JOURNAL OF CYBERSECURITY

Journal of Cybersecurity, 2(1), 2016, 15–28
doi: 10.1093/cybsec/tyw004
Advance Access Publication Date: 22 December 2016
Research paper

Research paper

Risky business: Fine-grained data breach prediction using business profiles

Armin Sarabi,[1,*] Parinaz Naghizadeh,[2] Yang Liu,[3] and Mingyan Liu[4]

[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA; [2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA; [3]Department of Computer Science, Harvard University, Cambridge, MA 02138, USA and [4]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA

*Corresponding author. E-mail: arsarabi@umich.edu.

Received 1 October 2015; revised 10 June 2016; accepted 21 July 2016

Abstract

This article aims to understand if, and to what extent, business details about an organization can help to assess a company's risk in experiencing data breach incidents, as well its distribution of risk over multiple incident types, in order to provide guidelines to effectively protect, detect, and recover from different forms of security incidents. Existing work on prediction of data breach mainly focuses on network incidents, and studies that analyze the distribution of risk across different incident categories, most notably Verizon's latest Data Breach Investigations Report, provide recommendations based solely on business sector information. In this article, we leverage a broader set of publicly available business details to provide a more fine-grained analysis on incidents involving any form of data breach and data loss. Specifically, we use reports collected in the VERIS Community Database (VCDB), as well as data from Alexa Web Information Service (AWIS), the Open Directory Project (ODP), and Neustar Inc., to train and test a sequence of classifiers/predictors. Our results show that our feature set can distinguish between victims of data breaches, and nonvictims, with a 90% true positive rate, and 11% false positive rate, making them an effective tool in evaluating an entity's cyber-risk. Furthermore, we show that compared to using business sector information alone, our method can derive a more accurate risk distribution for specific incident types, and allow organizations to focus on a sparser set of incidents, thus achieving the same level of protection by spending less resources on security through more judicious prioritization.

Keywords: data breach; resource allocation; risk assessment.

Introduction

Data are an important asset in every business; the valuable data of an organization may include private information such as medical records, credit card numbers, private customer data stored on the cloud, or even trade secrets, as well as public information such as the website of an online commerce company. Any incident involving such data, whether intentional (targeted attacks) or unintentional (internal errors), can disrupt a business and inflict damage on its assets and reputation. Therefore, a portion of an organization's resources should be dedicated to protecting itself from security incidents; preventive measures include maintaining regular backups, keeping software up to date, and employee education in order to reduce miscellaneous errors.

However, determining how to allocate resources in protecting one's assets, as well as choosing an optimal level of investment in each preventive measure, is not a trivial task, as there is a wide variety of ever-changing attack methods. To help identify common forms of data incidents, a number of projects have been created to collect information about incidents that involve some sort of data loss. Some of these projects, such as [1] and [2], focus exclusively on hacking attacks, while some (e.g. [3]) cover a broader range of incidents, including human errors, and physical loss of data due to theft.

Using these reports, organizations are able to identify prevalent incident vectors, and invest in self-protection in a more optimal way. However, a point that should not be overlooked is that not all businesses should be treated the same, as each business is prone to different forms of incidents. For instance, a cloud hosting company might be more likely to suffer from hacking or denial of service attacks, while a medical institution with a large number of personnel runs a relatively higher risk of data loss through human error.

In this article, we aim to better understand how information about a business is correlated with its risk of falling victim to different forms of data incidents. Determining the overall risk of experiencing any form of data incident will help organizations decide on an optimal level of security investment. Moreover, estimating the distribution of risk among multiple incident types will allow us to narrow down the recommendation on the most effective preventive measures, depending on the types of incidents the organization is most likely to face.

To this end, we use an incident dataset collected by the VERIS community [3] reporting a broad class of data incidents; these reports consist of detailed information about the incident itself (e.g. type of attack, assets involved), as well as the victim organization (e.g. business sector, number of employees). Furthermore, we select a set of nonvictim organizations by randomly selecting network domains from the Open Directory Project [4]. We combine these with statistics obtained from Alexa Web Information Service (AWIS) [5] about the websites of victim and nonvictim organizations, as well as information about network assets of an organization obtained from Neustar Inc. [6]. These features together constitute the "business details" of the organization. We then utilize this information to assess its overall risk of experiencing a data breach. We are able to identify, with 90% accuracy, victim organizations with the same attributes as companies that have previously experienced a breach, while maintaining a false positive rate of 11%. For victim organizations, we further estimate the conditional distribution of risk for specific incident types by considering three different categorizations for the incidents: (i) by type of data incident (e.g. error, hacking, etc.), (ii) based on the source of the incident (external, internal, or partner) and the motive behind it, and (iii) by considering the assets that were involved in the incident (e.g. media, server, etc.). Our results show that there is a clear correlation between each incident category and the victim's business details; this information can be used to provide guidelines on how an organization with limited budget for security should prioritize its security investment in allocating resources to different forms of self-protection.

In our earlier work [7], we examined the use of a different type of data, namely Internet measurement data on organizations' security posture (including malicious activities observed from hosts), to predict future cyber-security incidents. In the present study, we broaden our scope to include not only network/cyber incidents, but also noncyber data incidents such as physical theft and loss, miscellaneous errors, etc.

We note that while correlation studies to identify prevalent attack vectors have been done before, most notably see Verizon's annual Data Breach Investigations Report [8] using business sector information, our goal is to use additional business information to enable a more fine-grained study, whereby the incident type distribution is quantified not just for an entire business sector, but for specific individual businesses based on other features such as employee size, region of operation, etc. This allows us to generate sharper (more highly concentrated) incident type distributions; i.e., with more fine-grained definition of subsets within a sector, we are able to see incidents concentrated over a smaller number of types. An immediate consequence of this is that security investment and resource allocation decisions informed by such analysis are much more targeted and effective. We show that on average an organization can protect against 90% of all incidents by focusing on 70% of incident types; in some cases the latter can be significantly lower.

Our results are derived and presented in two parts. First, an unconditional prediction of an organization falling victim to a data incident, and second, prediction of the conditional distribution of risks over different incident types given that an incident occurs; the latter complements our estimation of the probability of an incident happening in the former. In practice, the absolute risk of experiencing an incident provides the organization with insight on the total amount of resources that should be allocated to self-protection, while the conditional risk can be used to decide the allotment of these resources to different forms of preventive measures. By combining these two results, one can also determine the absolute risk of a given incident type. In addition, the current study can guide better breach detection efforts. From this perspective, our study is aligned with the growing "assume breach" mentality in the security community [9]: everyone is a target hence all organizations should take measures to prevent, detect, and respond to incidents, in the most effective way. Last but not least, these findings can be used as guidelines in the emerging cyber-insurance market. A study of the distribution of risk among different forms of data incidents can help insurance providers better assess the potential amount of loss which in turn helps determine the contract terms, including premiums and coverage levels.

Non-goals: Note that our main goal in this study is to reveal attributes of a business that are correlated with experiencing a data breach incident, rather than to detect the manifestation of a breach. Examples of such attributes include, for instance, hacking attacks target entities in the information sector more often than other industries, and an organization with a large number of employees is inherently more prone to data breach through human error. Detecting or forecasting a hacking incident by finding security flaws would require probing an organization's internal network and devices, another nongoal of this study. Furthermore, predicting incidents such as internal error, or employee misuse through the vectors that cause them are even more challenging, due to the presence of human elements. Therefore, when using the terms "risk prediction," or "risk forecasting" we are referring to "risk assessment" by comparing an arbitrary organization's attributes to those of victims and nonvictims in our training samples, and not uncovering vulnerabilities that directly cause a breach. However, this does not imply that it is not possible to forecast cyber-security incidents without observing security flaws in how an organization is operating. A correlation study on the impact of business features such as sector and size on data breach incidents, can project how likely it is for an organization to be successfully targeted by an attacker, or suffer a data breach through human error, by determining how often similar entities have experienced data breaches in the past. Furthermore, even if a security flaw is detected in a system, the chance of it turning into a data breach by an attacker targeting said vulnerability, is partly determined by how the data breach can be monetized by the attacker, which is in turn influenced by the business features utilized in this study.

The rest of the article is organized as follows. In Section 2, we summarize existing work relevant to this study. In Section 3 we describe the datasets used in this article. In Section 4 we explain in detail how we build our risk assessment model, and we discuss and analyze the results in Sections 5 and 6. Section 7 concludes the article.

Related work

The main contribution of this study compared to existing literature is an in-depth and quantitative analysis of the risk distribution over security incident types for a given organization, which can help the latter more strategically allocate resources for prediction, prevention, and detection.

