CERIAS Tech Report 2009-28

The Association between the Disclosure and the Realization of Information Security Risk Factors

by Tawei Wang and Jackie Rees and Karthik Kannan
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

The Association between the Disclosure and the Realization of Information Security Risk Factors

Tawei Wang
College of Management
National Taiwan University
Taipei, 106 Taiwan
twang@ntu.edu.tw

Jackie Rees
Krannert Graduate School of Management
Center for Education and Research in Information Assurance and Security (CERIAS)
Purdue University
West Lafayette, IN 47907
jrees@purdue.edu

Karthik Kannan
Krannert Graduate School of Management
Center for Education and Research in Information Assurance and Security (CERIAS)
Purdue University
West Lafayette, IN 47907
kkarthik@purdue.edu

The Association between the Disclosure and the Realization of Information Security Risk Factors

Abstract

Firms often disclose information security risk factors in public filings such as 10-K reports. The internal information associated with disclosures may be positive or negative. In this paper, we are interested in evaluating how the nature of security risk factors disclosed, which is believed to represent the internal information regarding information security, is associated with future breach announcements. For this purpose, we build a decision tree model, which classifies the occurrence of future security breaches based on the textual contents of the disclosed security risk factors. The model is able to accurately associate disclosure characteristics with breach announcements about 77% of the time. We further explore the contents of the security risk factors using text mining techniques to provide a richer interpretation of the results. The results show that the security risk factors with action-oriented terms and phrases are less likely to be related to future incidents. We also conduct a cross-sectional analysis to study how the market interprets the nature of information security risk factors in annual reports at different time points. We find that the market reaction following the security breach announcement is different depending on the nature of disclosure. Thus, our paper contributes to the literature in information security and sheds light on how market participants can better interpret security risk factors disclosed in financial reports at the time when financial reports are released.

Keywords: information security, information security incident, risk factor, text mining

The Association between the Disclosure and the Realization of Information Security Risk Factors

1. Introduction

Firms often recognize the impact of information security risks and announce them publically. For example, Kohl's states in its 2006 annual report that "… [the company's] facilities and systems…may be vulnerable to security breaches… [which] could severely damage its reputation, expose it to the risks of litigation and liability, disrupt its operations and harm its business" (Kohl's 2007, p.8). In general, a firm discloses risk factors only after taking into consideration its internal information (e.g., Verrecchia 1983; Dye 1985; Skinner 1994; Kasznik and Lev 1995). Prior work has recognized two different types of internal information influencing disclosures. On one hand, firms may disclose information that is positively interpreted by the investors with the objective of improving the firm's valuation (e.g., Verrecchia 1983; Dye 1985). For example, a firm may disclose that it has already taken precautionary measures to mitigate the impact of negative events such as communication failures, hacker attacks, etc. On the other hand, firms may disclose information that is negatively interpreted by the investors with the aim of reducing future litigation costs associated with adverse events (e.g., Skinner 1994). The distinction between these two types of internal information is important for investors and debtors, when evaluating risks, and making financing and/or investment decisions (e.g., Firtel 1999; FASB 2008). The internal information is often reflected in the textual contents of the disclosure (e.g., Bettman and Weitz 1983; Abrahamson and Park 1994; Feng 2008). In this paper, we are interested in evaluating how the internal information is associated with the disclosures of information security risk factors.

Our study addresses the following two research questions: First, how is the nature of information security risk factors disclosed in annual reports associated with the occurrence of future security breach announcements? Second, how does the market interpret the textual

contents of disclosures both before and after information security breaches? We address these questions by drawing upon a diverse set of tools, and our study features both quantitative and qualitative measures. To answer the first question, we text mined the contents of the disclosed security risk factors; and developed a decision tree classification model to associate the mined textual content with security breach announcements. For the second question, we used the results from the decision tree model to perform a cross-sectional analysis and examined market reactions to the disclosed security risk factors. Note that the methodologies we adopt also relate to the discussions on the predictive and explanatory statistical models by Shmueli and Koppius (2009). Our decision tree classification model involves the predictive component, while the cross-sectional analysis is the explanatory component in that it helps relate the disclosure of information security to the disclosure theory literature.

Our analysis on disclosures is quite different from those in prior works. Both the finance and accounting literature have extensively studied market reactions to disclosures. They have primarily dealt with settings where there is not much of uncertainty associated between the disclosures and the realization of the events. For example, when firms disclose that they would miss their earnings estimates, the event (earnings miss) is often realized. However, with most operational risks, the uncertainty between disclosures and the realizations typically exists. To the best of our knowledge, none of the prior works have analyzed this uncertainty. In this paper, we study the uncertainty in the context of information security, which is an important component of operational risk (see, for example, Kohl's statement above). Specifically, we evaluate how the internal information held by a firm regarding information security relates to the realization of security breach announcements. Using the textual and cross-sectional analyses mentioned earlier, we provide important managerial insights regarding the nature of disclosure in the information security context.

The rest of the paper is organized as follows. We review the literature on the management

and the economics of information security and disclosures in Section 2. The data collection process is elaborated in Section 3. Next, in Section 4, we analyze the textual data of the disclosed information security risk factors. We further present the cross-sectional analysis in Section 5. In Section 6, we conclude with discussion of contributions, limitations and avenues for future research.

2. Literature Review

There are two major streams of literature that are directly related to our study. One is the research stream on management and the economics of information security. The other is the literature on disclosures in accounting.

2.1. Management and Economics of Information Security

There is a limited but growing body of knowledge in this stream of research. A few papers have analyzed security investment decisions while a few others have studied the management of information security policies and procedures. Gordon and Loeb (2002), Gordon et al. (2003), and Gal-Or and Ghose (2005) employ analytical frameworks to study security investment decisions. Tanaka et al. (2005) empirically analyze how vulnerabilities of the firm affect security investments. Goodhue and Straub (1991) show that security concerns vary by industry, company actions and individual awareness. Also, studies (e.g., Straub 1990; Siponen and Iivari 2006; Siponen 2006) demonstrate the critical role played by information security policies and standards in managing security risks. Often, such investment decisions, policies and actions are closely guarded by organizations in order to avoid exposing their vulnerabilities. By revealing security risk factors in annual reports, but not specific policies, firms convey their internal assessment of the risk factors to the market, as mentioned previously.

Research has also investigated the impact of information security breaches on a firm's business value. Based on different methodologies and different datasets, a few papers show that there exists a significant negative impact (e.g., Ettredge and Richardson 2003; Garg et al. 2003;

Cavusoglu et al. 2004; Acquisti et al. 2008), while others do not find such impact (e.g., Campbell et al. 2003; Hovav and D'Arcy 2003; Kannan et al. 2007). A part of our paper also seeks to address the differences in the findings even though our primary focus is the textual analysis of disclosures. As for the primary part of the analysis, we develop a classification model to investigate the association between the nature of information security risk factors and subsequent security incidents. Using insights regarding the associations, we also investigate how the nature of the disclosed security risk factors moderates the market reactions to security breach announcements.

2.2. Disclosures in Accounting

There is a rich body of literature in accounting that examines disclosures. When there is no disclosure cost, full disclosure exists because investors believe that non-disclosing companies have the worst possible information (e.g., Grossman 1981; Milgrom 1981). However, if disclosure costs or uncertainty exist, companies will disclose only when the benefits exceed the costs (e.g., Verrecchia 1983; Dye 1985). Disclosure may also be used to reduce ex post legal and reputation costs from bad news, or when the firm faces earnings disappointments (e.g., Skinner 1994; Kasznik and Lev 1995; Field et al. 2005). Specific to risk disclosures, one recent study by Jorgensen and Kirschenheiter (2003) formally models managers' decisions on voluntarily disclosing a firm's risks, and they find that firms with smaller future uncertainty will choose to disclose risk factors. Additionally, studies have focused on the quality and credibility of the disclosures (e.g., Lang and Lundholm 1993; Penno 1997; Stocken 2000), the usefulness of disclosures (e.g., Francis et al. 2002; Landsman and Maydew 2002), and other aspects of voluntary disclosures such as expectation adjustment, costs, analysts following, and signaling rationale (e.g., Lev and Penman 1990; Lang and Lundholm 1996).

