Hemantha S.B. Herath (Canada), Tejaswini C. Herath (Canada)

Copula-based actuarial model for pricing cyber-insurance policies

Abstract

Cyber-insurance is often suggested as a tool to manage IT security residual risks but the accuracy of premiums is still an open question. Thus, practitioners and academics have argued for more robust and innovative cyber-insurance pricing models. The paper fills this important gap in the literature by developing a cyber-insurance model using the emerging copula methodology. The premiums for first party losses due to virus intrusions are estimated using three types of insurance policy models. Our approach is the first in the information security literature to integrate standard elements of insurance risk with the robust copula methodology to determine cyber insurance premiums.

Keywords: cyber-insurance, copula, correlated risk, information security risk management.

Introduction

Reported financial losses due to information security breaches give us an overall glimpse of the severity of the information security problem. Security breaches result in losses of millions of dollars due to direct costs such as lost revenues, lost productivity, and lawsuits, as well as more intangible losses such as loss of customer goodwill, lost reputation, and lost business opportunities. The recent 2008 CSI Computer Crime and Security Survey (Richardson, 2008) notes that most organizations use security tools with almost 97% using antivirus software, 94% having firewalls, and 69% having intrusion detection systems. Despite this increased use of security measures, the security breaches and the losses due to these breaches remain high. It is difficult for security managers in any organization to know about and eliminate all the points of vulnerability in an IT system (i.e., create a foolproof system), and a hacker needs just one of these points of vulnerability to exploit (Anderson, 2001).

Recognizing that the total elimination of security breach risk is close to impossible, National Institute of Standards and Technology (NIST) recommends several risk mitigation techniques that are based on technical as well as non-technical controls. These techniques include risk assumption, risk avoidance, risk limitation, risk planning, research and acknowledgment, and risk transference (Stoneburner et al., 2002). Our article focuses on risk transference as a tool that minimizes some of the financial losses to firms (i.e., to transfer the risk by using other options to compensate for the loss, such as purchasing insurance). Both practitioners and academics have suggested using risk transference with insurance to absorb losses caused by security breaches and to supplement the existing set of tools used to manage IT security residual risk after IT security investments are made (see Gordon et al., 2003; Ogut et al., 2005 among others).

There are currently a variety of cyber-insurance products offered in the market. Cyber-insurance is a specialty insurance product that covers losses associated with a firm's information assets including computer generated, stored, and processed information. Traditional business insurance products cover tangible property but do not cover assets like data and information. Cyber-insurance is primarily designed to cover intangible business assets which traditional business insurance does not cover. Such insurance policies help protect against losses due to cyber attacks, employee breaches of network security, hackers, the associated liability of these events, the consequential expenses due to privacy breach, and liability over website content, among other potential losses (Oellrich, 2003). These insurance policies cover both first party business losses and third party liability (Betterly, 2007).

The cyber-insurance market evolved with the advent and dispersion of Internet use in commerce activities. Majuca et al. (2006) and Baer and Parkinson (2007) provide a nice discussion on the evolution of the cyber-insurance market. Although specialty coverage against computer crime first appeared in the 1970s, it was mostly an extension of traditional crime tied to electronic banking. It was in 1998, however, that the earliest cyber-insurance products were offered by technology companies which partnered with insurance companies (Majuca, 2006). These companies included ICSA TruSecure, Cigna Corp/Cisco Systems/NetSolve, J.S. Wurzler Underwriting, IBM/Sedgwick, Counterpane/Lloyd's of London, Marsh McLennan/AT&T, and AIG, and mostly offered first party coverage. Current cyber-insurance market includes many carriers such as ACE USA, American International Group (AIG), Chubb, AON, St. Paul Travelers, INSURETrust, SWBC, Allied World/Darwin, Aspen, The Hartford, Navigators, RLI, XL, and Zurich and retailers such as Digital Risk Managers/Lloyds, Euclid/ Hudson, and Safe-online/Lloyds. They offer both first party and third party coverage.

According to the Betterley (September 2010), cyber-insurance market is expected to grow due to widespread concern over data breaches, creative efforts by

hackers, and political support for regulatory action. While annual premium volume information is difficult to gather the U.S. annual gross premium revenue for cyber-insurance policies has grown from less than US$100 million in 2002 to US$300-350 million by mid-2006 (Baer and Parkinson, 2007). It is, however, forecast to be in the $600 million range by the end of 2010 (Betterley, June 2010). The market which primarily involved large businesses seeking coverage is now broadening to small and mid-sized companies which are becoming aware of the possibilities of liability (Betterley, June 2010). Furthermore, many carriers are reporting strong growth in premiums with several reporting growth of over 100%, and few reporting between 50-100%. Businesses are seeking coverage for the value of the data loss, lost revenue due to loss of data, lost revenue due to repair downtimes, legal expenses for damage to another party, cost of crisis management, notification, credit monitoring and restoration after a data breach, and regulatory fines and penalties (Betterley, September 2010). New trends also suggest privacy coverage to be driving the cyber-insurance market with new products such as Aspen's New Privacy Control Breach Response.

Premiums vary according to specific situation and the amount of coverage, and can range from a few thousand dollars for base coverage for small businesses (less than $10 million in revenue) to several hundred thousand dollars for major corporations desiring comprehensive coverage. Although insurance companies provide cyber-insurance products, the accuracy of the pricing and whether or not insurance providers are charging the right premiums is still an open question (Gordon et al., 2003). Premiums depend on the individual firm's security risk exposure and can vary substantially depending on the insurance provider. To address this, both practitioners and academics have argued for more robust and innovative cyber-insurance pricing models to stimulate increased growth in the cyber-insurance market (Baer and Parkinson, 2007; Betterly, 2007; Geer et al., 2003; Oellrich, 2003). In this article, we attempt to fill this important research gap by developing a cyber-insurance pricing model, where the premiums depend on the number of computers affected, the firm level dollar loss distribution, and the timing of the breach event.

The contribution of this article is threefold. First, we incorporate three elements of a standard insurance contract – the settlement amount that is paid, the occurrence of the event covered by the contract and the time when the settlement is paid into pricing cyber-insurance and explicitly model them in the context of information security. Second, the proposed model applies the copula methodology that allows for capturing of non-linear dependencies among the input pricing variables. Copulas do not place restrictions on the type of marginal distributions considered for the pricing variables. The novel integration of the copula methodology makes the modeling robust and offers a methodological contribution to information security research. The use of copulas is essential but relatively new to the cyber-security insurance industry. Finally, using empirical loss distributions based on publicly available ICSA survey data, we illustrate a copula-based Monte Carlo simulation model for pricing cyber insurance. More specifically, we use firm level data of the number of computers affected and the dollar losses due to virus incidents as a starting point to illustrate the assessment of the empirical joint loss distribution that is essential for premium pricing. We compute the premiums for first party losses using three types of insurance policy models: basic policies, policies with deductible and policies with deductible and co-insurance.

The paper is organized as follows. Section 1 provides a review of the related cyber-insurance literature. In Section 2, we provide a framework for cyber risk assessment that combines standard elements of an insurance contract with copula methodology from the perspective of information security. In this section, we also introduce the concept of copulas. Section 3 describes the cyber-insurance models and in Section 4, we illustrate the models using ICSA data. The final Section concludes this paper with a discussion of limitations, future research directions and managerial implications. An Appendix provides a primer on the copula methodology used in this article.