Data analysis

A relevant study to this article is Verizon's annual Data Breach Investigations Report (DBIR) [8]. The most recent report for 2015 contains detailed analysis on more than 79 000 security incidents from multiple sources including VCDB. The report contains a detailed analysis on statistics of the data including action types and vectors, actor types and motives, as well as victim demographics and industry. Moreover, starting from DBIR 2014 the authors identify nine patterns describing 92% (96% for DBIR 2015) of the incidents in their report. By categorizing the incidents into separate patterns, it is possible to analyze the distribution of incident varieties within each pattern and provide entities with more specific recommendations on how to invest in their security. The report also provides the spread of attack patterns within each industry, to narrow down the risk even more. For instance, it is pointed out that the main threat to organizations providing accommodation services is through Point-of-Sale (POS) intrusions, which describes 75% (91% for DBIR 2015) of the incident reports within this industry. Furthermore, Thonnard *et al.* perform a similar analysis on spear phishing targeted attacks in [10]. The authors identify risk factors at the organization level (industry sector and number of employees), and individual level (job level and type, location, and number of LinkedIn connections), that are positively or negatively correlated with the risk of experiencing targeted attacks.

As mentioned earlier, compared to the DBIR, we aim to provide a more fine-grained framework to give more specific guidance to organizations not only based on their industry, but utilizing a host of other features available to us. This includes demographic information, details about the size of the business and its popularity, and business sector information. Moreover, we couple our conditional risk distribution with the overall probability of breach in order to arrive at a more realistic sense of risk. For instance, even though a typical business in the accommodation sector is more prone to POS intrusions, their risk within that category might still be less than businesses in other sectors, given that their unconditional probability of breach is low.

Prediction of cyber incidents

The notion of predicting cyber incidents (rather than detection) has also enjoyed popularity recently. In [11], Soska *et al.* apply machine learning tools to predict the chance of a website turning malicious in the future, and show that their method can achieve 67% true positive and 17% false positive. In our previous study [7], we examine to what degree cyber-security incidents may be predicted by using a range of security posture data. Compared to the above studies, our goal in the present study is to consider a broader range of data incidents, including targeted and untargeted physical and cyber-attacks from both internal and external sources, and incidents due to error, while at the same time recognizing the difference between specific incident types by emphasizing the relative risk each incident type poses to a particular organization. Note that of the 2644 reports in the VCDB for 2013 and 2014, 981 are hacking and malware incidents (cyber incident), and the rest are nonnetwork related incidents (noncyber incident).

Other works related to this article include studies on the trends and costs associated with data breaches. In [12], Edwards *et al.* use Generalized Linear Models to uncover trends in data breaches, and conclude that the frequency and size of data incidents have not increased over the past decade. Furthermore, The 2015 Cost of Data Breach Study by Ponemon Institute and IBM [13], finds the average cost of a data breach to be $3.8 million, with $154 incurred for each lost or stolen record. The authors [14–18] conduct event-study analyses on the impact of data breach disclosures on market value, and conclude that there exists a negative and statistically significant correlation between the two. Moreover, in [19] Romanosky *et al.* provide an empirical analysis of data breach litigation, and in [20] discuss the impact of breach disclosure laws on identity theft.

Datasets

In this section, we illustrate the datasets used in our study, namely the VERIS Community Database (VCDB) [3], the Open Directory Project (ODP) [4], the AWIS [5], and the IP Intelligence service from Neustar, Inc. [6].

VERIS community database

The VCDB is currently composed of 5233 reports on publicly disclosed data breaches. The dataset includes incidents that occurred up to and including 2015, with 4961 entries corresponding to incidents after 2010. For our current study, we focus only on the 2013 and 2014 incidents, consisting of 1850 and 794 entries, respectively. The reports cover a wide variety of events, some examples of which are given in Table 1.

Each entry in the VCDB is reported using the Vocabulary for Event Recording and Incident Sharing (VERIS) [21]. The VERIS framework, as well as the VCDB, are initiatives by the Verizon RISK Team facilitating a unified approach to documenting and collecting security incidents. The VERIS fields for an incident are populated to answer "who did what to what (or whom) with what result?" [8]; details include the type of incident and the means by which it took place, the actor and motive, the victim organization, the assets which were compromised, timeline of the incident, and links to news reports or blogs documenting the incident. However, each entry might be only partially populated, since victim organizations tend to not disclose all the details regarding the incident.

We now explain the fields extracted from VCDB which are of interest in training and testing our classifiers. The first set is information regarding the type of attack, based on which each incident can be put in one of seven general categories: "environmental," "error," "hacking," "malware," "misuse," "physical," or "social." Each type may include additional fields that can help further differentiate incidents of the type. For instance, a "physical" incident might be further categorized as theft or loss, while a "hacking" incident might be identified as a SQL injection or a brute force attack. The second set identifies the actor responsible for the incident, falling in one of three types: "external," "internal," or "partner." The dataset may further include fields identifying the motive for each of these actor categories. The third set identifies the assets that were compromised during the incident. There are six possible asset types: "kiosk/terminal," "media," "network," "people," "server," and "user device."

We also extract three features about the victim organization from the existing VCDB fields as input for our classifiers: industry code, number of employees, and the region of operation of the victim organization. The industry code provided is the North American

Table 1. Incident examples from the VERIS Community Database

Time	Report summary
Apr 13	Hackers breach website of Hong Kong police force and publish nonpublic data, deface webpage.
Aug 13	A Lima, Ohio clinical psychologist is in the process of notifying clients that their office was robbed.
Sep 13	Pharmacy accidentally dumped hundreds of private medical records at a recycling depot.
Sep 13	Janitor is blackmailed into gathering documents from a court.
Sep 13	Parents of children at Hopkins Road Elementary Schools say their kids came home with sensitive data belonging to other students.
Dec 13	Multiple Brazilian government sites defaced by Anonymous in protest to upcoming FIFA World Cup.
Jan 14	Hacking group DERP launches DDoS against Xbox Live networks.
May 14	Someone hacked into an electronic traffic sign on Van Ness Avenue in San Francisco.
Jul 14	Anonymous takes down 1000 Israeli government and business websites for #OpSaveGaza.

Industry Classification System (NAICS) code [22] for the victim, which specifies the organization's primary economic activity. Although NAICS codes can extend to up to six digits, each further detailing the sector, we only extract the first two digits of the code for our incidents; this classifies the company as one of 25 different sectors. The employee count captures information about the size of the organization; this entry may be a numeric range (1–10, 11–100, 101–1000, 1001–10 000, 10 001–25 000, 25 001–50 000, 50 001–100 000, and over 100 000), or simply "small" or "large" (for approximately below or over 1000 employees, respectively) when an exact number is not available. Finally, we use the region of the organization as a feature by extracting the continent of operation for the victim. Note that any said features can be missing for a VCDB entry. In such cases, we generally add an additional "unknown" category.

AWIS

AWIS is a service offered by Amazon Web Services (AWS) [23] that provides information and statistics about websites; these include traffic volume, number of visitors, speed, number of pages linking to the website, and information about the organization that maintains the website, such as address, contact information, and stock ticker symbol.

We gather the following data from AWIS about the victim organization. We include the global and regional rank, and the number of pages linking in to the target website, as indicators of the popularity or familiarity of an organization. The regional rank of a website is extracted by finding the country which has the most contribution of page views to the website's traffic, and adding the rank in that country, as well as the country code, to our feature set. We also include the 30-day average and standard deviation of the website's global rank for a 1-month period before the incident, to identify recent trends in popularity. Other selected features include speed of the website (as a percentile compared to other websites), the age and locale of the website, the categories associated with it, and whether the underlying company is publicly traded in the stock market. We convert the number of pages linking in, and global, regional, and average historical rank to logarithmic scale, due to their large range of quantities. We further break each category, if possible, by separating the portion describing the region of the website. For instance, for "Regional/Caribbean/Barbados/Government," the general category is "Government," while "Caribbean/Barbados" is the regional category. For missing fields, we choose a reasonable default value, e.g. "unknown" for text fields, and ∞ for rank. The aforementioned attributes of an organization can provide further insight into its sector, region, familiarity, and size. By combining these with features obtained from our other datasets, we are able to build a detailed description of a business, which can in turn help identify its risk.