Prior work analyzing the textual contents of disclosures can be categorized into two groups. The first group includes papers that have analyzed the relation between disclosures and internal

information. For example, Abrahamson and Park (1994) demonstrate that the presence of outside directors, large institutional investors, and accountants reduce a firm's chances of not disclosing negative outcomes. As another example in this group, Feng (2008) shows that, when firms have lower earnings, their annual reports tend to be wordy. The second group includes papers that analyze disclosure and market reactions. For example, Balakrishnan et al. (2008) classify news articles into press- and firm-initiated articles and show that firm-initiated media has significant negative market reactions relative to press-initiated media. Tetlock et al. (2008) show that the textual contents in news articles provide qualitative information when estimating a firm's fundamental. Our paper is quite different from both these categories of papers in that we focus on the nature of disclosure and realization of the (information security) event.

2.3. Literature Combining Both Streams of Research

In this paper, we link both streams of research. To the best of our knowledge, Sohail (2006) is the only study that has also linked these two streams. In Sohail's paper, he demonstrates that the market values security disclosures, by showing that such disclosures are positively related to stock price at the time when financial reports are released. However, our paper has a different focus in that it develops a model to understand the relation between security risk factors disclosed in financial reports (10-K or 20-F for foreign firms) and information security breach announcements signaling the realization of the event. Specifically, we investigate how the nature of security risk factors disclosed in financial reports is associated with the possibility of future breach announcements. In addition, our paper analyzes how the market reaction to information security breach announcements is dependent upon the nature of disclosures.

3. Data Collection

Figure 1 provides the timeline which we use as a basis for our analysis. The disclosure at time *t* in Figure 1 is a list of information security risk factors that may adversely affect the firm's future performance, as reported in the annual report. See Appendix A for an example of a

security risk factor disclosed in an annual report. The announcements at time $t + 1$ in Figure 1 are the breaches reported in news articles. We expect the security risk factors disclosed at time t in Figure 1 to contain clues regarding the possibility of future security incidents occurring at time $t + 1$ as discussed above.

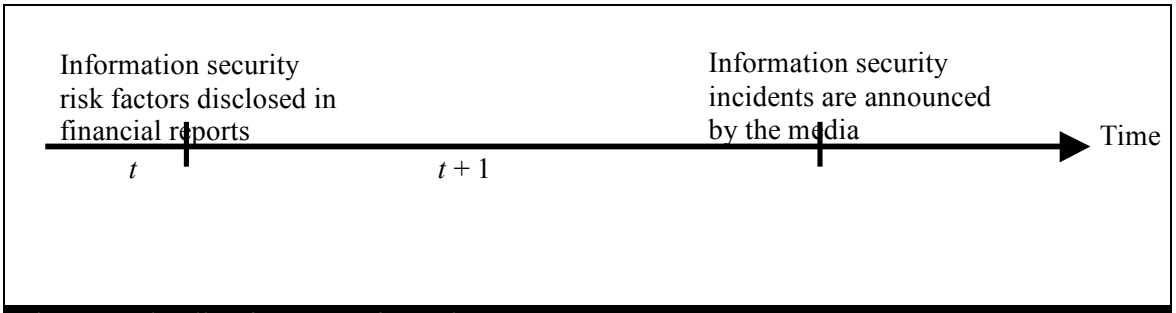

Figure 1. Timeline for Two Information Sets

We employ an endogenous stratified sampling method (e.g., Cosslett 1981; Cameron and Trivedi 2007) for our data collection. This method is commonly used when the event is rare (compared to nonevents), such as in international relations, wars, venture capital investments, and epidemiological infections (e.g., King and Zheng 2001; Sorenson and Stuart 2001). As we show below, information security breach announcements regarding publically traded firms are also rare. Estimations using an endogenous stratified sample are more efficient than using a full sample (e.g., Cosslett 1981; Imbens 1992).

Our data collection is a three-step process. First, we collected data from publically traded firms having breach announcements between 1997 and 2008 reported in major media outlets. We searched the *Wall Street Journal*, *USA Today*, the *Washington Post*, and the *New York Times* using the Factiva database as well as the *CNet* and *ZDNet* websites. We used the following search terms: (1) security breach, (2) hacker, (3) cyber attack, (4) virus or worm, (5) computer break-in, (6) computer attack, (7) computer security, (8) network intrusion, (9) data theft, (10) identity theft, (11) phishing, (12) cyber fraud, and (13) denial of service. These search terms were similar to those used in prior studies (e.g., Campbell et al. 2003; Garg et al. 2003; Kannan et

al. 2007). We screened the news articles and collected only those in which the breach announcement identified the specific date for the security incident, and the breached firm did not have any confounding events, such as earnings announcements, or mergers and acquisitions, around that date. The above process resulted in 101 firm-event observations from 62 firms. From this, it must be obvious that information security breach announcements regarding publically traded firms are rare. Note that these incidents correspond to events occurring at time $t + 1$ in Figure 1.

Second, for each event in the previous step, we gathered the information security risk factors disclosed in the breached firm's annual report (10-K or 20-F filings for foreign firms) published immediately *prior* to the breach announcement using EDGAR Online.[1] Note that some firms did not have any security risk factors disclosed in the annual report while others had several. Using this process, we collected 43 security risk factors, each corresponding to a breach announcement.[2] These disclosures correspond to period t in Figure 1.

Third, we need to collect security risk factors from firms that did not have any breach announcements (nonevents). However, one of the main questions with endogenous stratified sampling is how big should the sample size of nonevents be? There is considerable variation in the literature regarding how the total sample should be split between events and nonevents. Breslow and Day (1980) use a 20%-80% split of events and nonevents; Pinczowski et al. (1994) use a 30%-70% split; Rudolfer et al. (1999) use a 60%-40% split; and Steinberg et al. (2006) use a 50%-50% split. Lancaster and Imbens (1991) show that a 50%-50% split is optimal for

[1] http://www.sec.gov/edgar.shtml

[2] Suppose, in a particular year, if a firm has two events, we collected only the disclosure in the previous annual report and counted it as one disclosure in our dataset. Additionally, we counted each of the disclosures separately and ran our analysis, but our results were consistent.

estimation purposes. Consistent with their work, we also used a 50%-50% split. To check for robustness with respect to the splits, we also studied the performance of our decision tree when subjected to a progressive sampling method (e.g., John and Langley 1996; Frey and Fisher 1999; Morgan et al. 2003). The details of this and other robustness checks can be found later in Section 4.2.

For this third step, we randomly chose 62 firms without any breach announcement between 1997 and 2008. For each of these firms, we randomly picked the annual report from one of the years in the 12 year period (1997-2008) and collected information security risk factors in that annual report. We did not consider all 12 years because firms typically tend to add new risk factors to the earlier ones and, therefore, will lead to oversampling and biasing of our results. Through this process, we collected 34 risk factors. As before, not all firms had security risk factors in the annual report and a few firms had several.

From the above three steps, our dataset involves 124 (62 + 62) firms and 77 (43 + 34) information security risk factors. These firms are distributed across 28 different industries (two-digit SIC code). At the end of 2008, the firms had an average age of 22 years (standard deviation of 19 years), which was calculated based on the year range in Compustat, and average total assets of $2.8 billion (standard deviation of $8 million).