1. Related cyber-insurance literature

Cyber-insurance as a risk management tool gives rise to challenges that are typically not considered in traditional business insurance models. Issues related to pricing, adverse selection, and moral hazard are common to all forms of insurance. In addition, several technology-related characteristics make the pricing of cyber-insurance challenging. First, internet-related risks are unique in terms of location, degree, and visibility. Traditional policies do not comprehensively address the additional risks that firms face as a result of being part of the digital economy (Gordon et al., 2003). Second, the Internet is a shared medium. Firms use common software and interact with other firms. Cyber security involves many layers and requires attention through many stages of a system's lifecycle, from software design and system configuration to maintenance tasks such as patching to achieve final results. Thus, an important feature of cyber security is that it may need collaborative partnerships to attain these goals. These factors create interdependencies that make risk and vulnerability assessment difficult.

An understanding of cyber-risk issues is complicated but essential when designing insurance products. These unique challenges create a number of important research opportunities that need to be addressed from both – the point of view of insurance companies (supply side) and the insured (demand side). Pricing insurance products traditionally relies on actuarial tables constructed from historical records. However, unlike traditional insurance policies, cyber-insurance has no standard scoring system or actuarial tables for pricing premiums. The Internet is relatively new, and as such data about security breaches and losses does not exist or does so only in small quantities. This difficulty is further exacerbated by the reluctance of organizations to reveal details of security breaches due to loss of market share, loss of reputation, etc.

Recently, there has been a growing stream of research focusing on cyber-insurance. In one of the earliest articles proposing cyber-insurance, Gordon et al. (2003) discuss a framework for using cyber-insurance as a risk management technique. They describe the unique features of cyber-insurance and the problem of adverse selection and moral hazard which are common to all insurance markets. Bolot and Lelarge (2008) combine recent ideas from risk theory and network modeling in an economic approach to develop an expected utility insurance model. They investigate the interplay between self-protection and insurance. Their results show that using insurance is beneficial since it increases the security of the Internet. Ogut et al. (2005) investigate cyber-insurance explicitly from a moral hazard and adverse selection perspective. They show that the interdependence of IT security risk among different firms impacts a firm's incentive to invest in cyber-insurance products.

Recent literature in this area has recognized the value of copula methodology for modeling dependent risks (Böhme and Kataria, 2006; and Mukhopadhyay et al., 2006). Copulas, term coined by Sklar (1959), have been studied for over forty years. Copulas are functions that join or couple multivariate distribution functions to their one-dimensional marginal distribution functions. Alternatively, copulas can be described as multivariate distributions whose one-dimensional margins are uniform in the interval [0, 1] (Frees and Valdez, 1998; Nelsen, 1995). Copulas are of interest to statisticians for two main reasons: (1) as a way of studying scale-free measures of dependence; and (2) as a starting point for constructing families of bivariate distributions for simulation (Fisher, 1997). In the case of insurance, this implies modeling the non-linear dependencies in the pricing variables and using simulation to determine the premiums.

Mukhopadhyay et al. (2006) attempt to model cyber-insurance claims using copulas. They use a copula setting with the Bayesian Belief Networks (BBN) technique to quantify the e-risks associated with online transactions that would be affected by security breaches. They employ the multivariate normal copula to describe the joint distribution and the conditional distribution at each node on the BBN. They use the software FULLBNT to identify breach probabilities and make the following assumptions. The dollar losses at each node in the network are distributed binomially with assumed specific values and that cyber-insurance premiums are computed as a function of the expected value of the claim severity.

Our paper contributes to this recent but growing body of IS literature. We adopt an empirical approach using Archimedean copulas that is different to the process/utility based approaches used by Böhme (2005), Böhme and Kataria (2006), Mukhopadhyay et al. (2006) and Bolot and Lelarge (2008). The primary limitation of process/utility approach is that it cannot be used by practitioners since it is not based on an actuarial approach. We investigate cyber-insurance pricing using the emerging copula methodology for modeling dependent risks from an actuarial approach that is based on empirical distributions. In this paper we use two Archimedean copulas: Clayton and Gumbel.

2. Framework for assessing cyber risk

In this section, we develop a framework for assessing cyber risk that consider typical insurance pricing variables but from the perspective of IT security. More specifically, we discuss how to model loss function at the individual firm level from the population data using the number of affected computers as proxy for size.

2.1. Cyber-insurance risk elements. There are three elements of risk that are typically part of any insurance contract (Klugman, 1986): (1) the settlement amount that is paid; (2) the occurrence of the event covered by the contract; and (3) the time when the settlement is paid. In the proposed insurance model for pricing *first party business interruption due to security breaches*, the unknown random variable of interest is the amount (P) that is paid by the insurance company. The amount paid (P) will depend on the dollar loss amount ($\prod$) that is likely to be incurred by a firm with (q) number of affected computers. Suppose π is the observed dollar losses from available data (from, for example, the ICSA computer virus prevalence survey). We can model the loss ($\prod$) pertaining to a breach event at firm level as a function of both (q) and (π) given by $\prod = g(\pi, q)$. In the model, we assume that loss distribution for a

firm will depend on two random variables π and q. Since reliable data on $\prod$ is a scant, insurance companies can use the publicly available data such as the ICSA survey data to model $\prod$ using an appropriate copula[1]. The field of copulas in statistics is relatively new to the IS field. Therefore, for pedagogical reasons and for better understanding of the proposed model we provide a lucid introduction to copulas in Appendix.

Copulas are more appropriate to model the joint loss distribution $\prod = g(\pi, q)$ for two reasons. First, copulas allows combining any type of fitted marginal distributions for π and q. Thus, no restrictions are imposed on the type of marginal distributions for (q) and (π). This is quite useful in cyber-insurance given that a wide range of empirical distributions are possible. Second, copulas are ideal for investigating non-linear type dependencies that arise when non normal marginal distributions of the type for (q) and (π) are combined. The type of dependence between random variable; such as (q) and (π), is crucial in many respects to cyber risk management because the variables (q) and (π) are intertwined due to the partial dependence of the losses on the number of computers that are exposed to and affected by a security breach.

The second element of risk, the occurrence of the event that is covered by the contract, is modeled by a binary variable ω. It takes the value of 1 if the covered event has occurred or 0 otherwise. The third element of risk is the time until the settlement is paid (T), which is the time from issue of policy to when the claim is paid. This time period includes the time until the breach incident from the issuance of policy and the time until settlement after the incident takes place. It is reasonable to assume that the time between the incident and the settlement will be short (or equivalently zero). This assumption is valid since we are modeling first party business interruption (first party damage) and not liability insurance. If one is modeling cyber-insurance liability coverage, then the settlement time can be substantial due to legal proceedings and should be considered in the model. Thus, the parameter of interest for our model is time until breach incident. In information security, for random events such as virus intrusions, the Poisson distribution is widely used to model the arrival of intrusions per unit time (Conrad, 2005; Herath and Herath, 2009; Longstaff et al., 2000). One can use the Poisson intrusion rate process to determine the time until the IT system is breached. We use this approach in our paper to model the time until the breach incident or time until settlement (T).

The use of copulas implicitly assumes that the random events which caused the losses would be repeated. In the case of first party damage due to viruses, one can reasonably assume that the random breach event (virus intrusion) would likely follow a similar pattern. This may also be reasonably true for hacking if the event is random. However, it may not be the case if the hacker is particularly targeting a strategic organization such as a military establishment, NASA, etc. In this case, the pricing has to be tailored on a case by case basis by adding another layer of insurance on top of the general random events.