Other than age and historical traffic rank, AWIS only provides the most recent state of a webpage. Therefore, there is a relatively large time gap between our incidents (which happened in 2013 and 2014), and features obtained from AWIS (September 2015). Features such as main contributing country, locale, and category are related to the organization's region and sector of operation and are not expected to change over time. However global and regional rank, number of pages linking in, and whether the company is publicly traded can exhibit more dynamic behavior. For samples where both a global and historical rank was available,[1] the average mean absolute percentage error between the two was 6.5%. We therefore concluded that the order of a website's rank remains fairly static. Unfortunately, we could not procure similar measurements for other statistics of a webpage, since it involves caching results from AWIS and studying the changes over a long period. However, since regional rank and number of links also capture the popularity of a page, we expect them to show similar behavior.

ODP

ODP (also known as DMOZ) is the largest publicly available directory of the Web. Each entry includes a website URL, the title of the site and a short description, as well as the category of the website. By selecting random entries from this dataset, we can effectively choose random nonvictim organizations. For this study, we use a snapshot of ODP obtained on 19 September 2015 consisting of 3 771 141 entries, of which a random selection of 16 780 entries, which had not appeared in our victim dataset, is used in this study as nonvictim organizations. Note that our random selection may also capture victim organizations that were not reported in the VCDB. The portion of "tainted" samples in our nonvictim set is upper bounded by the overall rate of data incidents.

To elaborate more on the process of selecting nonvictim entities, we would first like to point out that an alternative way to select nonvictim organizations would be to choose random entries from a global business directory. However, since we do not have access to such a directory, websites are used as a proxy to identify organizations. In our earlier work [7], we have used a random selection of networks to identify organizations which matched our use of network security posture measurement data. However, this selection method would limit us to companies that own network assets of their own, and those who rely on hosting providers and content delivery networks would be excluded. In contrast, almost all

[1] Alexa provides global and historical rank, for the top 30 million and 1 million websites, respectively. This is also the primary reason we have included both types in our feature set.

organizations own a website and would be included in our current approach. Furthermore, incidents covered in the VCDB include those concerning large companies, as well as data breach reports on smaller entities such as personal webpages. Using a web directory allows us to include smaller entities in our selection, resulting in a more representative nonvictim group; this cannot be easily achieved by using a business directory.

IP Intelligence

IP Intelligence is a service offered by Neustar Inc. that includes geographic information, network characteristics, and ownership information over the IPv4 address space. More specifically, we use the ownership information in our study, consisting of the organization name that manages a given IP address, along with, where available, its corresponding NAICS code. This information allows us to identify the network responsible for maintaining a given website. Moreover, since VCDB only provides business sector information for victim organizations, the NAICS code included in the IP Intelligence dataset allows us to include this information in our overall risk prediction for both victim and nonvictim organizations. The snapshot used in this study was obtained on 22 May 2015. We include the name of the company listed as the owner of an IP address, its size (number of IP addresses owned by the same), and the NAICS code associated with it in our feature set. Doing so helps identify hosting providers with bad reputation, i.e. those with a higher than average presence in incident samples.

Pre-processing

To be able to combine these datasets for our study, we first have to match each incident report with the website of the victim organization. To obtain this information, we find the name of the victim organization through the "victim id" field in VCDB, and extract the first Google search result for the organization name. We then manually verify the results to ensure that the websites match the victim organizations. For ambiguous victim IDs (e.g. "Indian government website"), we further read the incident report provided by a news report or blog entry to find the website of the entity that suffered the data breach. For the 2644 incidents that occurred in 2013 and 2014, we extracted the website for 2062 of them. Note that of the 582 incidents that we dropped, 139 did not report the name of the victim organization, and the rest were not included in our study either because the victim name was too ambiguous (e.g. "Egyptian government" and "law firm in British Columbia"), or we could not find a website for the victim (e.g. "Ha Dinh primary school" and "Purple Cow gas station"). The mapping between a victim organization and its respective website will allow us to combine entries in the VCDB with data collected from AWIS. Note that for a given year, we omit duplicate incidents for each organization. As an example, there are over 200 entries in the VCDB corresponding to error incidents in the US Department of Veterans Affairs. We count all of these incident only two times, once in 2013 and once in 2014. If there are additional entries corresponding to other forms of data incidents (e.g. hacking), we include them as separate entries when assessing risk for specific incident types.

Note that statistics obtained from AWIS are often provided only for the top level domain of a website. For instance, domains such as "mail.google.com and maps.google.com" are redirected to the top domain "google.com." Subdomains are only regarded as separate entities when "they are identified as personal home pages or blogs" [24]. On the other hand, website details from ODP are generally more detailed, and can include any number of subdomains and subpages. Therefore to avoid inconsistencies, we replace URLs associated with victim organizations and our random selection of URLs from ODP with their respective domains from AWIS. We are able to map the URLs associated with our incident/victim and nonincident/nonvictim samples to 1606 and 16 254 unique domains, respectively.

The next step is to include features from Neustar Inc. We resolve the domain to obtain an IP address, and then look up the owner of that address. We augment our set of features with the name of the owner, its size (number of IP addresses listed under the same name), and the NAICS code associated with it. Out of the 17 860 domains from the previous step, we were able to map 17 772 of them to 5805 unique owners. Note that for 88 of the domains we were either not able to look up their IP address, or there was not any entry in the IP Intelligence dataset for that address. For these samples we list "unknown" under owner and NAICS code, and a size of zero.

Finally, we convert text fields to a set of binary features by tokenizing each distinct value. For categories and NAICS codes from IP Intelligence, we break each entry into multiple values with different levels of detail, and tokenize each separately. For example, "Business/E-Commerce/Consulting" is a subcategory of "Business, and Business/E-Commerce"; and a NAICS code of 51 720 (Wired Telecommunications Carriers) is a subsector of 51 (Information), and 517 (Telecommunications). To limit the total number of features, we ignore tokens that have been repeated less than 10 times in our samples.

Methodology

In this section, we will discuss the rationale behind the features selected for our model, followed by a detailed description of how to build a risk assessment model using the features and incident reports described in Section 3.

Feature set

In Section 3 we listed the features extracted from VCDB, AWIS, and IP Intelligence to be used in training our classifiers. We will now discuss our motivations for selecting these features, and why we expect them to be indicative of a company's risk of data breach. Note that while we provide simple examples for why a certain feature can be correlated with cyber-risk, our model can recognize more complex relationships within out feature set that can help the classifier make more accurate assessments. The first and foremost features are those that specify a company's sector of operation, namely the industry code extracted from VCDB, and the website category from AWIS. We expect an organization's industry to be strongly correlated with its risk of falling victim to different types of data breaches. A company's industry can provide insight into the types of records that can potentially be compromised (e.g. credit card information for retailers, or physical and digital records for health care), or motivations for targeted attacks (e.g. hacktivism for public administration entities). In addition, a business's sector can determine the value of data records to an attacker, which in turn influences the attacker's decision to launch an attack on said entity; this type of correlation also applies to other features used in our model, such as a business's size and region of operation. As we discussed in Section 2, DBIR [8] also uses industry information to give security recommendations to businesses within a sector.

The next set of features are those that specify the size of a company: employee count from VCDB, and whether a company is publicly traded, which is provided by AWIS. We expect the size of a company to be correlated with how often it is targeted by cyberattacks, since compromising a large company tends to be more profitable for an attacker. Furthermore, as we will see in Section 6.2, a larger employee count can increase the chances of breach through human error and employee misuse. Features like traffic rank (global, regional, or historical) and the number of links to a website are indicative of a website's popularity, and therefore correlated with the chances of a company being targeted.

We also expect a company's region of operation to be connected to its cyber-risk. The region is obtained from VCDB, as well as the website's top contributing country, locale, and regional category. Other features such as the age and speed of an organization's webpage, can provide more insight into its security posture. Older companies tend to be more experienced in protecting themselves against data breaches, and may have better policies in place to prevent them; a website's speed is an indication of how well it is being operated, which in turn can be associated with security posture.

Finally, our measurements from IP Intelligence will provide more details about an organizational network, which can be closely coupled with risk of network-related breach incidents. Note that an organization's website can be either hosted on the company's internal network, or by a hosting provider. The industry code from IP Intelligence can provide the classifier with the necessary information to distinguish between the two cases, since the NAICS code 518 (Data Processing, Hosting, and Related Services) can be associated with hosting providers. When an organization's website is hosted by a third party, the name of the hosting company and its size (in terms of number of IP addresses owned by the provider), can determine its reputation and how protected customers are. For self-hosted websites, the size of the organizational network will indicate the attack surface, and therefore the risk of breach through network incidents.

