Empirical Analysis of the Effects of Cyber Security Incidents

Ginger Davis, Alfredo Garcia and Weide Zhang

University of Virginia

June 27, 2007

Abstract

We undertake two types of complementary empirical analysis of the effects of cyber security incidents aimed at enabling a better understanding of the connection between cyber security risks and mitigation strategies. In our first type of analysis, we provide an empirical characterization of the reporting of cyber security incidents in specialized press and trade journals. We find that the likelihood of a cyber security incident being reported in specialized press increases with the total number of affected customers, the company breached being publicly traded and whether or not commercially sensitive information was lost. Armed with this characterization, we undertake an analysis of the time series associated with web traffic for a representative set of on-line businesses that have suffered widely reported cyber security incidents. We test for structural changes in these time series resulting from these cyber security incidents. Our results consistently indicate that cyber security incidents do not affect the structure of web traffic for the set of on-line businesses studied. We discuss various public policy considerations stemming from our analysis.

1 Introduction

According to Kart et al.(2006), the percentage of planned IT budgets assigned to cyber security in North America and Europe has varied from 7.93% in 2004 to 8.92% in 2005 and finally 7.75% in 2006. Whether these figures reflect an underlying rationale for cyber security risk assessment and mitigation is still an open question. Recent research estimating the effects of cyber security incidents (see Campbell et al (2003) and Goldfarb (2006)) has examined the effects of information security breaches on the stock market value of corporations and market share loss, respectively. For a wide variety of cyber security incidents, these studies have shown that the effects of cyber security incidents are short-lived. In the case of stock market valuations, confidence in the underlying economic fundamentals determining the value of the stock price is restored after a transient period of "noisy" trading following the news of a cyber security incident. Similarly, the market shares for Internet portals as measured by traffic levels quickly return to "normalcy" after denial of service attacks. While Andricic and Horowitz (2006) showed that cyber security incidents with long lasting effects (e.g. intellectual property theft) induce significant aggregate costs for the economy, the results reported by Campbell et al (2003) and Goldfarb (2006)) seem to indicate that there is no clear relationship between cyber security risks and the associated mitigation strategies pursued by individual corporations.

In this paper, we undertake two types of empirical analysis of the effects of cyber security incidents aimed at enabling a better understanding of the connection between cyber security risks and mitigation strategies.

In our first type of analysis, we provide an empirical characterization of the reporting of cyber security incidents. Given that the measurable effects of cyber security incidents seem to be either short-lived or negligible, the case for investing in cyber security could be argued on the grounds of adverse effects to a company's reputation. Granted, this is an "intangible" asset but one that may ultimately drive the final decision making for cyber security investments. A company's reputation is severely affected when a cyber security incident is widely reported in different media outlets. We find that the likelihood of a cyber security incident being reported in specialized press increases with the total number of affected customers, the company breached being publicly traded and whether or not commercially sensitive information was lost. A complete characterization of the reporting of cyber security incidents may prove to be an important first step in understanding why and to what scale different types of companies invest in cyber security, but this exercise is beyond the scope of this paper.

Instead, armed with the results of our first analysis, we focus our efforts on analyzing the effects that cyber security incidents may have on companies that predominantly conduct their businesses in an on-line fashion or alteratively, that provide on-line additional services to off-line costumers. The premise here is that cyber security incidents may prompt (security conscious) on-line customers to opt out and conduct their business elsewhere or at the very least, refrain from accessing on-line services. For companies relying almost exclusively on on-line channels, this presents an important business risk. In our analysis, we use time series associated with web traffic for a representative set of on-line businesses that have suffered widely reported cyber security incidents. We test for structural changes in these time series resulting from these cyber security incidents. Our results consistently indicate that cyber security incidents *do not* affect the structure of web traffic for the set of on-line businesses studied. There are potentially two explanations for this result. In the absence of reputation mechanisms (such as the ones implemented by Amazon and Ebay), customers engaged in infrequent transactions may simply remain unaware of cyber security incidents affecting the on-line portals they deal with. Alternatively, in the companies involved in sustained relationships (e.g. banks and other financial services) signficant "switching" costs may prevent a customer from changing providers even if he or she is aware of a potential cybersecurity risk exposure. Recent studies have provided significant empirical support for the existence of “switching” costs associated with information technologies (see Chen and Hitt (2002)).

Two types of public policy considerations stem from our analysis. Limited customer responsiveness to potential cybersecurity risks may be explained by a sort of “prisoner’s dilemma”. On the aggregate customers are better off punishing companies for negligent risk mitigation and therefore inducing more secure transactions. Individually, switching may prove too costly even when the costs of a cyber security breach are accounted for. The characterization of the likelihood that a cybersecurity incident affects a company’s reputation can be leveraged to construct simple on-line reputation systems which keep track of cybersecurity reports. This may enhance customers’ ability to select the more cybersecure firms for their transactions.

The structure of this paper is as follows. In section 2 we discuss our sample of cyber security incidents. In section 3, we present an empirical characterization of incident reporting. Section 4 contains our analysis on the structural effects of these incidents on web traffic. Finally, in section 5, we offer our conclusions.

2 A Sample of Cyber security Incidents

We are primarily interested in security breaches of businesses performed by outside intruders which have resulted in the breach of sensitive data. Since hacking activities are directly related to the reliability and trustworthiness of online portals where transactions are carried out and sensitive information is stored, we investigate how the breach would influence the number of visits to the website (number of transactions performed as well as other activities performed).

The security breach events are obtained from `www.attrition.org`. We obtain data concerning all reported breach events which took place between 2000 and 2007, including the following characteristics of the events: type of company, type of breach, data type breached, and number of people affected by the event. There were a total of 197 security breach events during this time frame. 76 of these security breaches resulted from online hacking or fraud activities to businesses. Our analysis concerning the effects of cyber security incidents on companies that predominantly conduct their businesses in an online fashion will pertain to these specific companies and events. Our analysis concerning the characterization of the reporting of cyber security incidents will consider all reported breach events which took place between 2000 and 2007.