2.2. Modeling the loss distribution $\prod$ using copulas. Copula methodology can be used effectively to model the joint dollar loss distribution due to cyber attacks at firm level. A key component of insurance pricing is to understand and model multivariate relationships. While linear regression may provide a basis for explaining the relationship between two (or more) variables, the model is based on normality assumptions and linear dependence. Linear regression would work if the marginal distributions are normal. However, the marginal distribution for the number of computers affected (q) and the dollar value of losses (π) may not be normal (as exemplified in case illustration). In the case of the pricing variable, the number of computers affected (q), the marginal distribution is likely to be of the type Pareto, Exponential, or Weibull, since a few viruses (15%-25%) account for (85%-75%) a large number of computers affected. Because the fitted marginal distributions are non-normal, the widely used classical Pearson's product moment correlation (ρ) cannot be used to model the dependency between the two variables. Correlation (ρ) measures the straight line association and the dependency is linear. Thus, in modeling a firm's loss distribution $\prod$ from the available empirical data, the copula dependency is more appropriate.

In the copula approach for modeling the firm's loss distribution, the first step (as described in Appendix, Section 3) is to identify the "appropriate copula" for modeling the non-linear dependence that explains the relationship between the two variables of interest, the number of computers affected (q), and the dollar value of losses (π). That is, we identify the joint distribution of (π, q) by the specific function say $\prod = g(\pi, q)$. Notice that we can now examine the firm's loss distribution of any known function of q and π. To determine the loss distribution, let l and m

[1] In 2008 when the financial crisis occurred, the copula methodology was widely criticized as it was used by Li (2000) to model risks associated with credit derivatives. It was later argued by Krugman (2008) that the primary cause of the credit crisis was not the use of copulas to model a single credit derivative but a lack of understanding regarding the aggregate risks caused by writing multiple derivatives contacts on the same instrument. Copula is the primary tool that allows marginal distributions to be combined when they are non-elliptical.

be the lowest and the highest limits of the number of likely computers affected. Assuming the dollar value of the likely losses can be prorated based on the number of computers affected (or exposed), say the following function captures the firm's specific loss distribution:

$$\Pi = g(\pi, q) = \begin{cases} a_1, & \text{if } q < l \\ a_2 + \left(\frac{q-l}{q}\right)\left(\frac{\pi}{10}\right), & \text{if } l \le q < m, \\ a_3 + \left(\frac{q-m}{q}\right)\left(\frac{\pi}{10}\right), & \text{if } q \ge m \end{cases} \quad (1)$$

where, a_i, $i = 1, 2, 3$ are constants. The jointly distributed values for a firm's loss distribution $\prod = g(\pi, q)$ can be computed using Monte Carlo simulation.

3. Copula-based cyber-insurance model

The probabilistic model for the cost of cyber-insurance for first party damage due to a breach based on fundamental risk elements of an insurance contract is:

$$C = \omega e^{-rT} P, \quad (2)$$

where ω is a binary variable, equal to one if the covered event occurs and zero otherwise, T is the time until the security breach incident, r is the discount rate, and P is the amount paid by the insurance company in the event of a breach. For simplicity, we assume that the covered event happens only once in the contract period and that the loss pertains to a single claim[1]. More specifically, T is the time to the first instance of a cyber security breach and we assume that the system fails after the first breach and the claim is paid only once. The contract period is up to the first breach event. In the case of cyber-insurance, the amount paid P and the time of issuance of the policy to the payment of the claim T can be reasonably assumed to be independent because P is not a function of T[2]. The net premium is given by:

$$\mathrm{E}(C) = \overline{\omega}\mathrm{E}(e^{-rT})\mathrm{E}(P), \quad (3)$$

where the probability of event occurring $\overline{\omega} = \mathrm{E}(\omega) = \mathrm{Prob}(\omega = 1)$. This net premium does not include the expenses and profits of the insurance company. In order to obtain the actual premium charged by the insurance company, an additional amount to cover the expenses and profits have to be added.

In this section we present three different types of cyber-insurance policy models. The first model is a basic first party damage policy with a zero deductible. The second model is a first party damage policy with a deductible. Finally, the third model considered is a first party damage policy with co-insurance and limit. For the first two models, suppose we consider a first party damage cyber policy that has a deductible (d), the observed random variable – an amount of loss ($\prod$), and (P) the amount paid which is the random of interest, then the policy relationships are modeled as:

Policy type 1. Basic first party damage policy with a zero deductible:

$$P = \Pi = g(\pi, q). \quad (4)$$

Policy type 2. First party damage policy with a deductible:

$$P = \begin{cases} 0, & \text{if } \Pi \le d \\ \Pi - d, & \text{if } \Pi > d \end{cases}. \quad (5)$$

Consider a third model with a deductible d, co-insurance of a, and a limit of k. The insurance pays nothing when the loss is below d, pays 100% $(1 - a)$ of excess losses over d, but never pays anything over k. The relationship between the random variable of interest P and the observed random variable $\prod$ can be modeled as:

Policy type 3. First party damage policy with co-insurance and limit:

$$P = \begin{cases} 0, & \text{if } \Pi \le d \\ (1-a)(\Pi - d), & \text{if } d < \Pi < d + \frac{k}{1-a}. \\ k, & \text{if } \Pi > d + \frac{k}{1-a} \end{cases} \quad (6)$$

In our approach, we can directly infer about P since the observed variable $\prod$ is known (it is estimated from the ICSA data)[3]. We do not have historical data of actual amounts paid (P) by insurance companies nor the number of polices to estimate the frequency of $\overline{\omega}$. In our approach, since we know $\prod$, we can assume that when $\prod > 0$ (a loss was incurred), then the event occurred and hence $\omega = 1$. However, since we do not have the data to estimate

[1] Klugman (1986) discusses several alternate ways to incorporate situations, where the covered event could happen several times during the contract period.

[2] If P is a function of T, one can use copulas to model appropriately the dependency and determine the joint distribution.

[3] Klugman (1986) provides an example where the observed variable is P (the amount paid by the insurance company) and the unknown loss distribution $\prod$ has to be inferred. In such instances, $\overline{\omega}$ the frequency (i.e., the number of times the covered event occurred) is found by the relative frequency of P (i.e., the number of times an actual payment of amount P was made for a firm incurring a loss of $\prod$/total number of polices observed).

the frequency and our main objective is to illustrate the copula approach in IT security context, we assume $\overline{\omega} = 1$. This is a simplifying assumption. Ideally, one should assume a fractional value for $\overline{\omega}$ or where a fitted distribution is available for the actual amounts paid for cyber policies $\overline{\omega}$ can be computed (see Klugman, 1986).

The other unknown variable is T – the time from when the policy is issued until the insurance settlement is paid. As explained in Section 2.1, since we are dealing with first party damage and not liability coverage, T can reasonably be assumed to be the time from the issuance of the policy until the breach incident or occurrence of the covered event, and can be modeled using a Poisson process. In a Poisson process with an expected arrival rate of λ intrusions per unit time, the number of arrivals in any time period t is λt and the numbers of arrivals in separate periods are independent of each other. Thus, the times between successive arrivals $A_i = t_1 - t_{i-1}$ are independent exponential random variables with mean $\frac{1}{\lambda}$. Assuming t_{i-1} has been determined, to generate the next arrival time t_i, the simulation procedure is as follows:

Step 1. Generate a uniform random variable $u \sim U(0,1)$ independent of any previous random variate.

Step 2. Return $t_i = t_{i-1} - \left(\frac{1}{\lambda}\right)\ln u$. Notice that $t_0 = 0$.