Construction of the classifiers

Our ultimate goal is to provide risk assessment for an arbitrary organization given its features, i.e. a distribution of risk over all incident types. This risk can be represented in two parts as follows:

$$\Pr(\text{Incidenttype}|t) = \Pr(\text{Incident}|t)\Pr(\text{Incidenttype}|\text{Incident}, t), \quad (1)$$

where t is the type of the organization in question, represented by its set of features. "Incident type" can be any of the available data incident types, e.g. physical theft. The first term will be referred to as the overall risk, with the second term the conditional risk. These two probabilities are estimated separately by constructing different classifiers.

Toward this end, we use Random Forest classifiers, an ensemble learning method that constructs multiple decision trees over the training data, and outputs the average of all individual trees' predictions [25]. Random Forest classifiers improve upon single decision trees by reducing over-fitting over the training set. For overall risk estimation [first term in the RHS of Equation (1)], we use our set of victim organizations coupled with a randomly selected set of nonvictim organizations to build a binary classifier; in this case all victim organizations no matter the type of incident are given a label "1."

To assess the conditional risk [second term in the RHS of Equation (1)], a naive way would be to take the incident signature (i.e. action, actor, and asset) of an entry as a class label, and the victim's features as input data for the classifier. However, given the large number of possible incident signatures, there are only a small number of samples per signature vector. Furthermore, as we have mentioned before, a significant number of incident entries provide only partial information about their corresponding incident. Ignoring such entries will leave us with even fewer samples.

Our solution to the above problem is to build multiple classifiers, each of them estimating a portion of the incident signature. This continues our previous use of the chain rule in probability. Assume that we want to estimate the risk factor for an organization of type t for experiencing a physical theft incident. We can break the conditional risk into multiple parts as follows:

$$\Pr(\text{Theft}|\text{Incident}, t) = \Pr(\text{Physical}|\text{Incident}, t)\Pr(\text{Theft}|\text{Physical}, t). \quad (2)$$

As a result, entries that cite a physical incident without specifying additional details will still be included for building and testing the first classifier [first term in the RHS of Equation (2)], but will be ignored when building the second classifier (i.e. theft). This method can be visualized as a tree as shown in Fig. 1, where each node represents a data breach type. The risk score at a node is the result of multiplying the risk at its parent node by the output of the classifier corresponding to said (child) node.

Note that the output of Fig. 1 is a conditional probability, conditioned on the event that an incident has occurred. To derive the absolute risk for the given breach type, we need to multiply the result by the overall probability of breach [first term in the RHS of Equation (1)]. In the remainder of this article, we will discuss and analyze the results on overall risk estimation and conditional risk for specific breach types separately. The rationale behind this separation is that the former serves as a forecast on security incidents. On the other hand, the point of the latter is not to make a single prediction on the type of incident that is going to happen, but to estimate the distribution of risk among multiple incident types; as we shall see, predictions for single incident types are significantly less accurate than overall risk estimations due to its density estimation nature. This point is further elaborated on in Section 5.2.

Overall risk prediction

To forecast the overall risk of breach, we assign labels zero and one to our nonvictim, and victim features, and train a Random Forest consisting of 50 trees over victim samples in 2013, and a random selection of 11 585 samples from nonvictim samples. We use features from AWIS and IP Intelligence for prediction, and omit features from VCDB since they have only been provided for victim organizations. We use the incident samples from 2014 and the rest of the nonvictim samples for testing.

Table A1 in the Appendix summarizes the importance of each feature in the final classifier. As is evident from the table, the most used

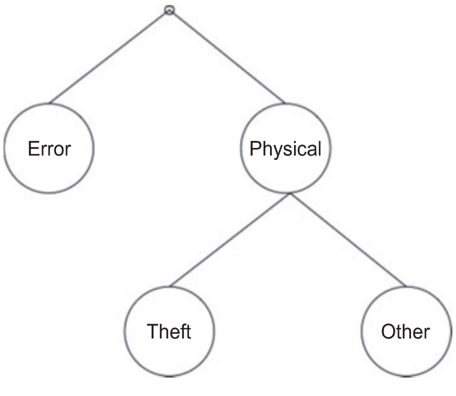

Figure 1. A sample risk assessment tree.

feature for this type of risk assessment is the general category of the website. For further elaboration on this point, we have shown the distribution of victim and nonvictim samples over the top level categories from AWIS in Fig. 2. While our incident samples have presence in most of the top level categories (excluding "adult," "maps," and "weather"), it is possible to identify categories that exhibit higher- or lower-than-average risk. For instance, the portion of incidents that belong to "health" and "government" are significantly larger than the global population, while the "world" and "arts" categories can be associated with low risk. Note that the "world" category describes webpages that are in languages other than English. The discrepancy in this case can also be due to underreporting, since VCDB tends to focus more on incidents that happened in the USA.

Note that inherent biases in our victim dataset may affect the output of our trained model. The most prominent examples are biases toward incidents in the USA, and also certain industries due to disclosure laws. For instance, businesses operating in retail are more likely to disclose data breaches due to concerns that customer information may have been compromised, while other industries might be underrepresented in publicly disclosed data breaches. Consequently, we may underestimate or overestimate a business's cyber-risk based on its region or sector. In other words, our model is estimating the risk of a publicly disclosed breach, which is not necessarily the same as risk for undisclosed data incidents. This issue may be alleviated by training models over specific groups of victims and nonvictims, e.g. training a classifier on the subset of samples that belong to a certain country, or industry, ensuring that samples are compared to organizations of the same type, and therefore with the same incident reporting rate. In this article, we do not train separate models for overall risk prediction since our goal is to provide a single assessment that can be used to compare organizational risk, regardless of region or sector. However, we will further explore this technique in Section 6 for assessing risk in different incident types.

Conditional risk prediction

Given the training and test samples (incidents belonging to 2013 and 2014, respectively), we first train a binary classifier for each node, using a Random Forest model consisting of 20 trees. To prevent over-fitting, we set the minimum number of samples at each leaf of the decision trees to 25. However, we may still experience some over-fitting due to the large number of features available to our classifier. To help alleviate this problem, we limit the number of features used for each Random Forest as follows: we always use the three features extracted from the VCDB, namely industry, employee count, and region. Out of the remaining 10 features, we select the most significant through cross validation, i.e. training multiple classifiers using different combinations of features, and selecting the one with the best performance. The list of features used for each classifier, as well as their importance in the resulting Random Forest classifiers, are also included in Table A1 in the Appendix.

Incident categorization

Using the classification method described above, we apply our risk assessment scheme separately to three parts of the incident signatures: action, actor, and asset. Each of these classifiers focuses on a separate aspect of an incident. If a single entry matches multiple incident categories, e.g. a hacking incident through misuse of privileges, we break it into multiple incidents that each belong to a single category.

Action type. The action type falls into one of the seven general categories discussed in Section 3.1. We omit "environmental" incidents, of which there are only four samples between 2013 and 2014. We further categorize "hacking" events into two subcategories: (i) hacking incidents that involve data breach through compromised credentials, including stolen credential, brute force, and backdoor attacks, and (ii) all other forms of hacking, 75% of which are SQL injection and Denial of Service attacks. We also divide "physical" incidents into two subcategories of (i) theft and (ii) everything else, 88% of which are due to tampering.

Knowing the action type can provide significant information on the types of preventive measures that can be used to reduce loss. For instance, the first group of "hacking" incidents can be prevented by setting strong passwords and changing them on a regular basis, as well as not storing unencrypted credentials at insecure locations. "Error" and "misuse" can be reduced by employee education, setting and enforcing internal regulations, and avoiding unnecessary access privileges for employees and/or business partners.

Actor type and motive. In addition to action types, we train our classifier based on the actor responsible for the incident. "Internal" actors are separated based on their motive into two subcategories of (i) financial motives, and (ii) other motives, including convenience,

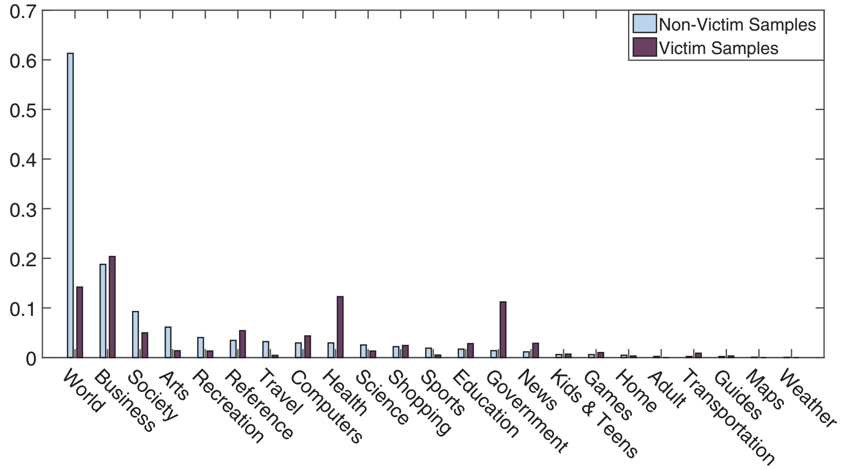

Figure 2. Distribution of all victim and nonvictim samples (both training and testing) over Alexa's top categories. Note that a website can belong to multiple or no categories.

espionage, grudge, ideology, and fun. “External” actors are similarly subcategorized into (i) financial, (ii) espionage, (iii) ideology, and (iv) fear, fun and grudge. Incidents due to “partners” are not further subcategorized due to insufficient samples.

