac{Var(N(t))}{t} = \frac{Var(W_1)}{(\mathbb{E}[W_1])^3},$$

and most importantly a central limit theorem can be shown to hold for the variance, stating in particular that $(Var(N(t))(\mathbb{E}[W_1])^{-3}\, t)^{-1/2}\,(N(t) - \lambda\, t)$ converges in distribution to $N(0,1)$ as $t \to \infty$ , thus allowing detailed probabilistic estimates for this quantity.

Thus, asymptotic results are achievable. For instance, the statistical quantities of $L(t)$ are estimated as

$$\mathbb{E}[L(t)] = \lambda\, t\, \mathbb{E}[C_1]\,(1 + o(1)),\;\; t \to \infty,$$
$$Var(L(t)) = \lambda\, t\, \{Var(C_1) + Var(W_1)\, \lambda^2\, (\mathbb{E}[C_1])^2\}\,(1 + o(1)),\;\; t \to \infty$$

Since the process $L(t)$ provides important information concerning the liability of the firm with respect to privacy related breaches, more information than just the moments will be welcome. For instance, within the context of the general renewal model, central limit type theorems may be proved for the distribution of $L(t)$. In particular,

$$P\left(\frac{L(t) - \mathbb{E}[L(t)]}{\sqrt{Var(L(t))}} \le x\right) \to \Phi(x),\;\; x \in \mathbb{R} \tag{1}$$

where $\Phi(x)$ is the cumulative normal distribution. Such results may provide detailed information concerning the probability of the total risk the firm is facing.

In the general case, the characteristic function for the total claim $L(t)$, $\phi(s;t) = \mathbb{E}[exp(\mathrm{i}\, s\, L(t))]$, where $s \in \mathbb{R}$ and i is the imaginary unit, may be calculated. Using the independence property of $N(t)$ and $C_i$ we obtain that

$$\phi(s;t) = \mathbb{E}[\exp(N(t)\ \ln(\phi_C(s))] = m_{N(t)}(\ln(\phi_C(s)) \qquad (2)$$

where $\phi_C(s)$ is the characteristic function for $C_i$.

For example the characteristic function when $N(t)$ is Poisson distributed with mean function $\mu(t)$ is given by

$$\phi(s;t) = exp(-\mu(t)(1 - \phi_C(s)))$$

for real $s$. Another choice of model for the claim arrival times may be for instance that the claims arrive with a geometric distribution with parameter $p \in (0, 1)$.

Well founded techniques from the theory of actuarial mathematics may be used for the analytical approximation of the total claim as well as its numerical simulation.

## 5 Inhomogeneity of the Population and Disclosure of More Than One Type of Data

In the above collective risk model we assumed that the population of clients that may claim compensation for a privacy violation is homogeneous, in the sense that they all share the same characteristics (income, level of computer literacy etc.). This may simplify the analysis but it is not a realistic assumption.

We will thus assume that the IT firm has a collection of clients, whose income is distributed by a probability distribution of income $F(y)$ and whose characteristics $z$ are distributed by a probability distribution $G(z)$. Then a possible claim will be a random variable which depends on parameters which are themselves random variables that follow some probability distribution which is either known objectively and treated as some sort of statistical probability, or can be thought of as a subjective belief concerning the composition of the population which may be treated using the methodology of Bayesian statistics.

If we then assume a logit or probit model with income $y$ and parameters $z$ then the possible claim will be a random variable $C$ such that $E[C \mid Y = y, Z = z] = C(y, z) \sim \mathit{Logit}(y, z)$ or $E[C \mid Y = y, Z = z] = C(y, z) \sim \mathit{Probit}(y, z)$ respectively.

This is valid for a single claim. We now wish to model the claims coming for compensation at different times as coming from different individuals (clients) with different characteristics. Therefore, the collective claim will be

$$L(t) = \sum_{i=0}^{N(t)} C(Y_i, Z_i)$$

where the random variables $Y_i$ and $Z_i$ represent draws from the distribution $F(y)$ and $G(z)$ respectively, at the times where the point process $N(t)$ takes the values $N(t) = i$, i.e. at the times where the claims occur.