4. Classification Model

In this section, we focus on mining the textual data to understand the information conveyed by security risk factors through a decision tree classification model. Text mining, in general, has proven to be a useful tool to extract information for finding nontrivial patterns and trends (e.g., Feldman and Sanger 2006). For example, text mining techniques have been used in different contexts, such as to classify news stories, detect fraud, and improve customer support (e.g., Han et al. 2002; Fan et al. 2006; Cecchini et al. 2007). In the information security context, we apply

text mining techniques to the contents of disclosed security risk factors so as to identify and categorize the elements of security risk factors that might associate with future incident announcements. These results serve as the inputs in the construction of a decision tree. We use a decision tree for the following two reasons. First, the inherent transparency and interpretability of decision tree models help users follow the path of the tree and understand the classification rules step by step (e.g., Baesens et al. 2003; Zhou and Jiang 2004). Second, studies such as Goto et al. (2008), Long et al. (1993), and Rudolfer et al. (1999) have shown that decision tree models have a better accuracy rate for our sampling method than logistic models. Moreover, studies such as Zadrozny (2004) and Fan et al. (2005), which compare various classification models, find that that decision tree models perform well even with biased samples. We also tested other classification models, such as neural networks, and the results are largely similar.

4.1. Decision tree classification model

Our decision tree classification model is the outcome of a three-step process, which is shown in Figure 2. Each of those steps is elaborated in detail below.

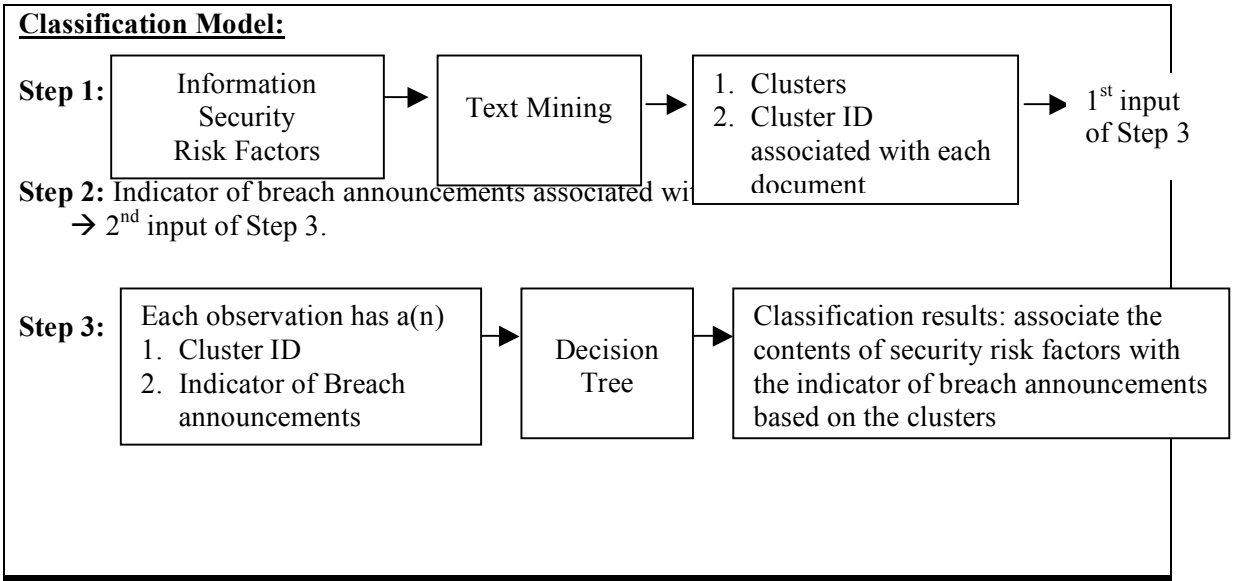

Figure 2. Process Flow for the Classification Model

In order to perform the analysis, we used the 77 security risk factors collected. In the first

step, we used SAS Text Miner to extract the terms and the related clusters from the textual contents of the disclosed information security risk factors (the cluster identification is a standard process and is detailed in Appendix B). This procedure resulted in each risk factor being associated with a cluster. We assigned the corresponding cluster ID to the risk factor. If a firm did not have any disclosed security risk factor, no cluster ID was assigned to it in our first analysis, which we refer to as the base case analysis. As we discuss later in Section 4.2, when we assigned a cluster ID to non-disclosures as well, our decision tree performed better than the base case.

The second step is quite straightforward. For our classification model, we introduced an indicator variable for each firm. The indicator variable was set to one if the firm had a breach announcement. Otherwise, it was set to zero. In this step, we continued to retain the association between each firm and its disclosed security risk factors.

In the third step, a decision tree was built to classify the indicator of breach announcements (from Step 2) based on the cluster ID (from Step 1). For this step, the following parameters were used. The dataset was partitioned into two subsets: 80% of the original dataset was used for training, and the other 20% for validation and testing. We initialized the prior probability of the classifier as the proportion of the number of related documents in the whole dataset. The software used for training, validating, and testing our decision tree classification model was SAS Enterprise Miner.

The performance of the classification model in Step 3 is dependent on the number of clusters generated from Step 1. Prior literature (e.g., Smyth 2000; Still and Bielek 2004; Tibshirani et al. 2001) recommends an iterative process to determine the optimal number of clusters. Consistent with this idea, we experimentally varied the number of clusters and repeated the three steps in Figure 2 until the error rate of the decision tree model in Step 3 was minimized. Eventually, we found that having four clusters in Step 1 resulted in the smallest error rate in Step 3. The

validation and testing results corresponding to the base case are presented below. The resulting decision tree always had two leaves from the root node.

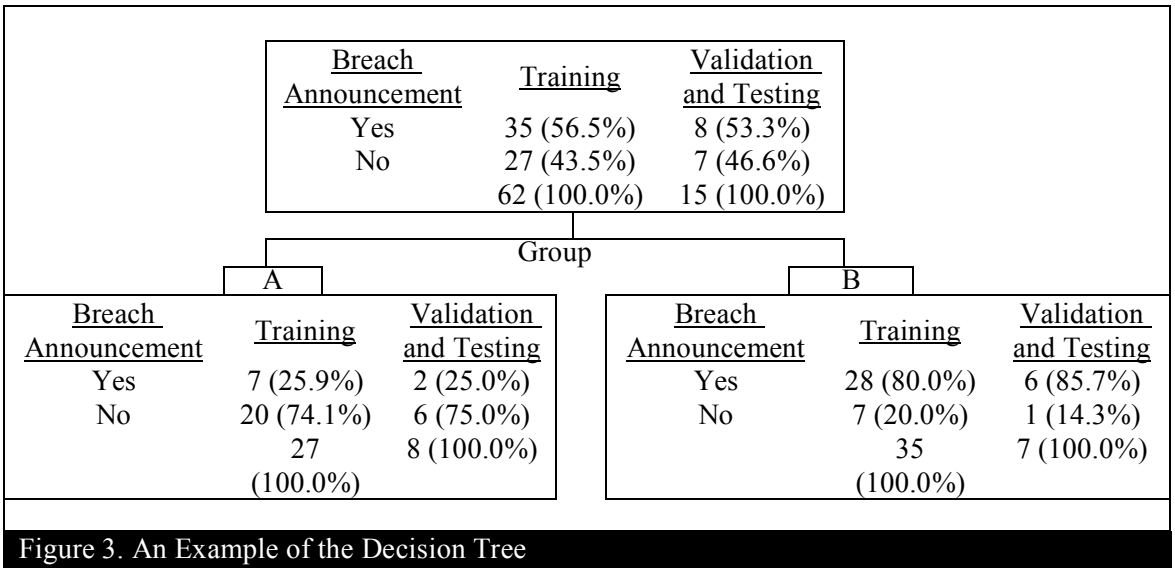

Figure 3. An Example of the Decision Tree

Figure 3 shows an example decision tree obtained from one set of random draws for the training, validation and testing datasets. The example used 62 and 15 documents for training, and validation and testing respectively. The classification model segregated all the disclosed security risk factors into two groups: A and B. Documents associated with clusters 1 and 2 from Step 1 were primarily associated with Group A, while clusters 3 and 4 with Group B. Interestingly, the association between breach announcements and the groups was also different. Notice that the majority of the documents (about 75%) in Group A, which corresponds to the left sub-tree in Figure 3, are associated with having "no breach announcement". In contrast, the majority of the

documents (about 80%) related to Group B in the right sub-tree are associated with "breach announcement".

