RC23672 (W0507-148) July 20, 2005
Mathematics

IBM Research Report

Modeling Operational Risks in Business Processes

Feng Cheng, David Gamarnik, Nitin Jengte, Wanli Min, Bala Ramachandran
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Modeling operational risks in business processes

Feng Cheng, David Gamarnik, Nitin Jengte, Wanli Min and Bala Ramachandran
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

July 18, 2005

Abstract

We propose an approach for modeling and analysis of operational risks in financial institutions. The importance of understanding and mitigating operational risks had been gaining a growing attention recently (1) and led to the necessity for providing systematic foundations for modeling such risks. In this paper, we propose a methodology for modeling operational risk based on business process models. By connecting the generation of a probabilistic network with the business process model, this approach enables changes in the operational risk model whenever different aspects of the business process in the financial institution changes. In addition, this can enable progress toward continual operational risk management, by automatically changing the parameters of the business process model based on monitoring the business process performance and cascading the change in the operational risk model, thereby synchronizing the model changes with the corresponding business process changes. We demonstrate this methodology with some examples including IT related risk, infrastructure and outsourcing related risks.

Keywords: Operational risk management, business processes, operational losses, risk modeling, probabilistic networks.

1 Introduction

The growing interest in operational risk management has been driven by a variety of factors, including recent high profile incidents such as those that occured in Barings Bank and the introduction of new regulations requiring businesses to measure and manage operational risk, such as the New Basel Capital Accord, known as Basel II (1). A prevailing definition of operational risk is given by the Basel Committee on Banking Supervision as "the risk of loss resulting from inadequate or failed internal processes, people or systems or from external events" (2). Financial institutions are endeavoring to develop approaches to measure, assess, monitor and manage their operational risks. This paper focuses on modeling approaches to assess and quantify operational risks.

Current approaches to operational risk modeling have to a large extent based on (a) statistical modeling of rare events and extreme value theory (3) (4), (5), (6)and (b) Bayesian networks (8). Commercial software is also available based on these techniques (see for example, (9) and (10)). In the statistical approach, operational loss events are collected in a loss database which is used to determine and fit appropriate frequency and severity distributions for loss events. These distributions are used in Monte-Carlo simulations to predict future operational loss distributions. This approach is useful for measuring operational risks and is a commonly used approach for economic capital allocation which is one of the key requirements for Basel II compliance. A key issue with the statistical approach is

that limited data is available on operational risk events and enterprises are seeking to address that either by initiating collection of loss event data and creation of operational risk loss databases based on intra-enterprise risk events and/or by using external loss databases from software vendors such as SAS, FitchRisk etc. Statistical approaches are however of limited utility for operational risk management since they do not provide insight as to how different factors relating to people, systems and processes can be modified to control and manage operational risk. Causal models such as Bayesian networks are useful in this context to assess and predict the effect of different causal factors on the overall operational risk. A technical issue with the Bayesian network approach is that the inferencing problem in Bayesian networks is in general a computationally hard problem, i.e. NP-hard problem, which implies that the computational effort grows exponentially as a function of input parameters such as risk events etc. (12), (13). Researchers are addressing this by developing more efficient algorithms for inferencing in Bayesian networks. A more fundamental limitation of this approach is that there is no systematic method known to construct these networks linked to business processes of a financial institution and hence require considerable effort to develop and maintain these models since the business processes and its supporting infrastructure comprised of people, systems etc. undergo constant change.

In this paper, we propose an alternative approach for operational risk modeling based on automatically developing a probabilistic network based on a description of the operational business processes in an enterprise, knowledge of its underlying resources, physical and logical infrastructure and the risks contained therein and subsequently solving the network. The advantage of this approach is that all enterprises seek to map and model their business processes. By connecting the generation of the probabilistic network with the business process model, this approach enables changes in the operational risk model whenever different aspects of the business process in the financial institution changes. In addition, this can enable progress toward continual operational risk management, by automatically changing the parameters of the business process model based on monitoring the business process performance and cascading the change in the operational risk model, thereby synchronizing the model changes with the corresponding business process changes. This methodology can further be used as a basis to evaluate different countermeasures for operational risk control and mitigation. A general methodology for risk control consists of three steps: identification of risks, quantitative analysis of identified risks and the construction of a plan to control the risks, given a risk tolerance level. The first step involves identifying the causes of operational risk, then estimating the frequency and severity of risk events. The second step includes analyzing various identified risk events and their impacts on business operations by a sound quantitative approach that will reveal the distribution of loss. It is at this step that different models enter. In the third step dominant risk events are identified and the cost-effectiveness of various risk countermeasures are estimated, on the basis of which an optimized risk control strategy is determined.

In Section 2 of the paper, we describe the overall proposed approach and provide a problem formulation. Sections 3 and 4 provide approaches for addressing task specific and flow specific risks and the operational risk computation methodology. Section 5 provides illustrative examples and describes how this method can be used for assessing countermeasures for operational risk management. Section 6 describes specific issues in using the proposed method for modeling outsourcing risk and illustrates it with an example. In Section 7, we make some concluding remarks and provide directions for further work.

2 Proposed Methodology

The meta-model underlying our proposed approach is outlined in Figure 1. In general, a business process can be defined using combinations of constructs, such as processes, sub-processes, tasks (alternatively referred to as activities), resources, forking/decisions, merge/joins, etc. Using these constructs, the

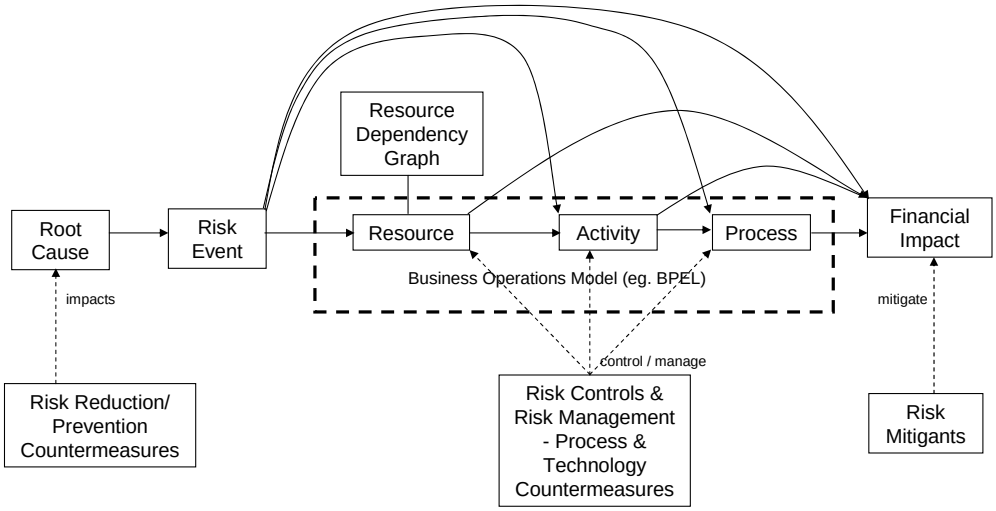
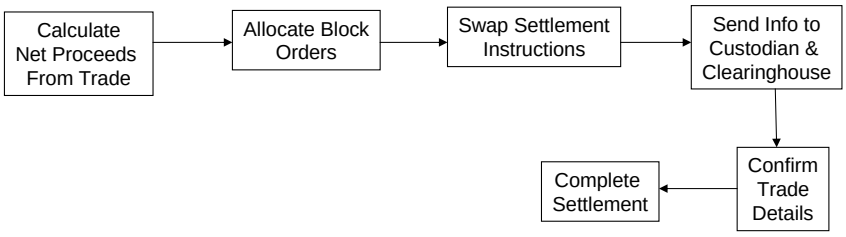


Figure 1: Meta-model for Operational risk modeling

process can be modeled in different layers of granularity in a hierarchical manner. We can incorporate the notion of operational risks with business processes. Operational risk events such as failure of an IT infrastructure component, natural disaster etc. can impact the business process resources. In general, risk events can also potentially impact activities or processes or even cause direct financial losses - for example employee fraud. The reader is referred to (2) for a categorization of different types of risk events. Risk events are described by an associated frequency distribution and severity distribution (for example, duration of an outage resulting from a risk event). Risk events may also have root causes associated with them. Operational risk can be managed by applying risk controls, countermeasures and mitigants. In the next subsection, we describe a formulation of this model with the objective of quantifying the operational risks. Figure 2 describes a simple business process for illustrative purposes that will also be used later in this paper to illustrate our methodology (this is a simplified version of a Broker / Dealer process). This sub-process has six tasks. Sending information to the custodian and clearinghouse requires a communication gateway which has a risk of failure. The number of transactions per unit of time (hour) is uncertain, but can be described by some probability distribution. A fixed revenue is generated for every completed transaction.