The simulation of the inhomogeneous population model, will give more realistic estimates on the possible distribution of claims.

A feasible way of modeling this situation is through the use of the heterogeneity model; according to which a firm may have $k$ different customers and each customer $k$ may be described by the pair $(\theta_k, \{C_{k,t}\}_t)$ where $\theta_k = (y_k, z_k)$ is the heterogeneity parameter which includes the characteristics of the customer, and $\{C_{k,t}\}_t$ is the sequence of claim sizes for customer $k$, over the time interval $[0, T]$ that the policy holds. We will assume that the $\theta_k$ are i.i.d. random variables, representing draws from the same distribution and that given $\theta_k$ the sequence $C_{k,t}$ is i.i.d. with known distribution, provided by the use of the RUM, say $F(\cdot, \mid \theta_k)$. Obviously, $P(C_{k,t} \leq x) = \mathbb{E}[F(x \mid \theta_k)]$.

The firm would like to estimate the claims expected from a particular customer type $k$, given the past claims this customer has asked for, i.e. given data $C_{obs,k} = \{C_1, C_2, \cdots, C_t\}$ for some $t \leq T_{obs}$, where $[0,\, T_{obs}]$ is some observation period, from the history of this customer. To this end, we may use the Bayes estimator, to obtain the best (in the sense of minimum $L^2$ error) estimate for the quantity under consideration. The reasonable quantities that enter this estimator will now be

$$\mu(\theta_k) = \mathbb{E}[C_{k,t} \mid \theta_k] = \int x \, dF(x \mid \theta_k)$$

$$Var(\theta_k) = \mathbb{E}[(C_{k,t} - \mu(\theta_k))^2 \mid \theta_k]$$

and the estimator will be

$$\hat{\mu} = \mathbb{E}[\mu(\theta_k) \mid C_{obs,k}]$$

The mean square error induced by this estimator will be

$$E = \mathbb{E}[Var(\theta_k) \mid C_{obs,k}]$$

These estimators depend only on the history of observed claims $C_{obs,k}$ for customer type $k$. To make further use of these estimators one must find the conditional density of $C \mid \theta$. This is provided for instance in the case of continuous distributions by

$$f_\theta(y \mid C = c) = \frac{f_\theta(y) f_{C_1}(c_1 \mid \theta = y) \cdots f_{C_1}(c_n \mid \theta = y)}{f_C(c)}$$

As an example consider the case where the claims are Poisson distributed, and the parameters are gamma distributed. In this case the Bayes estimator may be calculated exactly (see e.g. [24]) as

$$\hat{\mu}_B = \frac{\gamma + \sum_{i=1}^{n} C_i}{\beta + n}$$

where $\{C_i\}$ is the observed data and $\beta$ and $\gamma$ are the parameters of the gamma distribution, or in the equivalent representation

$$\hat{\mu}_B = (1 - w)\, \mathbb{E}[\theta] \, + \, w \, \bar{C}$$

where $\bar{C}$ is the sample mean for the customer $k$ and $w = \frac{n}{n+\beta}$ is a positive weight. Therefore the estimator can be expressed as a weighted average of the expected heterogeneity parameter and the sample mean.

The above formulae provide, in principle, an answer for the Bayes estimator, but they cannot in general provide easy to use analytic estimates in cases other than when special distributions, such as for instance those of the above example, are used. In the general situation, which is likely to arise in practice, one may focus on finding linear estimators that minimize the mean square error even though the Bayes estimator may be of a different form. In other words, in order to compromise between the accuracy of the exact Bayes estimator and the feasibility of its calculation we decide to look for estimators in a particular class of estimators of the form

$$\hat{\mu} = a_0 + \sum_{k=1}^{r} \sum_{t=1}^{n_k} a_{k,t}\, C_{k,t}$$

where $C_{k,t}$ are the observed claims and $\{a_0, a_{k,t}\}$ are constants to be estimated. The estimation procedure will take place by solving the minimization problem for the mean square error using standard techniques from linear model theory. One particular instance leading to an easy to use estimator is the Bühlmann model [25], which leads to estimators for $\mu(\theta_k)$ of the form

$$\begin{aligned} \hat{\mu} &= (1-w)\, \mathbb{E}[\mu(\theta_k)] + w\, \bar{X}_k \\ w &= \frac{n_k Var(\mu(\theta_k)}{n_k\, Var(\mu(\theta_k + \mathbb{E}[Var(C_{k,t} \mid \theta_k)]} \end{aligned}$$

The delicate nature of the claims, which often lead to the need of very refined statistical study, in the sense that even the same customer or class of customers may react differently in similar situations, may need heterogeneous models where heterogeneity is allowed within each policy. This will allow the treatment of different types of privacy breaches for the same type of customer $k$. This model will be treated in a separate publication.