Assessing risk associated with actor types can prompt organizations to determine policies for employee education and access to data (for “internal” types), guard their network periphery from “external” attackers, and perform due diligence when selecting “partners.”

Asset type. Finally, we look at the types of assets that were compromised during the incident. Asset types include “kiosk/terminal,” “media,” “people,” “server,” and “user device.” We have omitted “network” related assets due to insufficient number of samples. Knowing what asset types are more likely to be affected can significantly improve our ability to estimate the amount of potential loss following security incidents. This can guide insurance underwriters in designing more appropriate policies catered to specific client organizations. It can also be used to advice network administrators to keep regular backups when assets such as “media” and “server” are involved.

Comparison with DBIRs’ categorizations. Our choice of categorizations is consistent with the one adopted by Verizon in the 2008-13 DBIRs, but differs from the categorizations proposed in their latest 2014 and 2015 reports. DBIR 2014 uses hierarchical clustering to identify nine incident classification patterns (combinations of actions, assets, and actors) that can be used to describe 92% of all incidents. Examples of these patterns include cyber-espionage, point of sale intrusions, and insider misuse. Despite the effectiveness of this clustering method in accurately describing incidents in the dataset used by Verizon, an application to the subset available through VCDB would fail to provide a similar precision, see Table 2: due to lack of sufficient details, 18% of the VCDB data will not fit the nine proposed patterns (as opposed to only 6% in Verizon’s larger dataset). This is one of our main motivations for selecting three different categorizations based on VERIS primitives only, i.e., actions, actors, and assets.

Results

Overall risk

Figure 3a displays the Receiver Operating Characteristic (ROC) curve of our overall risk estimators, evaluated over the test samples. By identifying organizations with similar attributes to those that have previously experienced a data breach, we can achieve a 90% true positive rate in flagging organizations in our victim set as high risk, while keeping false positive rate at 11%. These numbers are comparable with our previous results in [7], where we were able to forecast cyber incidents with 90% true positive and 10% false positive rate. Figure 3b shows the distribution of the classifier output scores for victim and nonvictim test samples. There is a clear distinction between the two distributions, with victim samples having more bias toward higher scores, signifying more risk.

Moreover, Table 3 summarizes the accuracy of our model over Alexa’s top categories in Table 2, as well as the overall accuracy on all samples. Each row in Table 3 displays our model’s performance over the test samples in 2014 that belong to the corresponding category. We have removed categories where we have less than 20 victim samples. We have included the number of victims, and nonvictims in each category, as well as the true positive rate that is closest to 90%, along with its corresponding false positive rate. The Area Under Curve (AUC) metric displays the area under the ROC curve. Note that the AUC score is independent of the fraction of the test population in each class, making it a useful metric for evaluating performance on unbalanced datasets. The best accuracy belongs to the “Business” category, and “Reference” and “Government” perform the worst.

Risk distributions

Figure 4 shows our results on prediction of specific incident types. We have drawn ROC curves for three types each in the action, actor, and asset categorizations. Comparing to Fig. 3a, the accuracy of these classifiers is significantly lower, typically achieving a 80% true positive at 50–60% false positive rate, except for the asset type “kiosk” that achieves the same accuracy at 11% false positive (note that this asset type is only owned by a select few industries, which most likely contributes to the high accuracy observed here).

Table 2. VCDB data categorized using DBIR 2014 patterns

Incident type	Crimeware	Cyber Esp.	Ddos	Stolen	Cred.	Error Skimmers	PoS	Misuse	Web app	Else
No. of samples	67	16	106	326	333	66	19	272	399	356

Only 82% of the data can be described by the nine patterns.

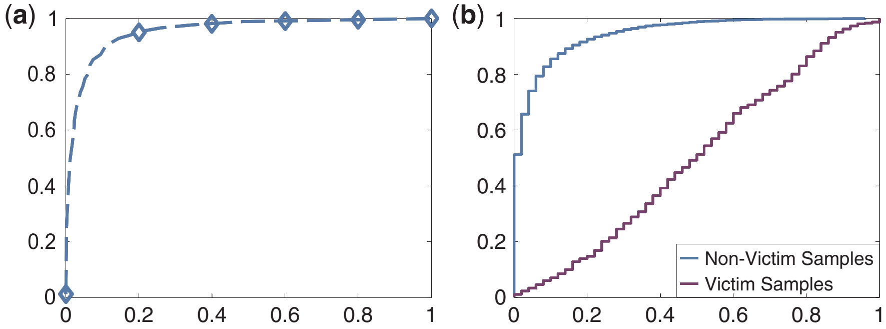

Figure 3. ROC curve for overall risk estimation (left), and cumulative distribution of risk (right).

To explain the difference between Figs 3a and 4, we will consider a model with n different incident types, and a sample entity with probability of breach of p. We will then analyze this example for different scenarios. In the first case, the absolute (unconditional) probability of breach for one incident type is equal to p, while other types have zero probability, and we will be able to predict with certainty the type of the data breach. In the second case, assume that all breach types are equally probable, and the conditional risk is a discrete uniform random variable. In this scenario, if our predictor outputs a label of one with probability q for all types, we will on average see q true positives, and $q(n-1)$ false positives, and the average true positive and false positive rates, averaged over all classifiers, will be equal to q. If the risk is equally distributed between k incident types, then for every true positive the predictor will be penalized by $k-1$ false positives, and the overall true positive and false positive rates will be q and $q(k-1)/(n-1)$, respectively. Note that regardless of the type of an organization, its risk will never be zero for breach types such as "error" and "misuse," and as long as it owns any form of network assets, it will be vulnerable to hacking incidents (e.g. through zero-day vulnerabilities). As a result, the main value of this risk distribution estimate is not as a forecast for a particular incident type, but rather as a prediction of how the overall risk is distributed over all incident types by combining the outputs of all classifiers. We will discuss how to interpret this conditional distribution in Section 6, and show that it can lead to a sparse or diverse range of risks.

To gain insights on how details about a business can affect their risk of experiencing various types of data breach, we start by deriving the distribution of risk over incident action types for each industry sector. The results for nine business sectors, as well as the overall distribution are included in Table 4; these results use only sector information in training the corresponding classifiers. Note that this is equivalent to simply measuring the distribution of incidents in each sector, since the Random Forest classifier is using only a single feature. There are a few observations on the risk distribution of different sectors. For instance, information companies are more prone to both types of hacking, and less likely to sustain damage due to physical incidents. In contrast, the health care industry has low risk in hacking but high risk in physical attacks, especially theft. These observations are intuitively to be expected, since information companies' most valuable assets are generally stored in nonphysical formats (e.g. on the cloud), while the health care industry may still use physical forms of archiving sensitive data such as patient information.

Table 3. Accuracy of overall risk estimation over Alexa's top categories

Category	Victims	Nonvictims	AUC	TPR (%)	FPR (%)
World	56	2204	0.928	91.1	18.2
Business	73	817	0.968	89.0	4.0
Society	20	325	0.922	90.0	20.0
Reference	21	134	0.841	85.7	43.3
Computers	25	119	0.939	88.0	16.8
Health	36	117	0.954	88.9	21.4
Government	58	42	0.876	89.7	35.7
Overall	482	4669	0.953	89.6	11.3

To highlight the additional gain we get by using more features than just industry sector information, we also show in Table 4a number of examples. In these cases, our classifiers can generate much more specific risk predictions. For instance, we can see that compared to a typical information company, Russian Radio has less risk in malware, social, and hacking through compromised credentials, but higher risk in error, misuse, and physical. Verizon and Macon-Bibb County exhibit a more uniform risk across the board. The higher risk for Verizon in error and misuse (also the lower risk of Macon-Bibb County in the same categories) can be attributed to their respective sizes. As the number of employees grows larger, so does the risk of data incidents due to human error and malevolent employees. These much more refined and targeted predictions would not be possible without using additional features. As we shall show later in Section 6.2, with proper thresholding the actual incidents in these organizations were also correctly identified.