2.1 Formulation

The essential elements of our operational risk model are



Resources & Events:
"Send Info to Custodian & Clearinghouse" task needs "Communication Gateway with Clearinghouse" resource
Associate "Failure" risk event with "Communication Gateway with Clearinghouse" resource

Figure 2: A simple Broker Dealer process

1. Events $E_1, E_2, \ldots, E_K$ which may cause failures. We consider the occurrence of failures during some fixed time period (say a month or a year). Event $E_i, 1 \leq i \leq K$ occurs L_i times during a period of interest, where L_i is a random variable which distributed according to a probability distribution $F_{L,i}(l), l \in \mathbb{N}$, independently for all i: $F_{L,i}(l) = \mathbb{P}(L_i = l)$. Each time an event E_i occurs, its severity is also a random variable D_i with the probability distribution $F_{D,i}(t), t \geq 0$. The severity can for instance represent the duration of a resource failure or even potentially the financial impact of specific risks like fraud.

2. A set of tasks, denoted henceforth as $T_1, T_2, \ldots, T_N$.

3. A set of resources, denoted henceforth as $r_1, r_2, \ldots, r_J$.

We assume that in normal state all the tasks $T_i, 1 \leq i \leq N$ are continually operated. The cost is incurred if one or several tasks are interrupted. In particular, we associate costs $C_1, C_2, \ldots, C_N$ with *non-execution* of tasks $T_1, T_2, \ldots, T_N$. Precisely, we mean that each C_i is a random function $C_i : \mathbb{R}_+ \to \mathbb{R}_+$. The value $C_i(t)$ represents the cost of not operating task T_i for t time units. The probability distribution of C_i is denoted by $F_{C,i}(x,t), t \geq 0$. Thus

$$F_{C,i}(t) = \mathbb{P}(C_i \leq x | \text{duration of failure of task } T_i = t).$$

A typical case of this is linear costs: $C_i(t) = C_i t$ for some random constants C_i. This is a realistic assumption when operational losses depend on the number of transactions that have not been processed, which is a typical case. But in general, this can take any functional form, such as a step function. In the linear case $F_{C,i}(x)$ represents the probability distribution of the cost of not operating task T_i for one time unit.

The three sets $\{E_i\}, \{T_i\}, \{r_i\}$ represent three layers of a network which we represent as a directed graph. The node set of this graph is the union $\{E_i\} \cup \{T_i\} \cup \{r_i\}$. There is a directed edge from event E_i to resource r_j if and only if the adversary event E_i causes a disruption of the resource r_j. Construct an $K \times J$ matrix $\mathcal{E}$ where (i,j) entry is one if the occurrence of the event E_i disrupts the resource r_j and zero otherwise. In other words the corresponding entry is one if there is a directed link from E_i into r_j and zero otherwise. e_i^m represents an m-dimensional unit vector with unity in i-coordinate. Also, let e^m denote the m-dimensional vector of all ones. $\text{sgn}(x)$ is the standard sign function that takes the value 1 or -1 depending on whether x is positive or negative and $\text{sgn}(x) = 0, x = 0$. When this function is applied to a matrix, it is applied to every element in the matrix. Then for every event E_i we can represent the collection of resources affected by event E_i as

$$r(E_i) \triangleq (e_i^K)^t \mathcal{E}. \tag{1}$$

In particular $r(E_i)$ is a J-dimensional zero/one vector with j-th entry equal to unity iff resource r_j is affected by event E_i. We also think of $r(E_i)$ as a subset of $\{r_1, \ldots, r_J\}$.

Similarly, we construct a directed link from r_i to T_j if the disruption of the resource r_i causes non-execution of task T_j. We denote by $\mathcal{R}$ a $J \times N$ zero/one matrix with (i,j)-th entry equal to one if resource r_i causes non-execution of task T_j (there is a directed link from r_i to T_j) and zero otherwise. Then for every resource r_i the collection of affected tasks can be written as

$$T(r_i) \triangleq (e_i^J)^t \mathcal{R} \tag{2}$$

Then we may combine (1) with (2) to conclude that for every event E_i the collection of tasks not executed if the event E_i occurs is

$$T(E_i) = \text{sgn}[(e_i^K)^t \mathcal{E}\mathcal{R}]. \tag{3}$$

The function sgn is utilized since the vector $(e_i^K)^t \mathcal{E}\mathcal{R}$ contains multiplicities – a given event can affect a given task via many resources. In addition

$$T = \mathrm{sgn}[\mathcal{E}\mathcal{R}], \tag{4}$$

gives the matrix whose (i,j)-th component is unity if event i affects task j, and is zero otherwise. We thus summarize our derivations as follows.

The three layers $\{E_i\}, \{T_i\}, \{r_i\}$ of our network as well the directed links between them, frequency F, severity (duration) D and the cost functions C constitute our Causal Network for Operational risk based on the business process.

3 Computation of operational losses

With the model constructed in the previous section, our next goal is to compute the risk exposure, given the primitives of the model. Specifically, our goal is to compute the probability distribution of the operational losses. While the model seems on the surface to be quite complex, it turns out that the required computations can be performed using fairly efficient computational procedures involving linear algebraic operations and convolution operation on probability distributions. This is particularly true in a special case when cost functions $C_i(t)$ are linear: $C_i(t) = C_i t$. Thus, throughout this subsection we assume that this is indeed the case. Then formula (4) implies that if an operational risk event E_i occurs once and has a duration of D time units, then the overall operational losses are found as

$$\sum_{j \in T(E_i)} C_j D.$$

The occurrence of E_i event L_i times with durations $D_1^i, D_2^i, \ldots, D_{L_i}^i$ leads to cost

$$\sum_{j \in T(E_i)} C_j \sum_{1 \le k \le L_i} D_k^i,$$

and the overall cost is found as

$$C_{\text{total}} \triangleq \sum_{1 \le i \le K} \Big(\sum_{1 \le k \le L_i} D_k^i \Big) \Big(\sum_{j \in T(E_i)} C_j \Big), \tag{5}$$

Applying independence of the random variables involved, we obtain expected overall operational loss as

$$\mathbb{E}[C_{\text{total}}] = \sum_{1 \le i \le K} \mathbb{E}[L_i]\mathbb{E}[D_i] \sum_{j \in T(E_i)} \mathbb{E}[C_j].$$

For computational purposes it is usually convenient to represent the expression above in matrix form. Thus let L, D denote the diagonal matrices corresponding to vectors $(\mathbb{E}[L_i]), (\mathbb{E}[D_i])$ respectively, and let C denote the vector with components $\mathbb{E}[C_j]$. Then we may write the total cost in matrix form as

$$\mathbb{E}[C_{\text{total}}] = (e^K)^t LDTC, \tag{6}$$

where the matrix T was defined in the previous section. Thus we obtain a very simple matrix form solution for the expected operational losses corresponding to our model.

Yet, the expectation is in many cases not a very relevant quantity, especially in the operational risk context. We need to determine the overall operational loss distribution that can be used to infer the

likelihood $\mathbb{P}(C_{\text{total}} > x)$ of the overall cost C_{total} exceeding a threshold x. These tails probabilities then can be used for finding an appropriate level x for which the exceedence probability $\mathbb{P}(C_{\text{total}} > x)$ is acceptably small (based on internal risk policies, or Basel Accord recommendations). Finding the tail probabilities for the expression (5) is a straightforward application of convolution operations, and the pseudocode for this is given in Fig ???

3.1 Computation of the loss function

The computation is carried out in the following steps.

Step 1. Compute $T(E_i)$ using Eq.(3) - the collection of tasks $(j = 1, \ldots, N)$ not executated if event $E_i, i = 1, \ldots, K$ occurs.

Step 2. Compute the probability distribution for the sum of durations of all event i occurrences,

$$\mathbb{P}(D^i = n) = \sum_{l \geq 0} \mathbb{P}(L_i = l)\mathbb{P}(\sum_{0 \leq k \leq l} D_k^i = n | L_i = l),$$

where $\sum_{0 \leq k \leq l} D_k^i = n$ is the probability distribution for the sum of durations of l occurrences of event i given by the l-fold convolution of $\mathbb{P}(D^i = n)$.

Step 3. Compute the probability distribution for the sum of durations of all event occurrences that cause non-execution of task j,

$$\mathbb{P}(D_j = n) = \mathbb{P}(\sum_{i: j \in T(E_i)} D^i = n),$$

which is obtained by convolution of the random variables D^i s.t. $j \in T(E_i)$.

Step 4. Compute the loss distribution due to non-execution of task j, which is given by

$$\mathbb{P}(S_j = nC_j) = \mathbb{P}(D_j = n).$$

Step 5. Rescale or resample the loss distribution $\mathbb{P}(S_j = nC_j)$ for all j. Choose $\Delta C > 0$ such that $k_j = C_j/\Delta C, j = 1, \ldots, N$, are integers. Then define

$$f_j(m) \triangleq \begin{cases} \mathbb{P}(S_j = m\Delta C) = \mathbb{P}(S_j = nC_j), & m = nk_j, \\ 0, & o.w. \end{cases}$$

Step 6. Compute the loss distribution due to non-execution of all tasks given by the convolutions of $f_j(m), j = 1, \ldots, N$, i.e.,

$$\mathbb{P}(S_{total} = m\Delta C) = f_1(m) * f_2(m) * \cdots * f_N(m).$$

Note that alternative methods could be used for Step 5.