To further validate our results above, we used a commonly-adopted procedure called 10-fold cross validation (e.g., Weiss and Kapouleas 1989; Kohavi 1995). Accordingly, we repeated Step 3 in Figure 1 ten times, each time with a different randomly chosen training, validation and testing dataset. The results were averaged across all ten trials and reported in Table 1 and Table 2. We found that they were largely similar to those in Figure 3. The results in Table 1 show that, about 75% of the documents in Group A are associated with "no breach announcement" and about 80% of the documents in Group B are associated with "breach announcement". In Table 2, the cells corresponding to accurate predictions are marked with a gray background for the sake of readability. The overall accuracy rate for the validation result is 77.42% (i.e., 45.16% + 32.26%). Note that we repeated this cross validation analysis a number of times and found that the average accuracy rate for the validations is about 76%.

Table 1. Cross Validation Results

		Group A		Group B	
Breach Announcement		Training	Validation and Testing	Training	Validation and Testing
Yes	Frequency	6.5	1	28.5	7
	Percentage	24.53	13.33	80.28	93.33
No	Frequency	20	6.5	7	0.5
	Percentage	75.47	86.67	19.72	6.67
Total	Frequency	26.5	7.5	35.5	7.5
	Percentage	100.00	100.00	100.00	100.00

The base case decision tree results demonstrate that there indeed exists a relation between the textual contents in the security risk factor disclosures and the realization of future incidents. Following up on this result, we executed two additional sets of analysis. First, we investigated how the textual contents of security risk factors disclosed are associated with Groups A and B. This analysis is presented in Section 4.3. Second, we studied if and how market reacts to the differences in the textual contents. Details of this analysis are provided in the cross-sectional

analysis section. Before we proceed with these two analyses, we check the robustness of the results presented in the base case.

Table 2. Confusion Matrix of the Cross Validation Results

Breach Announcement			Predict		
			Yes	No	Total
Actual	Yes	Frequency	28	7	35
		Percentage	45.16	11.29	56.45
		Row Percentage	80.00	20.00	
		Column Percentage	80.00	25.93	
	No	Frequency	7	20	27
		Percentage	11.29	32.26	43.55
		Row Percentage	25.93	74.07	
		Column Percentage	80.00	74.07	
	Total	Frequency	35	27	62
		Percentage	56.45	43.55	100.00

4.2. Robustness tests of the classification results

First, we investigated the sensitivity of the results to the settings of the text mining tool. In our base case analysis, as described in Appendix B, the terms were weighted differently during the clustering process depending on the frequency of their occurrences. The manner in which the terms were weighted can affect the clustering results and, therefore, the classification results also. It is possible that firms have implemented idiosyncratic policies regarding information security and have disclosed them. Such terms may be weighted less in the base case. So, to test the robustness of our result, we weighted all the terms equally. We observed that our classification model produced results similar to the base case.

Second, we accounted for double negative terms in the security risk factors (for example, "not incorrect" should be treated the same as "correct" when processing the textual contents). Without controlling for such terms, the text mining tool may interpret the security risk factor conversely. In our analysis, we manually took into account the double negative terms by reading through the disclosed risk factors and controlling for them in SAS Text Miner. Our results were consistent with the base case.

Third, as mentioned earlier, we did not include the firms that did not have any disclosed security risk factors in the base case dataset. We assigned a new cluster ID to the non-disclosing firms and re-performed the decision tree analysis. The result showed that Group A was now associated with three clusters: the new cluster, cluster 1 and cluster 2 while Group B was still associated with two clusters: cluster 3 and cluster 4. The results were consistent with those of the base case. In particular, Group A was associated with "no breach announcement" about 85% of the time and Group B was associated with "breach announcement" about 80% of the time. The accuracy rate of the classification model was 79.9%.

Fourth, we controlled for (1) industries, (2) the type of breach, i.e., confidentiality, integrity, or availability (e.g., Gordon et al. 2006), and (3) historical security risk factor disclosures and our results remained similar.

Fifth, papers such as Hughes and Pae (2004) and Bagnoli and Watts (2007) have argued that there exists a relation between mandatory (e.g., 10-K filings) and voluntary (e.g., risk factors) disclosures. For example, when the earnings performance disclosures in the 10-K filings – which are mandatory – are poor, firms may not disclose many security risk factors – which are voluntary – and vice-versa. Therefore, it could be argued that information security risk factors may depend on the earnings performance. However, we did not find any significant association between the earnings performance in the 10-K report and the number of information security risk factors disclosed.

Finally, we validated our classification model results using the progressive sampling method (e.g., Frey and Fisher 1999; John and Langley 1996; Morgan et al. 2003). This procedure involved building the decision tree model for different sample sizes. In our context, the sample size for the number of firms with breach announcements always remained the same. We varied the number of firms without breach announcements so that the total sample would be 100, 200, and 300 firms. Specifically, we randomly sampled 38, 138, and 238 non-breached firms from

Compustat. Consistent with observations in prior work, we noticed that the accuracy rate for our model increased with the total sample size but at a decreasing rate. Specifically, the accuracy rate increased from 75% to 78% and 79% as the sample size increased from 100 to 200 and 300. Also, we found that the classification results were largely similar to the base case, i.e., Group A was associated with "no breach announcement" about 75% of the time and Group B was associated with "breach announcement" about 80% of the time.

4.3. Comparison of the textual contents between the disclosure groups

In this section, we build up on the results from the base case. Specifically, we study how the textual contents differ between the two disclosure groups. For this analysis, we formed two pools of security risk factors, one corresponding to each group (Group A or Group B). Each pool was separately provided as an input to the text mining tool. The tool identified the terms and the associated textual contents that commonly occurred in that group and they are given in Table 3. In that table each row represents one cluster and each group has many clusters. Within each cluster, there are five terms with the highest calculated frequency in the cluster (see Appendix B for detailed information). A term with the plus (+) sign represents a group of equivalent terms. For example, both "ability" and "abilities" are considered equivalent. The percentage is the frequency of a set of terms divided by the total frequency. The root mean squared standard deviation (RMS Std.) measures how tight the clusters are. For cluster k, it is computed as Table 3, we calculated the percentage of roles played by the different terms in that cluster. For Group A, Cluster 1 has 40% (2 out of 5) of nouns ("resource" and "virus") and 60% of the terms are verbs. For Cluster 2, 60% (3 out of 5) of the terms are nouns ("customer", "disruption", and "process") and 40% of the terms are verbs. For Group B, Cluster 1 and Cluster 2 are composed entirely (100%) of nouns. 80% of Cluster 3 are nouns with only one adjective ("other"). Cluster 4 contains 80% nouns and one preposition ("with"). After calculating these percentages, we compare them across the two groups. Obviously, Group B does not have any verbs but has a

much higher percentage of nouns compared to Group A.

We further explore this result by comparing the terms across these two groups qualitatively. First, since these disclosures are about security risk factors, we observe several terms with negative meanings in the clusters within both groups. For example, terms such as "damage" and "disruption" appear in Group A, and terms "disaster", "failure", "loss", and "interruption" in Group B. Also, we observe terms about the type of incidents and the subjects or objects that could be affected because of the attacks, such as "virus", "customer", "revenue", "user", "business" and "telecommunication". After eliminating the common terms across these two groups, there are still several action related terms in the clusters of Group A which are not included in the clusters of Group B. Recall that Group A corresponds to the *no breach announcement group* while Group B is related to the *breach announcement group*. Therefore, the lack of terms about operations and actions such as "act", "prevent", and "implement" (bolded and italicized in Table 3) in Group B could possibly be associated with a negative interpretation of the disclosed risk factors. As a robustness check, we also explored all the possible clusters for both groups and found that the results are largely similar (see Table B.1 in Appendix B for the results).