3.1. Integrated copula-based simulation algorithm for pricing cyber-insurance. The integrated copula-based simulation algorithm for pricing a first party damage policy due to an IT security breach is given below. The steps are as follows:

Step 1. Fit a copula to the empirical data (q, π). Next, generate a sequence of bi-variate data (q_k, π_k) for the k^{th} iteration using the fitted copula (i.e., Clayton or Gumbel or any other copula). The procedures for generating data from Clayton and Gumbel are well summarized in Frees and Valdez (1998) and Nelson (1999), among others. We illustrate the simulation procedure for the Gumbel copula in detail in Section 4.

Step 2. For each sequence of bivariate data (q_k, π_k) in step 1 above, compute the losses $\prod^k = g(\pi_k, q_k)$ using equation 1 for a given (l, m).

Step 3. Model the first instance of a security breach or the time until incident using the simulation procedure for a Poisson process $T^k = t_1^k = t^k - \left(\frac{1}{\lambda}\right)\ln u^k$ discussed in Section 3.

Step 4. Compute the insurance premium for the k^{th} iteration as $C^k = \omega P^k e^{-\delta T^k}$. Notice that we have considered three types of first party damage policies in this paper.

Step 5. Compute the expected value and the standard deviation of the cyber-insurance premium as:

$$E[C] = \frac{1}{S}\sum_{k=1}^{S} C^k = \frac{1}{S}\sum_{k=1}^{S} \omega P^k e^{-\delta T^k}, \quad (7)$$

$$\sigma[C] = \sqrt{\frac{\frac{1}{S}\sum_{k}^{S}\left(C^k\right)^2 - \left[E(C)\right]^2}{S}}, \quad (8)$$

where S is the number of simulation runs.

4. Case illustration

In this section, we illustrate the copula approach for pricing cyber-insurance using data from the ICSA survey. Consider a hypothetical organization, firm A, with the firm level data pertaining to the number of computers affected q and the dollar losses π for each major computer virus encountered in 2003. As assumed in Conrad (2005), we consider the possibility of two ($\lambda = 2$) breaches a year. We assume that firm's IT system fails after the first breach and the claim is paid only once. The coverage is only for the first breach event and the contract period is till the occurrence of first IT breach.

We use the ICSA survey data for 2003 on actual computer virus incidences and the actual number of computers affected, modified and scaled down one hundred times to represent firm level. Notice that both the number of computers affected as well as the dollar value of losses are random events. That is, the number of computers affected will depend on the severity of the virus, the company's security posture, and the security policies in place. Similarly, the dollar value of losses will be random in the sense that, in the rare instance of the same number of computers being affected by two distinct viruses, the degree of loss will not be identical because it will depend on each virus's ability to penetrate and harm the computers. Also, it will depend on the computer type affected (i.e., the proportion of stand alone computers, servers, network computers, etc.). The population data is given in Table 1. In Figure 1, we provide a scatter plot of the logarithmic values of dollar losses verses the number of computers affected. The Pearson correlation coefficient of 0.976 and the scatter plot indicates a strong relationship among the two variables.

In order to price the cyber-insurance premium for firm A, based on the number of computers affected and the dollar value of losses pertaining to each virus incidence, we have to capture the non-linear dependence

using the best fit copula from the empirical data and simulate the bi-variate data (q_k, π_k). For the copula-based simulation, we need to identify the marginal distributions for the number of computers affected q and the losses π (see Table 1). We use ARENA software for identifying the marginal distributions. The marginal distribution for number of computers affected q and the losses π are Weibull distributions of the following form. The fitted marginal distributions are $q \sim 18 +$ Weibull (118, 0.586) and $\pi \sim 5340 +$ Weibull (38900, 0.586)[1]. Notice that both the distributions are shifted Weibull distributions. There are several reasons why in such a case one needs to use a copula-based model to simulate the pair of values (q_k, π_k). First, the marginal distributions are Weibull and not normal, thus one cannot use linear regression. Second, finding the joint distribution of the two variables is complex since they are shifted Weibull distributions. And finally, Pearson's product moment or linear correlation cannot be use since the marginals are non-normal. The copula approach, as illustrated below, enables us to identify the appropriate copula to determine a joint distribution that can be used with these non-normal marginal distributions. Furthermore, a copula approach allows the modeling of non-linear dependency, which is more appropriate for estimating the premiums.

Table 1. Surveyed computer virus data

	Virus	q # of computers	π $ losses
1	W32/Blaster	1291	$355 648.72
2	W32/Slammer	849	$339 832.66
3	W32/Sobig	238	$115 729.51
4	W32/Klez	140	$65 090.38
5	W32/Yaha	118	$45 402.25
6	W32/Swen	108	$66 053.73
7	W32/Dumaru	87	$39 182.88
8	W32/Mimail	70	$19 556.82
9	W32/Nachi	63	$20 087.13
10	W32/Fizzer	58	$20 465.35
11	W32/BugBear	50	$10 180.13
12	W32/Lirva	47	$11 769.29
13	W32/Sober	21	$6 944.48
14	W32/SirCam	21	$5 339.08
15	W32/Ganda	19	$7 547.77
Mean		212	$75 255
Standard deviation		363	$114 702

[1] We fitted marginal distributions using the ARENA software package, where the best fit marginal is the one with the minimum squared error. Weibull with squared error of 0.0226 for π and Weibull with squared error of 0.0257 for q were selected from among possible marginal distributions which included Exponential, Erlang, Gamma, Beta, Log-normal, Normal, Triangular and Uniform.