3 A Characterization of Cyber security Incident Reporting

In order to characterize the reporting of cyber security events, we first must classify what we consider to be a report. Because of the difficulty in collecting data from the traditional media such as newspapers or television programs, we use online news media as our primary sources. We consider the security breach reported if a description of the event appears in one of the news media sources we selected. The three categories of news media we consider as relevant include general, computer, and business information technology. In the general news media source category, we selected the top 7 sources ranked according to daily traffic volume as listed on Alexa.com. Table 1 lists these media sources. We excluded 3 of the top 10 ranked sources due to the fact that they were either portal news site (Yahoo news and Google news) or pure weather sites. Portal news sites were excluded because they are the news media of news media and thus irrelevant for determining which breach interests should receive news coverage. Additionally, we included business and finance media sites which were not rated as one of the top 10 sources based on traffic volume due to their relevance to security breaches of businesses. These are also included

News Media	URL
Yahoo News	news.yahoo.com
CNN	cnn.com
Google News	news.google.com
MSN News	msnbc.msn.com
NY Times	nytimes.com
BBC	bbc.com
Washington Post	washingtonpost.com
Business Week (Added)	businessweek.com
Reuters (Added)	reuters.com
Bloomberg (Added)	bloomberg.com
Forbes (Added)	forbes.com

Table 1: Top Ranked News Media (General)

News Media	URL
Slashdot	slashdot.org
PC World	pcworld.com
The Register	theregister.co.uk
Arstechnica	arstechnica.com
PC Mag	pcmag.com
The Inquirer	theinquirer.net
Computer World	computerworld.com
Information Week	informationweek.com
Extreme Tech	extremetech.com
Wired	wired.com

Table 2: Top Ranked News Media (Computer)

in Table 1. Additional news media sources selected are those top ranked news media for the general computer community (Table 2), business information technology (Table 3), and computer security (Table 4). Since the purpose of the news media is to propagate happenings within communities in which they focus, we assume that IT managers, who are primarily the decision makers in future security investment, might care about what is reported and will take corresponding actions if reported. The security breach is a good example of such an event. Although it must be an important investigation to study how IT managers react to the news coverage about security breaches which happened in their company, we will currently focus on what kind of security breach would more likely be reported and possibly explore the reasoning behind such likelihood.

News Media	URL
CNET	news.com.com
ZDNET	zdnet.com
Tech Republic	techrepublic.com.com
Internet	internet.com

Table 3: Top Ranked News Media (Business Info Tech)

News Media	URL
Security Focus	securityfocus.com
Zone-h	zone-h.org
Securiteam.com	securiteam.com
Tech Target	searchsecurity.techtarget.com
CERT	cert.org

Table 4: Top Ranked News Media (Computer Security)

For each security breach event, we determined whether it was reported by any of the sites above. We then create a variable, $report_i$, indicating whether or not breach event i was reported. If it was reported, then $report_i = 1$, and if it was not reported, $report_i = 0$. We use Google search extensively to accomplish this task. Different key word combinations are tried and all pages returned by Google search are explored. 135 of the 197 events are reported by at least one of the media sources selected.

3.1 Methodology

Empirical analysis regarding the characterization of the reporting of cyber security incidents is performed using a statistical classification methodology known as random forests (Breiman (2001)). We aim to characterize the decision making process of cyber security incident reporting based on incident characteristics such as company type, breach type, data type, and number of people affected. The basis of the random forest predictor is a tree-structured classifier (Breiman, Friedman, Olshen, and Stone (1998)). Tree-structured classifiers are constructed by first splitting the data into two descendant subsets based on some rule regarding one of the explanatory variables. This process is performed repeatedly by further partitioning each subset until no further partitioning is useful. The random forest predictor uses multiple tree-structured classifiers, each of which casts one vote for the prediction, and chooses the most popular prediction. In our case, we are classifying a breach event as one which is or is not reported by the media. We predict, or classify, each event according to some rules regarding the explanatory variables (company type, breach type, data type, and number of people affected.)

3.2 Results: Classification Trees

We use the classification tree methodology described above as a way of capturing the relationship between security breach events and the news reporting coverage of those events. The news reporting coverage is a binary variable indicating whether or not the security breach has been reported by at least one of the selected news media. As described above, the classification tree methodology is a data mining approach which classifies linear or nonlinear clusters within a data set given its observations. The classification tree is generated using the back-pruning method where a full tree (without any limit on its maximum depth) is produced according to classification accuracy and then pruned in order to scale the tree down to a reasonable size based on both accuracy and size. For a detailed description of classification tree models, refer to (Breiman (1998)). In our example, we use the news coverage observation as the dependent variable and the properties of a security breach as variables possibly used for classification. We also added another independent variable, “publicly traded,” which is an indicator of whether the associated company is publicly traded on the stock market. Our rationale for adding this variable is that the media might be more likely to report publicly traded companies.

The classification tree produces a tree model where each non-terminal node represents a splitting rule based on values of certain independent variable(s) and sends it to its left child or right child for further classification. The terminal node of the classification tree simply gives the binary news coverage result if the properties of a security breach event satisfy the splitting conditions along the path from tree root to this node.

Another advantage of the classification tree model is that it can handle cases with missing values on variables by using surrogate splits. It is also more robust than other linear classification models or regression models. When constructing the tree model, the trade off is made between the classification accuracy and the maximum tree size. Thus, the recommended way in (Breiman, 1998) is to first construct a full tree based on its accuracy metric and then prune back the tree taking both accuracy and tree size into consideration.

For the analysis of our data, most variables describing the properties of a security breach are categorical. Since there are multiple values for the data types associated with a certain security breach, we give a ranking on the data types breached and choose the one with highest rank to be the value for the breach data type variable used in constructing the classification models. The ranking is ACC > CCN > SSN > FIN > MED > PPN > EMA > DOB > MISC where ACC = Account Number, CCN = Credit Card Number, SSN = Social Security Number, FIN = Financial Information, MED = Medical Information, PPN = Private Personal Info, EMA = Email Address, DOB = Date of Birth, and MISC = Miscellaneous. Also, there are around 25% missing data in the variable denoting the total affected number of people by the breach event. Therefore, we generate two classification tree models–with and without using the variable denoting the total affected number of people by the breach event. Finally, for comparison purposes, we also used logistic regression on the same data. The results of the logistic regression are included in Appendix B.

We show the classification tree results in Figures 1 and 2. The figures include the classification rule for each branch of the tree along with the terminal nodes. Note that Figure 1 does not include the variable denoting the total affected number of people by the breach event while Figure 2 does include this variable. Tables 8-9 provide detailed summarized results and are included in Appendix A.