We also undertook additional analysis using concept links. Concept links provide context to the terms in the clusters which help us better explain the terms within each cluster. For example, if we observe that the terms "attack" and "denial" are often disclosed together, we are able to understand that the term "attack" in the cluster generally refers to the context of denial-of-service attacks. We checked the concept links for all the terms in all clusters for both groups. For Group B, 50% (2 out of 4) of the terms with concept links are general concepts, such as "breach" (see Figure 4 for an example), or specific subjects that might be affected such as "business". The remaining 50% are terms with negative meanings such as "disaster" and "interruption". That is, in the risk factors disclosed by Group B (corresponding to "breach announcement"), general

security terms or the subjects that might be affected play an important role in conveying information to the public (i.e., generally co-occurring with other phrases in security risk factors). However, for Group A, all the terms (2 out of 2) with concept links are action terms such as "implement" and "prevent" (see Figure 4 for an example). Thus, in the risk factors disclosed in Group A, action terms generally co-occur with other phrases in the risk factors. The results from the concept links confirm our findings that the major disclosure characteristic difference between these two groups is that Group A uses action terms to disclose security risk factors while Group B does not. In Appendix B, we also investigate how the disclosures change after a breach announcement.

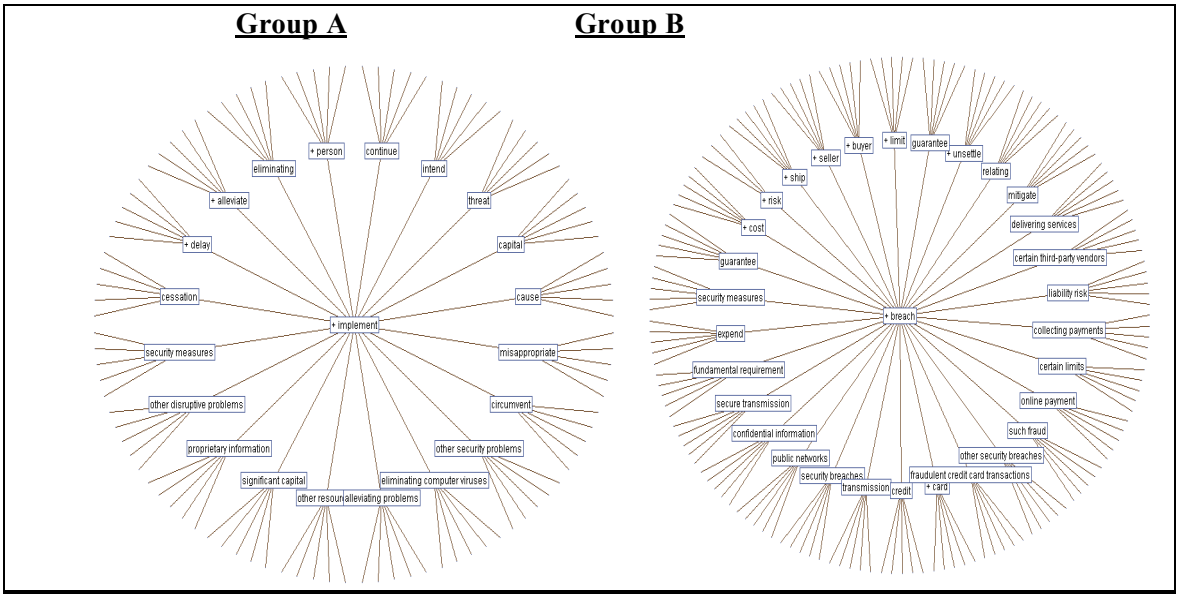

Figure 4. Examples of Concept Links

In summary, our classification model shows that disclosure characteristics are associated with the possibility of future uncertainties. Specifically, we find that when security risk factors involve action terms or terms about processes, the firms are less likely to be associated with future incidents. This result highlights the importance of our model to market participants. In particular, market participants can look for action terms or terms about processes and assess the firm's future uncertainty regarding information security differently than the firms with general

breach information disclosed in financial reports. We explore in the cross-sectional analysis section how the market interprets the nature of information security risk factors in annual reports.

5. Cross-Sectional Analysis

In this section, our primary focus is to investigate the association between market reactions to security incidents and security risk factors disclosed in financial reports based on the classification results presented above. By doing so, we are able to understand whether the investors' perception of the information conveyed by the security risk factors is different after the realization of security incidents.

We first assess how the market reacts to security risk factors at the time when the financial report is released. For this, we replicate Sohail's (2006) analysis but only with the smaller dataset used for our classification model. We do not control for the year or the industry as they did since we do not have enough observations. Consistent with Sohail's result but with a smaller magnitude, we find a positive association between stock price and security risk factors disclosed in financial reports at the time when the financial reports are released (0.94). However, our result is not significant ($p < 0.20$) which may be due to the smaller sample size. We also find an insignificant positive association when we limit our sample to those firms with breach announcements. Furthermore, the association is positive even after considering whether the security risk factors are disclosed with or without action oriented terms. These positive associations show that firms disclosing security risk factors are perceived to improve the valuation of the firm independent of the textual contents. We next consider how the market reaction differs with the disclosed security risk factors following a breach announcement.

6.1. Market reaction and disclosed security risk factors

For the rest of the analyses, we focus only on the firms with breach announcements. We use *Eventus* to compute the cumulative abnormal return (CAR) for each of the breach announcements

around the announcement date by applying the market model (details are given in Appendix C). The CAR estimates in our sample are used as the dependent variable in our estimation procedures. The average of the CARs across all events in our sample is -0.15% ($p < 0.10$) in the window (-1, +1), where -1 (+1) denote one day before (after) the breach announcement date.

Table 4. Variable Definitions

Variable	Definition
CAR	*CAR* is the market reactions to security breach announcements. Details are given in Appendix C.
Size	*Size* of a firm is the logarithm of the firm's total assets in the annual report before the breach announcement (data item AT in Compustat).
Industry	*Industry* with SIC code 73. We chose to control for the SIC code 73 because about 41% of the breached firms were in this industry category while the other 59% belongs to 20 different industry categories.
DNAct_Sec_Dis	*DNAct_Sec_Dis* is a dummy variable, equals 1 if the firm disclosed security risk factors without action-oriented terms, 0 otherwise.
DMatch	*DMatch* is a dummy variable representing whether the disclosed security risk factors are realized subsequently, equals 1 if there is a match, 0 otherwise
PMatch	*PMatch* measures the percentage of the disclosed security risk factors that are realized subsequently.
DSec_Dis	*DSec_Dis* is a dummy variable, equals 1 if the firm has security risk factors disclosed in financial reports, 0 otherwise.
Other_Dis	*Other_Dis* represents the risk factors disclosed in financial reports other than security risk factors. This variable controls for risk factor disclosing tendency of a firm and a firm's future uncertainty in general.

We investigate the association between a firm having action-oriented terms in disclosed security risk factors and the market reactions to security incidents (*CAR*). For this analysis, we focus only on firms that have disclosed security risk factors. Our variable of interest in this analysis is a dummy *DNAct_Sec_Dis*, which is set to 1 if no action-oriented term is present. To determine whether there is any action-oriented term in the disclosed security risk factor, we use the same feature provided by the text mining tool as mentioned in Section 4.3. If the tool indicates that a verb identified in Section 4.3, such as implement, prevent, act, is present in the disclosed security risk factor, we categorize the security risk factor as action-oriented; otherwise, it is categorized as not action-oriented. Column 2 in Table 5, identified as Model (1), shows the coefficient estimates for the dummy and all other control variables (all the variables used in our

regressions are defined in Table 4). Notice that the coefficient estimate for *DNAct_Sec_Dis* is significantly negative. It implies that when a firm discloses without action-oriented terms in security risk factors but has a breach announcement, there is a statistically significant negative market reaction. Thus, it appears that after the breach announcements, the market further takes into account the breached information and realizes the differences between the two types of internal information. As a robustness check, we also conduct an analysis, where the variable representing whether a disclosed security risk factor is action-oriented or not is based on the results from the classification model. If in the classification model a disclosed security risk factor is categorized as Group A, it is treated as an action-oriented disclosure in our estimation, but otherwise not. Even in this case, the coefficient of *DNAct_Sec_Dis* is significantly negative.