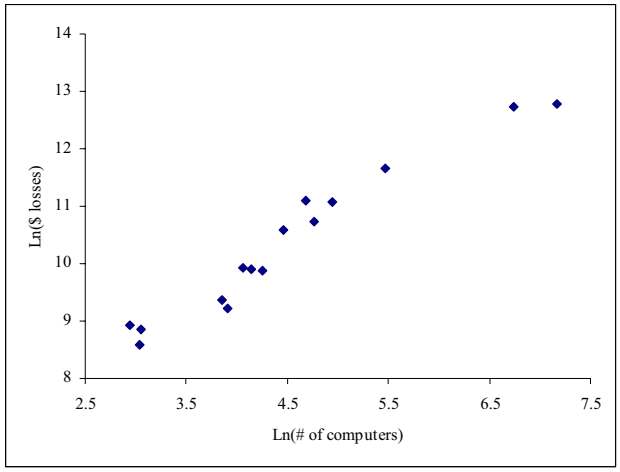

Fig. 1. Scatter plot $ losses verses # of computers on a logarithmic scale

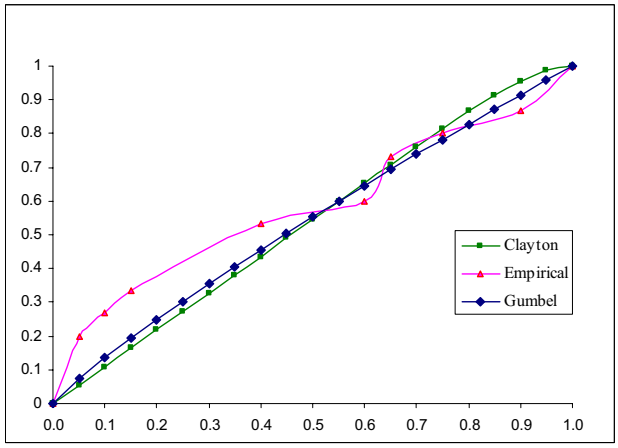

Fig. 2. Best fit copula

We use Kendall's Tau to measure the dependence between the number of computers affected (q) and the dollar losses (π) for the data in Table 1. We use the statistical software SPSS to compute Kendall's Tau, which is found to be 0.848. Next, we obtain θ values of 11.15789 and 6.578947 respectively using the equations in Appendix, Section 2 for Clayton and Gumbel copulas. In order to identify the appropriate copula, we follow the procedure outlined in Appendix, Section 3. The empirical distribution $K_E(z)$ and its parametric values $K(z)$ for Clayton and Gumbel copulas based on equations in Appendix, Section 3 (1) and (2) are shown in Figure 1. Frees and Valdez (1998) provides a nice exposition on fitting copulas including visual fit, the quantile-quantile (Q-Q) plots and more robust maximum likelihood approach. It is evident from Figure 2 that based on a visual fit, although both Clayton and Gumbel copulas are relatively close, the Gumbel copula provides the best fit. The Q-Q plots shown in Figure 3 also further confirm that Gumbel copula is the best fit. More robust statistical approaches pertaining to the specialized topic of copula selection methods are found in Frees and Valdez (1998), Huard et al. (2006) and Genest et al. (2009).

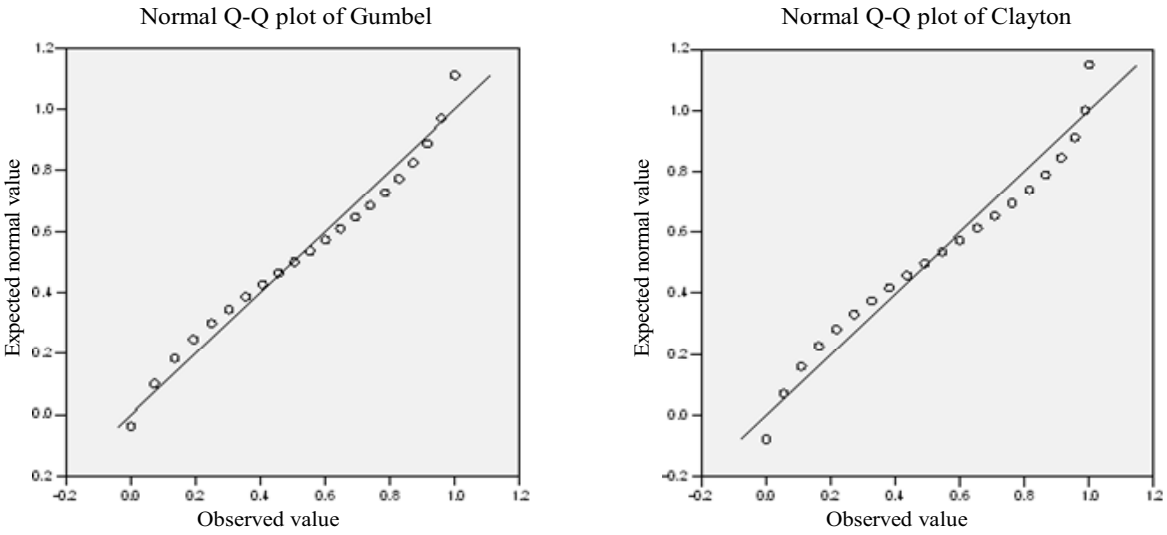

Fig. 3. Q-Q plots

In order to simulate bi-variate outcomes (q_k, π_k) using the Gumbel copula, we use the algorithm suggested by Marshall and Olkin (1988). For the Gumbel copula, the generator and Laplace transform are given in Appendix, Section 2. The inverse generator is equal to the Laplace transform of a positive Stable variate $\gamma \sim \mathrm{St}(\hat{\alpha},1,\Theta,0)$, where $\Theta = \left(\cos\left(\frac{\Pi}{2\theta}\right)\right)^{\theta}$ and $\theta > 0$.

Step 1. Simulate a positive Stable variate $\gamma \sim \mathrm{St}(\hat{\alpha},1,\Theta,0)$.

Step 2. Simulate two independent uniform [0, 1] random numbers u_1 and u_2.

Step 3. Set $q = F^{-1}(u_1^*)$ and $\pi = G^{-1}(u_2^*)$ where $u_i^* = \varphi\left(\frac{1}{\gamma}\ln u_i\right)$ and $\varphi(t) = \exp\left(-t^{\frac{1}{\theta}}\right)$ for $i \in [1, 2]$, where $F^{-1}(u_1^*) = 18 + 0.586\left(-\ln(u_1^*)\right)^{1/118}$ and $G^{-1}(u_2^*) = 5340 + 0.586\left(-\ln(u_2^*)\right)^{1/38900}$.

Nolan (2005) and Cherubini et al. (2004) suggest the following procedure to simulate a positive random variable $\gamma \sim \mathrm{St}(\hat{\alpha},1,\Theta,\delta)$:

Step 1(a). Simulate a uniform random variable: $\upsilon = U\left(\frac{-\Pi}{2},\frac{\Pi}{2}\right)$.

Step 1(b). Independently draw an exponential random variable (ε) with mean 1.

Step 1(c). $\theta_0 = \arctan\left(\tan\left(\frac{\Pi\hat{\alpha}}{2}\right)\right)\Big/\hat{\alpha}$ and compute

$$z = \frac{\sin\hat{\alpha}(\theta_0+\upsilon)}{(\cos\hat{\alpha}\theta_0\cos\upsilon)^{\frac{1}{\hat{\alpha}}}}\left[\frac{\cos(\hat{\alpha}\theta_0+(\hat{\alpha}-1)\upsilon)}{\varepsilon}\right]^{\frac{(1-\hat{\alpha})}{\hat{\alpha}}}.$$

Step 1(d). $\gamma = \Theta z + \delta$.

In order to determine the cyber-insurance premium for a firm with l = 10 and m = 500, the lowest and the highest limits of the number of computers likely affected, we ran the integrated copula-based algorithm given in Section 3.1 for S = 10 000 times. The net insurance premiums per computer related to the three cyber policies with a_1 = 400, a_2 = 125, and a_3 = 300 are given in Table 2.

Table 2. Premium per computer for cyber policy 1, 2, and 3

		Deductible (d)	0	500	1000	1500	2000	2500
Polycy 1		Average premium	$229					
		St. dev. (%)	62%					
Polycy 2		Average premium	$229	$224	$214	$214	$199	$184
		St. dev. (%)	62%	72%	71%	72%	83%	86%
Policy 3 k=25000	a = 5%	Average premium	$210	$204	$196	$195	$180	$168
		St. dev. (%)	57%	62%	65%	65%	70%	77%
	a = 10%	Average premium	$203	$192	$188	$188	$174	$167
		St. dev. (%)	59%	63%	65%	67%	73%	73%
	a = 15%	Average premium	$192	$181	$175	$177	$166	$164
		St. dev. (%)	61%	63%	68%	68%	71%	74%
	a = 20%	Average premium	$185	$178	$168	$160	$157	$151
		St. dev. (%)	61%	64%	67%	70%	72%	75%