## 6 Use of the Risk Model for the Insurance and Risk Management of an IT Business Dealing with Personal Data

The above collective risk model may be used for the insurance and risk management of an IT business that deals with personal data.

### 6.1 Insurance of an IT Business Handling Private Data

Consider that the IT business enters into an insurance contract with an insurance firm that undertakes the total claim $X = L(t)$ that its clients may ask for, as a consequence of privacy breaches, over the time $t$ of validity of the contract. This of course should be done at the expense of a premium paid by the IT business to the insurer. How much should this premium be?

This is an important question, that has been dealt with in the past in the literature.

One possible principle regarding premium calculation can be the following: In some sense, the premium should be such that the insurer is in the mean safe, meaning that it minimizes the risk of ruin of the insurer. Certain premium calculations principles along this line are $\pi(X) = (1 + a)\mathbb{E}[X]$ where $a$ is called the safety loading factor, $\pi(X) = \mathbb{E}[X] + f(var(X))$ where usual choices for $f(x)$ are $f(x) = ax$ or $f(x) = a\sqrt{x}$. Since the expectation and the variance of $X = L(t)$ can be calculated either explicitly or approximated within the context of the model presented in this work, the premium calculation is essentially done.

Other possible principles are utility based, studying the incentive of the insurer to accept the insurance contract for the firm. Clearly, if the insurer decides over uncertain decisions using an expected utility function $U$, the total premium $\pi$ demanded by the insurer will be such that the lottery $w - X + \pi(X)$ is indifferent to the certain wealth $w$, where $w$ is the initial wealth of the insurer. Therefore, the premium will be the solution of the algebraic equation

$$\mathbb{E}[U(w - X + \pi(X))] = U(w)$$

which clearly depends on both the distribution of $X$ as well as on the choice of utility function for the insurer. One possible choice would be the exponential utility function $U(x) = -\frac{1}{a}\exp(-\alpha\, x)$ which leads to the premium calculation

$$\pi(x) = \frac{1}{a} ln(\mathbb{E}[e^{aX}]).$$

The quantity $\mathbb{E}[e^{aX}]$ can be calculated for the risk model introduced above through the properties of the characteristic function (see equation (2)).

The methods of Section 5 may be used for a more accurate and fair calculation of the premium charged by the insurer, through the design of more personalized contracts, dealing with particular firms who interact with particular types of data and customers. The sophisticated techniques of Credibility Theory [25], may be used to estimate the correct individual premia for particular classes of customers. For instance, assuming a linear premium calculation principle, the correct individual premium for a firm treating customers in class $k$ would be proportional to $\mu(\theta_k)$. But in practice both $\theta$ and $\mu(\theta)$ may be unknown, so it is necessary to estimate them. Bayes estimators such as those of Section 5 may be used to estimate the correct individual premium, using observations of the claims from particular types of customers. Such Bayes estimators are often called the best experience estimators, for obvious reasons. The collective premium for a collection of different customer types may be estimated by taking the expectation of the individual premium over the distribution of different customer types $U(\theta)$.

### 6.2 Optimal Insurance Coverage of the Firm

Assuming oligopoly in the insurance business sector, the insurer decides on the levels of the premium to be charged per unit of coverage. This assumption is not unreasonable, since, for such specialized contracts, we expect that only a small number of insurance companies will be interested in offering them.