3.2 Computation of the losses in transform domain

It is at times useful to solve for the distribution of random quantities in transform domains. In a special case when the cost function C_j is not only linear but deterministic, the corresponding Laplace transform equation for C_{total} comes out again to be quite simple due to the simple linear structure of the equation and independence assumption on probability distributions. We introduce the moment generating function of our random quantities: $g_{L,i}(z) = \sum_{l\geq 0} z^l \mathbb{P}(L_i = l), g_{D,i}(z) = \int_0^\infty e^{zt} dF_{D,i}(t), g_{C,i} = \int_0^\infty e^{zt} dF_{C,i}(t)$. We assume that each cost rate C_j takes a deterministic value c_j and for simplicity we let

$$c(i) = \sum_{j \in T(E_i)} c_j.$$

Recall the following property of moment generating functions: if X is a continuous random variable with transform $f_X(z)$ and c is a constant, then cX has a transform $f_{cX}(z) = f_X(cz)$. Using standard properties of moment generating functions, we obtain

$$g_{C_{\text{total}}}(z) = \prod_{1\leq i\leq K} g_{L,i}\Big(g_{D,i}(c(i)z)\Big).$$

When C_j is not however a deterministic function, there is, unfortunately, no simple formula for the moment generating function. In this case the probability distribution of C_{total} needs to be computed directly.

4 Operational risks in business processes

It is possible that the cost functions C_j are not given directly, but rather come up from structural properties of the underlying business process. For example, if a server is not available for some time period several types of transactions (order completion, price quote checks, etc.) may not be completed depending on the applications deployed on the server, resulting in losses. These losses can be a combination of different types of losses depending on the type of processes the server is supporting. For example, the loss corresponding to failure of order completion is equal to revenue loss corresponding to the product not purchased. As another example, account opening process is associated with a stream of tasks like taking data, checking validity of the personal information, entering the data into the database, etc. Should there be a problem with any of these tasks the entire flow is interrupted.

In this subsection, we model explicitly the types of losses arising from various workflow processes. The computations are again fairly straightforward and are a combination of convolution and matrix algebraic operations. Business process modeling tools such as WBI Modeler provide a convenient framework for implementation of this approach.

We adopt exactly the model of the previous subsection, except for now we have a more detailed way for modelling the cost functions C_j. In particular, in addition to events E_i, tasks T_i and resources r_i, we introduce the notion of a *workflow*. A workflow F_j is associated with a particular ordered sequence of tasks $T_1^j, \ldots, T_{n(j)}^j$. The sequence of workflows is denoted by $F_1, F_2, \ldots, F_w$. Each workflow is also associated with an arrival rate λ_j. This is the rate with which the flow passes through the system. To each pair of flow F_j and task T_i^j which is a part of the flow F_j, we associate a queue (also referred to as a virtual buffer) B. We enumerate all the buffers as $B_1, B_2, \ldots, B_n$. In particular, $n = \sum_{1\leq j\leq w} n(j)$. We generate an n-dimensional vector λ, representing the external arrivals to the business process as follows. If buffer B_i corresponds to the first buffer of its flow F_j (that is the buffer corresponds to the task T_1^j),

then the entry is λ_j, otherwise the entry is zero. For example say we have two tasks T_1, T_2 and two flows F_1, F_2 with rates λ_1, λ_2. First flow F_1 goes through T_1 then T_2. Second flow goes through T_2 only. Then we have 3 buffers B_1, B_2, B_3 corresponding to pairs $(F_1, T_1), (F_1, T_2)$ and (F_2, T_2) respectively. The corresponding vector λ is $(\lambda_1, 0, \lambda_2)$.

In addition we generate an $n \times n$ routing matrix $P = (p_{i,j})$. Each entry $p_{i,j}$ represents the proportion of flow from buffer B_i which is routed into buffer B_j. If all of the flow to queue i is routed to queue j, then $p_{i,j} = 1$, else it is the fraction of flow of a particular class to queue i that is routed to queue j. In our earlier example, the flow from buffer B_1 goes entirely into buffer B_2, and the flow from buffer B_3 exits the network. So the corresponding matrix has values $p_{1,2} = 1$ and $p_{i,j} = 0$ for all the other values $1 \leq i, j \leq 3$. However, if only half of the flow from buffer B_1 goes into buffer B_2, then $p_{1,2}$ is set to $1/2$. It is sometimes the case that a single flow on the output generates several input flows. For example an online purchase completion generates one task corresponding to *record the transaction in a database*, one task corresponding to *generate the shipment instance* and one task corresponding to *send an email notification to the buyer* (and whatever additional transactions/tasks). In this case assuming that the buffer B_1 corresponds to task *online purchase*, and the buffers B_2, B_3, B_4 correspond to *record, generate shipment* and *email notification*, respectively, then the corresponding entries in the matrix P are $p_{1,2} = p_{1,3} = p_{1,4} = 1$.

How do we associate cost with our model? There are two typical cases corresponding to whether the cost is associated to a particular task or to the whole flow. We consider these cases separately as they lead to somewhat different modeling constructs.

4.1 Buffer and task specific cost structure

To each queue B_i we associate a cost rate c_i. An example of this case is - A task for processing securities that incurs a cost for processing each and produces a revenue. For simplicity, we assume that c_i takes deterministic (non-random) value. Thus the cost is a linear deterministic function of time. c_i indicates the cost per unit of time when the workflow passing through the queue B_i is interrupted. For example, if the flow F_j associated with queue B_i was interrupted for τ time units then the incurred cost is $c_i \lambda_j \tau$ (note the dependence on λ since the cost is really associated with the volume of tasks which were not processed during time interval of length τ, and this is given by $\lambda_j \tau$). We let C denote the diagonal $n \times n$ matrix with c_i-s on diagonal.

In case cost is naturally associated with entire task and not specifically buffers corresponding to a task, we instead construct a diagonal $N \times N$ matrix $\bar{C}$, with diagonal elements corresponding to cost rate per the corresponding task, where N is the number of tasks. Let also $\mathcal{T}$ denote the $N \times n$ matrix. We let the (i, j)-th component of $\mathcal{T}$ be one if task i contains buffer j and zero otherwise. It can be established using standard linear algebraic analysis, that the total flow passing through queue B_j is the j-th component of the $(n \times 1)$ vector $[I - P^t]^{-1}\lambda$, where I is the $n \times n$ identity matrix. This matrix representation of the flow passing through buffers is often used in queueing network literature in the analysis of queueing traffic flows passing through the buffers of processing stations, (see for example (11)). Then the total cost per unit of time "flowing" through buffer B_j is the j-th component of the vector $C[I - P^t]^{-1}\lambda$. When the cost is associated with a task, total cost "flowing" through any task T_i per unit of time is then the i-th component of the vector $\bar{C}\mathcal{T}[I - P^t]^{-1}\lambda$.

We now can tie this construct with our operational risk modeling framework. If a certain task T_i is interrupted for τ time units then the incurred cost is τ times the i-th component of $\mathcal{T}C[I - P^t]^{-1}\lambda$ ($\bar{C}\mathcal{T}[I - P^t]^{-1}\lambda$ for task specific cost structure). We write this as $(e_i^N)^t \mathcal{T}C[I - P^t]^{-1}\lambda\tau$ ($(e_i^N)^t \bar{C}\mathcal{T}[I - P^t]^{-1}\lambda\tau$), where e_i^N is the N-dimensional unit vector with unity in i-th position and zero everywhere else. Recall our notation for the frequency L_i and the frequency distribution F_{L_i} of event E_i, the duration

D_i and the duration distribution $F_{D_i}(t)$ of event E_i, and $R = \{r_1, \ldots, r_J\}$ - the list of resources. Recall also the matrices $\mathcal{E}$ and $\mathcal{R}$ connecting these constructs. We note again, that given an event E_i the collection of tasks affected by this event is $\operatorname{sgn}((e_i^K)^t\mathcal{E}\mathcal{R})$. Then if the duration of a particular event E_i is equal to $D_i = \tau$, then the total associated cost is given by $\operatorname{sgn}((e_i^K)^t\mathcal{E}\mathcal{R})\mathcal{T}C[I - P^t]^{-1}\lambda\tau$ $(\operatorname{sgn}((e_i^K)^t\mathcal{E}\mathcal{R})\bar{C}\mathcal{T}[I - P^t]^{-1}\lambda\tau)$. We now define

$$C(E_i) = \operatorname{sgn}((e_i^K)^t\mathcal{E}\mathcal{R})\mathcal{T}C[I - P^t]^{-1}\lambda \tag{7}$$

for buffer specific costs and

$$C(E_i) = \operatorname{sgn}((e_i^K)^t\mathcal{E}\mathcal{R})\bar{C}\mathcal{T}[I - P^t]^{-1}\lambda \tag{8}$$

for task specific costs. Then the total cost is found as in (5) to be

$$C_{\text{total}} = \sum_{1\le i\le K} C(E_i) \sum_{1\le j\le L_i} D_j^i. \tag{9}$$

Just like (6) the total expected cost can be represented purely in matrix form. As in the previous subsection let L and D be the diagonal matrices of expected frequency and duration of adversary events, respectively. Then

$$\mathbb{E}[C_{\text{total}}] = e^t LD \operatorname{sgn}(\mathcal{E}\mathcal{R})\mathcal{T}C[I - P^t]^{-1}\lambda, \tag{10}$$

for buffer specific costs and

$$\mathbb{E}[C_{\text{total}}] = e^t LD \operatorname{sgn}(\mathcal{E}\mathcal{R})\bar{C}\mathcal{T}[I - P^t]^{-1}\lambda, \tag{11}$$

for task specific costs.

