Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Tawei Wang, Karthik N. Kannan, Jackie Rees Ulmer,

To cite this article:
Tawei Wang, Karthik N. Kannan, Jackie Rees Ulmer, (2013) The Association Between the Disclosure and the Realization of Information Security Risk Factors. Information Systems Research 24(2):201-218. http://dx.doi.org/10.1287/isre.1120.0437

Information Systems Research
Vol. 24, No. 2, June 2013, pp. 201–218
ISSN 1047-7047 (print) | ISSN 1526-5536 (online)

http://dx.doi.org/10.1287/isre.1120.0437

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

Tawei Wang
School of Accountancy, Shidler College of Business, University of Hawaii at Manoa, Honolulu, Hawaii 96822, twwang@hawaii.edu

Karthik N. Kannan, Jackie Rees Ulmer
Krannert Graduate School of Management, Center for Education and Research in Information Assurance and Security (CERIAS), Purdue University, West Lafayette, Indiana 47907
{kkarthik@purdue.edu, jrees@purdue.edu}

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 evaluate how the nature of the disclosed security risk factors, believed to represent the firm's internal information regarding information security, is associated with future breach announcements reported in the media. 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 disclosed security risk factors with risk-mitigation themes are less likely to be related to future breach announcements. We also investigate how the market interprets the nature of information security risk factors in annual reports. We find that the market reaction following the security breach announcement is different depending on the nature of the preceding 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.

Key words: information security; information security incident; risk factor; text mining
History: Ram Gopal, Senior Editor; Ravi Bapna, Associate Editor. This paper was received September 2, 2010, and was with the authors for 7 months for 2 revisions. Published online in *Articles in Advance* October 5, 2012.

1. Introduction

Information technology (IT) plays a critical role in most organizations. Any event or action that could potentially threaten the security of IT resources, including the data and information within those resources, is of great concern. Firms recognize and often publicly announce the risks that information security threats pose to the smooth functioning of IT systems. For example, Kohl's states in its 2009 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 2010, p. 11). Our paper aims to study how the nature of disclosures regarding information security reflects a firm's security management strategy and is correlated with future realization of breach announcements.

The Basel II framework identifies three major types of risks faced by a firm: credit, market, and operational (BCBS 2001b). Operational risks are more internal to the firm whereas credit and market risks are influenced by factors external to the firm. IT and systems-related risks are classified as an operational risk type (BCBS 2001a). Prediction and management of market and credit risks are relatively easier because "data is plentiful" and "well established tools" are available (Schuermann 2005, p. 123). On the contrary, not only is the management of operational risk more involved but also has been drawing attention lately as businesses become more complicated (Jobst 2007). In fact, "...operational risk has a greater potential to transpire in greater and more harmful ways than many other sources of risk..." (Jobst 2007, p. 3). Moreover, because operational risks "...tend to be idiosyncratic to a particular institution..." (Herring 2002, p. 43), their assessment is also difficult. Given the increasing prominence of operational risks and the lack of a standard set of tools to deal with operational risks, investors are seeking to gain insights about the actual realization of these risks and how a firm manages them. Our paper demonstrates how insights about the link between the firm's risk-management strategy and the realizations of at least for one type of operational risk, the information security risk, may

be derived from public disclosures. Note that even in the information security domain, the assessment of the risks is recognized to be difficult (e.g., Straub and Welke 1998, GAO 1999).

In general, disclosures are consequences of a firm's internal information (e.g., Verrecchia 1983, Dye 1985, Skinner 1994, Kasznik and Lev 1995). On one hand, if the internal information is positive in nature, the firm may disclose so as to improve its valuation (e.g., Verrecchia 1983, Dye 1985). On the other hand, if the internal information is negative, the firm may still disclose risk factors with the aim of reducing litigation costs associated with possible future adverse events (e.g., Skinner 1994). In the information security context specifically, a firm may disclose positive internal information to indicate its preparedness in facing security threats, or negative internal information to reduce future litigation costs associated with future security breaches. The distinction between the two types of internal information is important for investors, debtors, and customers. However, it is not ex ante clear how the nature of disclosure associates with positive or negative internal information in the information security context. Given that prior literature has shown that internal information is often reflected in the nature of the disclosure (e.g., Bettman and Weitz 1983, Li 2008), this paper studies the relation between the nature of information security disclosures and the reported security incidents affecting the firm.

Our paper consists of two parts. The first part studies how the nature of information security risk factors disclosed in annual reports is associated with the occurrence of future security breach announcements, and the second part evaluates if the differences in the textual contents of disclosures explain the differences in market reactions to security breach reports. For the first part, we text mined the contents of security risk factors disclosed prior to breach announcements and developed a decision tree classification model to associate the mined textual content with security breach announcements. We also verified the decision tree results using econometric techniques that help overcome sample selection issues. For the second part, we used the results from the decision tree model to examine market reactions to the disclosed security risk factors. An interesting aspect of this paper is that we draw upon a diverse set of tools to address the research questions. Our choices of methodologies depend on whether we predict or explain the relations and are consistent with Shmueli and Koppius (2010).

We believe that our analysis of disclosures is quite different from those studied in the finance and accounting literatures. Studies in those literatures have mainly investigated the implications of textual contents of disclosures on firm performance (e.g., how the "risk sentiment" in financial reports affects stock returns). Thus, they can only indirectly infer about the impact of such internal information associated with the disclosures by using performance measures such as earnings. In our context, however, we directly infer about the realization of the risk (i.e., security breach) given the risk-management strategy revealed in the disclosures before breach announcements. Moreover, the realization of risks in general has not attracted the attention of researchers until recently. The prior lack of interest may be explained because the credit and market risks are the most common types of risks encountered by the firms and, as mentioned earlier, the tools to handle them are already well developed. On the contrary, management of operational risks is increasingly becoming important yet they are lacking well-established tools for management. Given these, and because operational risks are idiosyncratic to the firm, any systematic means of evaluating the realization of such risks is valuable. Our paper deals with realization of one type of operational risk, information security risk, that can be revealed from the textual contents of the disclosure that reflect a firm's security management strategy. To the best of our knowledge, this issue has not been studied previously.

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 §2. The data collection process is provided in §3. Next, in §4, we analyze the textual data of the disclosed information security risk factors. We present the results for the market reactions to reported breach announcments in §5. In §6, we conclude with a 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 the management and 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 research on the management and economics of information security. Papers have analyzed security investment decisions and also 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. Other 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. It is not only the design of the policies and investment decisions that are important to firms but also how they publicly communicate risk assessments of their activities.[1] It is this aspect that has received little prior attention and that we focus on in our paper.

Prior research has investigated the impact of information security breaches on a firm's market value. A few papers show that security breach reports lead to a significant negative market valuation (e.g., Ettredge and Richardson 2003, Garg et al. 2003, Cavusoglu et al. 2004, Acquisti et al. 2006), although others do not find any impact (e.g., Campbell et al. 2003, Hovav and D'Arcy 2003, Kannan et al. 2007). The second part of our paper is different from these prior works in that it seeks to explain if the variations in the market reactions can be explained by the nature of disclosures. It builds on our primary analysis, where we study the association between the nature of information security risk factors disclosed and subsequent security incidents as reported in the major media.

2.2. Disclosures in Accounting

There is a rich body of literature in accounting that examines disclosures (see Verrecchia 2001 and Dye 2001 for a comprehensive review and discussion). Verrecchia (2001) classifies the prior works on disclosures as association-, discretionary-, and efficiency-based research. The first type studies the association between the disclosure and metrics such as price and trading volume; the second type analyzes a firm's decision to disclose information; and the last type investigates the disclosure strategy without ex ante preferences. The main analysis of our paper (i.e., the classification model) is closer to the discretionary-based disclosures type research in that we examine the nature of disclosure. The analysis of market reactions to reported breach announcements, which acts as a support to the classification model, relates to the association-based research.

Early research on the motivation to disclose has shown that when there is no cost to disclose, full disclosure exists because investors believe that nondisclosing firms have the worst possible information (e.g., Grossman 1981, Milgrom 1981). However, when disclosure is costly, firms 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., Francis et al. 1994, Skinner 1994, Kasznik and Lev 1995, Field et al. 2005, Rogers et al. 2010). As we mentioned in the Introduction, we are interested in the distinction between the negative and positive internal information in the information security context.