Each node in every classification tree displays the following information: 1) predicted classification result (0 or 1), where "0" indicates the event was not reported and "1" indicates the event was reported, 2) total number of non-reported events in the sample reaching the node (left of the slash), and 3) total number of reported events in the sample reaching the node (right of the slash). Usually the terminal nodes (shown as rectangles) are used for classification and the role of non-terminal nodes is to split the events reaching different child nodes based on the splitting rule. For example, as you can see from Figure 1, at root, there are 62 events in the sample which are not reported while the remaining are reported. The sum of these partitioned events is the total sample size (197). The first splitting rule is whether or not a breach affected company is publicly traded, which divides the sample at root into 2 subgroups. An event case goes to the left child if the affected company is publicly traded or goes to the right child otherwise. Events reaching the left child total 100, 45 of which are not reported and 55 of which are reported. Similarly, events reaching the right child total 97, 17 of which are not reported and 80 of which are reported. For prediction purposes, an event case will be forwarded from the root to a terminal node based on the characteristics of the event and the splitting rules. Splitting rules are based on whether a company is publicly traded, breach type, business type, data type, and the total number of affected people. They are actually selected by the classification tree construction algorithm which maximizes the information gain. The breach type is the nature of a security breach which includes events including hacking activities, lost documents, stolen laptops, fraud, etc. The business type is the type of industry of the affected company. The data type is the kind of data breached including social security number, credit card information, account information, etc. Further information can be found at `www.attrition.org`. When interpreting the classification trees, note the following abbreviations. When splitting the observations based on business subtypes, MED = Medical, FIN = Financial, TECH = Technical, RETAIL = Retail, DATA = Data Broker, MEDIA = Media, IND = Industry, NFP = Not For Profit, ORG = Organization, INS = Insurance, and CITY = City. When splitting observations based on breach type, HACK = hacking incident, WEB = breach occurred over the web, LOST = lost disk drive, tape, document, media, or laptop, STOLEN = stolen disk drive, tape, document, media, or laptop, DISP = disposal of disk drive, tape, document, media, or laptop, FRAUD = fraudulent event, SNAIL = breach occurred by snail mail.

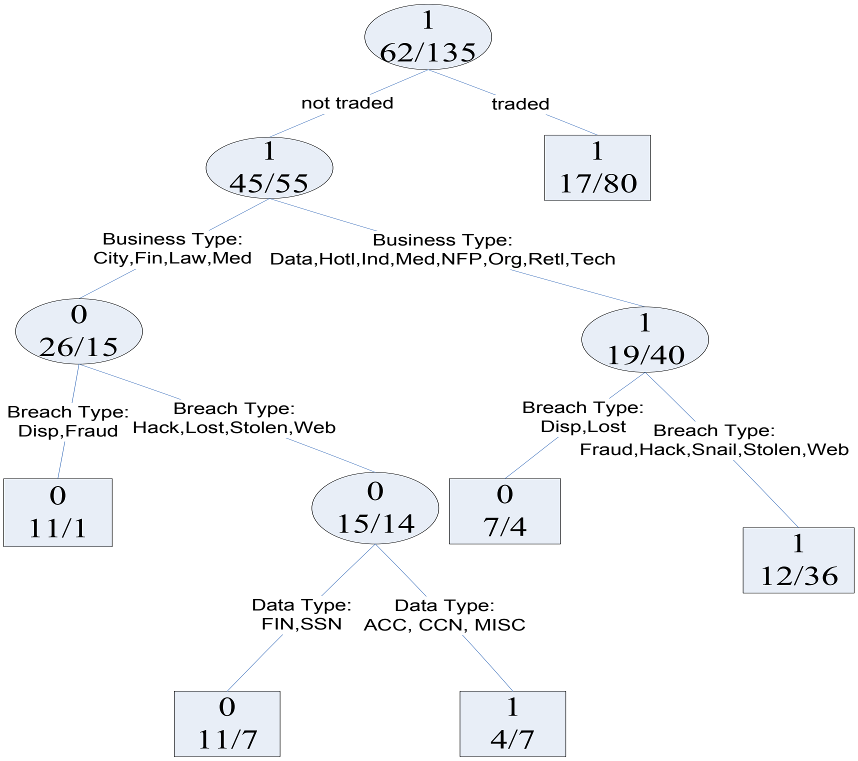

Figure 1: Classification Tree Without using Total Number of Affected People by the Security Breach

We can use these two classification tree models as a reference for finding: 1) which

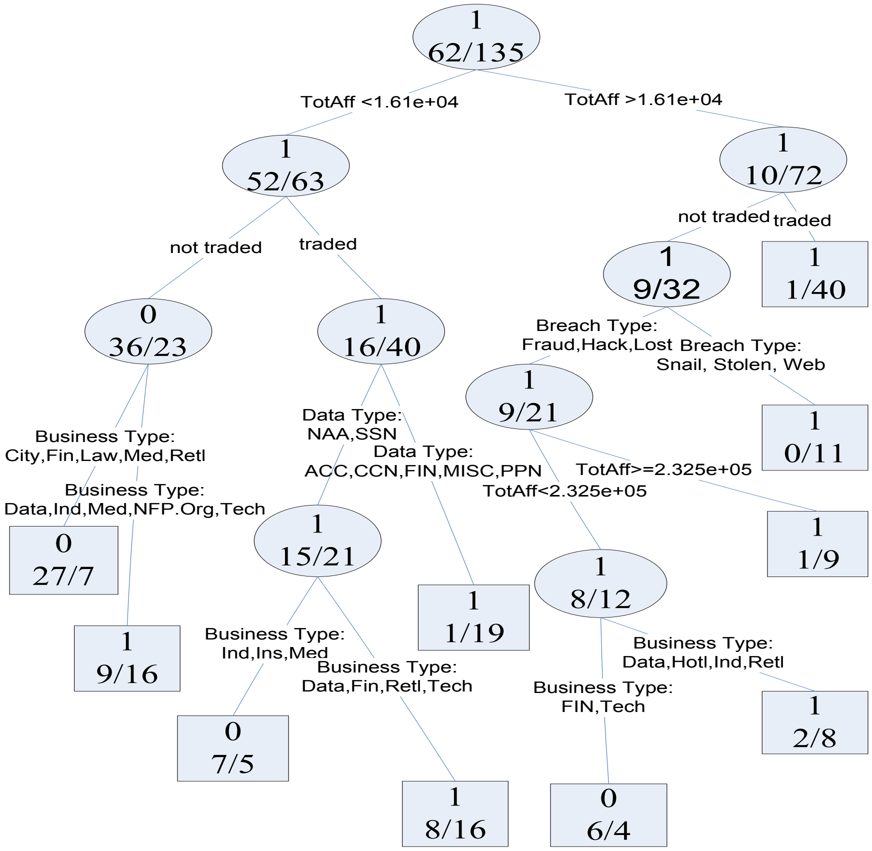

Figure 2: Classification Tree With using Total Number of Affected People by the Security Breach