Once the premium levels are set, using the principles provided in Section 6.1, the firm may decide on the optimal coverage that it will buy from the insurer. This may be done by considering an optimization problem. If the premium per unit of coverage is $\pi$, and the firm decides to cover itself for the total compensation $qX$ given that claims $X$ occur then it faces the lottery $w - X + qX - \pi qX$. It will choose the level of coverage $q$ by solving the maximization problem

$$\max_q \mathbb{E}[U(w - X + q\,X - \pi\,q\,X)]$$

given $\pi$. This will give the optimal coverage for the firm.

## 6.3 Risk Management of the IT Firm

For the risk management of an IT business handling personal data one may ask, what is the sum that is in danger at time $t$ for the business at some certainty level $\alpha$? This is the value at risk ($VaR$) for the IT firm which is defined through the quantile of the random variable $L(t)$. More precisely, the value at risk of the firm at time $i$ with confidence level $\alpha$ is $VaR(L(t);\alpha) = x$, where $P(L(t) > x) = \alpha$ for some $\alpha \in (0, 1)$. This corresponds to the largest sum that the firm is jeopardising at time $t$ with a confidence level $\alpha$. This quantity which is very important for the financial decisions of the firm can be calculated or approximated through our collective risk model.

For instance, in the case of the renewal model for the arrival of claims one may use the large deviation estimates given by (1) to provide estimates for the value at risk of the firm. In other, more complicated models, simulations may provide an answer.

## 6.4 Other Applications

The applicability of our collective risk model is by no means limited to the above applications.

Another alternative is its use for the design of contract structures between different firms handling sensitive data, or between such firms and insurers so as to allow for the optimal risk transfer and the best possible coverage. One may thus define the analogues of credit swaps or other credit derivatives that will effectuate the optimal risk transfer.

Another application of the proposed risk model is the study of an optimal insurance contract offering optimal coverage to two firms A and B, where A is assumed to be a contractor, subcontracting a project to B that is assumed to be of questionable credibility. As such, B may deliberately reveal private data of the clients of A for its own interest, thus exposing A to possible claims from its clients. One possible way of covering itself against this situation is to enter into a joint insurance contract so as to optimally cover its possible losses. In [26] we have studied the design of the optimal contract for this situation, taking for granted the possible loss $L$ for a security violation or privacy breech. The collective risk model proposed here, may be used within the context of [26] to refine the modeling of the subcontracting situation in the case of privacy.

## 7 A Simulated Example

In order to obtain a feeling of the practical aspects of the previous discussion we present a simple simulated example. We consider an IT firm that is worried about possible privacy violation claims by its clients. We make the following assumptions:

(1) Under no privacy violation incidents, the end of period wealth of the IT firm will be $W = 100\,000$.
(2) The IT firm has 100 clients
(3) Each client has income $y$ drawn from a Pareto distribution with mean 20 000, mode 10 000 and Pareto index 3
(4) Each client has personal characteristics described by a 10-dimensional vector $z$. Each coordinate of $z$ can take a value equal to either 1 or 0 with probability $0, 5$. A coordinate equal to 1 means that the client exhibits the corresponding characteristic, while a coordinate equal to 0 signifies that the client lacks the corresponding characteristic.
(5) Each client's level of satisfaction at each privacy state $i \in \{0, 1\}$, is described by a random utility function $u_i(y, z) + \epsilon_i$ with $u_i(y, z) = y + b_i z$, $b_0 - b_1 = (200, \cdots, 200)$ and error term $\epsilon = \epsilon_0 - \epsilon_1$ normally distributed with mean 0 and variance 1.
(6) The number of claims from these clients within the next period follows a Poisson distribution with parameter $\lambda = 50$.

The next figure illustrates the resulting distribution of the total claim that the IT firm may faces under the above assumptions. The mean of the distribution is approximately 50 000 and the standard deviation is about 7 500, while with 95% confidence the total claim will not exceed 63 000 .

Suppose now that the IT firm is considering to purchase insurance in order to cover against possible privacy violation claims. Let $\pi(X)$ denote the premium that an insurance company is charging in order to offer complete coverage to the IT firm for the next period. By the expression 'complete coverage', we mean that if the IT firm has to pay a compensation $x$ to its clients for privacy violations claims during the next period, then the insurance company will reimburse the IT firm the full amount $x$. However it may not be optimal for the IT firm to obtain complete coverage at the required premium $\pi(X)$ that the insurance company is charging. Therefore, the IT firm faces the problem of deciding what is the optimal percentage $q$ of coverage that should be bought from the insurance company; the IT firm will pay a premium $q\pi(X)$ to the insurance company and if during the next period the IT firm is required to pay an amount of $x$ to its clients because of privacy violation claims, then the insurance company will compensate the IT firm by an amount of $qx$.