Some disclosures are mandatory (e.g., the Sarbanes-Oxley Act of 2002 or SOX) and others are voluntary. Jorgensen and Kirschenheiter (2003) model voluntary disclosure decisions and find that firms with less future uncertainty will choose to disclose risk factors. We take the motivations for voluntary disclosure of risks into account when testing the robustness of our results.

Because our paper analyzes the textual contents of disclosures, we survey the related literature. Prior work in this area can be broadly categorized into two groups. The first includes papers that have analyzed the relation between disclosures and internal information. For example, Abrahamson and Park (1994) focus on the concealment of poor earnings performance in the letters to shareholders and demonstrate that a firm uses more negative words to explain when there is a larger decline of performance. Another example in this group, Li (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, Li (2007) counts the number of risk-related words in annual reports and suggests that investors can profit from the firms that have a significant change in the number of risk-related words from year to year. Davis et al. (2008) focus on the linguistic style (tone) and show that the stock market can react to different linguistic styles. 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. However, Loughran and McDonald (2011) build on Tetlock et al. (2008) and show that a different negative word list can better capture the tone in financial text. Kothari et al. (2009) investigate news, analyst reports, and annual reports and show that when the information is more positive, the stock price volatility is smaller. Our paper is quite different from prior work in inferring the nature of internal information based on realization of events and the nature of disclosure.

[1] Often, security investment decisions, policies, and actions are closely guarded by organizations in order to avoid exposing their residual vulnerabilities. This secrecy may seem counter to the disclosures in annual reports. Note that disclosures in annual reports are simply statements of risk and not specific policies.

2.3. Literature Combining Both Streams of Research

In this paper, we link both streams of research. To the best of our knowledge, Gordon et al. (2010) is the only study that also links these two streams. In their paper, they demonstrate 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 breaches reported in the media, i.e., 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 reports. In addition, our paper analyzes how the market reaction to reported information security breaches is dependent upon the nature of disclosures.

3. Conceptual Model and Data Collection

To develop theories for our analysis, we consider a three-staged overarching conceptual model, even though our analysis will pertain only to the last two stages. The first stage considers a firm's decision to disclose. As mentioned earlier, information security is recognized as a serious threat by firms. Hence, firms seek to avoid, mitigate, transfer, or accept risks arising from those threats consistent with the risk-management literature (e.g., Crouhy et al. 2006). For example, firms mitigate risks, by implementing technologies to protect digital assets, adopting stringent security policies, etc. They can transfer the residual security risks by purchasing insurance. Also, firms may realize that risks are part of today's business environment and accept them. In any case, firms do not publicize the specifics of how they manage their security risks for fear that revealed information may be exploited to inflict serious damage. This means that investors are also uncertain regarding the specific information security risk-management activities engaged in by firms. In such a scenario, prior literature (e.g., Verrecchia 1983, 2001; Gordon et al. 2010) argues that a firm's disclosures mainly help a firm reduce information asymmetry between the firm and investors, which lowers the discount rate (the denominator effect) as its uncertainty about handling of the risks decreases. Thus, in the information security context, if a firm has taken actions to mitigate risks, it has an incentive to disclose the positive internal information about its risk assessment. Even when a firm decides not to fully mitigate the risks (i.e., to accept the risks), it discloses the negative internal information so that it can avoid the future litigation costs. (See Appendix A, in the online appendix, for an example of a security risk factor disclosed in an annual report. The online appendix to this paper is available as part of the online version at http://dx.doi.org/10.1287/isre.1120.0437.) In both cases, i.e., when a firm indicates preparedness (positive internal information) or indicates a possibility of future litigation (negative internal information), disclosures should improve its valuation. Consistent with theory, Gordon et al. (2010) find the firm's valuation to improve after disclosure. Note that prior literature does not distinguish the market reaction with the type or motivation for disclosure. Also, given that the first stage is beyond the scope of our paper, our paper assumes that the disclosures are given and focuses on the second and the third stages.

The distinction between the two types of internal information may be recognized when the disclosed risks are realized, which we refer to as occurring in the second stage. In our data set, the risk realization corresponds to security breach announcements in news media.[2] In particular, we are interested in the correlation between the textual contents of the risk factors disclosed and the subsequent actual realizations. We expect disclosures that reflect a firm's active security management strategy will less likely be associated with breach announcements, whereas those that list security risk concerns or vulnerabilities, which may be motivated to limit litigation costs, i.e., accept risks, will more likely have a breach announcement.

The analysis in the third stage is motivated based on prior theory and is used to supplement our classification results. Following an event related to the disclosure, the beliefs may be revised because of new information, and the market reacts (e.g., Bamber et al. 1999). We expect the firm's valuation to primarily be impacted by the numerator effect (e.g., customers losing trust in the firm's ability to store data according to Casey 2004 and Pavlou et al. 2007). More specifically, we expect the numerator effect to interact with the nature of disclosure as follows. Compared to disclosures motivated by positive internal information, market reactions to breach announcements in the context of disclosures because of negative internal information should be more negative. This is because, immediately following a breach, the trust in a firm's ability to store data may be questioned more so given the lack of any action toward managing the risk in the prebreach disclosure; whereas, it may not be the case with positive internal information.

[2] In our data set, none of the articles were news releases from the firm. Therefore, our references to breach announcements and breach reports are independent articles appearing in the news media.

Our methodology choice for both stages follows Shmueli and Koppius (2010), who recommend the use of econometric analysis (such as regressions) for explaining the role of a variable, and predictive models (such as decision trees) for predictions and associations. Thus, we primarily employ a classification model to analyze the second stage (for the sake of robustness, we also present analyses using econometric techniques in §4.2).

"The loss distribution of operational risk [–of which information security is part of–] is heavy-tailed, i.e., there is a higher chance of an extreme loss event (with high loss severity) than the asymptotic tail behavior of standard limit distributions would suggest" (Jobst 2007, p. 4). In the information security setting, Wang et al. (2008) have shown similar characteristics to hold as well. When events occur rarely (compared to nonevents), the most suitable data collection procedure is the endogenous stratified sampling method (e.g., Cosslett 1981, Cameron and Trivedi 2007). Examples of scenarios where the method has been employed include wars, venture capital investments, and epidemiological infections (e.g., King and Zeng 2001, Sorenson and Stuart 2001). With rare events, econometricians have found that 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 process resulted in 101 firm-event observations from 62 firms. The maximum number of security risk factors disclosed in financial reports was 4 and the minimum was 0 with a mean of 0.6 and a standard deviation of 1.04. From the sample size, we can reasonably conclude that information security breach announcements regarding publically traded firms are rare.

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.[3] Note that some firms did not have any security risk factors disclosed in the annual report and others had several. Using this process, we collected 43 security risk factors, each corresponding to a breach announcement.[4]

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 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 predictive model, in this case a decision tree as discussed next, 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 §4.2.

Another decision in the third step is how to sample nonevents. We could either randomly sample nonbreached firms or collect a set of matching firms with no breaches and execute our analysis. We decided to present the results from random sampling of nonevents because it was conducive to study other splits of events and nonevents as well. We did not find any significant deviations in the results from using matched nonbreached firms. We collected 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,

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

[4] 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 data set. Additionally, we counted each of the disclosures separately and ran our analysis, and our results were consistent.

Table 1 Descriptive Statistics

Panel A. Firm characteristics

	Firms with breach announcements			Firms without breach announcements			Mean difference	Median difference
	Mean	Std. dev.	Median	Mean	Std. dev.	Median	Difference (t-stat)	Difference (χ^2)
Total assets	89,985.12	258,469.93	6,366.95	46,736.31	136,076.06	5,820.13	43,248.81 (0.932)	546.82 (0.079)
Debt/total assets	0.60	0.31	0.58	0.67	0.27	0.65	−0.07 (−0.163)	−0.07 (1.159)
Earnings per share	1.05	2.57	1.04	1.38	4.12	1.62	−0.33 (−0.000)	−0.58 (0.971)
Institutional ownership	0.57	0.26	0.64	0.57	0.21	0.59	0.00 (0.275)	0.05 (0.143)
ROA (return on assets)	0.03	0.16	0.05	−0.02	0.42	0.06	0.05 (0.524)	−0.01 (0.484)
Tangible ratio	0.43	0.29	0.39	0.46	0.29	0.47	−0.03 (−0.637)	−0.08 (0.541)

Panel B. Industry breakdown

Firms with breach announcements			Firms without breach announcements		
Two-digit SIC code	Description	%	Two-digit SIC code	Description	%
48	Communication	9.8	27	Printing and publishing	8.9
60	Depository institutions	8.2	35	Machinery	8.9
73	Business services	26.2	38	Measuring instruments	8.9
Other 18 industries		55.8	73	Business services	31.1
			Other 12 industries		42.2

Note. See http://www.osha.gov/pls/imis/sic_manual.html for detailed information about the description of industries.

not all firms had security risk factors in the annual report and a few firms had several.