Table 5. Results for the Cross-Sectional Analysis

Variables	Model (1)	Model (2)	Model (3)
Intercept	-0.0201	-0.0290	-0.0421
Size	0.0008	0.0013	0.0019
Industry	0.0053	-0.0075	-0.0076
DMatch		**-0.0513*****	
PMatch			**-0.0718****
DNAct_Sec_Dis	**-0.0387****		
DSec_Dis		-0.0048	-0.0109
Other_Dis	0.0011	0.0007	0.0006
Adj R^2	0.06	0.13	0.09
N	26	88	88

* significant at 10% ** significant at 5% ***significant at 1%
Note: Since the impacts of consecutive events are not clear, we exclude the observations of consecutive events and follow-up reports such as the denial-of-service attack in February 2000.

Our results are consistent with the market efficient hypothesis. The market appears to treat security risk factors disclosures as improving the valuation of the organization at the time when the financial reports are released. The improvement is independent of whether the nature of internal information is positive or negative. These observations appear to be consistent with the literature on disclosure theory. Our results also point out that, after breach announcements, the market appears to distinguish between the two types of internal information. Firms with non-

action oriented terms, which possibly indicate that disclosures were meant to avoid future litigation costs, suffered a significantly more negative market reaction than firms with action-oriented terms. This result can be attributed to the fact that, when disclosures are meant to reduce future litigation costs, the realization of the event indicates the realization of possible future litigation costs and reduction in the valuation of the firm.

Our classification model results suggested that the disclosure patterns imply the occurrence of future incidents. However, Model (1) only presented the results at the aggregate level but did not take into account the relation between the disclosed and the realized information security risks. Using Model (2) and Model (3), we evaluate the impact of a match between disclosed risk factors and the realized event. For these analyses, we use all the firms with breach announcements. Again, some of them have multiple disclosed security risk factors but others might not have any. In Table 5, we focus on two variables: *DMatch* and *PMatch*. The dummy variable *DMatch* was set to 1 if any of the disclosed security risk factors is realized while *PMatch* measures the percentage of the disclosed security risk factors that are realized subsequently. Column 3 and column 4 in Table 5 show the coefficient estimates for *DMatch* (-0.0513) and *PMatch* (-0.0513) and other control variables. The significant negative coefficients of *DMatch* and *PMatch* suggest that when the disclosed security risk factors are realized, there is a statistically significant negative market reaction. These results seem to indicate that the market reacts to breaches after taking into consideration the textual contents of the disclosed security risk factors.

6.2. Robustness tests for the cross-sectional analysis

We performed several robustness tests to control for additional factors that might affect our results and to validate our cross-sectional findings. First, since the average market reaction is not zero, as suggested by previous literature (e.g., Brown and Warner 1985; Fama and French 1993), we also used the Fama-French three factor model (see Appendix C) to estimate the market

reaction and perform the same set of analyses. Our results were largely the same.

Second, we controlled for the following variables that could potentially affect market responses to security incidents: attack history, incident types (namely, confidentiality, integrity, and availability type incidents), previous disclosure patterns, i.e., the number of security risk factors disclosed one year before the annual report we considered, and the time (in months) between annual report release date and breach announcements. Our results remained similar.

Third, according to Core (2001) and Leuz and Verrecchia (2000), the disclosure decision could be an endogenous variable. Accordingly, we performed a two-stage least square (2SLS) analysis to estimate the disclosure decision and to verify our results. We found that the results for Model (2) and (3) in Table 5 are similar but not for Model (1) in Table 5. However, given the number of observations we used in our analysis, this two-stage result needs to be interpreted with caution.

Last, we validated our results by verifying if our results also hold for other firms without any reported incidents (see, for example, Shadish et al. 2002). We determined, for every breached firm, one of its publicly-traded competitors that did not have any breach announcements from Yahoo! Finance and the Hoover's Database. We then performed the same analyses but did not find any significant results. Therefore, we can rule out other possible explanations and make sure that we have captured the relation between disclosed security risk factors and breach announcements.

6. Conclusions and Discussion

We often observe that firms disclose information security risk factors in their financial reports. However, as mentioned in the Introduction, it was not ex ante clear whether the disclosed security risk factors could show preparedness for such threats or indicate potential litigation/reputation costs. In order to clarify the issue mentioned above, our paper builds a classification model to investigate the relation between the textual contents of information

security risk factors disclosed in the financial reports and the possibility of future security incidents. The classification model demonstrates that the textual content of security risk factors is a good predictor of future breaches. Building on this, we further consider the characteristics of security risk factors. We find that firms, which disclose actionable information when they provide information security risk factors in annual reports, are less likely to be associated with security incidents.

We also examined how the market reacts to these security risk factors and if there are any differences due to the nature of disclosures. Our cross-sectional analysis found that, generally, the market reaction is positive immediately after the release of financial reports. However, after security breaches, market reaction varied with the nature of disclosure. As noted earlier, both these observations are consistent with the prior theories on disclosure. They are also consistent with our earlier findings from the classification model.

In summary, both sets of analysis show that market participants can determine the differences between the two types of internal information by focusing on the textual contents of the disclosed security risk factors at the time when the financial reports are released. By doing so, market participants can better evaluate the firm's performance and future uncertainty regarding information security. Our analysis is also useful to managers involved with disclosure decisions in that it provides insights regarding how the market treats the disclosures. One obvious question is: would it be in the interest of a manager to alter the firm's disclosures in light of these results? Specifically, would it be better for a manager to disclose action oriented terms even though its internal information is different? We do not believe that it is their interest to alter the disclosures particularly since the main purpose of the disclosures in such situations is to avoid future litigation costs.

Another contribution of the paper is the process employed in the paper to demonstrate the association between the disclosures and the realizations of the events. Even though we have

applied it to predict one kind of operational risk, which is the information security risk, the process may be used to analyze other types also. For example, it may be used for predicting the realization of an employee strike, etc.

Our paper is not without its limitations. First, in addition to a binary indicator of breach announcement as the classifier, we also considered using the textual contents of the breach announcements as the classifier. However, we did not find any distinct pattern across breach announcements which might result from the way how the media reported security breaches. Therefore, the textual content of the breach announcement cannot be used in our context. Second, our sample size for analysis is relatively small. Although we attempt to capture as large of a sample as possible, it is still problematic to collect a larger dataset based on our filtering processes. A larger dataset for security incidents might allow us to increase the prediction accuracy of the classification model and to have better estimates in the cross-sectional analysis section. Furthermore, many firms might suffer from information security incidents that are not disclosed to the public. Obviously, we are unable to incorporate this information into our sample. Third, we implicitly assume that the stock price truly reflects a firm's business value. Although the stock price for high-tech firms might be biased, we only look at the price change in a short time period. Thus, we believe that our results still hold even with this possibility that the high-tech firms' stock price is not fairly reflected. Fourth, we adopt a simple coding scheme for the disclosures. Although we believe that a more complicated coding scheme does not alter our main results, a finer coding scheme for all the disclosures about business risks that can be applied to different industries may provide more details than the present scheme. Lastly, our model for the cross-sectional analysis implicitly assumes that the disclosures affect CARs which is typical in the literature. However, the disclosures can affect the CARs and the CARs also affect a firm's subsequent disclosure decisions. Our model does not capture this interaction effect which is still an open question in the disclosure literature.

Possible future extensions are as follows. First, in our paper, we implicitly assume that the disclosures are creditable and truly reflect a firm's practices. However, some firms might disclose lots of information but invest little. On the other hand, some other firms might invest substantially in information security but refuse to disclose such investments to the public. Therefore, this anomaly is worth further investigation. Second, a larger dataset can be used to provide more meaningful text mining results for both information security risk factors and business risk factors. The text mining analysis of business risk factors can also provide a first glance on how these risks affect different businesses. Last, as different media becomes popular information sources for investors, we can further consider other media sources, such as blogs, to investigate the relation among different information sources, information security incidents, and stock price reactions.