Table 2 (cont.). Premium per computer for cyber policy 1, 2, and 3

		Deductible (d)	0	500	1000	1500	2000	2500
Policy 3 k=20000	a = 5%	Average premium	$195	$194	$181	$179	$174	$168
		St. dev. (%)	54%	54%	59%	63%	64%	68%
	a = 10%	Average premium	$185	$183	$174	$168	$169	$162
		St. dev. (%)	56%	58%	63%	65%	66%	70%
	a = 15%	Average premium	$177	$173	$168	$161	$159	$152
		St. dev. (%)	58%	60%	63%	68%	69%	72%
	a = 20%	Average premium	$169	$162	$165	$152	$152	$150
		St. dev. (%)	60%	61%	63%	67%	68%	70%
Policy 3 k=15000	a = 5%	Average premium	$165	$163	$159	$151	$148	$143
		St. dev. (%)	46%	49%	51%	57%	59%	64%
	a = 10%	Average premium	$159	$157	$154	$148	$145	$140
		St. dev. (%)	49%	49%	53%	58%	60%	64%
	a = 15%	Average premium	$158	$153	$149	$147	$143	$139
		St. dev. (%)	50%	53%	56%	58%	61%	65%
	a = 20%	Average premium	$154	$150	$149	$147	$142	$133
		St. dev. (%)	51%	55%	55%	58%	61%	68%
Policy 3 k=10000	a = 5%	Average premium	$123	$120	$117	$115	$112	$107
		St. dev. (%)	35%	39%	42%	46%	51%	55%
	a = 10%	Average premium	$122	$117	$113	$111	$108	$106
		St. dev. (%)	36%	41%	45%	49%	53%	56%
	a = 15%	Average premium	$120	$116	$115	$111	$107	$105
		St. dev. (%)	38%	42%	45%	50%	54%	57%
	a = 20%	Average premium	$116	$114	$112	$108	$106	$103
		St. dev. (%)	40%	43%	46%	51%	55%	58%

For Policy type 1 with no deductible, the average net annual premium per computer is $229. For Policy type 2, with a deductible ranging from $0 to $2500, the average net premium per computer reduces from $229 to $184. The reason for the reduction in average net premium per computer for the higher deductible is because with a higher deductible, a greater portion of risk is borne by the firm.

The deductible plays an important role in managing cyber security risk. For the insurance company, it is a way to lower its risk since the higher is the deductible, the lower their risk of paying out on a claim would be. Typically, cyber-insurance providers impose higher deductibles for firms with greater cyber security risks, for example, firms with consistently lower investment in cyber security, with poor security controls or with inadequate IT staff, among other factors. From a risk management perspective, for a firm, it is important to understand that deductibles affect the premiums. If the deductible is low, the net premium is higher, and vice versa. Firms, therefore, can decide on the deductible as a way to manage their annual cyber-insurance premium costs.

For Policy type 3, with a deductible, co-insurance, and limit, Table 2 provides the annual net premiums per computer for policies with the deductible ranging from $0-$2500, the co-insurance rate ranging from 5%-20%, and the limit ranging from $10000 to $25000. In Policy 3, the insurance company pays nothing when the loss is below the deductible, pays (1 – coinsurance rate) of excess losses over the deductible, but never pays anything over the limit. For a maximum limit of $25000 and a given deductible of $1500, the average net premium per computer reduces from $195 to $160 as the co-insurance rate increases from 5% to 20%. Thus, a higher co-insurance rate means that the firm bears a larger portion of the cyber risk, and hence the reduction in net premiums. The effect of the limit is as follows. It provides an upper bound on the amount of the average net premium paid by the insurance company. Therefore, as expected, when the limit is reduced, assuming the other variables hold constant, the average net premium is reduced. For example, with a deductible of $1000 and a co-insurance rate of 5%, the net premium per computer reduces from $196 to $117 when the limit is reduced from $25000 to $10000. The coinsurance is the amount of firm will bear above the annual deductible. The deductible plus the co-insurance will be the maximum out-of-pocket expense for the firm.

It is common practice to assume independence in pricing premiums. Thus, following Frees and Valdez (1998) we provide the ratios of dependence to independence cyber-insurance premiums to determine the extend of mispricing for a sample set of policies. In particular, we use the independent copula $C(u, v) = uv$ to compute the net insurance premiums per computer

for the independent case. A ratio below 1.0 indicate under valued premiums. As seen from Table 3, for Policy type 1 with no deductible and for Policy type 2 with deductible ranging from 500 to 2500, the undervaluation is the greatest. For Policy type 3, with a limit of $25000 and co-insurance rate ranging from 5% to 20% the premiums are still under valued but not as much due to the cap imposed by the limit. Therefore, in general one can argue that premium pricing errors can be substantial if independence is assumed making the case for considering non-linear dependence in cyber insurance pricing.

Table 3. Ratios of dependence to independence premiums

			Cyber Policy 1, 2, and 3					
Deductible (d)			0	500	1000	1500	2000	2500
Policy 1		Mean	0.218					
Policy 2		Mean	0.218	0.221	0.216	0.217	0.203	0.196
Policy 3 k=25000	a = 5%	Mean	0.861	0.859	0.856	0.884	0.831	0.778
	a = 10%	Mean	0.837	0.820	0.813	0.834	0.790	0.783
	a = 15%	Mean	0.835	0.796	0.804	0.817	0.784	0.798
	a = 20%	Mean	0.832	0.810	0.795	0.783	0.771	0.754

Discussion

This paper develops a cyber-insurance pricing model explicitly considering the three primary risk variables in pricing insurance policies, namely the occurrence of the covered cyber breach event, the time when the insurance is paid, and the amount paid. We illustrated three different types of cyber-insurance policies: a basic policy with no deductible, a policy with a deductible, and a policy with a deductible, coinsurance, and limit. To the best of our knowledge, cyber-insurance literature to date has not explicitly considered these aspects of insurance which are fundamental for appropriately modeling the risks associated with a cyber-insurance contract. Another important aspect of this article is the use of existing data, specifically the publicly available ICSA survey data, for developing and illustrating an actuarial approach to cyber-insurance pricing.

An important point indicated by ICSA data is that the marginal distributions may be non-normal. The dependence between the number of computers affected and the dollar losses is correlated, but not in the typical linear fashion due to the non-normal marginals. In this regards, the usual linear dependence measure, which is the Pearson's product moment correlation, breaks down due to the existence of non-normal or non-elliptical marginals. This problem is severe and especially important as cyber losses tend to be distributed in a Pareto fashion with few viruses or few hacking incidents resulting in large losses and affecting a large number of computers.

In order to appropriately price the risk at firm level, we illustrated the emerging copula approach for modeling dependent risks. This is a more robust technique to obtain the joint loss distributions for two main reasons. First, it allows consideration of non-linear dependencies for correlated risks. Secondly, it permits simulating from a copula model without explicitly having to determine the joint distribution for the two given marginals. Thus, this approach is quite versatile since any type of marginal distributions can be used.

Limitations and avenues for future research. One of the primary constraints in pricing cyber-insurance, as identified by Gordon et al. (2003) and Baer and Parkinson (2007), is the paucity of data on e-crimes and related losses. The data problem is further exacerbated due to the fact that firms do not reveal details concerning security breaches (Geer et al., 2003; Gordon et al., 2003). The paucity of data is a limitation of the proposed copula approach since the approach uses data to determine the appropriate copula and to price the average net annual premiums. While firm-level data may not be available, in our approach we used the publicly available ICSA survey data as a sample of the population of interest and modified it for firm level by using the lower and upper bound number of computers affected as a proxy for firm size. Thus, we used the best available population data to infer about the firm level. As such, one can argue that our model suffers from the data paucity problem. Another limitation is the quality of the data. While we proxy for firm level using the number of computers, it is more likely that π will depend on the number of vulnerabilities, which can be expected to depend on the security precautions taken by the firm (security posture). For example, daily monitoring and updating of virus signatures is less risky than weekly virus signature updates. Similarly, if the firm has resources and can afford hourly monitoring and updates, the firm's IT system is likely more safe than with daily or weekly monitoring. In the event of an incident, the dollar losses will depend on the type of computer affected and the user environment.