From these three steps, our data set involves 124 $(62 + 62)$ firms and 77 $(43 + 34)$ information security risk factors. The descriptive statistics of our sample (event and nonevent firms) are given in Table 1. We also compared the mean and the median between these two groups but did not find the groups (statistically) significantly different. In addition, we performed the Kolmogorov-Smirnov test regarding the industry distribution of the firms across the two groups but again did not find differences (p-value was 0.766).

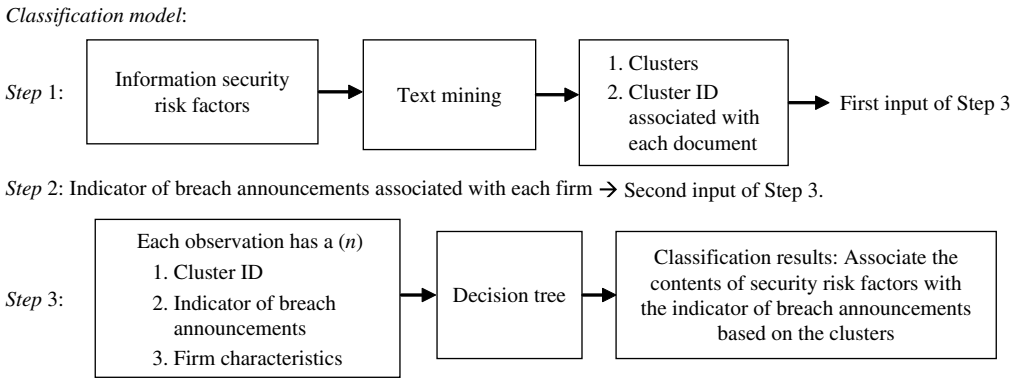
4. Text-Mining and Classification Model

For the second-stage analysis, we combine text-mining techniques with a decision tree model. We mine the textual data so as to understand the information conveyed by disclosed security risk factors. In general, text-mining 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 to classify news stories, detect fraud, and improve customer support (e.g., Han et al. 2002, Fan et al. 2006, Cecchini et al. 2010). Our paper employs text-mining techniques to categorize the elements of security risk factors that relate to future incident announcements. We establish the relation between the text categories and future incident announcements by constructing a classification model, specifically, a decision tree model. We chose the decision tree model for the following reasons. 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). Moreover, 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 the stratified sampling method than logistic models. Furthermore, 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

The steps to obtain our decision tree model are shown in Figure 1 and expanded next. For building the model, we use the 77 security risk factors collected. As the first step, the risk factors are fed as input

Figure 1 Process Flow for the Classification Model



to the SAS Text Miner to form clusters based on the textual contents of the disclosed security risk factors. For forming the clusters, we employed the expectation-maximization (EM) algorithm. The key idea in the algorithm is that it maximizes the possibility of cluster membership of the overall (the sum of the) Gaussian distribution of all the clusters by grouping text with similar contents into the same cluster using the weighted frequency of the terms. More details about the algorithm are available in Appendix B.1 in the online appendix. This procedure resulted in every firm's disclosure being assigned a cluster ID. If a firm did not disclose any security risk factors, no cluster ID was assigned to it in the analysis presented later. When we assigned a cluster ID to nondisclosures for robustness, our decision tree performed even better.

The second step in Figure 1 sets an indicator variable for each firm and is meant to capture if the firm had a breach announcement. The outputs from the first two steps are fed as inputs to the third step.

The third step builds a decision tree to classify the indicator for breach announcements (from Step 2) based on the cluster ID (from Step 1). As stated later in the robustness tests, we considered other variables, such as firm size in Step 3 but those variables do not result in any new branches (see also our discussion on the two-stage least square procedure). For building the decision tree, our original data set was partitioned into two subsets for building the decision tree: 80% was used for training, and the other 20% for validation and testing. 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 in 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. Hence, we experimentally varied the number of clusters and repeated the three steps in Figure 1 until the error rate of the decision tree model in Step 3 was minimized. To compute the errors, we employed a commonly adopted procedure called 10-fold cross validation (e.g., Weiss and Kapouleas 1989, Kohavi 1995), which is similar to boot-strapping methods employed for econometric analysis. Accordingly, we repeat Step 3 in Figure 1 10 times, each time with a different randomly chosen training, validation and testing data set.

We executed the entire procedure many times and found the results to be robust. A sample set of results from the 10-fold cross-validation are shown in Table 2 and the overall accuracy rate is 77.42% (i.e., 45.16% + 32.26%). Across a number of iterations, we found the average accuracy rate for the validations is about 76%. We also performed a chi-square test to take into account the effect of type I and type II errors. The test (chi-square of 18.87) shows that our prediction does not occur by normal random chance; the corresponding p value is < 0.01.

Corresponding to the smallest error rate in Step 3, it turns out that there are four clusters in Step 1. The resulting decision tree involves two main branches: whether the disclosure belongs to cluster 1 or 2, or to cluster 3 or 4. Disclosures belonging to clusters 1 and 2 were typically associated with no breach

Table 2 Confusion Matrix of the Cross Validation Results

			Predict		
Breach announcement			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

Table 3 Cross Validation Results by Clusters

		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Breach announcement	(%)	Training	Validation and testing	Training	Validation and testing	Training	Validation and testing	Training	Validation and testing
Yes	Freq	3	0	3.5	1	16.5	3	12	3
		23.08	0.00	25.92	18.18	84.62	85.71	75.00	100.00
No	Freq	10	3	10	4.5	3	0.5	4	0
		76.92	100.00	74.07	81.82	15.38	14.29	25.00	0.00
Total	Freq	13	3	13.5	5.5	19.5	3.5	16	3
		100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

announcement, and clusters 3 and 4 with breach announcements. Table 3 shows the association with each cluster on average across the 10 iterations of our cross-validation procedure. About 75% of the documents in clusters 1 and 2 together are associated with "no breach announcement" and about 80% of the documents in clusters 3 and 4 together are associated with "breach announcement."

We next studied how the textual contents differ between the two disclosure groups: clusters 1 and 2 together, which we hereafter refer to as group A, and clusters 3 and 4 together, group B. We used the SAS Text Miner tool to identify co-occurring terms within each group. Table 4 shows the most frequently co-occurring terms within each group along with their frequency percentages (as percentages of all other sets of words) and the root mean squared standard deviations (which measures how close the setoff words are. See Appendix B.1 in the online appendix for more details). In the table, a term with the plus (+) sign represents a group of equivalent terms; for example, both "ability" and "abilities" are considered equivalent.

When we compare the roles (e.g., verb, noun, or an adjective) played by the terms in each group, we observe a striking difference between the two groups. For group A, the first set of words contains 40% nouns and 60% verbs, and the second set contains 60% nouns and 40% verbs. For group B, the first two sets are all (100%) nouns, 80% of the words in the third set are nouns and 20% adjectives, and the last set contains 80% nouns and 20% prepositions. Note that group B does not have any verbs but has a higher percentage of nouns compared to group A. Recall that group A corresponds to the no breach announcement group, while group B to the breach announcement group. Thus, consistent with the conceptual model, it appears that firms in group A have taken actions to avoid, mitigate, or transfer risks, but firms in group B have chosen to accept risks.

Table 4 Text Mining Results of Information Security Risk Factors

Co-occurring group	Terms	Percentage	Root mean Square Std.
	Group A (clusters 1 and 2 together)		
1	+resource, +virus, +require, +implement, +damage	55.7	0.1113
2	+prevent, +customer, +disruption, +process, +act,	44.3	0.1127
	Group B (clusters 3 and 4 together)		
1	+disaster, +telecommunication, +interruption, +failure, +loss	26.0	0.1496
2	+revenue, +virus, +security, information, +breach	26.0	0.1555
3	+number, +user, +business, +security, other	28.0	0.1508
4	+event, with, +result, +business, +loss	20.0	0.1472

To gain further understanding about the clusters from the decision tree, we employ the content analysis methodology. The content analysis literature, such as Krippendorff (2003), recommends employing both qualitative and quantitative analyses in order to reach appropriate conclusions. The quantitative analysis generally starts with word counts and transforms the textual contents into statistical data, sometimes based on dictionaries (e.g., Krippendorff 2003, Tetlock 2007, Tetlock et al. 2008). On the contrary, the qualitative analysis requires the researcher to transform the textual data into subjective interpretations/patterns or sense-making meanings (e.g., Patton 2002, Hsieh and Shannon 2005). In our case, for the quantitative analysis we used the General Inquirer software and for the qualitative analysis, we interpreted the textual contents. For each group/cluster, we present our insights (see Appendix C in the online appendix for additional details regarding the quantitative measures).

4.1.1. Cluster 1 (part of group A). These disclosures can be described as technical (specific to security activities), brief, and action oriented. They seem to correspond to risk-mitigation activities by disclosing firms. This cluster had 16 disclosures, with an average of 123.5 words per disclosure, and was the shortest of all four clusters in terms of number of words per disclosure, of all four clusters. The quantitative analysis revealed that Cluster 1 disclosures had the highest percentage of interpretive verbs (IAV) at 10.05%,

and also high percentages of strong (Strng) words (14.24%), active (Actv) words (12.06%) as well as negative (Neg) words (8.38%). The disclosures were primarily oriented toward the availability of IT resources to customers, employees, and vendors, as well as confidentiality and integrity concerns. They focus on the type of attack or problem that might occur, and what, in general, the firm has deployed to prevent such problems from occurring.

4.1.2. Cluster 2 (part of group A). These disclosures can be best described as emphasizing the project management challenges relevant to IT and how these challenges affect prevention, detection, and recovery activities within the overall context of business growth. The theme of these disclosures is risk mitigation but with some degree of risk acceptance. There were 19 disclosures in this cluster, with an average of 149.37 words per disclosure and is the second shortest cluster. The quantitative analysis revealed that these disclosures have the highest percentage of strong (Strng) words at 17.15% as well as the highest percentage of active (Actv) words of all of the clusters at 12.86%. This was also the most negative cluster as measured by negative (Neg) words (9.43%) and interestingly the most positive (Pos) words at 6.17% as compared to the other clusters. There were also a relatively high percentage of interpretive verbs (IAV) at 9.95%. These disclosures are more focused on concerns about systems implementation, migration, and integration projects. Thus, they also mention about availability concerns, business disruptions, and liability concerns.

4.1.3. Cluster 3 (part of group B). These disclosures can be described as indicating problems that have either occurred or indications of how a firm intends to respond to a future attack, as opposed to indicating what defenses are employed by firms. The disclosures in this cluster have the theme of risk acceptance, from both past and potential events. This cluster had 19 disclosures and an average of 299.16 words per disclosure, the longest among the clusters. The formal content analysis shows that the cluster has high percentages of strong (Strng) words at 14.03% and active (Actv) at 10.73%, but lower than clusters 1 and 2. This cluster has the highest percentage of economic (Econ) words of any cluster at 7.59% and the highest percentage of means words (e.g., "access," "by," "sale," etc.) (6.11%) and our/us/we (Our) words at 7.43% of all of the clusters. Many of them mention previous security problems, such as data breaches. Many also point out that the company may be lacking sufficient backup/recovery systems or they are underinsured for potential losses. The disclosures also consistently mention that the firms might need to spend more money in the future because of experienced or forecasted security problems.

4.1.4. Cluster 4 (part of group B). These disclosures seem to describe external threats (such as from vendor/supply chain partners) and can be considered as risk accepting. There were 23 disclosures, with an average of 196.87 words per disclosure, which is the second longest among the clusters. The quantitative analysis of the disclosures shows the lowest percent of strong (Strng) words (12.89%) of all four clusters. Availability concerns are mentioned in 18 out of 23 disclosures. There is also a theme of mentioning the risk from activities of vendors, supply chain partners, and other third parties and the impact on the firm (15 out of 23 disclosures). There are also frequent mention of locations, and how specific locations contain unique risks, such as fire, earthquakes, etc.

Clusters 1 and 2 (which correspond to group A) from Step 1 seem to be associated with disclosures that are process, implementation, or actions oriented; and clusters 3 and 4 (group B) seem to be mainly associated with disclosures about occurrences of problems. Interestingly, consistent with the disclosure theory, clusters in group B are associated with breach announcements in the decision tree model.

4.2. Robustness Tests of the Classification Results

We test the robustness of the analyses in this subsection. Broadly, we address concerns (in the following order) regarding the settings with which the original classification results were generated, the lack of econometric techniques that can account for biases in stratified sampling, and possible endogeneity problems arising from missing variables or because of the choice of nonevents in our data set.

The following five sets of analyses were executed to test the sensitivity of our original classification model to various settings employed in building our classification model. First, we investigated the sensitivity of the results to the settings of the text-mining tool. In our base case analysis, the terms were weighted differently during the clustering process depending on the frequency of their occurrences, as described in Appendix B.1 in the online appendix. 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 improperly interpret

the security risk factor. 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, the firms that did not have any disclosed security risk factors were excluded in the original data set. We assigned a new cluster ID to the nondisclosing firms and constructed the decision tree. Group A was now associated with three clusters: the new cluster, cluster 1, and cluster 2, and 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 improved to 79.9%.

Fourth, we controlled for (1) firm size; (2) industries; (3) the type of breach, i.e., confidentiality, integrity, or availability (e.g., Gordon et al. 2006); and (4) historical security risk factor disclosures; but our results remained similar.

Fifth, 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, respectively. Specifically, we randomly sampled 38, 138, and 238 nonbreached 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.

Next, we considered econometric techniques to address biases in the stratified sampling procedures. Note that although techniques for correcting biases in the stratified samples have existed for some time, only recently there have been developments to deal with discrete dependent variables. Given that our dependent variable (whether or not a firm has a reported security breach) is Boolean we employ techniques from King and Zeng (2001) and Ramalho and Ramalho (2006) for our analysis. Both papers are similar in that they correct the sampling bias in endogenous sampling by reconsidering the unbiased distribution of the data. King and Zeng (2001) focus on logistic regression in the application context of wars, and the Ramalho and Ramalho (2006) paper provides a formal and general proof of the approach as applicable to normal linear and probit models. Following King and Zeng (2001) (also see Imai et al. 2008, 2009 for the estimation procedures), we estimated the rare event logistic regression (the bias corrected logistic regression). Consistent with our classification result, the bias-corrected model suggests that membership in group B increases the possibility of a reported security breach by 1.5304 ($p < 0.01$) relative to membership in group A. We also implemented the generalized method of moments (GMM) correction from Ramalho and Ramalho (2006) in our context to correct the bias. Specifically, we used a probit model to investigate the association between the clusters from the textual content (namely, group A and group B) and information security breach by using the GMM estimation process. The result also shows that membership in group B increases the possibility of a reported security breach by 0.369 ($p < 0.01$) compared to membership in group A.

We investigated if there is a relation between mandatory (e.g., 10-K filings) and voluntary (e.g., risk factors) disclosures, as Hughes and Pae (2004) and Bagnoli and Watts (2007) note. For example, when the earnings performance disclosures in the mandatory 10-K filings are poor, firms may not disclose many voluntary security risk factors, and vice versa. Therefore, it could be argued that information security risk factors may depend on the earnings performance. So, we analyzed the earnings performance in the 10-K report and the number of information security risk factors disclosed but we did not find any significant association between them.

Lastly, we account for possible endogeneity in our classification model results. One could argue that we missed firm-specific characteristics, which affect both disclosures and realized events. In response, as mentioned earlier, we executed the classification model where the nonevent samples included firms with no security breach reports but were also matched with breached firms on factors such as the same SIC (standard industrial classification) code and the most similar market capitalization. In particular, first, we collected a list of possible control firms within the same industry as the firms in the experiment group, i.e., the same four-digit SIC code industry. We ruled out control firms that had any breach announcement in our sample period (because these were the primary confounding issue for our analysis on the nature of disclosure). Following that, we chose a control firm that had a similar size as its corresponding firm in the experiment group. The size measure we used was

market capitalization. As an additional test, we also used total assets as a measure of size but find the results similar. Next, we present the analysis based on the control firms using market capitalization as the size measure when performing the matching. See Appendix D in the online appendix for the cross-validation result of the matched sample. The result shows that 74.1% (43 out of 58) observations are correctly predicted, and thus our earlier results were qualitatively validated. To further address the endogeneity issue, we conducted the analysis using a two-stage estimation model as well as propensity score matching.

In our two-stage estimation model, we follow Larcker and Rusticus (2010) and consider endogenous disclosure decisions. They suggest that the instrument variables be decided based on prior works. So, for our analysis, we consider Gordon et al. (2010), which argues that a firm's information security disclosure decision is dependent on firm size, return on assets, industry, long-term assets percentage, liquidity, volatility of stock returns, analyst followings, and institutional ownership. From these variables, we identify two instruments that affect a firm's disclosure decisions of action or nonaction oriented information, but also that are least likely to be related to breach announcements: return on assets (*ROA*), and long-term assets divided by total assets (*Tangible*). We account for *ROA* because firms with higher *ROA* should have better corporate governance mechanisms that lead to different security management strategies. However, it is not clear that because of a higher *ROA*, a firm becomes a target of security breach. Similarly, firms with a higher tangible assets ratio tend to have smaller uncertainty in general compared to the firms with a high percentage of intangible assets (Field et al. 2005) and so, the variable *Tangible* accounts for a firm's ability to allocate resources to operationalize their security strategies. Again, it is not clear that firms with more property, plants, and equipment and more long-term investments are more or less likely to have reported breach announcements. Although firm size and industry are also likely to affect the disclosure decision in our context, they are likely to be associated with the reported breach announcements as well. We consider them as control variables. Specifically, our models are given in Equations (2SLS-1) and (2SLS-2-1):

$$\begin{aligned} DAct\_Sec\_Dis_{it} = {} & \alpha_0 + \alpha_1 ROA_{it-1} + \alpha_2 Tangible_{it-1} \\ & + \alpha_3 Size_{it-1} + \alpha_4 Industry_{it-1} \\ & + \alpha_5 SOX_{it-1} + \varepsilon_{it}. \end{aligned} \tag{2SLS-1}$$

$$\begin{aligned} Breach_{it} = {} & \beta_0 + \beta_1 DAct\_Sec\_Dis^*_{it-1} \\ & + \beta_2 Size_{it-1} + \beta_3 Industry_{it-1} \\ & + \beta_4 SOX_{it-1} + \mu_{it}, \end{aligned} \tag{2SLS-2-1}$$

where the subscript i denotes for firm i, time t is at the end of the fiscal period, and time $t-1$ is at the beginning of the fiscal period. The variable $DAct\_Sec\_Dis_{it}$ is a dummy used to represent whether the disclosure is action oriented or not. The control variables are: $Size_{it-1}$ is the logarithm of total assets because larger firms could easily become a target of attacks, $Industry_{it-1}$ is a dummy variable indicating whether a firm is a Web-based firm (SIC code 73), and SOX_{it-1} to represent whether the observation is after the Sarbanes-Oxley Act was enacted, i.e., 2004. We account for SOX_{it-1} because breach announcements may become less likely as firms improve their internal controls in response to SOX. The variable, $Breach_{it}$, is a dummy indicating whether there is a breach announcement or not, while $DAct\_Sec\_Dis^*_{it-1}$ is the predicted probability of whether a disclosure is action oriented or not.

Before we present our estimates, we present some evidence supporting our two-stage model. At first glance, the Hausman test ($F = 2.49$, $p = 0.105$) shows that the two-stage model is not more preferable than the one-stage model. However, Larcker and Rusticus (2010) argue that the Hausman test itself is not sufficient to test for the appropriateness of the instrument variables. Our partial R^2 (0.34) and the F-statistic (14.44) from the first stage are reasonably high, which indicate acceptable instruments (Larcker and Rusticus 2010, Stock et al. 2002, Stock and Yogo 2005). Also, because we use multiple instruments, we test for over identification but find no support for it (chi-square value of 1.2; $p > 0.10$). The results of estimates from our two-stage model are given in Table 5. Consistent with the results in the main analysis, disclosing with action-oriented terms can reduce the possibility of security breaches (-1.74 and -0.59, $p < 0.05$ and $p < 0.01$) by using both regular logistic regression and rare event bias-corrected logistic regression.

Next, we describe the procedure we conduct for propensity score matching. For this procedure, we use a logistic model to estimate the tendency (propensity score) of being in disclosure group A or disclosure group B. The dependent variable of this model is a dummy variable indicating whether a firm belongs to the disclosure group A (i.e., cluster 1 or 2) or disclosure group B (i.e., cluster 3 or 4). The independent variables are firm size (total assets), industry, profitability (ROA) and tangible asset ratio (*Tangible*). Then we perform a standard nearest-neighbor matching algorithm based on the propensity scores (Painter 2004). This algorithm identifies the control firm ("nearest neighbor") for each firm in the experimental group using the propensity scores. The resulting control group is verified

Table 5 Results for the Two-Stage Model

	Regular model		Bias-corrected model	
	First stage	Second stage	First stage	Second stage
Intercept	−**0.87***	2.15	−**0.87***	**1.13****
	(−**1.83**)	(1.13)	(−**1.83**)	(**2.22**)
*DAct_Sec_Dis**		−**1.74****		−**0.59*****
		(−**2.28**)		(−**2.87**)
Instruments				
ROA	−**2.59*****		−**2.59*****	
	(−**5.31**)		(−**5.31**)	
Tangible	−**0.42****		−**0.42****	
	(−**2.13**)		(−**2.13**)	
Control variables				
Size	**0.06*****	−0.09	**0.06*****	−0.03
	(**3.01**)	(−1.04)	(**3.01**)	(−1.15)
Industry	0.05	**0.78****	0.05	**0.28****
	(0.43)	(**2.03**)	(0.43)	(**2.25**)
SOX	**0.27***	−0.14	**0.27***	−0.10
	(**1.74**)	(−0.24)	(**1.74**)	(−0.58)
N	62	62	62	62
Partial R^2	0.34		0.34	
Partial *F*-statistic	14.44		14.44	
Over-identifying restriction test	$\chi^2 = 1.20$ ($p = 0.273$)		$\chi^2 = 1.20$ ($p = 0.273$)	
Hausman test	$F = 2.49$ ($p = 0.105$)		$F = 2.49$ ($p = 0.105$)	

Note. *t*-statistics are in parentheses.
*Significant at 10%, **significant at 5%.

to be similar to the experimental group.[5] Again, we then use the results from the first stage to reperform the classification analysis and the market reactions to reported security breaches. For the classification analysis, we use a regular and a bias-corrected logistic regression to investigate the association between the occurrence of breach announcements and the predicted disclosure group. The result is given in Table 6 and the variables are defined as earlier. As shown in the table, *DAct_Sec_Dis** is statistically significant, indicating that having action-oriented terms is negatively associated with the possibility of breach announcements.

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, the firms are less likely to be associated with future incidents. This result is important for market participants to understand the internal information related to information security. We next study if and how the market reacts to the differences in the textual contents, and this result appears in §5.

[5] After the matching, the mean difference of *ROA*, *Tangible*, *Size*, *Industry* across groups are 0.08 ($p = 0.09$), 0.03 ($p = 0.80$), −0.17 ($p = 0.86$), −0.30 ($p = 0.11$).

5. Market Reactions to Breach Announcements Given Disclosed Security Risk Factors

The analysis in this section corresponds to the third stage of our conceptual model. Our interest in this analysis is to explore how well the differences in the nature of disclosure shed light on the market reactions to security incidents. This analysis is also useful in providing validity to the previous results regarding the nature of disclosure.

As mentioned earlier, in the first stage, following security related disclosures, Gordon et al. (2010) find that market valuations improve. There is no theoretical background pointing to one type of internal information being better than the other. Consistent with that, we did not find any difference in the market valuation improvements at the time of disclosure between the two groups from the classification model. However, in the third stage, after a breach is reported, theory predicts the revision of market valuations because of the new information regarding the breach (Bamber et al. 1999). Specifically at this stage, we expect to observe differences because of the nature of information. That is, given the nature of the disclosed security risk factors, investors have some ideas about a firm's security risk-management postures and the corresponding concerns. The news

Table 6 Results of the Propensity Score Matching

Dependent variable: Breach announcement	Regular logistic regression (chi-squared statistics in parenthesis)			Bias-corrected logistic regression (z statistics in parenthesis)		
Intercept	−3.29 (−2.50)	5.72 (0.93)	5.72 (0.93)	**−2.79**** (**−2.12**)	4.81 (0.78)	4.81 (0.780)
*DAct_Sec_Dis**	**−1.39**** (**−2.32**)	**−1.48**** (**−2.10**)	**−1.48**** (**−2.10**)	**−1.16*** (**−1.95**)	**−1.14*** (**−1.67**)	**−1.14*** (**−1.67**)
Size		−0.40 (−0.15)	−0.40 (−0.15)		−0.33 (−1.17)	−0.33 (−1.17)
SOX			NA			NA
N	35	35	35	35	35	35
AIC	40.79	40.47	40.47	40.79	40.47	40.47

Note. We are not able to estimate the coefficient for *SOX* because of the variability of the data.

(i.e., the breach announcement) provides additional information about the realization of the security risks that would affect the market's valuation. There is a general consensus that breaches cannot be completely prevented. Given that, investors should have greater confidence in firms that take preventive action. Hence, we expect the market responses to be more negative for the nonaction oriented disclosures than the action-oriented ones.

5.1. Market Reaction and Disclosed Security Risk Factors

For the rest of the analyses, we focus only on the firms with reported breaches. 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 E in the online appendix). 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.[6] We investigate the association between a firm having action-oriented (risk-mitigating) terms in disclosed security risk factors and the market reactions to security incidents (*CAR*). We studied the impact of disclosures on CAR across other windows and found the results to be qualitatively valid.

Our variable of interest in this analysis is a dummy *DAct_Sec_Dis*, which is set to one if an action-oriented term is present. To determine whether there is any action-oriented term in the disclosed security risk factor, we use the results from the classification model. If in the classification model a disclosed security risk factor is categorized as group A (i.e., cluster 1 or 2), it is treated as an action-oriented disclosure in our estimation, but otherwise not. We used the variables mentioned in Table 7 in the regression. Column 2 in Table 8 shows the coefficient estimates for the first set of regression we ran. Notice that the coefficient estimate for *DAct_Sec_Dis* is significantly positive. It implies that when a firm discloses with action-oriented terms in security risk factors but has a breach announcement, the market reaction is less negative. Thus, it appears that after the breach announcements, the market further takes into account the breach information and realizes the differences between the two types of internal information. As a robustness check, we also conducted an analysis by using the feature provided by the text-mining tool discussed in §4.1. If the tool indicates that a verb (such as implement, prevent, or act) identified in §4.1, is present in the disclosure, we categorized the security risk factor as action oriented, and otherwise not. Even in this case, the coefficient of *DAct_Sec_Dis* is significantly positive. Thus, our results are found to be consistent with our expectation that the market appears to distinguish between the two types of internal information.

Model (1) only presents the results at the aggregate level but does 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 reported breaches. As before, some of the firms disclose multiple security risk factors but others do not disclose any. In Table 8, we focus on two variables: *DMatch* and *PMatch*. The dummy variable *DMatch* is set to one if any of the disclosed security risk factors is realized, and *PMatch* measures the percentage of the disclosed security risk factors that are realized subsequently. Model (2) and model (3) in Table 8 show the coefficient estimates for *DMatch* (-0.051 and -0.053) and *PMatch* (-0.072 and -0.080) 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 indicate that the market reacts

[6] We also investigated longer windows, such as $(-3, +3)$, $(-5, +5)$, and $(-7, +7)$ and found the results to be consistent.

Table 7 Variable Definitions

Variable	Definition
CAR	*CAR* is the market reactions to security breach announcements. Details are given in Appendix E.
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.
DAct_Sec_Dis	*DAct_Sec_Dis* is a dummy variable, equals one if the firm disclosed security risk factors with action-oriented terms, zero otherwise.
DMatch	*DMatch* is a dummy variable representing whether the disclosed security risk factors are realized subsequently, equals one if there is a match, zero 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 one if the firm has security risk factors disclosed in financial reports, zero 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.

Table 8 Results for the Market Reaction Given the Nature of Security Disclosures

Variables	Model (1)		Model (2)		Model (3)	
Intercept	−0.043	(−0.97)	−0.029	(−0.69)	−0.015	(−0.35)	−0.042	(−0.98)	−0.023	(−0.53)
Size	0.002	(1.34)	0.001	(0.77)	0.001	(0.49)	0.002	(1.12)	0.001	(0.77)
Industry	−0.005	(−0.51)	−0.008	(−0.82)	−0.006	(−0.61)	−0.008	(−0.80)	−0.005	(−0.56)
DMatch			**−0.051*****	**(−3.12)**	**−0.053*****	**(−3.53)**				
PMatch							**−0.072****	**(−2.26)**	**−0.080*****	**(−2.77)**
DAct_Sec_Dis	**0.023***	**(1.76)**			0.015	(1.18)			0.019	(1.50)
DSec_Dis			−0.005	(−0.34)			−0.011	(−0.77)		
Other_Dis	−0.001	(−1.02)	0.001	(1.12)	0.000	(0.411)	0.001	(0.94)	−0.000	(−0.18)
Adj R^2	0.04		0.13		0.15		0.09		0.11	
N	88		88		88		88		88	

Notes. *t* statistics are in parenthesis. Because 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.

*Significant at 10%, **significant at 5%, ***significant at 1%.

to breaches after taking into consideration the textual contents of the disclosed security risk factors.

5.2. Other Robustness Tests for the Market Reaction

We perform several robustness tests to control for additional factors that might affect our results and to validate our findings in §5. First, because the average market reaction is not zero, as suggested by previous literature (e.g., Brown and Warner 1985, Fama and French 1992), we also use the Fama-French three factor model (see Appendix E in the online appendix) to estimate the market reaction and perform the same set of analyses. Our results are largely similar.

Second, we control 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 remain similar.

Third, we validate our results by verifying if our results also hold for other firms without any reported incidents (see, for example, Shadish et al. 2002). For every breached firm, we identified from Yahoo! Finance or the Hoover's database one of the firm's publicly traded competitors that did not have any reported breach, after accounting for confounding events. We then performed the same analyses but did not find any significant deviations.

Lastly, here again, we account for endogeneity by performing a two-stage least square (2SLS) analysis as well as the propensity score matching. For the two-stage model, we retain the first stage as in Equation (2SLS-1) and the second stage model as in Equation (2SLS-2-2):

$$CAR_{it} = \gamma_0 + \gamma_1 DAct\_Sec\_Dis^{*}_{it-1} + \gamma_2 DSec\_Dis_{it-1} + \gamma_3 Other\_Dis_{it-1} + \zeta_{it}, \quad \text{(2SLS-2-2)}$$

where CAR_{it} is the three-day cumulative abnormal return around the breach announcement day as stated earlier; $DAct\_Sec\_Dis^{*}_{it-1}$ is the predicted disclosure decision from Equation (2SLS-1); $DSec\_Dis_{it-1}$ is a dummy to capture whether the firm has security risk

Table 9 Results for the Two-Stage Model (Market Reaction)

Dependent variable: CAR	
Intercept	0.002 (0.157)
*DAct_Sec_Dis**	**0.030* (1.718)**
DSec_Dis	**−0.026* (−1.981)**
Other_Dis	0.000 (0.197)
N	62
Adj. R^2	0.06

Note. t statistics are in parenthesis.

Table 10 Results for the Propensity Score Matching (Market Reaction)

Dependent variable: CAR	
Intercept	−0.021 (−0.951)
*DAct_Sec_Dis**	**0.013 (1.938)*****
DSec_Dis	−0.053 (−1.597)
Other_Dis	0.000 (0.206)
N	35
Adj. R^2	0.12

Note. t statistics are in parenthesis.

factors disclosed in financial reports; and $Other\_Dis_{it-1}$ represents the number of risk factors disclosed other than security risk factors.

The results in Table 9 demonstrate that the market reaction is smaller for the firms disclosing with action oriented terms (0.030, $p < 0.10$). Again, given the number of observations we used in our analysis, this two-stage result needs to be interpreted with caution. We also performed the propensity score matching for the market reactions to reported security breaches and found the results to be qualitatively similar as given in Table 10.

5.3. Comparison of the Security Risk Factors Disclosed Before and After Breach Announcement

Following a breach, some firms do change the security risk factors disclosed. So, we also investigate if there is any general pattern of change between the risk factors disclosed before and after the breach announcement. For this purpose, we only focused on firms with breach announcements. As expected, the co-occurring groups of words in the risk factors disclosed are different before and after the breach (see the two sets of co-occurring groups of words in Table 11). Following the breach, the co-occurring terms appear to indicate that firms disclose risks in more generic terms, such as "business," "computer," "disruption," and "attack." Furthermore, our analysis on how the words are linked with each other (by studying the "concept-links," see Appendix B.2 in the online appendix for details) indicates that the security risk factors become more diversified after breach announcements.

Table 11 Characteristics of Information Security Risk Factors Before and After Breach Announcements

Co-occurring group	Terms	Percentage	RMS std.
	Before security breach announcements		
1	+ disaster, + telecommunication, + interruption, + failure, + loss	26.0	0.1496
2	+ revenue, + virus, + security, information, + vulnerability	26.0	0.1555
3	+ number, + user, + business, + security, other	28.0	0.1508
4	+ event, with, + result, + business, + loss	20.0	0.1472
	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

6. Conclusions and Discussion

Operational risks are increasingly becoming important, yet they are quite difficult for investors to form an assessment of how the risks are handled by firms. Our paper focuses on one type of operational risk, the information security risk, and investigates the information conveyed in the related disclosures. Ex ante, from the disclosures, it was not clear whether the security risk factors indicated positive or negative internal information. To investigate this, our paper develops a classification model that relates the textual contents of information security risk factors disclosed in financial reports to the possibility of future security incidents. We find that the textual content of security risk factors is indeed a good predictor of future reported breaches. More specifically, we find that firms that disclose actionable (risk-mitigating) information are less likely to be associated with security incidents.

We also examined if there are any differences in how the market reacts because of the nature of disclosures. We find no significant difference in the market reaction to the nature of disclosure immediately following the release of financial reports. However, after security breach announcements, the market reaction varied with the nature of disclosure. These observations are consistent with the prior theories on disclosure. They also appear to support our findings from the classification model.

In summary, we show that market participants can determine differences between the two types of internal information by focusing on the textual contents of the disclosed security risk factors. 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. Our findings suggest that firms taking proactive action have an incentive to truthfully disclose their posture to information security. One obvious follow up 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 in their interest to alter the disclosures particularly because 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 also be used to analyze other types of risk. 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 patterns across breach announcements that might result from the way 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 data set based on our filtering processes. A larger data set for security incidents might allow us to increase the prediction accuracy of the classification model and to have better estimates of market reactions to security breaches. 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. Fourth, our model for the market reactions to security breaches 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. Last, it is possible that there are many other factors affecting information security related disclosures. Because of the data access limitation and the availability of supporting prior literature, we do not consider most of the possible factors.

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 in security risk mitigation activities. On the other hand, some other firms might invest substantially in information security but refuse to disclose such investments to the public. Therefore, this potential anomaly is worth further investigation. Second, a larger data set 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 an initial glimpse into how these risks affect different businesses. A larger data set can also allow us to perform a longitudinal analysis to further understand the nature of security-related disclosures. Last, as different types of media, such as social media and blogs, becomes more popular information sources for investors, we can extend our analysis to investigate the relation among different information sources, information security incidents, and stock price reactions.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/isre.1120.0437.

Acknowledgments

The authors would like to thank the review team for their comments and suggestions. The authors also thank the participants of the Big Ten Information Systems Symposium 2007, INFORMS 2007, 9th Annual Information Security Symposium, WEIS 2008, 2008 AMCIS doctoral consortium, 2008 ICIS doctoral consortium, workshops at Purdue University, Georgia Institute of Technology, National Chenchi University, and National Taipei University in Taiwan for feedback on the preliminary versions of this paper. Specific thanks go to Kemal Altinkemer, Sarah Rice, Ramanath Subramanyam, Susan Watts, and Pei-Chang Liao for their valuable comments. The authors would like to thank General Inquirer for granting us permission to use their software. The authors are grateful for the financial support from Purdue Research Foundation, National Taiwan University, I3P, and the Ministry of Education in Taiwan.

References

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

Acquisti A, Friedman A, Telang R (2006) Is there a cost to privacy breaches? An event study. *The Fifth Workshop on the Econom. Inform. Security (WEIS), Robinson College, University of Cambridge, London*.

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

Bagnoli M, Watts SG (2007) Financial reporting and supplemental voluntary disclosures. *J. Accounting Res.* 45(5):885–913.

Balakrishnan K, Ghose A, 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. Seventh Workshop on the Econom. Inform. Security* (*WEIS* 2008, *Hanover, New Hampshire*).

Bamber L, Barron OE, Stober TL (1999) Differential interpretations and trading volume. *J. Financial Quant. Anal.* 34(3):369–386.

Basel Commitee on Banking Supervision (BCBS) (2001a) *Operational risk*. Supporting Document to the New Basel Capital Accord. Bank for International Settlement, retrieved November 3, 2010 from http://www.bis.org/publ/bcbsca07.pdf

Basel Commitee on Banking Supervision (BCBS) (2001b) *Overview of the new Basel Capital Accord*. Bank for International Settlement, retrieved November 3, 2010 from http://www.bis.org/publ/bcbsca02.pdf

Bettman JR, Weitz BA (1983) Attributions in the board room: Causal reasoning in corporate annual reports. *Admin. Sci. Quart.* 28(2):165–183.

Breslow NE, Day NE (1980) The analysis of case-control studies. *Statistical Methods in Cancer Research* Chap. 1. (IARC Scientific Publications, Lyonm, France).

Brown S, Warner J (1985) Using daily stock returns: The case of event studies. *J. Financial Econom.* 14(1):3–31.

Cameron AC, Trivedi PK (2007) *Microeconometrics: Methods and Applications* (Cambridge University Press, New York).

Campbell K, Gordon LA, Loeb MP, Zhou L (2003) The economic cost of publicly announced information security breaches: Empirical evidences from the stock market. *J. Comput. Security* 11(3):431–448.

Casey E (2004) Reporting security breaches: A risk to be avoided or responsibility to be embraced? *Digital Investigation* 1(3):159–161.

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

Cecchini M, Aytug H, Koehler GJ, Pathak P (2010) Detecting management fraud in public companies. *MIS Quart.* 56(7):1146–1160.

Cosslett SR (1981) Maximum likelihood estimator for choice based samples. *Econometrica* 49(5):1289–1316.

Crouhy M, Galai D, Mark R (2006) *The Essentials of Risk Management* (McGraw Hill, New York).

Davis A, Piger J, Sedor L (2008) Beyond the numbers: Managers' use of optimistic and pessimistic tone in earnings press releases. *AAA Financial Accounting and Reporting (FARS) Mid-Year Meeting, Phoenix, AZ*.

Dye RA (1985) Disclosure of nonproprietary information. *J. Accounting Res.* 12(1):123–145.

Dye RA (2001) An evaluation of "essays on disclosure" and the disclosure literature in accounting. *J. Accounting Econom.* 32(1–3):181–235.

Ettredge ML, Richardson VJ (2003) Information transfer among Internet firms: The case of hacker attacks. *J. Inform. Systems* 17(2):71–82.

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

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

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

Feldman R, Sanger J (2006) *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data* (Cambridge University Press, UK).

Field L, Lowry M, Shu S (2005) Does disclosure deter or trigger litigation? *J. Accounting Econom.* 39(3):487–507.

Francis J, Philbrick D, Schipper K (1994) Shareholder litigation and corporate disclosures. *J. Accounting Res.* 32(2):137–164.