References

Abrahamson, E., C. Park. 1994. Concealment of negative organizational outcomes: an agency theory perspective. *Academy of Management J.* **37**(5) 1302-1334.

Acquisti, A., A. Friedman, R. Telang. 2008. Is there a cost to privacy breaches? an event study. Working Paper, Carnegie Mellon University.

Baesens, B., R. Setiono, C. Mues, J. Vanthienen. 2003. Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Sci.* **49**(3) 312-329.

Bagnoli, M., and Watts, S. G. 2007. Financial reporting and supplemental voluntary disclosures. *J. of Accounting Res.* **45**(5) 885-913.

Balakrishnan, K., Ghose, A., and Ipeirotis, P. 2008. The impact of information disclosure on stock market returns: the Sarbanes-Oxley act and the role of media as an information intermediary. *Proc. of the Seventh Workshop on the Econom. of Information Security* (WEIS 2008).

Bettman, J. R., B. A. Weitz. 1983. Attributions in the board room: causal reasoning in corporate annual reports. *Administrative Sci. Quart.* **28**(2) 165-183.

Cameron, A. C., P. K. Trivedi. 2007. *Microeconometrics: methods and applications*, Cambridge University Press, New York.

Campbell, K., L. A. Gordon, M. P. Loeb, L. Zhou. 2003. The economic cost of publicly announced information security breaches: empirical evidences from the stock market. *J. of Computer Security* **11** 431-448.

Cavusoglu, H., B. Mishra, S. Raghunathan. 2004. The effect of internet security breach announcements on market value of breached firms and internet security developers. *Internat. J. of Electronic Commerce* **9**(1) 69-105.

Cecchini, M., H. Aytug, G. J. Koehler, P. Pathak. 2007. Detecting management fraud in public companies. Working Paper, University of South Carolina.

Core, J. E. 2001. A review of the empirical disclosure literature: discussion. *J. of Accounting and Econom.* **31**(1-3) 441-456.

Cosslett, S. R. 1981. Maximum likelihood estimator for choice based samples. *Econometrica* **49**(5) 1289-1316.

Dye, R. A. 1985. Disclosure of Nonproprietary Information. *J. of Accounting Res.* 12(1) 123-145.

Ettredge, M. L., V. J. Richardson. 2003. Information transfer among internet firms: the case of hacker attacks. *J.of Inform. Systems* **17**(2) 71-82.

Fama, E., K. French. 1992. The cross-section of expected stock returns. *J. of Finance* **47**(2) 427–465.

Fan, W., I. Davidson, B. Zadrozny, P. S. Yu. 2005. An improved categorization of classifier's sensitivity on sample selection bias. *5th IEEE Internat. Conf. on Data Mining*, Houston, Texas, USA.

Fan, W., L. Wallace, S. Rich, Z. Zhang. 2006. Tapping the power of text mining. *Comm. of the ACM* **49**(9) 77-82.

FASB. 2008. *Statement of financial accounting concepts No.2*, Financial Accounting Standard Board.

Feldman, R., J. Sanger. 2006. *The text mining handbook: advanced approaches in analyzing unstructured data*, UK: Cambridge University Press.

Feng, L. 2008. Annual report readability, current earnings, and earnings persistent. *J. of Accounting and Econom.* **45**(2-3) 221-247.

Field, L., M. Lowry, S. Shu. 2005. Does disclosure deter or trigger litigation? *J. of Accounting and Econom.* **39** 487-507.

Firtel, K. B. 1999, Plain English: a repraisal of the intended audience of disclosure under the securities act of 1933. *Southern California Law Review* **72** 851–897.

Francis, J., K. Schipper, L. Vincent. 2002. Expanded disclosures and the increased usefulness of earnings announcements. *The Accounting Rev.* **77**(3) 515-546.

Frey, L., D. Fisher. 1999. Modeling decision tree performance with the power law. *Proc. of the 7th Internat. Workshop on Artificial Intelligence and Statistics*, San Francisco, CA 59-65.

Gal-Or, E., A. Ghose. 2005. The economic incentives for sharing security information. *Information Systems Res*. **16(**2) 186-208.

Garg, A., J. Curtis, H. Halper. 2003. Quantifying the financial impact of IT security breaches. *Inform. Management & Computer Security* **11**(2) 74-83.

Goodhue, D. L., D. W. Straub. 1991. Security concerns of system users: a study of perceptions of the adequacy of security. *Information & Management* **20**(1) 13-27.

Gordon, L. A., M. P. Loeb. 2002. The economics of information security investment. *ACM Transac. on Inform. and System Security* **5**(4) 438-457.

Gordon, L. A., M. P. Loeb, W. Lucyshyn. 2003. Sharing information on computer systems security: an economic analysis. *J.of Accounting and Public Policy* **22**(6) 461-485.

Gordon, L. A., M. P. Loeb, W. Lucyshyn, T. Sohail. 2006. The impact of the Sarbanes-Oxley Act on the corporate disclosures of information security activities. *J. of Accounting and Public Policy* **25** 503-530.

Goto, M., T. Kawamura, K. Wakai, M. Ando, M. Endoh, Y. Tomino. 2008. Risk stratification for progression of IgA nephropathy using a decision tree induction algorithm. *Nephrology Dialysis Transplantation* **24**(4) 1242-1247.

Grossman, S. J. 1981. The information role of warranties and private disclosure about product quality. *J. of Law and Econom.* **24**(3) 461-483.

Han, J., R. Altman, V. Kumar, H. Mannila, D. Pregibon. 2002. Emerging scientific applications in data mining. *Comm. of the ACM* **45**(8) 54-58.

Hovav, A., J. D'Arcy. 2003. The impact of denial-of-service attack announcements on the market value of firms. *Risk Management and Insurance Rev.* **6**(2) 97-121.

Hughes, J., S. Pae. 2004. Voluntary disclosure of precision information. *J. of Accounting and Econom.* **37** (3) 261–289.

Imbens, G. 1992. An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica* **60**(5) 1187-1214.

John, G., P. Langley. 1996. Static versus dynamic sampling for data mining. *Proc. of the 2nd Internat. Conf. on Knowledge Discovery and Data Mining*, Portland 367-370.

Jorgensen, B. N., M. T. Kirschenheiter. 2003. Discretionary risk disclosures. *The Accounting Rev.* **78**(2) 449-469.

Kannan, K., J. Rees, S. Sridhar. 2007. Market reactions to information security breach announcements: an empirical study. *Internat. J. of Electronic Commerce* **12**(1) 69-91.

Kasznik, R., B. Lev. 1995. To warn or not to warn: management disclosures in the face of an earnings surprise. *The Accounting Rev.* **70**(1) 113-134.

King, G., L. Zeng. 2001. Logistic regression in rare events data. *Political Analysis* **9**(2) 137-163.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. of the 14th Internat. Joint Conf. on Artificial Intelligence*, Montréal, Québec, Canada 781-787.

Kohl's. 2007. Annual report for the year ended February 3, 2007. Retrieved November 30, 2008 from http://www.kohlscorporation.com/InvestorRelations/pdfs/10k.pdf.

Krippendorff, K. 2003. Content analysis: an introduction to its methodology, Sage Publications, Inc.

Lancaster, T., G. Imbens. 1991. Choice based sampling: inference and optimality. Working Paper, Department of Economics, Brown University.

Lang, M. H., R. J. Lundholm. 1993. Cross-sectional determinants of analyst ratings of corporate disclosures. *J. of Accounting Res*. **31** 216-271.
Lang, M. H., R. J. Lundholm. 1996. Corporate disclosure policy and analyst behavior. *The Accounting Rev.* **71**(4) 467-492.