Dealing with rare events

Looking at Table 4, there is an imbalance in the overall frequency at which different incident types appear in our dataset. Social incidents occur rarely as compared to error and hacking incidents. It is indeed possible that social incidents are rare events, and therefore should not be a priority when determining security policies. However, an important challenge in building a risk assessment model is underreporting of security incidents by victims. Data breach reports are largely undisclosed, as organizations tend not to expose their security posture information unless necessary. Our dataset, VCDB, is a collection of publicly disclosed breaches; these incidents have either been detected by external sources (e.g. website defacement) or are incidents which an organization is obligated to report due to the compromise of private customer information (e.g. payment information or health records). Thus, not only incidents are commonly underreported, but it is also safe to assume the existence of selection bias in the data: each incident type is represented differently as a

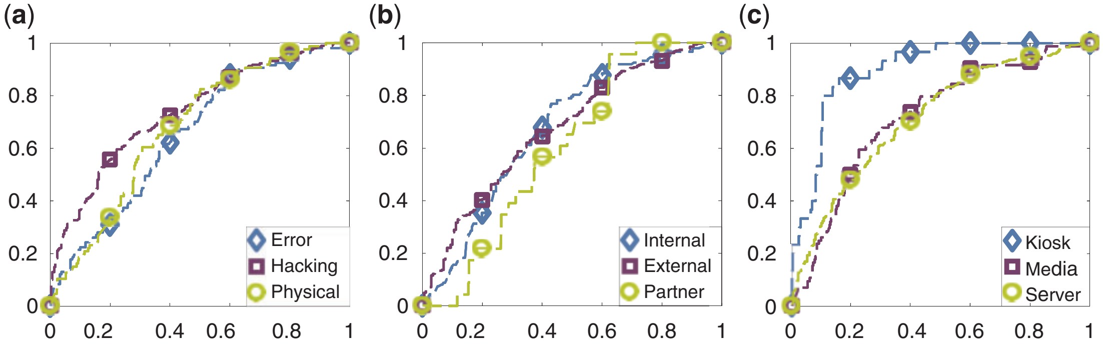

Figure 4. ROC curves for action (left), actor (middle), and asset (right) classifiers.

Table 4. Conditional risk distribution by business sector, and for sample organizations (highlighted rows)

		Hacking				Physical		
	Error	Comp. Cred.	Other	Malware	Misuse	Theft	Other	Social
Manufacturing	0.08	0.09	0.33	0.13	0.22	0.13	0.00	0.02
Retail trade	0.15	0.26	0.11	0.19	0.09	0.09	0.11	0.02
Information	0.09	0.28	0.41	0.07	0.04	0.03	0.01	0.07
Russian Radio	0.14	0.16	0.40	0.02	0.10	0.10	0.03	0.03
Verizon	0.28	0.17	0.22	0.08	0.19	0.06	0.05	0.05
Finance and insurance	0.25	0.09	0.11	0.05	0.12	0.10	0.19	0.07
Pro., Sci. and Tech. Svcs	0.16	0.09	0.56	0.04	0.13	0.09	0.00	0.02
Educational Svcs	0.30	0.13	0.21	0.06	0.11	0.14	0.00	0.05
Health care and social asst	0.25	0.08	0.03	0.02	0.23	0.38	0.02	0.01
Accommodation and food Svcs	0.08	0.37	0.00	0.18	0.16	0.11	0.11	0.00
Public administration	0.27	0.09	0.29	0.03	0.17	0.10	0.01	0.03
Internal revenue service	0.21	0.08	0.15	0.06	0.17	0.09	0.02	0.03
Macon-Bibb County	0.20	0.13	0.23	0.07	0.14	0.23	0.04	0.04
Overall	0.22	0.12	0.21	0.06	0.15	0.14	0.04	0.04

result of both availability and variation of detection methods, and the corresponding industries' disclosure policies. This bias could cause a tendency toward flagging and protecting from incidents that are reported more often, in turn resulting in poor protection against less commonly reported incidents.

One way to address this issue is to ignore the frequency at which incident types are reported. In other words, rather than looking at each row in Table 4, we could base our decisions on the distribution of risk within each column. For instance, we can make the observation that finance and insurance companies exhibit higher than average risk in social incidents, even though the absolute risk in this category is the second lowest in its respective row. By having different standards, or thresholds, of what signifies high risk in each category, we can alleviate the impact of potential underreporting and reporting bias in the dataset and prevent the tendency of ignoring rare events by ensuring equal protection among all incident types. Specifically, after training our classifiers and obtaining risk outputs on the input data, we specify thresholds for each incident type separately, such that the reduction in risk is consistent among all types; this is detailed in the next section. Note that this "normalization" of risk scores is possible mainly due to the fact that we are constructing separate classifier for each incident type.

Interpreting the classifier output

After estimating an organization's risk in each category by feeding its features into our classifier, the next step is to interpret these scores by determining what range of values indicate heightened risk. Based on our discussion in the previous section, this is achieved by computing the ROC curve for each binary classifier on the training set, and choosing the point that corresponds to a predefined true positive rate. We will use the family of thresholds corresponding to these points to determine risky incident types for any arbitrary organization, hereafter referred to as the "risk profile." Selecting a more conservative set of thresholds (i.e. higher true positive rate) will tighten the business's security by advising it to invest in a larger set of self-protection methods. This selection represents the trade-off between the amount of resources an organization allocates to self-protection, and the reduction in incidents it desires to attain. From this point on when referring to "thresholds" used for deriving the risk profile, we simply mean the family of thresholds acquired for a specific true positive rate. We find these thresholds by looking at the ROC curve of each classifier, and finding the point that corresponds to a specific accuracy (e.g. 80% true positive rate), this is explained in Section 6.1. Note that these thresholds are specific to our incident source (VCDB), through its reporting rates on different incident types. Therefore, an incident dataset with different reporting rates would yield a new set of thresholds.

Evaluation

For evaluation, we first obtain the risk profiles of organizations in our test samples, for various sets of thresholds. We then calculate the accuracy of our risk assessment model, by counting the number of incidents which belong to one of the risky types forecasted by the risk profiles. An important advantage of our model is in reducing the number of risky types predicted for each organization; achieving the same accuracy by advising organizations to focus on a smaller set of incident types will help achieve the same level of protection by spending less resources on security.

Figure 5a, c, and e summarize our results over action, actor, and asset types, respectively. Each point in the plot denotes the accuracy of risk profiles obtained from a particular set of thresholds, versus the average number of risky types forecasted by these profiles. To illustrate the improved performance of using our extended set of features, we have also included the accuracy curve of a predictor using industry information alone (Table 4). For action, actor, and asset types we can correctly forecast 90% of the incidents in our dataset by flagging, on average, 5.6 (70% of incident types), 4.0 (67%), and 3.5 (70%) incident types, respectively. In other words, we can achieve this accuracy by eliminating at least 30% of all incident types. Using only business sector information, the numbers increase to 6.5 (81%), 4.8 (80%), and 3.6 (72%). The distinction is more visible when predicting over action and actor types.

Note that for a given point in the plot, the number of risky types in the risk profile can vary across organizations. Figure 5b, d, and f demonstrate the distribution of organizations over their predicted number of risky types, corresponding to the 80% accuracy point in the top plots. Looking at Fig. 5b we can see that using all features, there are organizations whose risk profiles only consist of 1 or 2 incidents types, while others include up to seven types.