variables are used for generating splitting rules and 2) which properties certain breach events have that lead to increased likelihood of news media reporting. For example, in Figure 1, if an affected company is publicly traded, then 80 out of 97 are actually reported. In Figure 2, if the total affected number is larger than 1.61×10^4 and the affected company is publicly traded, then 40 out of 41 are reported. When interpreting these results, it is also important to note that not all of the classification rules have low classification error. By combining the results from both models, we can get some overlapping properties from both models which lead to the same classification result with lower error rate. For example, if, for a breach event, the associated company is not publicly traded and its business type is Data, Hotl, Ind or Retl and $1.61 \times 10^4 \leq$total affected number $\leq 2.325 \times 10^5$, then the probability for correct classification for a randomly chosen tree would be $\frac{0.75+0.8}{2} = 0.775$. We take the sample correct classification rate as the probability of correct classification on the terminal node 2 (Table 1) and terminal node 4 (Table 2).

4 Testing the Effects of Cyber security Incidents on Web Traffic

Markets in which trading takes place "off-line" rely significantly on the trust created by repeated interaction between buyers and sellers. Markets that operate mostly through on-line channels tend to be more anonymous. This may explain the emergence of reputation mechanisms. For example, Amazon and E-bay have a rating system through which customers provide feedback on the overall quality of the transaction. This helps new customers better assess the available options for undertaking new transactions. Livingston (2005) provides empirical evidence for significant returns to sellers' reputation on E-bay.

Another dimension of quality of service for on-line channels pertains to the level of cyber security implicit in the transaction. To the extent that a firm's on-line channel is subject to cyber security incidents, customers may prefer other more secure channels. Recently, companies whose main retail channel is "off-line" have also started providing on-line additional services to their regular costumers. In both cases, our working hypothesis is that cyber security incidents may prompt (security conscious) on-line customers to opt out and conduct their business elsewhere or at the very least, refrain from accessing on-line services. In this sense, cyber security incidents may significantly alter a firm's overall business through a reputation effect similar to one reported by Livingston (2005). Specifically, we undertake a structural test of the time series associated with on-line portal traffic. A test for structural change is an econometric test to determine whether the coefficients in a regression model are the same in separate subsamples (Chow (1960)). Since our interest is to test the effects of cyber security incidents on web traffic our choice of subsamples comes from different time periods: before and after the event.

To illustrate consider time series associated with the the web traffic for *Choicepoint* (see Figure 3). If we run simple linear regressions for time windows pre- and post- the reporting of the cybersecurity breach it appears as if there is a significant change in the structure of web traffic. This would lead to the conclusion that the reporting of the cybersecurity incident did indeed affect negatively web traffic. However, as we shall see in what follows this conclusion is likely to be incorrect due to significant serial correlation and volatility inherent to the series.

Let y_t denote the daily traffic volume of a website for a particular company on day t, for $t = 1, \ldots, T$. We model y_t as a segmented deterministically trending and heteroskedastic autoregressive model as in (Wang and Zivot (2006)):

$$y_t = a_t + b_t t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots \phi_r y_{t-r} + s_t u_t, \tag{1}$$

for $t = 1, \ldots, T$, where $u_t | \Omega_t \sim$ *i.i.d.* $N(0, 1)$ and Ω_t denotes the information set at time t. Additional assumptions are that the parameters a_t, b_t, and s_t are subject to $m < T$ structural changes, m initially known, with break dates $k_1, \ldots, k_m$, $1 < k_1 < k_2 < \cdots < k_m < T$, so that the observations can be separated into $m + 1$ regimes. Let $\boldsymbol{k} = (k_1, k_2, \ldots, k_m)'$ denote the vector of break dates. For each regime i $(i = 1, 2, \ldots, m + 1)$, the parameters a_t, b_t, and s_t are given by

$$a_t = \alpha_i, b_t = \beta_i, s_t = \sigma_i \geq 0 \tag{2}$$

for $k_{i-1} < k_i$ with $k_0 = 1$ and $k_{m+1} = T + 1$. This model is termed a partial structural change model since the autoregressive parameters are assumed to be constant across regimes.

The number of lags, r, is selected by choosing the last partial autocorrelation which is statistically different from zero. The model is estimating using the Bayesian MCMC method of Gibbs sampling. For details, we refer to (Wang and Zivot (2006)). For the regression parameters $B = \{a_1, \dots a_{m+1}, b_1, \dots b_{m+1}, \phi_1, \dots \phi_r\}$ we use the natural conjugate multivariate normal prior: $N(0, \Sigma_B)$, where Σ_B is a diagonal matrix with 1000 on the diagonal. For each σ_i, we use the natural conjugate inverted gamma prior $IG\,(v_0 = 2.001, \lambda_0 = 0.001)$. We generate 2000 iterations of the Gibbs sampler and use the first 500 iterations as burn-in. Model selection is performed by estimating several models, each with a different number of break points, $m = (0, 1, 2, 3)$. The final model selected will be the one with the lowest BIC value, since this criterion provides a consistent estimate for the true number of break points in the model after the conditional posterior distribution is collected for each set of parameters (see Yao (1988)). The way we obtain this estimate is through replication. We use 100 repetitions of each model, and the final model selected is the one whose BIC is lowest in the highest proportion of the replicated runs. The other estimators such as a_t, b_t and σ_t are calculated as the mean of their conditional distribution, and the break dates (if they exist) are chosen as the ones with the maximum number of occurrences in the selected model.