Security breaches often pose numerous types of losses: (1) lost productivity; (2) lost revenue; (3) clean up costs; and (4) financial performance impact, to name a few. These costs depend on the type of computers

that are breached. For example, if the virus has crippled the administrative PCs, it will impose employee productivity losses as well as clean up costs related to the attack. If the computer involved is a web-server that is used mostly in e-business types of activity, in addition to clean up costs there is likely to be lost revenues. The current paper thus provides a launching pad for a plethora of research. There is abundant scope for future research, including topics such as the development of cyber-insurance policies based on product diversity (hacking, malware, etc.) and security postures, the integration of additional correlated risks using the process approaches with the actuarial approach, etc.

A slow growth in the cyber insurance industry can be partially attributed to the fact that losses from security breaches are highly correlated because of the Internet. The issue of high correlation among the insured in cyber insurance is contrary to the principle of portfolio balancing in other types of insurance services. This issue, however, is endogenous and cannot be avoided since the Internet infrastructure is a globally shared medium. While our model may be affected to some degree by the above general limitation, the actuarial approach we adopt is data driven and a good set of data would minimize this limitation.

Managerial implications. The actuarial approach, discussed in this article, increases the awareness that it is important to collect data on security breaches for negotiating lower premiums on cyber-insurance products. In addition to demonstrating a sound methodology, this paper illustrates how one can work with the currently available data to get a better idea of the premium pricing. This is much better than using ad-hoc models, which is the current practice. As pointed out by many IT security researchers, one of the main problems that the cyber-insurance industry currently is facing is the disparity in premiums charged for cyber-insurance products and the lack of innovative quantitative models. Most traditional insurance products use historically collected data for determining insurance premiums. In a similar way, the insurance companies can collect the cyber risk related data over time and subsequently modify the models. The proposed approach provides a starting point. It is reasonable to speculate whether cyber-insurance will be helpful or harmful in encouraging IT security due to the issue of moral hazard (Gordon et al., 2003). However, cyber-insurance combined with adequate IT security investments allows firms to better manage cyber risks.

Conclusion

In this paper, we developed a copula-based simulation approach for determining the annual net premiums for three different types of first party damage cyber-insurance policies. This paper makes a significant contribution to the literature in cyber-insurance risk modeling and pricing since it considers the fundamental insurance contract variables in an actuarial approach and illustrates a copula methodology for modeling cyber risks which allows the combining of non-normal risk distributions. Despite the limitations, the proposed copula-based cyber-insurance model makes a significant methodological contribution to the cyber security area as it provides a theoretically sound modeling perspective using an actuarial approach to assess the insurance premiums for cyber-insurance products. The proposed approach is the first in the information security literature to integrate standard elements of insurance risk with the robust copula methodology.

Acknowledgements

We wish to acknowledge valuable comments from participants at the Seventh Annual Forum on Financial Information Systems and Cyber-security: A Public Policy Perspective, January, 2011, Robert H. Smith School, University of Maryland, Maryland, USA, the INFORMS-Canadian Operations Research Society (CORS) Annual Conference, Toronto, Ontario, June, 2009, and the Sixth Workshop on the Economics of Information Security (WEIS), Carnegie-Mellon University, Pittsburgh, Pennsylvania, June, 2007. We also wish to sincerely thank an anonymous referee for helpful comments which considerably helped improve an earlier version of this article. Dr. Hemantha Herath acknowledges financial support from the Social Sciences and Humanities Research Council (SSHRC) of Canada (Grant #: 410-2009-1398). Dr. Tejaswini Herath acknowledges financial support from the Social Sciences and Humanities Research Council (SSHRC) of Canada (Grant #: 410-2010-1848). The usual disclaimer applies.

References

1. Anderson, R. (2001). Why information security is hard: an economic perspective, 17th Annual Computer Security Applications Conference (ACSAC). New Orleans, LA.
2. Baer, W.S., Parkinson, A. (2007). Cyber-insurance in IT security management. IEEE Security and Privacy 5, pp. 50-56.
3. Betterley, R.S. (September 2010). The Betterley report. Understanding the cyber risk insurance and remediation services marketplace: a report on the experiences and opinions of middle market CFOs.
4. Betterley, R.S. (2010). The Betterley report. Cyber risk and privacy market 2010: one of the hottest new P&C products ever attracts numerous insurers, June.
5. Betterly, R.S. (2007). The Betterley report. Cyberrisk market survey 2007.
6. Böhme, R. (2005). Cyberinsurance revisited, Workshop on the Economics of Information Security (WEIS). Harvard University, USA.

7. Böhme, R., Kataria, G. (2006). Models and measures for correlation in cyber-insurance, Workshop on the Economics of Information Security (WEIS). Cambridge University, UK.
8. Bolot, J., Lelarge, M. (2008). Cyber-insurance as an incentive for Internet security, Workshop on the Economics of Information Security (WEIS). Hanover, NH, USA.
9. Cherubini, U., Luciano, E., Vecchiato, W. (2004). Copula methods in finance. West Sussex: John Wiley and Sons.
10. Conrad, J.R. (2005). Analyzing the risks of information security investments with Monte-Carlo simulations, Workshop on the Economics of Information Security (WEIS). Harvard University.
11. Fisher, N.I. (1997). Copulas, in: Kotz, S., Read, C.B., Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences. New York: John Wiley and Sons, pp. 159-163.
12. Frees, E.W., Valdez, E. (1998). Understanding relationships using copulas, *North American Actuarial Journal*, 2, pp. 1-25.
13. Geer, D., Jr., Hoo, K.S., Jaquith, A. (2003). Information security: why the future belongs to the quants, IEEE Security and Privacy 1, pp. 24-32.
14. Genest, C., Remillard, B., and Beaudoin, D. (2009). Goodness-of-tests for copulas: a review and power study. *Insurance: Mathematics and Economics*, 44, pp. 199-213.
15. Genest, C., Rivest, L. (1993). Statistical inference procedures for bivariate Archimedean copulas, *Journal of the American Statistical Association*, 88, pp. 1034-1043.
16. Gordon, L.A., Loeb, M.P., Sohail, T. (2003). A framework for using insurance for cyber risk management, Communications of the ACM 46, pp. 81-85.
17. Herath, H.S.B., Herath, T.C. (2009). Investments in information security: a real options perspective with Bayesian postaudit, *Journal of Management Information Systems*, 25, pp. 337-375.
18. Huard, D., Evin, G., and Favre, A. (2006). Bayesian copula selection, *Computational Statistics and Data Analysis*, 51, pp. 809-822.
19. Klugman, S. (1986). Loss distributions. *Proceedings of Symposia in Applied Mathematics*, 35, pp. 31-55.
20. Longstaff, T.A., Chittister, C., Pethia, R., Haimes, Y.Y. (2000). Are we forgetting the risks of information technology? *IEEE Computer*, 33, pp. 43-51.
21. Majuca, R.P., Yurnick, W., and Kesan, J.P. (2006). The evolution of cyber-insurance, ACM Computing Research Repository (CoRR), Tech report cs.CR/0601020
22. Marshall, A.W., Olkin, I. (1988). Families of multivariate distributions, *Journal of the American Statistical Association*, 83, pp. 834-841.
23. Mukhopadhyay, A., Chatterjee, S., Saha, D., Mahanti, A., Sadhukhan, S.K. (2006). E-Risk management with insurance: a framework using copula aided Bayesian belief networks, 39th Hawaii International Conference on System Sciences. Hawaii.
24. Mukhopadhyay, A., Saha, D., Chakrabarti, B.B., Mahanti, A., Podder, A. (2005). Insurance for cyber-risk: a utility model, Decision 32.
25. Nelsen, R.B. (1995). Copulas characterization, correlation and counterexamples, *Mathematics Magazine*, 68, pp. 193-198.
26. Nelsen, R.B. (1999). An Introduction to Copulas: Springer-Verlag New York, Inc.
27. Nolan, J.P. (2005). Multivariate stable densities and distribution functions general and elliptical case, Department of Math, American University.
28. Oellrich, H. (2003). Cyber-insurance update. CIP Report 2, pp. 9-10.
29. Ogut, H., Menon, N., Raghunathan, S. (2005). Cyber-insurance and IT security investment: impact of independent risk, Workshop on the Economics of Information Security (WEIS). Harvard University, Cambridge, MA.
30. Richardson, R. (2008). 2008 CSI computer crime and security survey, Computer Security Institute.
31. Schweizer, B., Wolff, E.F. (1981). On non-parametric measures of dependence for random variables, *The Annals of Statistics*, 9, pp. 879-885.
32. Sklar, A. (1959). Functions de repartition a n dimensions et leurs merges. Publ. Inst. Statist. Univ. Paris 8, pp. 229-231.
33. Stoneburner, G., Goguen, A., Feringa, A. (2002). Risk management guide for information technology systems, National Institute of Standards and Technology (NIST).