We present a number of these samples in Table 5, whose risk scores have already been discussed in Table 4. The first two examples in the table belong to the information sector, and the last two are public administration organizations. We have included the risk profiles for these sample organizations using our extended

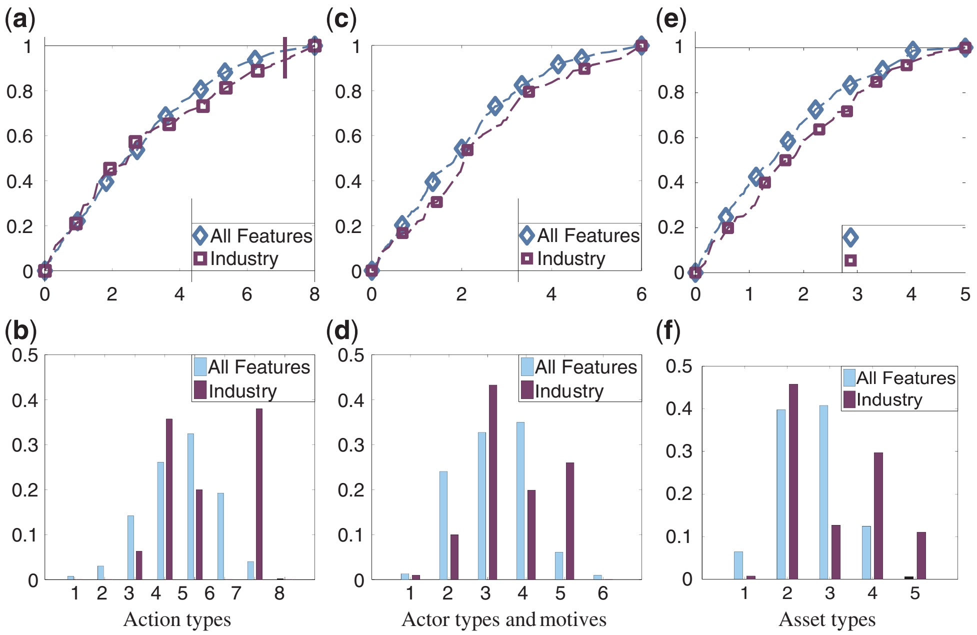

Figure 5. Detection rate vs. average number of risky types (top), and distribution of organizations over the number of types in their risk profiles (bottom).

Table 5. Risk profiles for sample organizations, and their corresponding industries' profiles

Organization	Error	Hacking		Malware	Misuse	Physical		Social
		Comp. Cred.	Other			Theft	Other	
Information		▨	▨	▨				▨
Russian Radio		▨	▨ ×					
Verizon	▨	▨	▨ ×	▨	▨		▨	▨
Public administration	▨	▨	▨	▨	▨	▨		▨
Macon-Bibb County	▨ ×	▨	▨		▨	▨	▨	▨
Internal revenue Service	▨				▨ ×			

Gray cells signify incident types with high risk, and crosses indicate the actual incident.

feature set, as well as the risk profile using only industry. For the information sector, the latter recommends focusing on both types of hacking, as well as social incidents, whereas for public administration it deems all but the second type of physical incidents risky. In contrast, using our extended feature set, we are able to eliminate malware and social incidents as likely threats for Russian Radio, and still provide an accurate risk profile. Similarly for the Internal Revenue Service, we are able to narrow down the list of threats to two types without losing accuracy. Macon-Bibb County and Verizon are assessed to have a broad range of risks, more so than their respective industry average would suggest; this highlights that for these organizations they may be attacked on multiple fronts, which may call for a different type of resource allocation strategy. The point is that this type of fine-grained prediction is much more specific to an organization itself rather than using the industry average as a proxy. We also note that in all these cases our risk profile correctly captured the actual incident occurrences (as indicated by an "×").

It is worth noting that the gray cells in Table 5 not marked with an "×" are incident types deemed likely by our classifier but unrealized in reality (not observed in our dataset). These should not be viewed as discrepancy; rather, the relationship between a predicted risk profile and actual incident occurrence is analogous to that between a dice with a certain probability of turning up each side and the outcome of tossing the dice in a particular random trial. In other words, in the example of the Internal Revenue Service, even though misuse is the only incident that actually occurred, the result suggests that an error event could just as well have happened. This is because in essence our classification constructs risk profiles by extracting details about a business and examining actual incidents that have occurred to other, "similar" companies. In this case, for organizations that share the same business model as the Internal Revenue Service, error and misuse constitute the majority of data breach reports; thus given the information available to us, both incident types are regarded risky.

To close this section, we display the average risk profile over action types of all organizations, as well as average risk profiles over action types for different industry sectors and sizes in Table 6. Each number in the table represents the percentage of organizations, for whom the respective incident type is deemed risky. For instance, 61.9% of all organizations have high risk in hacking incidents due to compromised credentials. However, for 100% of organizations in the information sector this type of hacking poses a high threat. The risk profiles are obtained for the 80% accuracy point in Fig. 5a.

Table 6. Average risk profiles by business sector and size

		Hacking				Physical		
Industry (number of samples)	Error	Comp. Cred.	Other	Malware	Misuse	Theft	Other	Social
Manufacturing (39)	30.8	97.4	51.3	89.7	33.3	28.2	76.9	41.0
Retail trade (63)	34.9	100.0	46.0	76.2	42.9	9.5	68.3	23.8
Information								
Small (49)	22.5	100.0	100.0	65.3	12.2	8.2	38.8	59.2
Large (41)	36.6	100.0	80.5	70.7	36.6	0.0	51.2	87.8
Finance and insurance								
Small (53)	66.0	62.3	18.9	75.5	18.9	34.0	75.5	60.4
Large (91)	64.8	41.8	29.7	31.9	67.0	49.4	86.8	75.8
Pro., Sci. and Tech. Svcs (44)	54.6	72.7	27.3	50.0	27.3	45.5	36.4	43.2
Educational Svcs								
Small (27)	81.5	44.4	14.8	63.0	40.7	92.6	25.9	33.3
Large (46)	89.1	34.8	2.2	19.6	41.3	82.6	41.3	26.1
Health care and social asst								
Small (97)	59.8	28.9	7.2	22.7	54.6	95.9	46.4	10.3
Large (97)	93.8	10.3	3.1	7.2	96.9	96.9	42.3	24.7
Accommodation and food Svcs (33)	72.7	6.1	15.1	48.5	87.9	78.8	54.6	9.1
Public administration								
Small (41)	95.4	85.4	24.4	22.0	63.4	51.2	9.8	19.5
Large (96)	97.9	32.3	10.4	2.1	93.8	67.7	0.0	55.2
Overall (1426)	61.6	61.9	37.5	32.4	56.9	51.5	38.1	38.9

We highlight a number of trends in Table 6. As discussed previously, large companies tend to have higher risk in error and misuse. Sectors that are more prone to error include large health care, and both small and large public administration. Large health care and large public administration companies also run a high risk of misuse. Incidents of error exhibit a substantial presence in all business types, the minimum being 21.2% for information companies. Note that overall, all of the incident types are flagged for at least 30% of our samples, even though their occurrence rate is widely different as evidenced in the last row of Table 4. This is due to our choice of ignoring the a priori distribution of incidents, as explained in detail in Section 6.

Comparing Tables 5 and 6 can help provide some insight on how having additional features has helped eliminate (or introduce) possible risks for those sample organizations. For instance, small information companies tend to have lower risk in social incidents, and this has helped us eliminate this category as a possible threat for Russian Radio. We can also see that small public administration and large information companies have a more uniform risk among all types, attributed to the risk profiles for Macon-Bibb County and Verizon, respectively. The Internal Revenue Service, a large public information company, is expected to have less risk in the second type of physical incidents, as well as hacking and malware. Note that one cannot completely explain the generated risk profiles by only looking at business sector and size information alone, as they are a result of analyzing the dataset's distribution over all the features in Table A1. For instance, large public administration organizations tend to have higher risk in social events than small ones, even though this incident type has been flagged for Macon-Bibb County and not the IRS. In this case, other features of the IRS have contributed to its lower risk.

Conclusion

Our results demonstrate how, and to what extent, can business details about an organization help forecast its overall risk of data breach, as well as the relative risk of experiencing different types of data incidents. We observe that it is possible to forecast future security incidents with high accuracy. However, even though there is notable correlation between organization features and the incident signatures in our dataset, it is impossible to assert with certainty the types of incident an organization is likely to face. We acknowledge the fact that there is an inherent randomness in incidents suffered by organizations: no business is prone to a single type of incident. As observed in our results, while risk in incidents such as hacking and theft may vary largely across sectors, any organization is likely to experience incidents due to miscellaneous errors. Nonetheless, feeding further information into our classifiers may help construct more accurate risk profiles. The feature set used in this article provides only high level information about the organization itself, and not its security posture. Even though these features are the easiest to obtain, as they all are publicly available, further information indicative of an organization's security policies will undoubtedly help narrow down its risk profile. Externally observable signals, such as the ones used in [7], as well as inside information, may be used to infer a business's security posture.

Note that our model's output is as good as the labels that our incident dataset provides. VCDB reports publicly disclosed data breaches, and therefore our model's output is essentially assessing the risk of publicly disclosed data breaches. Whether these results can generalize for data breaches that were not reported, depends on how representative our incident samples are. There are a number of biases in self-reporting of incidents, and those that are externally detected by a third party. For example, incidents that involve customer information such as credit card numbers are more frequently reported, and attacks such as website defacement can be easily detected externally. However, we might not have a representative sample of incidents that result in the theft of trade secrets, and proprietary information.