Landsman, W., E. Maydew. 2002. Has the information content of quarterly earnings announcements declined in the past three decades? *J. of Accounting Res.* **40**(3) 797-807.
Leuz, C., R. E. Verrecchia. 2000. The economic consequences of increased disclosure. *J. of Accounting Res*. **38**(3) 91-124.
Lev, B., S. H. Pennman. 1990. Voluntary forecast disclosure, nondisclosure, and stock prices. *J. of Accounting Res*. **28**(1) 49-76.
Long, W. J., J. L. Griffith, H. P. Selker, R. B. D'agostino. 1993. A comparison of logistic regression to decision-tree induction in a medical domain. *Computers and Biomedical Res.* **26** 74-97.
Milgrom, P. R. 1981. Good news and bad news: representation theorems and applications. *Bell J. of Econom*. **12**(2) 380-391.
Morgan J., R. Daugherty, A. Hilchie, B. Carey. 2003. Sample size and modeling accuracy with decision tree based data mining tools. *Academy of Information and Management Sci. J.* **6**(2) 77-92.
Penno, M. 1997. Information quality and voluntary disclosure. *The Accounting Rev.* **72**(2) 275-284.
Pinczowski, D., A. Ekbom, J. Baron, J. Yuen, H. Adami. 1994. Risk factors for colorectal cancer in patients with ulcerative colitis: a case-control study. *Gastroenterology* **107**(1) 117-120.
Rudolfer, S. M., G. Paliouras, I. S. Peers. 1999. A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome. *Computers and Biomedical Res.* **32** 391-414.
SAS Institute Inc. 2004. *Getting started with SAS® 9.1 text miner*. Cary, NC: SAS Institute Inc.
Shadish, W. R., T. D. Cook, D. T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. NY: Houghton Mifflin Company.
Shmueli, G., O. Koppius. 2009. The challenge of prediction in inofmraion systems research. Working Paper, University of Maryland, currently under 3rd round review at *MIS Quarterly*.
Siponen, M. 2006. Information security standards focus on the existence of process, not its content. *Comm. of the ACM* **49**(8) 97-100.
Siponen, M., J. Iivari. 2006. Six design theories for IS security policies and guidelines. *J. of the AIS* **7**(7) 445-472.
Skinner, D. J. 1994. Why firms voluntarily disclose bad news. *J. of Accounting Res.* **32**(1) 38-60.
Smyth, P. 2000. Model selection for probabilistic clustering using crossvalidated likelihood. *Statistics and Computing* **10** 63-72.
Sohail, T. 2006. *To tell or not to tell: market value of voluntary disclosures of information security activities*. Unpublished doctoral dissertation, University of Maryland, Maryland.
Sorenson, O., T. Stuart. 2001. Syndication networks and the spatial distribution of venture capital investment. *The American J. of Sociology* **106**(6) 1546-1588.
Steinberg, G. D., B. S. Carter, T. H. Beaty, B. Childs, P. C. Walsh. 2006. Family history and the risk of prostate cancer. *The Prostate* **17**(4) 337-347.
Still, S., W. Bialek. 2004. How many clusters? an information-theoretic perspective, *Neural Computation* **16** 2483-2506.
Stocken, P. 2000. Credibility of voluntary disclosure. *RAND J. of Econom.* **31**(2) 359-374.
Straub, D. W. 1990. Effective IS security: an empirical study. *Information Systems Res.* **1**(3) 255-276

Tanaka, H., K. Matsuura, O. Sudoh. 2005. Vulnerability and information security investment: an empirical analysis of e-local government in Japan. *J. of Accounting and Public Policy* **24**(1) 37-59.

Tetlock, P., Saar-Tsechansky, M., and Macskassy, S. 2008. More than words: quantifying language to measure firm's fundamentals. *J. of Finance* **63** 1437-1467.

Tibshirani, R., G. Walther., T. Hastie. 2001. Estimating the number of clusters in a dataset via the Gap statistic, *J. of the Royal Statistical Society B* **63**(2) 411-423.

Verrecchia, R. E. 1983. Discretionary disclosure. *J. of Accounting and Econom.* **5**(3) 179-194.

Verrecchia, R. E. 2001. Essays on disclosures. *J. of Accounting and Econom.* **32**(1-3) 97-180.

Weiss, S. M., L. Kapouleas. 1989. An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. *Proc. of the 11th Internat. Joint Conf. on Artificial Intelligence*, Detroit, Michigan, 781-787.

Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. *Proc. of the 21st Internat. Conf. on Machine Learning*, Banff, Canada, 903-910.

Zhou, Z., Y. Jiang. 2004. NeC4.5: Neural Ensemble Based C4.5. *IEEE Transac. on Knowledge and Data Engineering* **16**(6) 770-773.

Appendix A. An Example of Security Risk Factor

"System Interruption and the Lack of Integration and Redundancy in Our Systems May Affect Our Sales

Customer access to our Web sites directly affects the volume of goods we sell and thus affects our net sales. We experience occasional system interruptions that make our Web sites unavailable or prevent us from efficiently fulfilling orders, which may reduce our net sales and the attractiveness of our products and services. To prevent system interruptions, we continually need to: add additional software and hardware; upgrade our systems and network infrastructure to accommodate both increased traffic on our Web sites and increased sales volume; and integrate our systems.

Our computer and communications systems and operations could be damaged or interrupted by fire, flood, power loss, telecommunications failure, break-ins, earthquake and similar events. We do not have backup systems or a formal disaster recovery plan, and we may have inadequate insurance coverage or insurance limits to compensate us for losses from a major interruption. Computer viruses, physical or electronic break-ins and similar disruptions could cause system interruptions, delays and loss of critical data and could prevent us from providing services and accepting and fulfilling customer orders. If this were to occur, it could damage our reputation."

Excerpt from Amazon's annual report for year 2000, retrieved on Apr.23, 2007
Source: http://www.sec.gov/Archives/edgar/data/1018724/000103221001500087/0001032210-01-500087.txt

Appendix B. Cluster Analysis, Concept Links, and Clusters before and after Breach Announcements

The cluster analysis is performed as follows using SAS® 9.1 Text Miner. First, text parsing decomposes the sentences into terms and creates a frequency matrix as a quantitative representation of the input documents. When decomposing the docum	12.7%	0.1410

ents, we choose to rule out definite as well as indefinite articles, conjunctions, auxiliaries, and interjections since these terms do not help provide meaningful results in our context. This matrix also shows the weight for the terms. The weight for term *i* in document *j* (w_{ij}) is the multip		

lication of the frequency weight (L_{ij}) and the term weight (G_i). In our study, the frequency weight is the logarithm of the frequency (f_{ij}) of term i in document j plus one, i.e., $L_{ij} = \log_2 (f_{ij} + 1)$. The term weight of term i (G_i) is calculated as *1+*system			
7	adversely, code, +program, +sale, +store	3.8%	0.1314
8	+assurance, fraud, +internal controls, +policy, +statement	3.8%	0.1092
9	+business, +cause, identity, +risk, +theft	2.5%	0.1137
After Security Breach Announcements			
1	+business, information, not, security, +service	45.3%	0.177

2	+computer, +experience, +failure, +interruption, +result	25.0%	0.171
3	+disruption, +interruption, +loss, +telecommunication, +system	23.4%	0.164
4	+attack, + harm, + have, other, + type	6.3%	0.152

Appendix C. Stock Price Reactions to Information Security Incidents

In our study, the market model is used to capture the impact of security incidents.

$$R_{it} = \beta_0 + \beta_1 R_{mt} + \varepsilon_{it} \qquad \text{(C-1)}$$

where R_{it} denotes company i's return of the common stock at period t which equals to $(p_t - p_{t-1}) / p_t$. Dividends and stock splits are excluded here because (1) they are rare events and (2) we have already considered confounding events. Thus, stock return of a certain company equals to the change in stock price or the capital gain. R_{mt} stands for the corresponding market return at period t and is estimated by the CRSP equally weighted index. The CRSP equally weighted index is the average of the returns of all trading stocks in NYSE, AMEX and NASDAQ. β_0 and β_1 are the parameters and estimated in a 255-day periods ending at 45 days before the estimation window we choose by ordinary least square (OLS) method. We calculate the abnormal return (AR) from the market model: