e-Risk Management with Insurance : A framework using Copula aided Bayesian Belief Networks

Arunabha Mukhopadhyay
Indian Institute of Management Calcutta
arunabha@ iimcal.ac.in

Samir Chatterjee
Claremont Graduate University
Samir.Chatterjee@cgu.edu

Debashis Saha
Indian Institute of Management Calcutta
ds@iimcal.ac.in

Ambuj Mahanti
Indian Institute of Management Calcutta
am@iimcal.ac.in

Samir K Sadhukhan
Indian Institute of Management Calcutta
samir@iimcal.ac.in

Abstract:

e-business organizations are heavily dependent on distributed 24X7 robust information computing systems, for their daily operations. To secure distributed online transactions, they spend millions of dollars on firewalls, anti-virus, intrusion detection systems, digital signature and encryption. Nonetheless, a new virus or a clever hacker can easily compromise these deterrents, resulting in losses to the tune of millions of dollars annually. To cope up with the problem, in this work we propose to further enhance their security management by investing in e-risk insurance products as a viable alternative to reduce these individual financial losses. We develop a framework, based on copula aided Bayesian Belief Network (BBN) model, to quantify the risk associated with online business transactions, arising out of a security breach, and thereby help in designing e-insurance products. We have simulated marginal data for each BBN nodes. The Copula model helps in arriving at the joint probability distributions from these marginal data. From the joint distribution data, we arrive at the conditional distribution tables for each node. This is input to the Bayesian Belief Network model. The output is frequency of occurrence of an e-risk event. Frequency of loss multiplied with the expected loss amount, provides the risk premium to be charged by insurance companies.

Keywords: e-commerce, security breach, e-risk, Bayesian Belief Network (BBN), copula, premium

1. Introduction

E-risk is defined as the possibility of an electronic event, whose occurrence causes loss to e-businesses. These include compromise of network security components (such as firewall, proxy servers, anti virus), the compromise of the organization web server, and incorrect or indecent material posted on the web site (commonly called graffiti), service providers (i.e., Application Service Provider (ASP) or Internet Service Provider (ISP)) failing, identity theft (i.e., confidential customer information is hacked from an organizational database; example, pin numbers of credit cards from a bank), attacks by disgruntled employees, cyber-extortion, Denial of Service (DoS) by making malicious calls to the router, attack by wireless devices (such as PDAs ,mobile phones etc). CSI/FBI 2004 report [1] states that the most vital e-risk in USA is virus attack (loss of $55 Million). It is followed by DoS attack (loss of $26 Million), and theft of proprietary information (loss of $11 Million). These losses do not include other intangible losses due to customer denied access, churn, loss of loyalty, lost business opportunity etc.

Organizations spend millions annually for deployment of sophisticated technical defenses (such as encryption, access control and firewalls) [2] and intrusion detection systems to guard against malicious attacks. CSI/FBI report [1] states that almost 99% of organizations in USA have Antivirus software, 98% have firewalls, 71% have proxy servers, 68% have intrusion detection systems. Yet security breaches are very common.

Anderson [3] opines that the chance of a clever hacker breaking into the system is much higher than the chance that the CTO would detect it. Say, there are n weak points in an organizational network. The probability that the hacker would find one is $1/n$. In contrast, the CTO has to be aware of all the n weak points to protect the system from malicious attacks. Schneier [4] notes that a new virus can easily comprise the perimeter security devices, as there is no signature available to the anti-virus engine to track it down. The CSI/FBI report [1] corroborates this fact, as loss due to virus attack in USA alone was $55 Million in 2004.

To supplement the existing security measures and to reduce the monetary loss, an effective alternative mechanism is insuring [5, 6, 7] against these risks. This would help reduce the financial burden on the organizations, as the insurance company would indemnify the loss. In effect, the organizations risk is being passed on to another party at the cost of a premium. This reduces the companies concern about "self insuring" (i.e., keeping aside huge amount for contingency purposes). This, in turn, is a good corporate

strategy, as huge amounts are not locked away for contingency provisions and security breaches.

The contribution of this paper lies in quantifying the expected loss incurred by an e-business organization due to a security breach and suggesting a premium amount that an insurance company can charge to hedge this loss. This is in contrast to Gordon's [2] work, which is focused on finding out an optimal investment on security apparatus to reduce security breaches. Our paper supplements Gordon's [5] work by assessing and quantifying e-risk, which he considers, as the most important component of cyber-risk management by e-business organizations. We propose that insurance companies too can use this model to evaluate the probable vulnerabilities of an e-business organization and use it for e-insurance product design.

We have developed a representative causal diagram (using Bayesian Belief Networks (BBN)) elucidating the probable reasons for a security breach and populated it with expert opinions. For our study we customized a BBN tool called FULLBNT [8] developed in Matlab and used it to simulate various security breach scenarios and noted the probabilities (i.e., loss frequency) at each of the causal node(s). To make the model robust and generalized, we have simulated marginal probability data at each of the causal nodes and used Copula model to arrive at the joint probability distribution for the entire network. From this joint probability data we have arrived at the conditional probability tables needed as basic inputs to the BBN network. We assume that the marginal data can be obtained from the log files of organizational computer network. For simplicity we have assumed that each of the causal nodes have a similar binomial loss amount distribution of the form Binomial (1000, 0.2). Expected Loss (or risk premium) is defined as the product of loss frequency and loss amount distributions. Finally, we arrive at the office premium for the e-risk insurance product by adding the risk premium with expense loading (10% of the risk premium) and contingency loading (10% of standard deviation) factors respectively.

This paper comprises 6 sections. In Section 2 we present a rationale for considering e-risk as an operational risk and thereon describe the prescribed methods for quantifying operational risk as described in literature. Section 3 provides a brief overview on Bayesian belief networks and Copula. In Section 4, we use BBN for security breach analysis. In section 5, we provide a model for premium calculation that an insurance company can use for e-insurance product design. Section 6 provides the concluding remarks.

2. e-Risk

Most banks world over provides features like ATM and Internet banking. Such online transaction cost much less (customer transaction through a branch costs about $1USD in the U.S, but an online transaction cost just $0.02 USD [9]) and also increases customer satisfaction. The retail business and services sector today are very much dependent on Internet based online transactions. Forester [10] reports that online retail sales (B2C) in USA in 2003 was $95.7 billion, and projects it to be $229.9 billion USD by 2008. The success of an automated supply chain management or a virtual online organization (such as amazon.com, e-bay) is solely dependent on the Internet/network channels. But, the communication channel or the Internet till date is prone to virus attacks, snooping, sniffing and DoS by malicious intruders. A hacker or a virus attack can affect the top line of an organization and in the long run have effects on the bottom line too.

The three approaches for quantifying operational risk [11] are as listed in Table1.

Table 1: Operational Risk Quantification Techniques

Process approach	Factor approach	Actuarial approach
Causal Bayesian belief networks	Risk Indicators	Empirical Loss distributions
Fuzzy logic Statistical quality control and reliability analysis	Capital Asset pricing Model (CAPM)	Explicit distributions parameterized using historical data.
Connectivity System dynamics	Predictive models	Extreme-Value Theory

The *process approach* focuses on the modeling chain of activities that comprise an operation or transaction and finding out the exact risk for each process. The *factor approach* aims to determine a mathematical equation that relates the level of operational risk for institution or business or process to a set of factors as follows:

$$\text{Operational risk} = a_0 + \sum_{i=1}^{n} a_i * \text{factor}_i \quad (1)$$

Methods of regression and discriminant analysis are used to determine the values of a_i. The *actuarial approach* aims to identify the loss frequency and the loss amounts distribution associated with an event using past data.

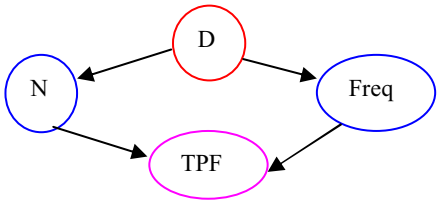
In this paper, we have used Bayesian Belief Networks to model e-risk for online organizations. We choose the process approach, as it is easy to visualize the chain of activities that mostly lead to a security failure, based on past data and expert opinion. Further BBN, has the ability to train itself based on observed data supplied to it, and identify conclusively the factors that are responsible for a security breach. Modeling e-risk through, the factor approach is difficult, as identifying the causal factors isn't easy. While actuarial approach is constrained by availability of data points, in case of e-risk modeling.

3. Bayesian Belief Networks

Bayesian belief network (BBN) [12] is a graphical relationship between causal variables. BBNs enable reasoning under uncertainty and combine the advantages of an intuitive visual representation with a sound mathematical basis of Bayesian probability. A BBN graph is directed acyclic graph. Nodes are random variables and represent an uncertain event. The arcs indicate causal relationship between the variables. Each node has a probability table associated with it. The root node(s) have only marginal probability associated with it, while the child nodes have conditional probabilities tables associated with it. Initially, each of the nodes is populated using a belief (i.e., either an expert view or results from empirical study). If an event in the BBN graph is observed with certainty, it causes each of the probabilities at the nodes to be recalculated, thereby providing a better understanding of the process being modeled. The full joint probability is joint distribution over all the values in the domain and is computed as a product of conditional probabilities for every node-parent combination as follows in (2):

$$P(X_1,X_2\ldots\ldots\ldots X_n)=\prod_i P(X_i \mid Parents(X_i)) \quad (2)$$

For example, say we have a BBN as shown in Figure 1.



D=Daily, N=Nature of monitoring, Freq=Frequency of update, TPF=Techno Policy Failure.

Figure 1: A BBN

The joint probability of all the nodes is as follows:

$$P(D, N, Freq, TPF) = P(D)*P(N|D)*P(Freq|D) *P(TPF|Freq, N)\ldots\ldots\ldots\ldots \quad (3)$$

Let us assume that each of the 4 nodes has 2 states each (i.e. true or false) .Some sample conditional probabilities can be calculated as follows:

$$P(N = True \mid TPF = True) = \frac{P(N = True, TPF = True)}{P(TPF = True)}$$

$$= \frac{\sum_{d,freq} P(D = d, N = True, Freq = freq, TPF = True)}{P(TPF = True)}$$

. (4)

Similarly,

$$P(Freq = True \mid TPF = True) = \frac{P(Freq = True, TPF = True)}{P(TPF = True)}$$

$$= \frac{\sum_{d,n} P(D = d, N = n, Freq = True, TPF = True)}{P(TPF = True)}$$

(5)

In both these cases,

$$P(TPF = True) = \sum_{d,n,freq} P(D = d, N = n, Freq = freq, TPF = True)$$

(6)

3.1 Copula

The essence of the copula approach [13] is that a joint distribution of random variables can be expressed as a function of marginal distributions. This is explained by Sklar's Theorem [13]. The theorem states that given a joint cumulative distribution $F(x_1,x_2, \ldots\ldots\ldots\ldots,x_n)$ for random variables X_1 , X_2, ……, X_n with marginal cumulative distribution functions (CDFs) $F_1(x_1)$, $F_2(x_2)$,……, $F_n(x_n)$, F can be written as a function of marginals:

$$F(x_1,x_2, \ldots\ldots\ldots\ldots,x_n) = C[F_1(x_1),F_2(x_2),\ldots,F_n(x_n)] \quad (7)$$

where $C(u_1,u_2\ldots\ldots\ldots\ldots u_n)$ is a joint distribution function with uniform marginals.

Moreover, if each F_i is continuous, then C is unique, and if each F_i is discrete, then C is unique on the Ran (F_1) X Ran (F_2)……X Ran(F_n) , where Ran(F_i) is the range of F_i. The function C is called a copula.

Given that F_i and C are differentiable, the joint density $f(x_1,x_2, \ldots\ldots\ldots\ldots,x_n)$ can be written as

$$f(x_1,.,x_n) = f_1(x_1)..\times f_n(x_n)\times c[F_1(x_1),...F_n(x_n)] \quad (8)$$

where $f_i(x_i)$ is the density corresponding to the cumulative distribution function $F_i(x_i)$ and c is the copula density.

From (8) it is clear that the copula density c encodes information about dependence among X_is. c is also called the dependence function.

The correlation amongst the random variables is captured in pairs, using measures of dependence or association. Two common measures are Spearman (ρ) and Kendall (τ) correlation coefficient. The correlation matrix so obtained is termed R*. The final R matrix is obtained by using the transformations $r_{ij}=\sin(\Pi\tau_{ij}/2)$ and $r_{ij}=\sin(\Pi\,\rho_{ij}/6)$ respectively for Kendall and Spearman. The correlation matrix (R*) is obtained from the data or a domain expert.

Incorporating the R matrix (8) can be rewritten as follows:

$$f(x_1,..,x_n \mid R) = f_1(x_1) \times ..... \times f_n(x_n) \times c[F_1(x_1),...F_n(x_n) \mid R^*] \quad (9)$$

Assuming multivariate normal copula we can rewrite (9) as follows:

$$f(x_1,x_2,\ldots..,x_n \mid R^*) = f_1(x_1) \times f_2(x_2),\ldots..\times f_n(x_n) \times \frac{e^{\{-0.5*(\Phi^{-1}[F_1(x_1)],.....\Phi^{-1}[F_n(x_n)]) \times (R^{-1}-I) \times (\Phi^{-1}[F_1(x_1)],.....\Phi^{-1}[F_n(x_n)])^T\}}}{|R|^{1/2}} \quad (10)$$

Similarly, assuming multivariate normal copula the conditional probabilities are obtained by matrix partitioning method as shown below:

$$R = \begin{bmatrix} R_{n-1} & 1 \\ r^T & 1 \end{bmatrix}$$

and $y = (y_{n-1} \quad y_n)$ where $y_{n-1} = (y_1 \ldots\ldots y_{n-1})$ and R_{n-1} is the (n-1) X (n-1) correlation matrix for $(Y_1 \ldots\ldots Y_{n-1})$

The conditional probabilities are arrived by (11)

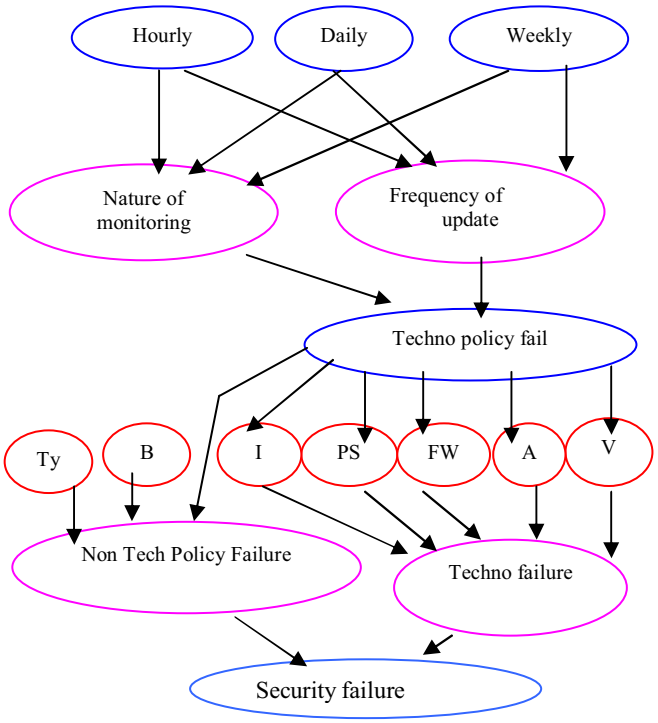
$$f(x_n \mid x_1,\ldots..,x_n,R) = f_n(x_n) \times e^{\{-0.5*[\frac{\Phi^{-1}[F_n(x_n)]-r^T R^{-1}_{n-1} y_{n-1}}{1-r^T R^{-1}_{n-1} r} - (\Phi^{-1}[F_n(x_n))^2]\}} \times (1-r^T R^{-1}_{n-1} r)^{-1/2} \quad (11)$$

where pdf of the normal distribution is given by ϕ and Φ denotes the cdf.

There are numerous families of copulas like Frank's, Archimedean [13], Multivariate normal copula etc. The choice of copula depends on the nature of the problem. [13].

4. Vulnerability Analysis Using BBN

Figure 2 is the BBN of the security failure of an organization. We assume that a security failure can occur primarily due to technology failure or improper non - technology policy implementation. Technology failure occurs if either the IDS or the proxy or the firewall or authentication is compromised or there is a virus attack. The compromise of the technology can be mainly ascribed to improper technology policy implementation (i.e. which port of the firewall to be left open etc). The effectiveness of a techno policy depends on frequency of update and nature of monitoring. The techno policy issues need to be reviewed or monitored on an hourly, daily or weekly basis. An effective non-technology policy (i.e. availability of budget and type of policy (i.e. centralized or local) reduces security breaches too. Improper technological policy (i.e., grant of wrong user privileges) can also lead to non-techno policy failure (i.e., an authorization failure) and in turn a security breach.



Ty=type, B=Budget availability, I=IDS compromised, PS=Proxy-Server compromised, FW=Firewall compromised, A=-Authentication compromised, V= Virus attack.

Figure 2: A BBN of security failure

We first supplied the directed acyclic graph (DAG) comprising of the 12 nodes i.e. hourly, daily, weekly, nature of monitoring, frequency of update, I, PS, FW, A, V and techno failure to FULLBNT. For this study, we have kept ourselves restricted to 4 nodes only (techno policy failure, FW, V and Techno failure). Each node has 5 states each (very low, low, medium, high, very high). We simulated the marginal probability for each of these 4 nodes, using a uniform distribution function. We populated the correlation matrix (R), as required by copula model, based on the DAG. We have assumed R to be symmetrical and diagonals all having ones. Then using Copula model we arrived at the

joint distribution of the 4 nodes. From this joint distribution we generated the conditional distribution for each of the nodes, as required by BBN.

We then ran the FULLBNT tool to investigate the following cases.

Case I: P (FW=very high| Techno failure =very high) =0.201609.

Case II: P (Virus attack=very high | Techno failure =very high). =0.201989

From this analysis, it is clear that virus attack was the main cause for the system we had simulated. Similarly, we can find out the causal effect for any of the 12 nodes present in Figure 2. Thus BBN provides an effective mechanism for vulnerability analysis.

5. Premium Calculation

We now discuss a premium calculation model [14, 15, 16], for an insurance company, using the Collective Risk Model [17] to arrive at the expected loss. The loss frequency and loss amount of an e-risk are both stochastic variables. The expected loss or claim severity is the product of claim size and claim amount.

We assume that E (N) denotes the expectation of the loss frequency distribution. While E(X) is the expectation of the loss amount distribution. We would obtain the data about the claim frequency from the BBN. We assume all the nodes have a similar binomial loss amount distribution of the form Binomial (1000, 0.2).The mean loss (E(X)) is 200 and its variance (V(X)) is 160. The expected loss or claim severity (E (S_N)) and its variance (Var(S_N)) are obtained using (12).

$$E(S_N)=E(N)\ *E(X)$$
$$Var(S_N)=E(N)*Var(X)+\{E(X)\}^2*Var(N) \quad (12)$$

Finally the premium is arrived at by using (13):

$$Premium= (1+q)*\ E(S_N)+\ k*\ \sqrt{Var(S_N)} \quad (13)$$

Where q is the loading factor (profit an overhead charges and k is the contingency loading.

We will now illustrate the model using the following example. We assume an expense loading of 10% of the risk premium and a contingency loading, of 10% of standard deviation, for premium calculation.

Case I: P (FW=very high| Techno failure =very high) =0.201609

Here the expected downtime (E (N)) and variance (V (N)) of the FW is 854.55 and 740,448 respectively. The expected claim severity is obtained using (12) as follows:

Mean = E (S) =expected loss= 854.55 *200=$170,910 USD.
Variance = Var (S) = 854.55*160 + $(200)^2$*740,448
=29,618,056,728

The premium the insurance company would charge the e-business organization for insuring its Firewall and techno policy failure is obtained using (13). The premium is (1+0.10)* 170,910+ 0.10*√29,618,056,728= $205,211 USD.