4.2 Flow specific cost structure

Suppose now costs are associated with flows $F_j, 1 \le j \le w$. In addition we assume that the routing matrix P consists of zeros and ones only. Physically, this corresponds to the case when the entire flow from one buffer flows either into some other buffer or leaves the system. In particular, the branching, merging, joint and forking constructs are not allowed. The reason is that in the presence of such constructs the notion of a particular flow is not well defined as the flow can be partitioned into several flows or be a joint of several flows. Then it is not clear whether cost can be associated with flows in any meaningful way.

Suppose a cost rate c_j is associated with a flow F_j. Let C be the corresponding diagonal matrix. Construct an $N \times w$ matrix $\mathcal{F}$ where (i, j) entry is one if flow F_j contains task T_i and zero otherwise. Construct also a w-vector $\bar{\lambda}$ with i-component equal to λ_i. That is $\bar{\lambda}$ is simply the vector of flow rates corresponding to the given collection of flows. Note the difference with λ where the flow rates were associated with buffers. Now, if a particular task T_i is interrupted for τ time units then the corresponding cost is found as $e_i\mathcal{F}C\bar{\lambda}\tau$. Then the cost associated with a particular event E_i with duration τ is found as

$$C(E_i) = \operatorname{sgn}((e_i^K)^t\mathcal{E}\mathcal{R}\mathcal{F})C\bar{\lambda}\tau.$$

Then the total cost is again as in (5)

$$C_{\text{total}} = \sum_{1\le i\le N} C(E_i) \sum_{1\le j\le L_i} D_j^i. \tag{12}$$

The corresponding matrix representation of the expected cost is then

$$\mathbb{E}[C_{\text{total}}] = e^t LD\,\text{sgn}(\mathcal{ERF})C\bar{\lambda}. \tag{13}$$

Whether the cost is associated with a task, buffer or a flow depends on the underlying structure of the risk model. Often the cost is incurred when a particular sequence of tasks (flow) is not executed. In this case it makes sense to associate costs with flows. In other cases, however charges are incurred when a particular resource or task in the network is not operating (for example charges associated with service level agreements in an outsourcing context). In this case it makes sense to associate costs with tasks. Note, that we could associate costs with resources as well, the derivation being pretty much the same, thus the details are omitted.

5 Illustrative Examples

We have developed a prototype for operational risk modeling using WBI Modeler. This is a general tool for Business Process Modeling that enables the modeling of business processes using combinations of predefined constructs. Some of the key constructs are defined below:

- Process: This is a group of tasks and other processes, thus enabling hierarchical decomposition of a business process into lower level processes and tasks.
- Task: This describes an activity in the business process and can have multiple characteristics such as costs, time to completion, task logic, resources required to execute the task etc. A business process describes an orderly flow between the tasks within it and is represented by directed connections between tasks.
- Information Artifacts: These describe items that will flow through the process at different stages.
- Forking / Decisions: A fork is the branching of an incoming connection to multiple outgoing connections. An incoming token is replicated for each outgoing branch. A decision is like a fork except that the selection of the outgoing branch is conditional either on the result of an expression (an if-then-else type of decision) or on a random selection. A decision may have multiple outgoing connections for each incoming connection.
- Merge / Join: This is the converse of branching, where multiple input flows come together to pursue a common output flow. In joining, tokens arriving through the multiple flows all have to arrive before the common output connection is triggered (AND-logic). In merging, the output flow is triggered whenever a token arrives through any of the input connections (OR-logic).
- Resource: Entities which tasks require to perform their function and these can be perishable or non-perishable. Resources can be assigned calendars describing their availability, e.g. an employee with availability of 9am-5pm, Monday-Friday except holidays. Resources can have various costs associated with them such as cost per quantity or units of measure.
- Loops: This specifies a repetitive invocation of some other process. The types of repetitions are analogous to the FOR loops, WHILE-DO loops and DO-WHILE loops found in various programming languages.

A screenshot of this prototype is shown in Figure 3.

We now illustrate the approach proposed in the earlier sections using the simple Broker-Dealer process described in Figure 1, Section 2. Our goal is to quantify the operational risk - the probability distribution of lost revenues.

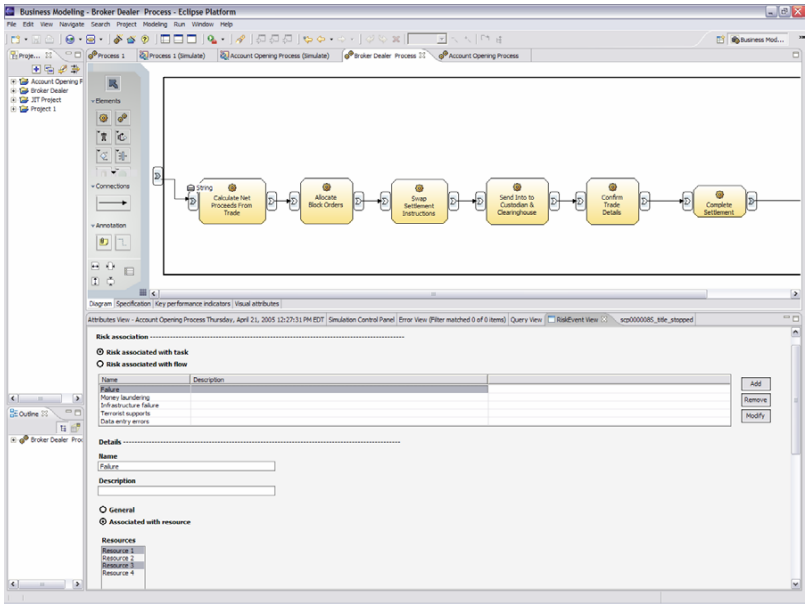


Figure 3: Operational risk modeling prototype using WBI Modeler

5.1 The distribution of losses resulting from a singular event

In order to perform the lost revenue (cost) analysis we need to identify the elements of the 3-layers of our operational risk model architecture. For our first example assume that there is only one type E of adversary event: communication gateway interruption. Assume that the duration of the typical interruption can take two values: $D = D_1$, with probability (w.p.) p_1 (**event I**) and $D = D_2$, w.p. $p_2 = 1 - p_1$ (**event II**). The associated unique resource for our model is r–"communication gateway". There is only one task T which may be interrupted by failure of the resource r. It is generically referred to as "Transaction" task. Interrupting task T for τ time units results in losses described as lost revenues. The revenue flow per unit of time is assumed to be distributed as a Poisson random variable with parameter λ_a and the revenue per unit of time is H dollars in commission. We first ignore the issue of frequency of interruptions (random variable L in our model formulation) and assume that interruption happens exactly once. In other words we estimate the losses due to a single interruption event. The amount of losses C due to communication gateway interruption D is then computed as follows.

$$P(C = Hn) = \sum_{i=1}^{2} p_i \exp(-\lambda_a D_i)(\lambda_a D_i)^n / n!, \quad n = 0, 1, \cdots \tag{14}$$

$$P(C > \Theta) = \sum_{n > \Theta/H}^{\infty} P(C = Hn) \tag{15}$$

$$E(C) = H\lambda_a E(D) = H\lambda_a \sum_{i=1}^{2} p_i D_i \tag{16}$$

Assuming $\lambda_a = 10000/day$, H=\$ 10, $p_1 = p_2 = 1/2, D_1 = 1/8, D_2 = 1/16, \Theta =$ \$12000, the numbers are: $P(C > \Theta) = 0.459, E(C) =$ \$9375. This completes the task of computing the probability distribution of the losses due to the resource interruption, but more can be done. For example, we would like to understand what types of events are most likely to be responsible for large losses? It is not hard to see that the first type has a more dramatic effect, simply because its duration is twice as large and probability of occurring is the same. But the difference is quite significant as the following computations show.

To identify the event(s) that give rise to $P(C > \Theta) > 0.90$. We refer to the Figure 4. Given $\Theta =$ \$12000, P(potential losses due to **event I** $> \Theta) = 0.92$, P(potential losses due to **event II** $> \Theta) \approx 0$. So in this simple example only **event I** gives rise to potential losses that exceed the threshold \$12000 with over 90% probability, while the second event is relatively insignificant at this threshold.

5.2 Distribution of losses over a time period

We now consider a more complicated model where we take into account the frequency of communication failures during a specific time interval of interest (month or a year), consider a more complicated outage distribution and simultaneously address the issue of countermeasures which can be adopted which will reduce the typical frequency and duration of down-times (service interruptions). Assume that during a year service interruption can occur k times with probability $p_k, \sum_k p_k = 1$. Each occurrence causes a "down" period of D which has a Gamma distribution with parameters α, λ_d, ie. $D \sim$ Gamma(α, λ_d). The arrival of trade orders is again modeled as a homogenous Poisson process with rate λ_a, each of which brings in commission fee of H dollars.

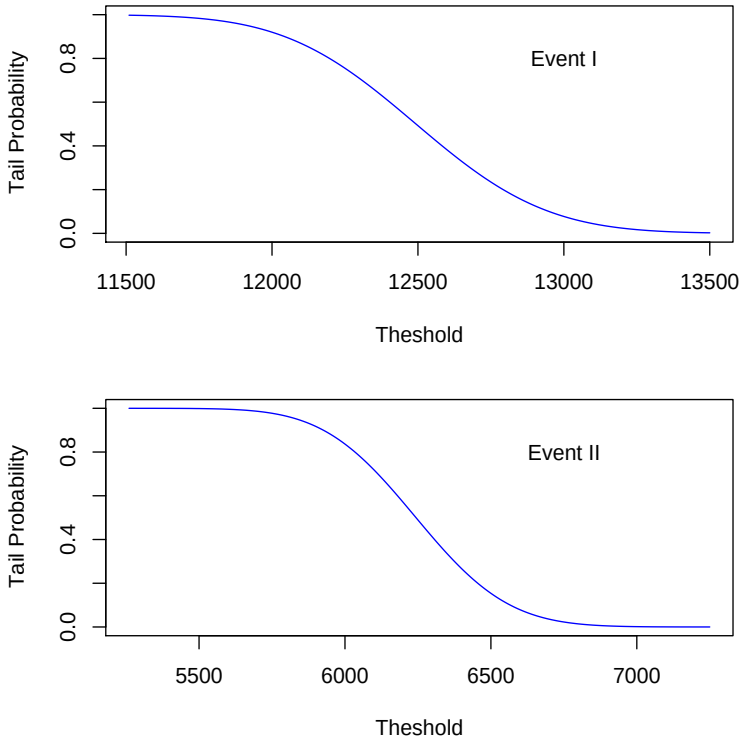


Figure 4: Tail probability $P(C > \Theta)$ of the potential losses caused by **event I** and **event II**. Horizontal axis is the threshold Θ in the unit of dollars.

$$
\begin{aligned}
P(L_I = Hn) &= \sum_k p_k P(L(D_1) + \cdots + L(D_k) = Hn) \\
&= \sum_k p_k E[P(L(D_1) + \cdots + L(D_k) = Hn | D_1, \cdots, D_k)] \\
&= \sum_k p_k E[\exp(-\lambda_a \sum_{i=1}^{k} D_i)(\lambda_a \sum_{i=1}^{k} D_i)^n / n!] \qquad (17)
\end{aligned}
$$

Eq. (17) is obtained by noticing that $L(D_1), \cdots, L(D_k)$ are independent Poisson random variables with mean $\lambda_a D_1, \cdots, \lambda_a D_k$ conditioning on $D_1, \cdots, D_k$. Since $\sum_{i=1}^{k} D_i$ has a gamma distribution with parameters αk and λ_d whose density function is $f(u) = e^{-\lambda_d u}(\lambda_d u)^{\alpha k - 1}\lambda_d/(\alpha k - 1)!$, Eq. (17) can be further simplified:

$$
\begin{aligned}
R.H.S(17) &= \sum_k p_k \int \frac{e^{-\lambda_a u}(\lambda_a u)^n}{n!} \frac{e^{-\lambda_d u}(\lambda_d u)^{\alpha k - 1}\lambda_d}{(\alpha k - 1)!} \, \mathrm{d}u \\
&= \sum_k p_k \frac{(n + \alpha k - 1)!}{n!(\alpha k - 1)!} \frac{\lambda_a^n \lambda_d^{\alpha k}}{(\lambda_a + \lambda_d)^{n + \alpha k}} \qquad (18)
\end{aligned}
$$

Let $\lambda_a = 10000$ per day, $H = \$10$. There are three types of events: **event I, II, III**. Losses due to each type of event has a distribution given in Eq. (18). Now, we assume that countermeasures (for

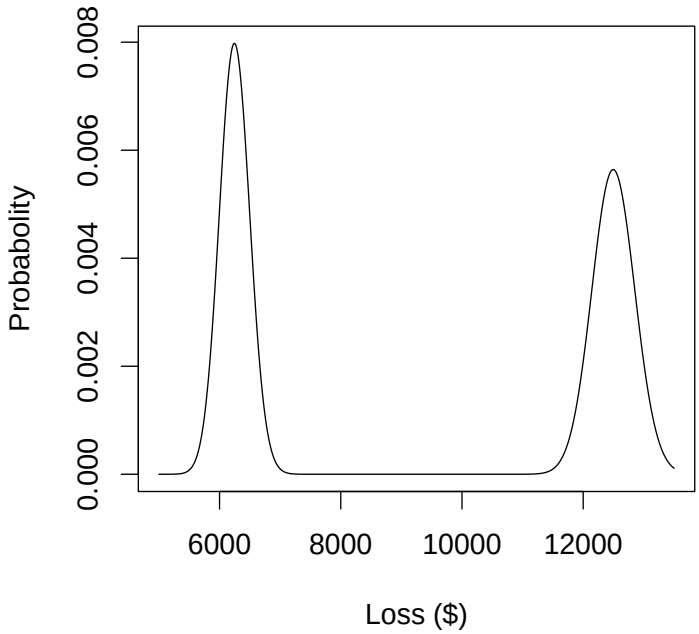


Figure 5: Probability distribution of losses C in one cycle, in the unit of dollars.

example server upgrade or introducing multiple servers in place of one) can reduce the probability of multiple events as well as the duration of "down" period per occurrence. The corresponding parameters are summarized in Table 1.

Table 1: Parameters for three types of events. $P = (p_1, p_2, \cdots)$.

	no countermeasures			with countermeasures		
	P	α	λ_d	P	α	λ_d
event I	(0.9, 0.1)	5	5	(0.95, 0.05)	3	4
event II	(0.5, 0.3, 0.2)	3	5	(0.7, 0.2, 0.1)	2	5
event III	(0.4, 0.3, 0.2, 0.1)	3	6	(0.5, 0.4, 0.1)	2	6

The distribution of losses (L) is a convolution of losses (L_I, L_{II}, L_{III}) due to the three types of events, ie.

$$P(L = Hn) = \sum_{n_1+n_2+n_3=n} P(L_I = Hn_1)P(L_{II} = Hn_2)P(L_{III} = Hn_3) \tag{19}$$

The distribution of losses L is presented by Fig. 6. Clearly with counter measures, the distribution of losses shifts to smaller amount. In particular, the 90% percentile changed from \$456030 to \$286810.

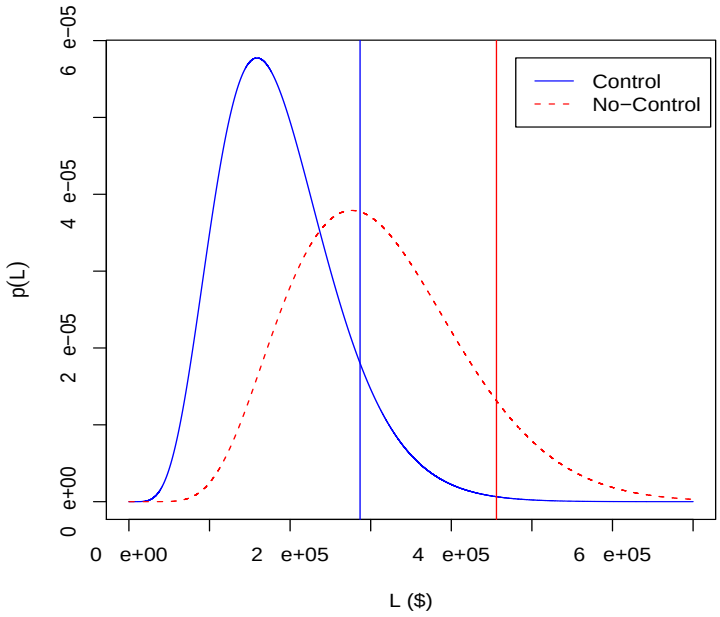


Figure 6: Probability distribution of losses in two scenarios: with countermeasures (blue) vs. without countermeasures (red). The vertical lines correspond to the 90% percentiles.

5.3 Allocation of Countermeasure Resources

Our proposed model suggests a direct way for creating countermeasures to mitigate the operational risk, provided that the impact of countermeasures is known (reduced time to recover systems, lower frequency of adversary events, etc.) However, often the number of possibilities for allocating the budget for countermeasures can be quite substantial and budget limitations prevent mitigating all the possible adversary events. Thus it is desirable to have an automated procedure for computing the optimal allocation of countermeasure resources subject to the budget constraints. This problem can be formalized as the optimization problem subject to the linear (budget) constraints. Our decision variables are $\eta_1, \cdots, \eta_d$: the proportions of total budget that is allocated to control adverse events $1, \cdots, d$, $\eta_1 + \cdots + \eta_d = 1$. Let $C(\eta_1, \cdots, \eta_d)$ be the cost after countermeasure with allocation proportion $\eta_1, \cdots, \eta_d$. The problem reduces to:

$$\min_{\eta_i \geq 0, \eta_1 + \cdots + \eta_d = 1} g(C(\eta_1, \cdots, \eta_d)) \tag{20}$$

where $g(\cdot)$ is an objective function determined by our criteria. For instance, we may take $g(\dot{)}$ to be the expectation: $E[C(\eta_1, \cdots, \eta_d)]$. To obtain the distribution of $C(\eta_1, \cdots, \eta_d)$, we need information on the effect of countermeasures on reducing the impact caused by respective adverse events.

We now demonstrate how optimal counter-measure can be determined by our model. We consider again the broker/dealer process, assuming λ_A the arrival rate of orders and H the value of each order. Assuming two types of independent adverse events $\mathbf{I}, \mathbf{II}$ whose probability of number of occurrence are $P_0 = .95, P_1 = .05, P_2 = 0$ for the type $\mathbf{I}$ events and $P_0 = .92, P_1 = .06, P_2 = .02$ for the type

II events. Each occurrence has a duration follows Gamma distribution with various means. Assume counter-measures can be taken for **I**, **II** to reduce the mean duration of respective occurrence while P_0, P_1, P_2 remains the same. The more money spend on the counter-measures, the less is the mean duration of each occurrence afterwards. The Tables 2, 3 has a summary of these counter-measures.

Table 2: Various counter-measures with expenses and mean durations after the corresponding counter-measure on adverse event **I** enforced.

Expense	0	0.25H	0.45H	0.65H	0.8H	H
Mean duration(day)	$\frac{2}{3}$	$\frac{2}{4}$	$\frac{2}{5}$	$\frac{2}{6}$	$\frac{2}{7}$	$\frac{2}{9}$
SD of duration(day)	$\frac{\sqrt{2}}{3}$	$\frac{\sqrt{2}}{4}$	$\frac{\sqrt{2}}{5}$	$\frac{\sqrt{2}}{6}$	$\frac{\sqrt{2}}{7}$	$\frac{\sqrt{2}}{9}$

Table 3: Various counter-measures with expenses and mean durations after the corresponding counter-measure on adverse event **II** enforced.

Expense	0	0.22H	0.5H	0.83H
Mean duration(day)	$\frac{2}{4}$	$\frac{2}{5}$	$\frac{2}{7}$	$\frac{2}{10}$
SD of duration(day)	$\frac{\sqrt{2}}{4}$	$\frac{\sqrt{2}}{5}$	$\frac{\sqrt{2}}{7}$	$\frac{\sqrt{2}}{10}$

Here we adopt the objective function to be the 99% percentile of value-at-risk (VaR). Assume $\lambda_A = 10^6$ per day and $H = \$10^6$. The Fig. 7 is the plot of VaR vs. counter-measures expenses. Notice the far upper-left point corresponding to the case of "as-is". We make two observations from this plot. First, subject to a given budget for counter-measures expenses, we can clearly identify the optimal allocation of the resources to control adverse events **I**, **II**. For instance, if the budget is $\$800,000$, then the optimal combination will be the one with minimum VaR whose total expenses is less than the budget, as illustrated by the figure. The optimal choice in this case is: allocating $0.25H = \$250,000$ to control risk due to event **I** and $0.5H = \$500,000$ to control risk due to event **II**. Second, we can draw an efficiency curve along the boundary of the points in the scatter plot, the points on the steepest descent boundary correspond to the most efficient combinations of counter-measures in the sense of the ratio of reduced VaR to total expenses.

5.4 Multiple adversary events, resources and tasks. Network example

The overall network structure of the broker/dealer model discussed in the previous subsections is fairly primitive: one resource and one task. We now enrich the model somewhat and again refer to Figure 2. We assume that successful completion of the overall transaction described on this figure consists of the tasks represented by the rectangles starting with the task "Calculate Net Proceeds from the Trade" and ending with the task "Complete Settlement". In the terminology of Subsections 4.1 and 4.2 we have a workflow process F_1 associated with this set of tasks. Typically all of these tasks require some resources. Assume here that only three resources are engaged. The first resource r_1 is referred to simply as "server" and is required by the first task "Calculate Net Proceeds ...". The second resource r_2 is communication gateway associated with the task "Send Information to ...". The last resource r_3 is a database required by the last task "Confirm Trade Details". In addition, suppose there is a different workflow F_2 corresponding to some "Account Status Change" process which only consists of the existing task "Send Information to Custodian and Clearinghouse", which requires communication

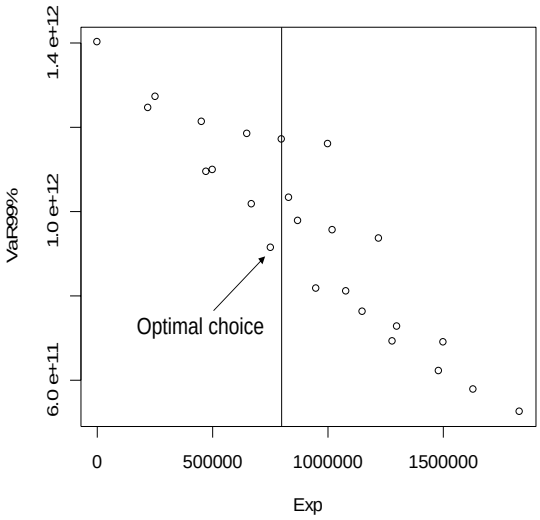


Figure 7: Value-at-Risk (VaR) after different combinations of counter-measures are taken.

gateway resource r_2. Successful completion of the flow F_1 of tasks brings $C_1 = 10$ dollars of revenue. The flow of tasks F_2 does not bring revenue, but non-completion of this flow induces cost $C_2 = 2$. The arrival rates for the flows F_1, F_2 are assumed to have Poisson distribution with parameter $\lambda_1 = 10000$ per day and $\lambda_2 = 20000$ per day, respectively. The corresponding column vector of arrival rates is $\lambda = (\lambda_1, \lambda_2)^t$. To complete the model we need to identify the types of adversary events and their impact on resources. We assume two types of adversary events: power outage E_1 which affects all of the three resources, and security breach E_2 which affects only communication gateway resource r_2. Event E_1 (E_2) has a deterministic duration $D_1 = 2$ ($D_2 = 5$) time units and frequency given by a probability vector $p_1 = (.5, .3, .2)$ ($p_2 = (.7, .3)$). Namely, power outages will not occur during the period of interest with probability 0.5, and will occur once and twice with probabilities 0.3 and 0.2 respectively. Similar explanation for p_2. Thus the expected power outage frequency is $\mathbb{E}[L_1] = .3 + 2 \times .2 = .7$ and expected security breach frequency is $\mathbb{E}[L_2] = .3$.

This completes the description of 3-layer elements of the model and now we proceed to computations. First we identify the matrices connecting adversary events to resources, tasks and flows. The matrix $\mathcal{E}$ is 2×3 since we have two adversary events and 3 resources. Matrix $\mathcal{R}$ connecting resources to tasks is 3×6 since we have six tasks, and the matrix $\mathcal{F}$ connecting tasks to flows is 6×2 since we have two flows. The diagonal matrices of durations D, frequencies L and costs C are all 2×2. The corresponding

entries of these matrices are

$$\mathcal{E} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix},\ \mathcal{R} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},\ D = \begin{pmatrix} 2 & 0 \\ 0 & 5 \end{pmatrix},$$

$$\mathcal{F} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix},\ L = \begin{pmatrix} .7 & 0 \\ 0 & .3 \end{pmatrix},\ \mathcal{C} = \begin{pmatrix} 10 & 0 \\ 0 & 2 \end{pmatrix},\ \lambda = \begin{pmatrix} 10000 \\ 20000 \end{pmatrix}$$

We now apply formula (13) to obtain the total expected cost.

$$C_{\text{total}} = e^t LD\,\text{sgn}(\mathcal{E}\mathcal{R}\mathcal{F})C\lambda = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} .7 & 0 \\ 0 & .3 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 5 \end{pmatrix} \times$$

$$\times\,\text{sgn}\Big[\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \Big] \begin{pmatrix} 10 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 10000 \\ 20000 \end{pmatrix}$$

$$= 460,000.$$

This completes the computation of the expected overall cost for the broker/dealer example. The probability distribution can be also obtained using formula (9) in a straightforward way, but the computations are very involved and we omit them.

6 Modeling Outsourcing Risk

The earlier sections in this paper were primarily devoted to modelling operational risk when the corresponding cost structure had a simple linear form. Of course in reality this is not always the case and the cost associated with operational failures can have a more complicated form. In this section we consider an example of computing operational risk with a more complicated cost structure and illustrate them with examples. An important motivation in this regard is assessment of outsourcing risk, which is an important component of the overall operational risk. As financial institutions are outsourcing several aspects of their operations, in particular information technology services, they are required to assess and manage operational risks from outsourcing (15). Moreover, financial institutions are expected by regulators to have appropriate Business Continuity plans for its critical outsourced activities (7). Outsourcing arrangements are typically managed using Service-Level Agreements (SLA). Consider the case where a financial institution outsources its IT infrastructure to a IT service provider. The SLA contain charges related to operational failures of the hardware configuration. Some typical costs in SLA's include penalty charges $C_1 = c_1$ if the total outage D_1 during a fixed time period (say a month or a year) exceeds a specified threshold d_1^*. For example, an SLA may charge $c_1 = 3$ points (with some equivalent dollar amount) if the total outage duration exceeds $d_1^* = 100$ minutes during a 3-month period. If the total outage duration is less than d_1^* then the incurred cost (charges) is $C_1 = 0$. Another typical cost in SLA's is the maximum duration of any particular outage. Specifically, if any of the hardware units fails for more than d_2^* time units during the same time period, then a charge of $C_2 = c_2$ is applied, and if all of the outages had duration shorter than d_2^* then no charges are incurred.

6.1 Max/sum type piece-wise linear cost of SLA

In this subsection we give a complete treatment of the model with max/sum type cost function just described. In order to complete the details of the model, we need to identify the elements of the 3 layers of our operational risk architecture: risk events, resources and tasks. We consider a fairly simple such model so that we may concentrate on the non-linearity aspect of the operational losses. Thus we assume that there exists two types of resources: N_{new} "new" and N_{old} "old" resources. There are two types of risk events: the new resources are impacted by the event $E_{\text{new,i}}, i = 1, \ldots, N_{\text{new}}$ "new hardware unit fails", and the old resources are impacted by the event $E_{\text{old,j}} j = 1, \ldots, N_{\text{old}}$ "old hardware unit fails". The frequency L_{new} and the duration D_{new} of events E_1 is given by their corresponding probability distributions $F_{L_{\text{new}}}$ and $F_{D_{\text{new}}}$.

Since the cost function is not linear, the development in the previous section cannot be applied. Yet we can still compute the probability distribution of the losses directly from our threshold assumptions on the cost structure. It turns out that a particularly convenient approach is using Laplace transforms. We first compute the Laplace transform $g_D(s)$ of the overall duration of the failures D_{total} of all the units. We decompose it as the sum of $D_{\text{total,new}}$ and $D_{\text{total,old}}$ corresponding to the total duration of outages corresponding to new and old type hardware units respectively. The Laplace transform of each is found as follows.

$$g_{D_{\text{total,new}}}(s) = (g_{L_{\text{new}}}(g_{D_{\text{new}}}(s)))^{N_{\text{new}}}$$
$$g_{D_{\text{total,old}}}(s) = (g_{L_{\text{old}}}^{N_{\text{old}}}(g_{D_{\text{old}}}(s)))^{N_{\text{old}}}.$$

And the overall duration of the failure is found as $g_{D_{\text{total}}}(s) = g_{D_{\text{total,new}}}(s) g_{D_{\text{total,old}}}(s)$. The probability distribution of the cost corresponding to the total failure, which only takes values 0 and c_1 is then found as follows. The cost c_1 is incurred with probability $\mathbb{P}(C_1 = c_1) = \mathbb{P}(D_{\text{total}} > d_1^*)$ which is found as a sum of the coefficients of the polynomial $g_{D_{\text{total}}}(s)$ which correspond to powers exceeding d_1^*; and no cost is incurred with probability

$$\mathbb{P}(C_1 = 0) = 1 - \mathbb{P}(D_{\text{total}} > d_1^*).$$

We now demonstrate this using a numerical example. In our example, we set $d_1^* = 100$, $c_1 = 3$, $d_2^* = 50$ and $c_2 = 2$. We have $N_{\text{new}} = 3$ new resources and $N_{\text{old}} = 2$ old resources. We assume that $L_{\text{new}} = 1$ and $= 2$ with probabilities .05 and .03 respectively, and $L_{rmnew} = 0$ with the remaining probability .92. The duration is assumed to be $D_{\text{new}} = 10$ minutes and 55 minutes with probabilities .9 and .1 respectively. Respectively, assume that $L_{\text{old}} = 1$ and $= 2$ with probabilities .15 and .10 respectively, and $L_{\text{old}} = 0$ with the remaining probability .75. The duration is assumed to be $D_{\text{new}} = 15$ minutes and 60 minutes with probabilities .7 and .3 respectively. We find

$$\begin{aligned}
g_{D_{\text{new}}}(s) &= .9s^{10} + .1s^{55} \\
g_{D_{\text{old}}}(s) &= .7s^{15} + .3s^{60} \\
g_{L_{\text{new}}}(s) &= .92 + .05s + .03s^2 \\
g_{L_{\text{old}}}(s) &= .75 + .15s + .1s^2 \\
g_{D_{\text{total}}}(s) &= \left(.92 + .05(.9s^{10} + .1s^{55}) + .03(.9s^{10} + .1s^{55})^2\right)^3 \times \\
&\quad \times \left(.75 + .15(.7s^{15} + .3s^{60}) + .1(.7s^{15} + .3s^{60})^2\right)^2.
\end{aligned}$$

We have performed MATLAB based computations of the these polynomials (they are based on a straightforward application of the convolution "conv.m" command). Using these polynomials we found that the

probability of incurring a charge $\mathbb{P}(C_1 = c_1)$ is 3.95%. We also found that the probability of incurring a charge due to failure one of the old hardware units is 2.96% while the probability of incurring a charge due to a failure of one the new units is only 0.12%.

We now compute the probability distribution of the second cost type C_2. We denote by $D_{\max,\text{new,j}}, 1 \leq j \leq N_{\text{new}}(D_{\max,\text{old,j}}, 1 \leq j \leq N_{\text{old}})$, the maximum duration of an outage of any of the new (old) hardware units. Then

$$\begin{aligned}\mathbb{P}(C_2 = 0) &= \Big(\prod_{1\leq j\leq N_{\text{new}}} \mathbb{P}(D_{\max,\text{new,j}} \leq d_2^*)\Big)\Big(\prod_{1\leq j\leq N_{\text{old}}} \mathbb{P}(D_{\max,\text{old,j}} \leq d_2^*)\Big) \\ &= \Big(\sum_{n\geq 0} \mathbb{P}(D_{\text{new}} < d_2^*)^n \mathbb{P}(L_{\text{new}} = n)\Big)^{N_{\text{new}}} \Big(\sum_{n\geq 0} \mathbb{P}(D_{\text{old}} < d_2^*)^n \mathbb{P}(L_{\text{old}} = n)\Big)^{N_{\text{old}}} \\ &= g_{L_{\text{new}}}^{N_{\text{new}}}(\mathbb{P}(D_{\text{new}} < d_2^*)) g_{L_{\text{old}}}^{N_{\text{old}}}(\mathbb{P}(D_{\text{old}} < d_2^*)).\end{aligned}$$

This formula then can be used directly for computing the probability $\mathbb{P}(C_2 = 0)$ of no charges incurred. For our numerical example we find that $\mathbb{P}(D_{\text{new}} < d_2^*) = \mathbb{P}(D_{\text{new}} < 50) = .9, \mathbb{P}(D_{\text{old}} < d_2^*) = \mathbb{P}(D_{\text{old}} < 50) = .7$ We have already computed the moment generating functions of L_{new} and L_{old}. Using this

$$\mathbb{P}(C_2 = 0) = (.92 + .05 \times .9 + .03 \times .9^2)^3 (.75 + .15 \times .7 + .1 \times .7^2)^2 = .79,$$

and

$$\mathbb{P}(C_2 = c_2) = \mathbb{P}(C_2 = 2) = 1 - \mathbb{P}(C_2 = 0) = .21.$$

6.2 A more complicated SLA cost structure

In the original treatment of CLS problem of hardware failure with piece-wise constant cost function, we take a summation over individual hardware component's breakdown time. In reality these hardware components maybe supporting the same business process and the same cost (the cost of downtime) is incurred when one or several components fail. Thus our previous approach suffers potentially from overcounting problem as downtime of different components maybe overlapping. Under certain conditions the overlapping of breakdown period might not be negligible, therefore we take a further look at this problem and compute more accurately the cost of downtime under some additional modeling assumptions. Since the overlapping is inherently a dynamical problem, we need to propose a non-static model and we propose the use of finite state Markov processes, as the simplest model which encompasses dynamics and uncertainty at the same time.

Suppose there are H hardware components supporting one business process. Suppose these components operate independently, each one alternates within states "normal" and "failure". Assuming the length of "normal" and "failure", follows respective exponential distribution with rate $\lambda_i, \mu_i, i = 1, \ldots, H$, further assume the length of "normal" and "failure" period independent. In the derivations below we obtain the expression for probability distribution of "normal" $S_0(T)$ time for the entire system during a time period T. This can be used for computing the probability distribution of the cost associated with SLA. For example if the cost is c dollars when the total "failure" time exceeds d time units during time period $[0, T]$ and 0 dollars otherwise, then, just as in previous subsection, the probability distribution of the cost is c with probability $\mathbb{P}(S_0(T) > d)$ and 0 with the remaining probability $1 - \mathbb{P}(S_0(T) > d)$.

In order to compute the distribution function $\mathbb{P}(S_0(T) \leq s), s \geq 0$, introduce $X_t = (X_1(t), \cdots, X_H(t))$ – the indicator for the H components' status at time t where $X_i(t) = 0$ if the ith component is in "normal" condition and otherwise $X_i(t) = 1$. Clearly X_t is a Markov process with finite state space. In fact,

it can be described as a birth-death (**BD**)process. The system is "failure" at time t if X_t has at least one component equal to 1.