Frey L, Fisher D (1999) Modeling decision tree performance with the power law. Heckerman D, Whittaker J, eds. *Proc. 7th Internat. Workshop on Artificial Intelligence and Statist., Fort Lauderdale, FL*, 59–65.

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

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

Goodhue DL, Straub DW (1991) Security concerns of system users: A study of perceptions of the adequacy of security. *Inform. Management* 20(1):13–27.

Gordon LA, Loeb MP (2002) The economics of information security investment. *ACM Transac. Inform. System Security* 5(4):438–457.

Gordon LA, Loeb MP, Lucyshyn W (2003) Sharing information on computer systems security: An economic analysis. *J. Accounting and Public Policy* 22(6):461–485.

Gordon L, Loeb M, Sohail T (2010) Market value of voluntary disclosures concerning information security. *MIS Quart.* 34(3):567–594.

Gordon LA, Loeb MP, Lucyshyn W, Sohail T (2006) The impact of the Sarbanes-Oxley act on the corporate disclosures of information security activities. *J. Accounting and Public Policy* 25(5):503–530.

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

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

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

Herring RJ (2002) The basel 2 approach to bank operational risk: Regulation on the wrong track. *J. Risk Finance* 4(1):42–45.

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

Hsieh H, Shannon SE (2005) Three approaches to qualitative content analysis. *Qualitative Health Res.* 15(9):1277–1288.

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

Imai K, King G, Lau O (2008) Toward a common framework for statistical analysis and development. *J. Computational and Graphical Statist.* 17(4):892–913.

Imai K, King G, Lau O (2009) Zelig: Everyone's statistical software. Accessed November 3, 2010, http://gking.harvard.edu/zelig.

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

Jobst A (2007) Operational risk—The sting is still in the tail but the poison depends on the dose. *J. Operational Risk* 2(2):3–59.

John G, Langley P (1996) Static versus dynamic sampling for data mining. Simoudis E, Han J, Fayyad U, eds. *Proc. 2nd Internat. Conf. Knowledge Discovery and Data Mining, Portland, OR*, 367–370.

Jorgensen BN, Kirschenheiter MT (2003) Discretionary risk disclosures. *The Accounting Rev.* 78(2):449–469.

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

Kasznik R, Lev B (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, Zeng L (2001) Logistic regression in rare events data. *Political Anal.* 9(2):137–163.

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

Kohl's (2010) Annual report for the year ended January 30, 2010. Retrieved August 17, 2010 from http://www.sec.gov/Archives/edgar/data/885639/000119312510061795/d10k.htm#tx88612_3

Kothari S, Li X, Short J (2009) The effect of disclosures by management, analysts, and financial press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Rev.* 84(5):1639–1674.

Krippendorff K (2003) *Content Analysis: An Introduction to Its Methodology* (Sage Publications, Thousand Oaks, CA).

Lancaster T, Imbens G (1991) Choice based sampling: Inference and optimality. Working paper, Department of Economics, Brown University, Providence, RI.

Larcker D, Rusticus T (2010) On the use of instrumental variables in accounting research. *J. Accounting Econom.* 49(3):186–205.

Li F (2007) Do stock market investors understand the risk sentiment of corporate annual reports? Working paper, University of Michigan.

Li F (2008) Annual report readability, current earnings, and earnings persistent. *J. Accounting Econom.* 45(2–3):221–247.

Long WJ, Griffith JL, Selker HP, D'Agostino RB (1993) A comparison of logistic regression to decision-tree induction in a medical domain. *Comput. Biomedical Res.* 26(1):74–97.

Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries and 10-Ks. *J. Finance* 66(1):35–65.

Milgrom PR (1981) Good news and bad news: Representation theorems and applications. *Bell J. Econom.* 12(2):380–391.

Morgan J, Daugherty R, Hilchie A, Carey B (2003) Sample size and modeling accuracy with decision tree based data mining tools. *Acad. Inform. Management Sci. J.* 6(2):77–92.

Painter J (2004) SPSS macro for propensity score matching. Accessed November 3, 2010, http://ssw.unc.edu/VRC/Lectures/index.htm.

Patton MQ (2002) *Qualitative Research and Evaluation Methods* (Sage Publications, Thousand Oaks, CA).

Pavlou PA, Liang H, Xue Y (2007) Understanding and mitigating uncertainty in online exchange relationships: A principal–agent perspective. *MIS Quart.* 31(1):105–136.

Pinczowski D, Ekbom A, Baron J, Yuen J, Adami H (1994) Risk factors for colorectal cancer in patients with ulcerative colitis: A case-control study. *Gastroenterology* 107(1):117–120.

Ramalho EA, Ramalho JJS (2006) Bias-corrected moment-based estimators for parametric models under endogenous stratified sampling. *Econom. Rev.* 25(4):475–496.

Rogers J, Van Buskirk A, Zechman S (2010) Disclosure tone and shareholder litigation. *AAA Financial Accounting and Reporting (FARS) Mid-Year Meeting, San Diego, CA*.

Rudolfer SM, Paliouras G, Peers IS (1999) A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome. *Comput. Biomedical Res.* 32(5):391–414.

Schuermann T (2005) A review of recent books on credit risk. *J. Appl. Econometrics* 20(1):123–130.

Shadish WR, Cook TD, Campbell DT (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Houghton Mifflin Company, NY).

Shmueli G, Koppius O (2010) The challenge of prediction in information systems research. Working paper, University of Maryland.

Siponen M (2006) Information security standards focus on the existence of process, not its content. *Comm. ACM* 49(8):97–100.

Siponen M, Iivari J (2006) Six design theories for IS security policies and guidelines. *J. AIS* 7(7):445–472.

Skinner DJ (1994) Why firms voluntarily disclose bad news. *J. Accounting Res.* 32(1):38–60.

Smyth P (2000) Model selection for probabilistic clustering using crossvalidated likelihood. *Statist. Comput.* 10(1):63–72.

Sorenson O, Stuart T (2001) Syndication networks and the spatial distribution of venture capital investment. *Amer. J. Sociol.* 106(6):1546–1588.

Steinberg GD, Carter BS, Beaty TH, Childs B, Walsh PC (2006) Family history and the risk of prostate cancer. *The Prostate* 17(4):337–347.

Still S, Bialek W (2004) How many clusters? An information-theoretic perspective. *Neural Comput.* 16(12):2483–2506.

Stock JH, Yogo M (2005) Testing for weak instruments in linear IV regression. Stock JH, Andrews DWK, eds., *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*, Chap. 5. (Cambridge University Press, UK), 80–108.

Stock JH, Wright JH, Yogo M (2002) A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econom. Statist.* 20(4):518–529.

Straub DW (1990) Effective IS security: An empirical study. *Inform. Systems Res.* 1(3):255–276.

Straub DW, Welke R (1998) Coping with systems risk: Security planning models for management decision making. *MIS Quart.* 22(4):441–469.

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

Tetlock P (2007) Giving content to investor sentiment: The role of media in the stock market. *J. Finance* 62(3):1139–1168.

Tetlock P, Saar-Tsechansky M, Macskassy S (2008) More than words: Quantifying language to measure firm's fundamentals. *J. Finance* 63(3):1437–1467.

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. *J. Royal Statist. Soc. B* 63(2):411–423.

United States General Accounting Office (GAO) (1999) Information security risk assessment: Practices of leading organizations. Accessed November 3, 2010, http://www.gao.gov/special.pubs/ai00033.pdf

Verrecchia RE (1983) Discretionary disclosure. *J. Accounting Econom.* 5(3):179–194.

Verrecchia RE (2001) Essays on disclosures. *J. Accounting Econom.* 32(1–3):97–180.

Wang J, Chaudhury A, Rao HR (2008) A value-at-risk approach to information security investment. *Inform. Systems Res.* 19(1):106–120.

Weiss SM, Kapouleas L (1989) An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. Sridharan NS, ed. *Proc. 11th Internat. Joint Conf. Artificial Intelligence, Detroit*, 781–787.

Zadrozny B (2004) Learning and evaluating classifiers under sample selection bias. Brodley CE, ed. *Proc. 21st Internat. Conf. Machine Learn., Banff, Canada*, 903–910.

Zhou Z, Jiang Y (2004) NeC4.5: Neural ensemble based C4.5. *IEEE Transac. Knowledge and Data Engrg.* 16(6):770–773.