Appendix. An introduction to copula methodology

1. Definition: Sklar (1959) theorem. Let X and Y denote continuous random variables (lower case x, y represent their values) with bivariate distribution function $H(x, y)$ and marginal distribution function $F(x)$ and $G(y)$. Let $F^{-1}(.)$ and $G^{-1}(.)$ be the inverse of F and G. Then for any uniform random variables U and V with values $u, v \in [0, 1]$ (i.e., make the probability transformation of each variate $U = F(X)$ and $V = G(Y)$ to get a new pair of variates $U \sim U(0, 1)$ and $V \sim U(0, 1)$, exist a copula C such that for all $x, y \in R$:

$$H(x,y) = C(F(x),G(y)) = C(u,v). \quad (1)$$

If F and G are continuous, then C is unique. An important feature of copulas is that any choice of marginal distributions can be used. Copulas are constructed based on the assumption that marginal distribution functions are known.

Copulas allow us to study the dependence or association between random variables. There are several ways to measure dependence. The most widely used measures are the Spearman's Rho and Kendall's Tau. Copulas precisely account for

the interdependence of random variables. For example, between two random variables X and Y, the dependence properties of the joint distribution (the manner in which X and Y move together) are precisely captured by the copula for strictly increasing functions of each variable. The two standard non-parametric dependence measures expressed in copula form are as follows:

Kendall's Tau is given by: $\tau = 4\iint_{I^2} C(u,v)\mathrm{d}C(u,v) - 1$, (2)

and Spearman's Rho is given by: $\rho = 12\iint_{I^2} C(u,v)\mathrm{d}u\mathrm{d}v - 3$. (3)

The expressions for Kendall's Tau and Spearman's Rho for some known families of copulas are presented in Section 1. Copulas provide a way to study scale-free measures of dependence. In empirical applications, where data is available, we can use the dependence measure to specify the form of copula. Genest and Rivest (1993) provide a procedure for identifying a copula when bivarate data is available. Once the appropriate copula is identified, it can be used to simulate random outcomes from dependent variables.

2. Review of Gumbel and Clayton copula. In this paper, we review two one-parameter bivariate Archimedean copulas adopted from Frees and Valdez (1998) and Nelsen (1999). Nelsen (1999, pp. 94-97) lists 22 one parameter families. Archimedean copulas have nice properties and are easy to apply. The parameter θ in each case measures the degree of dependence and controls the association between the two variables. When $\theta \to 0$ there is no dependence, and when $\theta \to \infty$ there is perfect dependence. Schweizer and Wolff (1981) show that the dependence parameter θ which characterizes each family of Archimedean copulas can be related to Kendall's Tau. This property can be used to determine empirically the applicable copula form.

1. Clayton copula (1978):

Generator: $\varphi_\theta(t) = (t^{-\theta} - 1)$;

Bivariate copula: $C_\theta(u,v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-\frac{1}{\theta}}$;

Laplace transform: $\varphi(t) = \varphi_\theta^{-1}(t) = (1-t)^{-\frac{1}{\theta}}$;

Kendall's Tau $\tau_\theta = \dfrac{\theta}{\theta+2}$.

2. Gumbel copula (1960):

Generator: $\varphi_\theta(t) = (-\ln(t))^\theta$;

Bivariate copula: $C_\theta(u,v) = \exp\left\{-\left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{\frac{1}{\theta}}\right\}$;

Laplace transform: $\varphi(t) = \varphi_\theta^{-1}(t) = \exp(-t^{\frac{1}{\theta}})$;

Kendall's Tau $\tau_\theta = 1 - \theta^{-1}$.

3. Identifying a copula form. The first step in modeling and simulation is identifying the appropriate copula form. Genest and Rivest (1993) provide the following procedure (fit test) to identify an Archimedean copula. The method assumes that a random sample of bivariate data (X_i, Y_i) for i = 1, 2,…, n is available. Assume that the joint distribution function H has an associated Archimedean copula C_θ; then the fit allows us to select the appropriate generator φ. The procedure involves verifying how close different copulas fit the data by comparing the closeness of the copula (parametric version) with the empirical (non-parametric) version:

Step 1. Estimate the Kendall's correlation using the non-parametric or distribution-free measure:

$$\tau_E = \binom{n}{2}^{-1} \sum_{i<j} \mathrm{Sign}\left[(X_i - X_j)(Y_i - Y)_i\right].$$

Step 2. Identify an intermediate variable $Z_i = F(X_i, Y_i)$ having a distribution function $K(z) = Pr(Z_i \le z)$. Construct an empirical (non-parametric) estimate of this distribution as follows:

$$Z_i = \frac{\text{number}\left\{(X_i, Y_j) \text{ such that } X_j < X_i \text{ and } Y_j < Y_i\right\}}{n-1}.$$

The empirical version of the distribution function $K(z)$ is $K_E(z)$ = proportion of $Z_i \le z$.

Step 3. Construct the parametric estimate of $K(z)$. The relationship between this distribution function and the generator

is given by $K(z) = z - \frac{\varphi(z)}{\varphi'(z)}$, where $\varphi'(z)$ is the derivative of the generator and $0 \le z \le 1$. The following is a specific form of $K(z)$ for the two Archimedean copulas reviewed in this paper.

1. Clayton copula $$K(z) = \frac{z(1+\theta - z^{\theta})}{\theta}. \quad (4)$$

2. Gumbel copula $$K(z) = \frac{z(\theta - \ln z)}{\theta}. \quad (5)$$

Repeat Step 3 for several different families of copulas, i.e. several choices of $\varphi(.)$. By visually examining the graph of $K(z)$ versus z, or using statistical measures such as minimum square error analysis, one can choose the *best* copula. This copula can then be used in modeling dependencies and simulation.