4.1 Traffic Volume Data

Daily traffic volume data were obtained from `www.Alexa.com`, at which further discussion can be found. It is typically called the daily reach of that web site. This data consists of the number of unique visitors to that site per million of total internet users. The data is gathered for a period of 5 months (the month when a security breach is reported plus 2 months worth of data before and after the reported breach). We are interested in examining these data as they relate to the 31 security breaches mentioned above, but we were only able to obtain data for 16 of these security breaches due to data incompleteness.

	Constant Term	Trending Term
Before Breach Date	45.563941 4.475568	0.207776 0.141589
After Breach Date	52.322693 3.985507	−0.137645 0.036019

Table 5: Regression Coefficient for ChoicePoint Series(Before and After Breach Report Date)

A time series plot of the web traffic volume data for ChoicePoint.com from 7/2005 to 3/2006 is shown below in Figure 3. ChoicePoint has mean reach 42.32 and standard deviation 14.93. ChoicePoint provides restricted access to online data services serving the data needs of businesses of all sizes, as well as federal, state and local government agencies. A security breach is reported to the public on 9/16/2005. The pink dot in Figure 3 indicates the breach report date. Also, the dashed lines represent 2 regression lines fitting the two subsequences of the time series (one before and one after the breach report date). The regressors are just a constant and a trending term. The coefficients are displayed in Table 6. They are significantly different leading to the belief that there is likely a structural break. However, by using the structural break detection methodology and model selection based on the Bayesian Information Criteria, we conclude that there is no structural break for traffic volume data during that period.

4.2 Results: Structural Break Detection

The structural break detection method described above is applied to the time series of daily web traffic volume for online sites with security breaches reported to the public. Let y_t as in Equation 1 denote web traffic volume on day t. We are interested in detecting a structural break around the time of a reported security breach. The Gibbs sampling algorithm presented above is employed for the estimation of the model. Sixteen daily traffic volume time series for different online sites are tested using the Gibbs sampler method. Table 6 shows the 16 companies under investigation along with some characteristics of the company and security breach: the dates for the daily traffic volume data which were

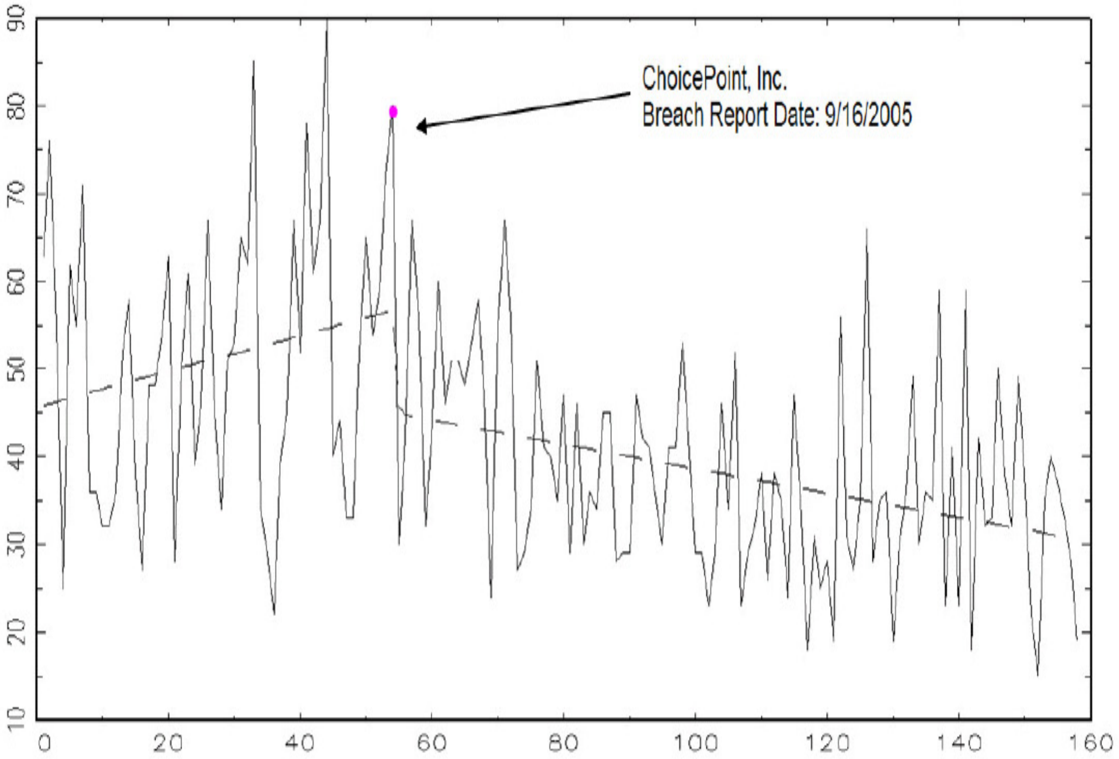

Figure 3: Traffic Volume Data for ChoicePoint from 7/2005 - 3/2006

collected, business type, what data are breached, the type of security breach, what date the event is publicly reported, total affected number of consumers, whether the company is publicly traded, the number structural breaks detected according to the estimated model, and how strong the result is in terms of the BIC dominating proportion. As you can see from Table 6, the companies under study are mostly in the financial and retailing (FIN and RETAIL) sectors, but also include one online data provider (DATA) and one media provider (MEDIA). We think sites which base most of its service purely online will be more affected by security breach events. Sites of this kind in our sample are ChoicePoint, Equifax, Lexis Nexis, and TransUnion. For each event, there are also different kinds of data being breached such as social security number (SSN), credit card number (CCN), names and address (NAA) and account information (ACC). As shown in Table 6, there are no structural breaks for 15 of the 16 events. The result is strong but a little bit surprising. We expected to see some difference in detecting structural breaks between companies providing pure online services and those offering goods and services both online and offline. However, not only are there no structural breaks detected for pure online service providers (ChoicePoint, Equifax, Lexis Nexis and TransUnion), but also the conditional distributions drawn from the Gibbs Sampler is more informative (lower average standard deviation). The estimates for the parameters are calculated as the mean of the Gibbs Sampler data and the break point is taken to be the mode of the distribution. DSW Shoes has 2 structural breaks detected directly after the breach has been publicly reported. DSW Shoes is an online shoes retailing company selling shoes both online and off line. The trend parameter is insignificant before the first structural break, positive and significant between the first and second structural break, and insignificant after the second structural break. The mean volatility also increases short-term between the first and second break points. This finding is the only evidence that security breaches have an effect on web traffic. The effect is a temporary increase in traffic trend and traffic volatility.