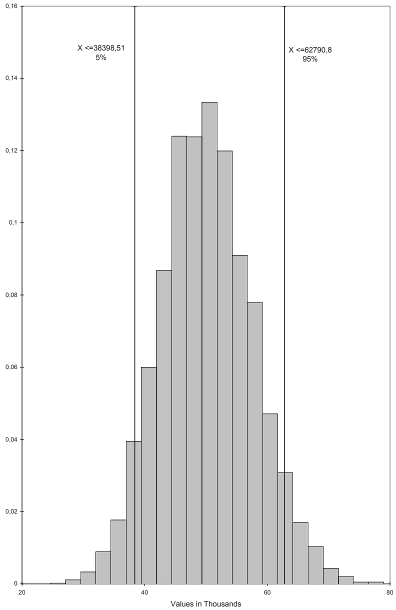

**Fig. 1.** The distribution of the total privacy violation claims under assumptions 1-6. The expected value is equal to approximately 50 000 and the standard deviation is about 7 430. With 95% confidence, the total claim will not exceed 63 000.

Assume moreover that the IT firm, deciding on the basis of expected utility, has preferences with regard to end of period wealth that are described by a utility function $U(x)$. Then, given the insurance premium $\pi(X)$, the IT firm has to solve the maximization problem

$$\max_q \mathbb{E}[U(W - X + q\,X - q\,\pi(X)]$$

To make things more concrete, let us consider two cases of utility functions of the IT firm, an exponential one given by $U(x) = 1 - \exp(-0,003\%\,x)$ and a logarithmic one given by $U(x) = \ln(x)$. It turns out that, no matter which utility function is used, if the premium demanded by the insurance company is $\pi(X) = 55\,000$ then $q = 0$, i.e. the IT firm is not willing to obtain any insurance at all for such a price. If however $\pi(X) = 50\,000$ (which is the expected value of the total claim) then, no matter which utility function is used, it turns out that $q = 100\%$, i.e. at this price the IT firm is willing to obtain full coverage. Finally, for a price $\pi(X) = 50\,200$ it turns out that $q = 80\%$ in the case of exponential utility, while $q = 71\%$ in the case of logarithmic utility.

Suppose now that, after further investment in infrastructure and strengthening of security procedures, the IT firm revised its model about the number of privacy incidents claims arriving within the next period and has estimated that it follows a Poisson process with parameter $\lambda = 10$. Then the distribution of the total claim exhibits a mean of approximately 10 000 and a standard deviation of about 3 300, while with 95% confidence the total claim will not exceed 16 000. This kind of analysis may complement a cost benefit analysis of the IT company well with regard to the level of security related investments.

## 8 Conclusions

Management of personal data is today becoming a crucial need for many users, applications and IT firms. In this paper a risk model which models the risk that an IT firm is exposed to, as a result of privacy violation and possible disclosure of personal data of her clients, has been proposed. The basis of the model is a RUM, which aims at capturing the subjective nature of the privacy value. A collective risk model has also been proposed, modeling the exposure of the firm over a certain time period, for homogeneous and inhomogeneous client populations. The model has been used for designing and valuating insurance contracts that optimally cover the firm or for risk management purposes. The model may be utilized in the framework of many other interesting applications.