We further assume $\lambda_i = \lambda, \mu_i = \mu, i = 1, \cdots, H$, ie. H identical components. This assumption simplifies notation of X_t in our discussion but still illustrates the main idea. In this case the number of components in "failure" follows a Markov process, denoted by X_t (a little abuse of notation). The state space is $\{0, 1, \cdots H\}$. The transition rate of the **BD** process is easily obtained: $q_{i,i+1} = (H-i)\lambda, q_{i,i-1} = i\mu, i = 1, \cdots, H$. We assume the initial state being 0, ie. there is no failed components at the beginning. To obtain the distribution of occupation time of state 0 in a finite time interval $[0, T]$, we use the technique of "uniformization". This techniques essentially construct an independent Markov process with the same state space, but the mean time spent in each state has an exponential distribution with same rate ν, which we choose to be $\nu = H(\lambda + \mu)$. The new process in addition allows a fictitious transition into same state i, the transition probability is:

$$P_{i,i} = 1 - \frac{(H-i)\lambda + i\mu}{\nu},\ P_{i,i+1} = (H-i)\lambda/\nu,\ P_{i,i-1} = i\mu/\nu \tag{21}$$

The original Markov process's transition probability $P^*_{i,j}(t)$ is $\sum_{n=0}^{\infty} P^n_{i,j} e^{-\nu t} \frac{(\nu t)^n}{n!}$. Let $S_0(T)$ be the total occupation time of state 0 in $[0, T]$ with initial state $X(0) = 0$. That is $S_0(T)$ is the total "normal" time during $[0, T]$ - the quantity of interest. We use the uniformized version of $\{X_t, t \geq 0\}$ and conditioning on $N(T)$, the number of transitions by time T.

$$P(S_0(T) \leq s) = \sum_{n=1}^{\infty} e^{-\nu T} \frac{(\nu T)^n}{n!} P(S_0(T) \leq s | N(T) = n) \tag{22}$$

Notice $[0, T]$ is partitioned into $(0, X_{(1)}), (X_{(1)}, X_{(2)}), \cdots, (X_{(n)}, T)$ where $X_{(i)}$ are the ordered arrival time of n transitions by T. The lengths of these intervals $Y_{(i)} = X_{(i)} - X_{(i-1)}, i = 1, \cdots, n+1$ with $X_{(0)} = 0, X_{(n+1)} = T$ are exchangeable. Therefore, assume k transitions are into state 0, ie. k entries of $Y_1, \cdots, Y_{n+1}$ corresponding to occupation of state 0, then $S_0(T)$ equals the sum of these entries, which has the same distribution as $Y_1 + Y_2 \cdots + Y_k = X_{(k)}$ due to "exchangeability". It's known that $P(X_{(k)} \leq s | N(T) = n) = \sum_{i=k}^{n} \binom{n}{i} (s/T)^i (1 - s/T)^{n-i}$. This leads to

$$P(S_0(T) \leq s | N(T) = n) = \sum_{k=1}^{n} P(k-1, n-1) \sum_{i=k}^{n} \binom{n}{i} (s/T)^i (1 - s/T)^{n-i} \tag{23}$$

where $P(k, n)$ is the probability of having k transitions into state 0 out of the n transitions. $P(k, n)$ can be derived from a system of recursion relationships as follows. Let $P_i(k, n)$ be the probability of having k transitions into state 0 given n transitions and the last visited state is i. Conditioning on the state j visited by the $(n-1)$th transition, we have:

$$P(k, n) = \sum_{i=0}^{H} P_i(k, n) \quad 0 \leq k \leq n \tag{24}$$

$$P_i(k, n) = P_i(k-1, n-1) P_{i,i} + \sum_{j \in \{i-1, i+1\}} P_j(k, n-1) P_{j,i} \quad i, j \in \{0, \cdots, H\} \tag{25}$$

The initial values are: $P_0(1,1) = P_{0,0}, P_1(0,1) = P_{0,1}, P_0(1,2) = P_{0,1}P_{1,0}, P_1(1,2) = P_{0,0}P_{0,1}, P_0(2,2) = P^2_{0,0}$. Combining Eq. (25), (24), (23), (22), we obtain the $P(S_0(T) \leq s)$. In practice, the summation of

Eq. (22) is truncated at sufficiently large n. This completes the expression for the distribution of the total "normal" time during $[0,T]$ as well as the probability distribution of the SLA cost.

In case the distribution of normal and failure times does not have an exponential form, coming up with the analogous expression is not tractable. But by way of approximation we can compute the average (expected) normal time during a time period $[0,T]$ when T is large using the theory of renewal processes. Denote by X, Y the period of "normal" and "failure" of one component, and $\phi_X(s), \phi_Y(s)$ their moment generating functions. Let $P(T)$ be the probability that the component is in "normal" at time T, and its moment generating function is $\phi(s)$. By results of alternating renewal theory, we have

$$\phi(s) = \frac{1-\phi_X(s)}{s(1-\phi_X(s)\phi_Y(s))} \tag{26}$$

$P(T)$ is obtained by inverting moment generating function of $\phi(s)$. For the system of H independent components, the probability of being in "normal" at time T is $\prod_{h=1}^{H} P_h(T)$ where $P_h(T)$ is the probability obtained from Eq. (26) for the hth component. By ergodic property of stationary stochastic processes we have that the average normal time during $[0,T]$ is asymptotically $P_h(T)$, provided that T is sufficiently large.

The expression for the total normal time just obtained, while simple, is of limited use since it is only applicable for large time periods $[0,T]$ and the ergodic theory is, unfortunately, of limited use for controlling the accuracy of approximations.

7 Concluding Remarks

We have described a framework for operational risk modeling in this paper, based on a description of the business process in a financial institution, its human, physical and logical infrastructure and the risks contained therein. This methodology is advantageous for operational risk assessment and management, compared to existing approaches, since changes to the business process operational models can automatically get translated to changes in the operational risk models.

We believe researchers have only started to look at the operational risk management area and there exists a wide scope for further research. A taxonomy needs to be developed that can associate operational risk event types with potential root causes and relate them to controls and countermeasures that can be deployed to better manage the operational risk. Current libraries of operational risk events only contain information on historical losses. These need to be enriched with information relating to root-causes. Furthermore, benchmarks need to be gathered for risks related to different root-causes. Firms model operational losses based on information contained in loss databases to determine the loss distribution, from which the capital allocation is identified by computing the appropriate quantile. The modeling fallacies need to be researched for high confidence level quantiles accompanied with sparse data, which is the typical case for operational risk events. The utilization of quantitative analysis methods as a decision support mechanism for real-time or near real-time operations is a rich area that merits research attention. This could pave the way for discovering operational risks and managing them in a pre-emptive mode thus avoiding operational losses, rather than being limited to observing loss events and laying aside capital to manage operational risks.

Acknowledgements. We gratefully acknowledge many enlightening discussions with Jonathan Rosenoer, Chonawee Supatgiat and Chris Kenyon.

References

[1] Basel Committee on Banking Supervision: The New Basel Capital Accord, Bank for International Settlements, April 2003.

[2] Basel Committee on Banking Supervision: Working Paper on the Treatment of Operational Risk, Bank for International Settlements, Sept. 2001.

[3] Advances in Operational Risk: Firm-wide issues for Financial Institutions, Risk Books, Second Edition, 2003.

[4] M.G.Cruz: "Modeling, Measuring and Hedging Operational risk", John Wiley & Sons, Sec 2003.

[5] M.Cruz, R.Coleman and G.Salkin: "Modeling and Measuring Operational risk", *The Journal of Risk*, Vol. 1, Iss. 1, pg 63, 1998.

[6] S.Ebnother, P.Vanini, A.McNeil and P.Antolinez: "Operational Risk: A Practitioner's View", *The Journal of Risk*, Vol. 5, Iss. 3, pg 1, 2003.

[7] Federal Financial Institutions Examination Council, "FFIEC IT Examination Handbook - Business Continuity Planning Booklet", March 2003.

[8] J.King: "Operational Risk- Measurement and Modeling", Wiley Publishers, 2001.

[9] SAS - http://www.sas.com/industry/fsi/oprisk/index.html

[10] Agena - http://www.agena.co.uk/corporateagenarisk/operationalriskinfinancialservices.html

[11] H. Chen and D.D. Yao, *Fundamentals of queueing networks: Performance, asymptotics and optimization*, Springer Verlag, New York, 2001.

[12] D.M.Chickering, D.Geiger, D.Heckerman: Learning Bayesian Networks is NP-hard, Technical Report MSR-TR-94-17, Microsoft Research, 1994.

[13] P.Dagum, M.Luby :Approximating Probabilistic Inference in Bayesian Belief Networks is NP-hard, *Artificial Intelligence*, 60 (1), pg 141-153, 1993.

[14] WBI Modeler - http://www-306.ibm.com/software/integration/wbimodeler/advanced/

[15] Basel Committee on Banking Supervision: "Consultative Document on Outsourcing in Financial Services", August 2004.