Furthermore, the discrepancy in reporting rates of different incident types, might lead to underestimation or overestimation of risk in our assessments. While we alleviate this issue in our treatment of rare events in Section 6 for estimating risk distributions, the problem remains for overall risk assessments. Moreover, we only focus on discrepancies for specific incident types (by action, actor, and asset types), and do not take other factors into consideration. Other variables that may impact the reporting of incidents include region (VCDB mainly focuses on US incidents), and business sector. A more comprehensive source of data breaches can improve our assessments, and allow us to use more sophisticated machine learning methods in order to find factors that influence cyber-risk.

It is worth noting that incident types are often too ambiguous to act upon for a security unaware business operator, hence the need for explicit, actionable security recommendations. Note that there indeed exist frameworks providing such recommendations. For example, the SANS institute's critical security controls [26] is composed of 20 categories of security controls, each describing a specific action or policy that can be implemented by a business in order to raise its security levels. Verizon uses this framework to provide general security recommendations in its annual Data Breach Investigations Report, and the SANS institute offers a partial mapping between these controls and the VERIS incident categorizations. Translating our risk profiles into actionable security recommendations is a direction for future work. Furthermore, our current dataset does not contain information on the monetary impact of each incident type. Obtaining such information, and combining it with the cost of protection for each incident type, will allow us to provide more economically-informed recommendations.

Funding

This work was supported by the National Science Foundation (NSF) [CNS-1422211]; and by the Department of Homeland Security (DHS) Science and Technology Directorate, Homeland Security Advanced Research Projects Agency (HSARPA), Cyber Security Division (DHS S&T/HSARPA/CSD), BAA 11-02 [contract number HSHQDC-13-C-B0015].

References

1. The Web Application Security Consortium. *Web Hacking Incident Database*. http://projects. webappsec.org/w/page/13246995/Web-Hacking- Incident- Database (30 September 2015, date last accessed).
2. Passeri P. Hackmageddon. http://hackmageddon.com (30 September 2015, date last accessed).
3. VERIS Community Database (VCDB). http://vcdb.org (30 September 2015, date last accessed).
4. DMOZ. The Open Directory Project. http://www.dmoz.org (30 September 2015, date last accessed).
5. Amazon Web Services. Alexa Web Information Service. http://aws. amazon. com/awis (30 September 2015, date last accessed).
6. Neustar. IP Intelligence. https://www.neustar.biz/services/ip-intelligence (30 September 2015, date last accessed).
7. Liu Y, Sarabi A, Zhang J *et al*. Cloudy with a chance of breach: forecasting cyber security incidents. In: 24th USENIX Security Symposium, USENIX Association, 2015.
8. Verizon Enterprise. Data Breach Investigations Reports (DBIR). http://www.verizonenterprise. com/DBIR (30 September 2015, date last accessed).
9. Hines C. *Why Companies Must Adopt the "Assume Mentality" When It Comes to Breaches*. https://blog.cloudsecurityalliance.org/2015/02/27/why-companies-must-adopt-the-assume-mentality-when-it-comes-to-breaches (27 February 2015, date last accessed).
10. Thonnard O, Bilge L, Kashyap A *et al*. Are you at risk? Profiling organizations and individuals subject to targeted attacks. In: *Financial Cryptography and Data Security*, Springer, 2015.
11. Soska K, Christin N. Automatically detecting vulnerable websites before they turn malicious. In: 23rd USENIX Security Symposium, USENIX Association, 2014.
12. Edwards B, Hofmeyr S, Forrest S. Hype and heavy tails: a closer look at data breaches. In: Workshop on the Economics of Information Security (*WEIS*), 2015.
13. IBM. 2015 Cost of Data Breach Study. http://www-03.ibm.com/security/data-breach (30 September 2015, date last accessed). 2015.
14. Campbell K, Gordon LA, Loeb MP *et al*. The economic cost of publicly announced information security breaches: empirical evidence from the stock market. *J Comput Secur* 2003;**11**:431–48.
15. Cavusoglu H, Mishra B, Raghunathan S. The effect of internet security breach announcements on market value: capital market reactions for breached firms and internet security developers. *Int J Electron Commer* 2004;**9**:70–104.
16. Acquisti A, Friedman A, Telang R. Is there a cost to privacy breaches? An event study. *ICIS 2006 Proc* 2006;1563–80.
17. Kannan K, Rees J, Sridhar S. Market reactions to information security breach announcements: An empirical analysis. *Int J Electron Commer* 2007;**12**:69–91.
18. Gordon LA, Loeb MP, Zhou L. The impact of information security breaches: has there been a downward shift in costs? *J Comput Secur* 2011;**19**:33–56.
19. Romanosky S, Hoffman D, Acquisti A. Empirical analysis of data breach litigation. *J Empir Leg Stud* 2014;**11**:74–104.
20. Romanosky S, Telang R, Acquisti A. Do data breach disclosure laws reduce identity theft? *J Policy Anal Manag* 2011;**30**:256–86.
21. The VERIS Framework. http://veriscommunity.net (30 September 2015, date last accessed).
22. NAICS Association. http://www.naics.com (30 September 2015, date last accessed).
23. Amazon Web Services (AWS). http://aws.amazon.com (30 September 2015, date last accessed).
24. Amazon Web Services. Alexa Top Sites https://aws.amazon.com/alexa-top-sites (30 September 2015, date last accessed).
25. Ensemble Methods http://scikit- learn.org/stable/modules/ensemble.html (30 September 2015, date last accessed).
26. SANS Institute. Critical Security Controls. https://www.sans.org/critical-security-controls (30 September 2015, date last accessed).

Appendix

Table A1. Features and feature importances for all classifiers

	Industry	Employee Count	Region	Rank	Local Rank	Rank History	Links In	Website Age	Speed	Locale	Traded	Category	Network Size	Network Name	Network Industry
Overall															
Overall	x	x	x	7.6	11.9	12.6	6.3	3	5.4	4.5	0.2	36.2	3.3	3.6	5.3
Action															
Error	21.4	25.2	18.8	x	x	x	x	9	x	x	5.7	x	19.9	x	x
Hacking	27.8	9	29.2	8.7	x	x	10.5	8.1	x	x	x	x	6.8	x	x
Comp. Cred.	0	25	8.3	x	x	x	x	16.7	25	x	x	8.3	16.7	x	x
Other	0	17.4	33.3	x	10.7	11.7	x	4.6	10.6	x	x	x	11.6	x	x
Malware	20.5	8.2	4.2	x	13	33	x	x	7.7	1.8	x	x	11.5	x	x
Misuse	17.4	9.7	6.9	24.2	x	x	19.5	9.3	x	11.4	1.6	x	x	x	x
Physical	11.3	3	7.6	x	x	33.1	6.1	x	5.6	x	0.4	33	x	x	x
Theft	26.4	0.5	2	x	x	38.7	x	6.4	6.9	x	1.9	x	17.2	x	x
Other	24.9	9.6	4.1	x	x	x	16.1	24.9	x	x	x	x	20.4	x	x
Social	14.2	21.4	18.9	x	18.8	x	x	x	26.8	x	x	x	x	x	x
Actor															
External	28.9	7.1	11.7	15.4	x	x	x	6.1	x	17.4	1.8	x	11.6	x	x
Financial	12.7	13.3	27.9	2.1	30.9	9.8	3.2	x	x	x	x	x	x	x	x
Ideology	18.7	38.5	25.8	6.8	x	x	4.1	x	6	x	x	x	x	x	x
Other	13.6	4.1	40.6	x	33.5	x	x	x	8.2	x	x	x	x	x	x
Internal	28.3	16.6	40.8	x	x	x	x	12.2	x	x	2	x	x	x	x
Financial	17.4	0	0	x	x	x	x	12.5	18	x	x	37.8	14.3	x	x
Other	8.3	0	0	x	x	37.4	18.3	x	x	x	x	20.4	15.6	x	x
Partner	19.3	11.8	12.2	x	x	x	x	16.5	x	5.3	x	22.3	12.6	x	x
Asset															
Kiosk/Terminal	13.3	11.7	5.1	x	x	9.9	1.9	x	2.9	x	0.9	54.4	x	x	x
Media	10.4	8.3	10.6	x	7.8	x	3.9	x	3.1	x	0.6	55.2	x	x	x
People	19.7	15.4	24.7	x	x	x	x	x	28.9	10.5	0.7	x	x	x	x
Server	15.7	3.1	17.6	x	13	11.9	x	3.7	x	2.2	1.6	27.2	3.9	x	x
User Device	8.3	5.5	7.2	14.3	17.3	38.9	x	6	x	2.3	0.2	x	x	x	x

Crosses indicate features that have not been used in training the corresponding classifier.