Company	Traffic Volume Data	Business Type	Breach Type	Data Breached	Breach Report Data	Number Affected	Publicly Traded	Struct. Breaks	Prop. of BIC
Playboy	9/2001 - 1/2002	MEDIA	Hack	CCN	11/20/2001	unknown	Yes	0	94%
BJ's Wholesale Club	1/2004 - 5/2004	RETAIL	Hack	CCN	3/19/2004	unknown	Yes	0	99%
Lexis Nexis	2/2005 - 6/2005	DATA	Hack	SSN, NAA	4/12/2005	310000	No	0	94%
Equifax Canada Inc	4/2005 - 8/2005	FIN	Hack	NAA, ACC	6/17/2005	605	Yes	0	85%
DSW Shoes	2/2005 - 6/2006	RETAIL	Hack	CNN	4/18/2005	1496000	Yes	2	99%
Scottrade	9/2005 - 1/2006	FIN	Hack	ACC, NAA	11/26/2005	140000	No	0	97%
Sam's Club	10/2005 - 2/2006	RETAIL	Hack	CCN	12/12/2005	unknown	No	0	95%
Cooks Illustrated	11/2005 - 3/2006	RETAIL	Hack	CCN	1/30/2006	unknown	No	0	100%
Ross-Simons	2/2006 - 6/2006	RETAIL	Hack	SSN,CCN	4/12/2006	32000	No	0	98%
Vystar Credit Union	1/2006 - 9/2006	FIN	Hack	SSN,NAA	5/31/2006	34000	No	0	89%
MoneyGram	11/2006 - 3/2007	FIN	Hack	NAA,ACC	1/12/2007	79000	Yes	0	100%
Choice Point	7/2005 - 3/2006	DATA	Fraud	SSN,NAA	9/16/2005	163,000	Yes	0	98%
Polo Ralph Lauren	2/2005 - 6/2005	RETAIL	Fraud	CCN	4/15/2005	180000	Yes	0	96%
Bank of America	3/2005 - 7/2005	FIN	Fraud	ACC	5/23/2005	676000	Yes	0	91%
HSBC	4/2006 - 8/2006	FIN	Fraud	ACC	6/27/2006	20	Yes	0	78%
TransUnion	9/2006 - 1/2007	FIN	Fraud	SSN, FIN	11/30/2006	1700	No	0	56%

Table 6: Breach Event Descriptions. Traffic volume data = Dates of traffic volume data analyzed; Business type = business type, where FIN = financial, RETAIL = retailing, DATA = online data provider, and MEDIA = media provider; Breach type: HACK = hacking event, FRAUD = fraudulent event; Data breached: SSN = social security number, CCN = credit card number, NAA = names and address, ACC = account information; Breach report date = date the breach was publicly reported; Number affected = total number of people affected by the breach event; Publicly traded = indicator of company being traded publicly; Struc. breaks = number of structural breaks estimated in the model; Prop. of BIC = proportion of models which had the lowest BIC estimate for the number of structural breaks indicated

5 Conclusions

In this paper, we have presented the results of two complementary lines of empirical analysis for assessing the effects of cyber security incidents.

In our first type of analysis, we provide an empirical characterization of the reporting of cyber security incidents. Given that the measurable effects of cyber security incidents seem to be either short-lived or negligible, the case for investing in cyber security could be argued on the grounds of adverse effects to a company's reputation. Granted, this is an "intangible" asset but one that may ultimately drive the final decision making for cyber security investments. A company's reputation is severely affected when a cyber security incident is widely reported in different media outlets. A complete characterization of the reporting of cyber security incidents constitutes an important first step in understanding why and to what scale different types of companies invest in cyber security. We find that the likelihood of a cyber security incident being reported in specialized press increases with the total number of affected customers, the company breached being publicly traded and whether or not commercially sensitive information was lost.

In our second type of analysis, we focus on cyber security incidents affecting companies that predominantly conduct their businesses in an on-line fashion. Using time series associated with web traffic for a representative set of on-line businesses that have suffered widely reported cyber security incidents, we test for structural changes resulting from these cyber security incidents. Our results consistently indicate that cyber security incidents do not affect the structure of web traffic for the set of on-line businesses studied. There are potentially two explanations for this result. In the absence of reputation mechanisms (such as the ones implemented by Amazon and Ebay), customers engaged in infrequent on-line transactions may simply remain unaware of cyber security incidents affecting the on-line portals of their choosing. Alternatively, potential "switching" costs for customers engaged in long-term relationships (i.e. banks and other financial services) may deter them from changing providers even if they are fully aware of potential cybersecurity risk exposures.

Two types of public policy considerations stem from our analysis. Limited customer responsiveness to potential cybersecurity risks may suggest the ocurrence of a sort of "prisoner's dilemma". On the aggregate customers are better off punishing companies for negligent risk mitigation and therefore inducing more secure transactions. Individually, switching may prove too costly even when the costs of a cyber security breach are accounted for. A better understanding of the likelihood a cybersecurity incident is reported is important in that it indicates which types of companies may be more senstitive to cybersecurity concerns. Simple reputation systems keeping track of cybersecurity reports can be developed to help customers choose the more cybersecure firms.

Appendix A: Classification Tree Results

Tables 8-9 provide detailed results of the classification rules for the classification trees. They include the characteristics on certain breach events and whether, with some probabilistic accuracy, it would be reported by the news media listed above. The first column shows the assigned index number for the terminal node on the tree. The second column describes the properties of certain breach events while the third column lists the classification result if a specific case possesses these properties. The rightmost column provides the classification error rate for cases having the described properties in our sample. The numerator is the number in the sample not correctly classified by the terminal node and the denominator is the total number in the sample reaching the node.

Appendix B: Logistic Regression Results

We used logistic regression initially for predicting the probability of media reporting for the security breaches. We include them here for comparison with the results from the classification tree models. We used the forward, backward, and stepwise model selection criteria to select variables for inclusion in the logistic regression. These results are included in Table 10.