## References

1. Warren, S.D., Brandeis, L.D.: The rights to privacy, Harvard Law Review, vol. 5(1), pp. 193–220 (1890)
2. Westin, A.F.: Privacy and Freedom. Atheneum, New York (1967)
3. Gritzalis, S.: Enhancing Web privacy and anonymity in the digital era. Information Management and Computer Security 12(3), 255–288 (2004)
4. Phelps, J., Nowak, G., Ferrell, E.: Privacy Concerns and Consumer Willingness to Provide Personal Information. Journal of Public Policy and Marketing 19(1), 27–41 (2000)
5. Fox, S.: Trust and privacy online: Why Americans want to rewrite the rules, Tech. rep. The Pew Internet & American Life Project, Washington D.C (2000)
6. Culnan, M.J., Milne, G.R.: The Culnan-Milne Survey on Consumers and Online Privacy Notices: Summary of Responses (December 2001), `http://www.ftc.gov/bcp/workshops/glb/supporting/culnan-milne.pdf`
7. Hoffman, D.L., Novak, T.P., Peralta, M.A.: Building Consumer Trust Online. Communications of the ACM 42(4), 80–85 (1999)
8. Milberg, S.J., Smith, H.J., Burke, S.J.: Information Privacy: Corporate Management and National Regulation, Organization Science, vol. 11(1), pp. 35–57 (2000)
9. Smith, H.J.: Information Privacy and Marketing: What the U.S. Should (and Shouldn't) Learn from Europe, California Management Review 43(2), 8–33 (2001)
10. Smith, J., Milberg, S., Burke, S.: Information Privacy: measuring individuals' concerns about organizational practices. MIS Quarterly 20, 167–196 (1996)
11. Dhillon, G.S., Moores, T.T.: Internet privacy: Interpreting key issues. Information Resources Management Journal 14(4), 33–37 (2001)
12. Cranor, L.F., Reagle, J., Ackerman, M.S.: Beyond concern: Understanding Net Users's Attitudes About Online Privacy, AT&T Labs -Research Technical Report TR 99.4.3 (1999), `http://www.research.att.com/library/`

13. Wang, H., Lee, M.K.O., Wang, C.: Consumer Privacy Concerns about Internet Marketing. Communications of the ACM 41(3), 63–70 (1998)
14. Ackerman, M.S., Cranor, L.F., Reagle, J.: Privacy in e-commerce: examining user scenarios and privacy preferences. In: Proceedings of the First ACM Conference on Electronic Commerce, pp. 1–8 (1999)
15. Spiekermann, S., Grossklags, J., Berendt, B.: E-privacy in 2nd generation e-commerce: privacy preferences versus actual behavior. In: Proceedings of the 3rd ACM Conference on Electronic Commerce, pp. 38–47 (2001)
16. Hann, I., Hui, K.L., Lee, T.S., Png, I.P.L.: Online information privacy: Measuring the cost-benefit trade-offs. In: Proceedings of the Twenty-Third International Conference on Information Systems, Barcelona, Spain, pp. 1–10 (2002)
17. Chellappa, R.K., Sin, R.: Personalization Versus Privacy: An Empirical Examination of the Online Consumer's Dilemma. Information Technology and Management 6(2-3) (2005)
18. Acquisti, A., Grossklags, J.: Losses, gains, and hyperbolic discounting: An experimental approach to information security attitudes and behavior. In: 2nd Annual Workshop on Economics and Information Security (WEIS) (2003)
19. Acquisti, A., Grossklags, J.: Privacy and Rationality in Individual Decision Making. IEEE Security and Privacy 3(1), 26–33 (2005)
20. Westin, A.F.: Privacy and American Business Study (1997), http://www.pandab.org
21. Faja, S.: Privacy in E-Commerce: Understanding user trade-offs. Issues in Information Systems VI(2), 83–89 (2005)
22. Laudon, K.C.: Markets and Privacy. Communications of the ACM 39(9), 92–104 (1996)
23. Acquisti, A., Friedman, A., Telang, R.: Is there a cost to privacy breaches? an event study. In: Workshop on the Economics of Information Security (WEIS) (2006)
24. Mikosh, T.: Non-life insurance mathematics: An introduction using stochastic processes. Springer, Heidelberg (2006)
25. Buhlmann, H., Gisler, A.: A course on credibility theory and its applications. Springer, Heidelberg (2005)
26. Gritzalis, S., Yannacopoulos, A.N., Lambrinoudakis, C., Hatzopoulos, P., Katsikas, S.K.: A probabilistic model for optimal insurance contracts against security risks and privacy violations in IT outsourcing environments. International Journal of Information Security 6(4), 197–211 (2007)