Case II: P (Virus attack=very high | Techno failure =very high). =0.201989

Here the expected downtime (E (N)) and its variance (V (N)) due to anti virus failure are 847.48 and 769,957 respectively. The expected claim severity is obtained using (12) as follows:

Mean = E (S) =expected loss=847.48*200=$169,496.USD
Variance = Var (S) = 847.48 *160 + $(200)^2$*769,757
=30,790,415,597

The premium of e-insurance product for insuring against virus attack is obtained from (13) is (1+0.10)* 169,495.46+ 0.10*√30,790,415,597 = $203,993 USD.

6. Conclusion

The study was aimed to identify the vulnerable point in the network security of an online organization, and subsequently quantify the risk associated with online transactions. We have developed a BBN elucidating the causal relationship amongst the various vulnerable points that result in security breach. To generalize the calculations, we have assumed marginal distributions at each of the nodes. Using Copula we calculated the joint distribution of the entire DAG and finally, from it arrived at the conditional distribution for each of the nodes. The BBN provides us with the frequency of loss for each of the nodes, given certain evidence. We then come up with the expected loss amount by multiplying the expected loss amount with the probability of occurrence of an event. Thus we have a model for quantifying e-risk in monetary terms in case of failure of online systems. Insurance companies can use this model to set premium for e-insurance products.

7. References

[1] Lawrence A.Gordon, Martin P.Loeb, William Lucyshyn and Robert Richardson, CSI/FBI Computer Crime and Security Survey, GoCSI.com,2004

[2] Gordon and Loeb, The economics of information security investment, ACM Trans on Inf Sys Sec, 5,4, , 438-457, Nov 2002

[3] Anderson, R., Why information security is hard: An economic perspective, In *Proceedings of 17th Annual Computer Security Applications Conference (ACSAC)* ,New Orleans, La. Dec.10–14,2001.

[4] Schneier, B. Secrets & Lies.: Digital security in a Networked World (2nd edition), Wiley, New York, 2000.

[5] Lawrence A.Gordon, Martin P.Loeb, Tashfeen Sohail, A framework for using Insurance for Cyber-risk management**,** Communications of the ACM, Vol.46, No.3 81, March 2003,

[6] Torsten Grzebiela, Insurability of Electronic Commerce Risks, Proceedings of 35th HICSS, 2002.

[7] Rowen R Hillary, Avoiding Bad days for Risk managers insurance for Internet ,www.thelenreid.com/articles/article/art_70_idx.htm

[8] FULLBNTsoftware,http://www.ai.mit.edu/~murphyk/Software/BNT

[9] Kellermann Tom, Mobile risk management: e-finance in the Wireless Environment, World Bank Group,www1.worldbank.org/finance/assets/images/Mobile_Risk_Management.pdf , May2002.

[10] Burner E.Rick, Forrester Sees Strong Ecommerce GrowthContinuing, www.marketingvox.com/archives/2003/08/05/forrester_sees_strong_ecommerce_growth_continuing,5 Aug, 2003.

[11] Smithson Charles, Song Paul, Quantifying operational risk, www.risk.net, July 2004

[12] Jensen F V. Bayesian Networks and decision Diagrams, Springer 2001.

[13] Clemen T Robert , Reilly Tereence , Correlations and Copulas for decision and risk Analysis , Management Science ,Vol45 ,No2 , 208 –224 ,Feb 1999.

[14] Mukhopadhyay A , Chatterjee S , Saha D , Mahanti A , Chakrabarti B B , Podder A K ,Mitigating Security breach losses in e-commerce through Insurance ,Proceedings of 4th Security Conference ,Las Vegas ,Nevada ,March 30-31 ,2005

[15] Mukhopadhyay A , Chatterjee S , Saha D , Mahanti A , Podder A K ,e-risk :A case for insurance ,Proceedings of the Conference on Information Systems and Technology Management ,New Delhi ,July 23-26 ,2005

[16] Mukhopadhyay A , Saha D , Mahanti A , Chakrabarti B B , Podder A ,Insurance for cyber-risk: A Utility Model ,Decision, Vol 32 ,No 1, 153-170., June 2005.

[17] Hossack B I, Pollard J,Zehnwirth B ,Introduction to Statistics with applications to general insurance , Cambridge University Press,1983.