According to Table 10, we choose the variables "total affected number by breach" and "publicly traded" as the two independent variables in the logistic regression. The cases with a missing value for the total affected number are ignored, resulting in around 75% of the 197 cases being used for estimating the logistic regression model. Thus, we have a total of 134 observations: 93 of which are security breaches with reporting by at least one of the selected news media and 41 of which are not reported by any of the selected media. After examining estimation results included in Tables 11-13, we find that the variables included in the model are statistically different from 0. The reason for the coefficient estimate for “publicly traded” variable being than 0 is that “publicly traded” is used as the reference (when equal to 0), then the coefficient reflects the effect when the associated company is not publicly traded. It can also be seen from the odds ratio value of the variables (Traded = 0 vs Traded = 1) as 0.208. If a company is publicly traded, it is more likely to be reported. We thus see the importance of these variables in explaining whether or not a security breach will be reported.

The following classification tables list the error rates for a given threshold value for the logistic regression prediction, that is, how many news reported breach events are correctly classified and how many are not, along with their corresponding false positive and false negative rates. The false positive rate is calculated as incorrect positive/(correct positive+incorrect positive). The false negative rate is calculated accordingly.

The threshold value is compared against the value produced by the logistic regression and the logistic model classifies it as “positive” (reported) if its value is above the threshold and as “negative” (not reported) otherwise. Table 14 is an excerpt of the classification table. These values are included since the false positive and negative rates are comparatively lower than cases with other probability level values.

Both the classification tree and logistic regression models could be used for classification and predicting. The advantage of using logistic regression is that it can quantify the effect of variables on predicting the dependent variable. However, it might suffer from the quasi-complete separation problem and also will not work well if the dependent variable is not linearly related with the independent variables. The reason we focused more on the classification tree model results is because they are more robust and flexible. It offers us more information on the classification effect for all of the variables in our data set and it handles cases with missing values well.

References

[1] Andrijcic, E. and Horowitz B. (2006) “A Macro-Economic Framework for Evaluation of Cyber Security Risks Related to Protection of Intellectual Property”, Risk Analysis, Vol. 26 No. 4 pp 907-923.

[2] Breiman, L., Friedman J., Olshen R., Stone C. (1998), *Classification and Regression Trees*, Chapman & Hall, New York, pp. 20-22.

[3] Breiman, L. (2001) “Random Forests,” Machine Learning, Vol. 45 pp. 5-32.

[4] Campbell, K., Gordon L., Loeb M., and Zhou L. (2003) “The Economic Cost of Publicly Announced Information Security Breaches: Empirical Evidence from the Stock Market,” Journal of Computer Security, Vol. 11, No. 3, pp. 431-448.

[5] Chen, P. and Hitt L. (2002) “Measuring Switching Costs and Their Determinants in Internet Enabled Businesses: A Study of the Online Brokerage Industry” Information Systems Research Vol 13. No 3. pp 255-276.

[6] Chow, Gregory C. (1960) “Tests of Equality Between Sets of Coefficients in Two Linear Regressions,” Econometrica, Vol. 28 pp. 591-605.

[7] Goldfarb, A. (2006) “The Medium-Term Effects of Unavailability” Quantitative Marketing and Economics Vol 4, No. 2 pp. 143-171.

[8] Inclán,C. (1993) “Detection of Multiple Changes in Variance Using Posterior Odds”, Journal of Business and Economic Statistics, Vol. 11, No. 3 pp. 289-300.

[9] Kerk, K., Penn J., Atwood C. and Albornoz J. (2006)“The State of Information Security Spending”, Forrester Research Inc.

[10] Livingston J. (2005) “How Valuable Is a Good Reputation? A Sample Selection Model of Internet Auctions” The Review of Economics and Statistics, Vol. 87, No. 3, pp. 453-465.

[11] Wang, J and Zivot E. (2006) “A Time Series Model of Multiple Structural Changes in Level, Trend, and Variance”, mimeo.

[12] Yao,Y-C. (1988) “Estimating the number of change points via Schwarz’ criterion”, Statistics and Probability Letters, Vol. 6, pp. 181-189.

Company	AR Lags	a_t	b_t	ϕ	σ_t	Structural Break Date
Playboy	2	39.927702 29.505161	0.249747 (0.266900)	0.659740 (0.076636) 0.286526 (0.076334)	144.047060 (8.632660)	*NA*
BJ's Whole Sale	3	44.840565 (7.648868)	−0.091453 (0.035663)	0.221921 (0.082257) −0.046332 (0.084013) 0.025756 (0.081242)	16.431521 (0.991511)	*NA*
Lexis Nexis	1	83.342371 (18.302988)	0.070280 (0.119002)	0.626703 (0.077359)	36.069868 (2.721002)	*NA*
Equifax Canada Inc	1	104.506567 (23.763041)	0.052433 (0.137608)	0.582161 (0.087791)	41.510943 (3.352297)	*NA*
DSW Shoes	2	25.487487 (8.679060)	0.119988 (0.092142)	0.161068 (0.221029) −0.026973 (0.081421)	16.678751 (3.485094)	4/18/2005
		0.717864 (31.656043)	1.753378 (0.967551)	same	47.049394 (69.951417)	4/22/2005
		34.488430 (17.186950)	−0.043201 (0.138468)	same	18.968294 (5.374838)	
Scottrade	1	51.394453 (32.096700)	0.303954 (0.386272)	0.913278 (0.046858)	83.003035 (7.116663)	*NA*
Sam's Club	2	41.961976 30.327283	0.625002 (0.480327)	0.550148 (0.091324) 0.326283 (0.090892)	159.544877 (11.745347)	*NA*
Cook's Illustrated	3	50.807211 (11.527743)	0.005213 (0.068632)	0.238953 (0.082047) 0.079675 (0.083779) 0.097070 (0.081607)	35.729403 (2.170634)	*NA*
Ross Simons	3	38.148269 (8.475537)	−0.085680 (0.047408)	0.243629 (0.084408) 0.086971 (0.086425) 0.097161 (0.083985)	21.426224 (1.332300)	NA
Vystar Credit Union	4	27.175072 (5.529731)	−0.063418 (0.019770)	0.120797 (0.074860) 0.159764 (0.074664) 0.058654 (0.074441) 0.057014 (0.074102)	11.096993 (0.601305)	*NA*
MoneyGram	2	36.380562 (7.208054)	−0.070379 (0.043243)	0.087911 (0.096981) 0.201361 (0.097155)	11.798004 (0.870793)	*NA*
ChoicePoint	4	40.090801 (7.656994)	−0.110078 (0.030275)	0.183845 (0.079285) 0.017415 (0.080125) −0.064836 (0.079717) 0.109797 (0.078429)	12.991664 (0.771088)	*NA*
Polo Ralph Lauren	3	76.345152 (20.907682)	0.018977 (0.081487)	0.397295 (0.080105) 0.035053 (0.086153) 0.210669 (0.080105)	42.582355 (2.674826)	*NA*
Bank of America	1	5.695791 (31.634118)	2.215841 (1.642280)	0.988020 (0.011004)	475.455814 (33.184019)	*NA*
HSBC	2	28.332526 31.652389	0.648074 (0.590894)	0.567357 (0.090514) 0.373945 (0.090188)	191.486143 (13.762807)	*NA*
TransUnion	1	69.851846 (10.631970)	−0.046977 (0.066426)	0.448030 (0.071576)	36.596018 (2.156526)	*NA*

Table 7: Estimated parameters and structural break dates for time series model of structural change in daily web traffic around security breach events. Refer to Equation 1.

No	Properties of Breach Events	Classification	Error Classification Rate (Sample)
1	publicly traded	Reported	17/97 = 0.17
2	1. not publicly traded 2. business type: Data, Hotl, Ind, Medi, NFP, Org, Retl, Tech 3. breach type: Fraud, Hack, Snail Mail, Stolen, Web	Reported	12/48 = 0.25
3	1. not publicly traded 2. business type: City, Fin, Law, Med 3. breach type: Hack, Lost, Stolen,Web 4. breached data: ACC,CCN,MISC	Reported	4/11 = 0.36
4	1. not publicly traded 2. business type: Data, Hotl, Ind, Medi, NFP, Org, Retl, Tech 3. breach type: Dispose, Lost	Not Reported	4/11 = 0.36
5	1. not publicly traded 2. business type: City, Fin, Law, Med 3. breach type: Dispose, Fraud	Not Reported	1/12 = 0.08
6	1. not publicly traded 2. business type: City, Fin, Law, Med 3. breach type: Hack, Lost, Stolen, Web 4. breached data: FIN, SSN	Not Reported	7/18 = 0.38

Table 8: Classification Tree Rules Without Total Affected Number

No	Properties of Breach Events	Classification	Error Classification Rate (Sample)
1	1. total affected number $\geq 1.61 \times 10^4$ 2. publicly traded	Reported	$1/41 = 0.02$
2	1. total affected number $\geq 1.61 \times 10^4$ 2. not publicly traded 3. breach type: Snail, Stolen, Web	Reported	$0/11 = 0.00$
3	1. $2.325 \times 10^5 \leq$total affected number 2. not publicly traded 3. breach type: Fraud, Hack, Lost	Reported	$1/10 = 0.10$
4	1.$1.61 \times 10^4 \leq$total affected number $\leq 2.325 \times 10^5$ 2. not publicly traded 3. breach type: Fraud, Hack, Lost 4. business type: Data, Hotl, Ind, Retl	Reported	$2/10 = 0.20$
5	1. total affected number $\leq 1.61 \times 10^4$ 2. publicly traded 3. data breached: ACC, CCN, FIN, MISC, PPN	Reported	$1/20 = 0.05$
6	1. total affected number $\leq 1.61 \times 10^4$ 2. publicly traded 3. data breached: NAA, SSN 4. business type: Data, Fin, Retl, Tech	Reported	$8/24 = 0.33$
7	1. total affected number $\leq 1.61 \times 10^4$ 2. not publicly traded 3. business type: Data, Ind, Med, NFP, Org, Tech	Reported	$9/25 = 0.36$
8	1.$1.61 \times 10^4 \leq$total affected number $\leq 2.325 \times 10^5$ 2. not publicly traded 3. breach type: Fraud, Hack, Lost 4. business type: Fin, Tech	Not reported	$4/10 = 0.40$
9	1. total affected number $\leq 1.61 \times 10^4$ 2. publicly traded 3. data breached: NAA, SSN 4. business type: Ind, Ins, Med	Not reported	$5/12 = 0.41$
10	1. total affected number $\leq 1.61 \times 10^4$ 2. not publicly traded 3. business type: City, Fin, Law, Med, Retl	Not reported	$7/34 = 0.20$

Table 9: Classification Tree Rules With Total Affected Number

Selection	Variable(s) Chosen
forward	publicly traded
backward	total affected number, publicly traded
stepwise	publicly traded

Table 10: Logistic Regression Model selection methods and variables selected

Test	Chi-Square	DF	Prob > χ^2
Likelihood Ratio	34.8078	2	$< .0001$
Score	15.7860	2	0.0004
Wald	17.7028	2	0.0001

Table 11: Testing Global Null Hypothesis: Logistic Regression Coefficients =0

Effect	DF	Chi-Square	Pr> χ^2
TotalAff	1	5.9926	0.0144
Not Publicly Traded	1	12.0017	0.0005

Table 12: Analysis of Effects in Logistic Regression

Parameter	DF	Coefficient Estimate	Standard Error	Wald Chi-Square	Prob> χ^2
Intercept	1	1.2638	0.3824	10.9219	0.0010
TotalAff	1	$8.619E-6$	$3.521E-6$	5.9926	0.0144
Not Publicly Traded	1	-1.5696	0.4531	12.0017	0.0005

Table 13: Logistic Regression Estimation Results

Prob Level	Correct Positive	Correct Negative	Incorrect Positive	Incorrect Negative	False Positive %	False Negative %
0.46	73	24	17	20	18.9	45.5
0.48	72	26	15	21	17.2	44.7
0.50	71	26	15	22	17.4	45.8
0.52	69	27	14	24	16.9	47.1
0.54	68	29	12	25	15.0	46.3
0.56	67	29	12	26	15.2	47.3
0.58	67	29	12	26	15.2	47.3
0.60	67	29	12	26	15.2	47.3
0.62	67	29	12	26	15.2	47.3
0.64	67	29	12	26	15.2	47.3

Table 14: Logistic Regression Threshold Classification Results