A New Approach to Quantifying, Reducing and Insuring Cyber Risk: Preliminary Analysis and Proposal for Further Research

Neil Gandal (Tel Aviv University)
Michael Riordan (Columbia University)
Shalom Bublil (Kovrr)

February 26, 2020

Abstract

Few would dispute that cyber risk is a very serious problem for the global economy and for society. But there is a "disconnect" between acknowledgement of the problem and action to address the problem. What is the relationship between vulnerabilities, preventive measures, and security incidents, like the leaking of sensitive data (say credit card information) to the web? To the best of our knowledge, little if anything is known about the relationship among these variables and no one has examined this issue empirically at the micro level, that is, at the level of the firm.

In this paper, we put together a remarkable and unique cross-sectional data set at the firm level that includes information on vulnerabilities, attempted email attacks, incidents (breaches), precautions (security measures.) and firm characteristics. The data set contains slightly under 1000 small and medium firms in the U.K. We empirically examine the data and show that there are meaningful correlations among incidents and the other variables. Finally, we estimate a reduced form model with incidents as the dependent variable to illustrate the potential from employing such data.

1. Introduction

We are moving into a world in which assets are primarily digital and not physical. In 2018, physical assets such as buildings and equipment accounted for only 16 percent of the value of S&P 500 firms.[1] Digital assets are increasingly subject to cyber risks. At the "micro" level, such risks include falling prey to ransomware, viruses, and phishing attacks. Cyber-attacks can shut down ecommerce websites, steal assets from bank accounts, and shut down key government websites. Cyber-attacks can also result in huge or even catastrophic losses for two reasons: (i) correlated risk and (ii) interdependent risk. Examples of such risks and attacks include:

- An employee of a Pennsylvania city missed some software updates in 2019. The employee then clicked on a phishing email (some of which are quite sophisticated) and the malware spread through the system. (The software updates apparently would have blocked the malware.) The attack shut down the city government's computers for more than a week. It took more than 1 Million USD to rid the system of the malware.[2]
- The use of common infrastructure providers creates correlated risks. In February 2019, hackers successfully breached the servers of the email provider VFEmail. The hackers re-formatted all the disks on every VFEmail virtual machine (VM) server, essentially destroying all of them including all backup servers. As a result, all US customers of VFEmail lost up to twenty years of data in the process.[3] VFEmail was a small provider. The email service provider market is highly concentrated with Microsoft, Google, and Yahoo among the dominant providers. A successful attack on one of industry leaders would have devastating consequences.
- Connected networks create huge interdependent risk. The "WannaCry" ransomware attack in May 2017 is an example of such an attack. The attack targeted computers running an the Microsoft Windows operating system. The initial infection was through a vulnerability in the operating system. (A patch, or update had been available for quite some time, but not every organization installed the update.) Within a day the code was reported to have infected more than 230,000 computers in over 150 countries. The attack is called a network worm because it had a "transport mechanism" that automatically spread it throughout the network without

[1] See https://ipcloseup.com/2019/06/04/21-trillion-in-u-s-intangible-asset-value-is-84-of-sp-500-value-ip-rights-and-reputation-included/

[2] See Ransomware Attacks Are Testing Resolve of Cities Across America, by New York Times, By Manny Fernandez, David E. Sanger and Marina Trahan Martinez, available at https://www.nytimes.com/2019/08/22/us/ransomware-attacks-hacking.html.

[3] See https://krebsonsecurity.com/2019/02/email-provider-vfemail-suffers-catastrophic-hack/. By Brian Krebs, 12 February 2019, "Email provider VFEmail suffers 'catastrophic' hack

any human interaction. The "worm" scanned for vulnerable operating systems and gained access to such operating systems through the vulnerability; it then installed and executed a copy of itself. The attack was estimated to have caused billions of U.S. dollars in damages.

Few would dispute that cyber risk is a very serious problem for the global economy and for society. But there is a "disconnect" between acknowledgement of the problem and actions taken to address the problem:

- One might expect that greater awareness in recent years about viruses and malware would have resulted in nearly all firms and consumers protecting and insuring their websites, computers, and their digital assets. That is not the case. Surprisingly, even in 2016, nearly 25% of the personal computers in the world had no basic protection from viruses and malware. There are currently (2019) more than 4.4 billion users worldwide! Hence, the magnitude of the problem is substantial – and the large number of unprotected users provides strong incentives for cyber-criminals. The New York Times reported (on August 23, 2019) that more than forty local governments in the United States suffered cyber-attacks in 2019.[4]
- Given the explosion of cyber-security risk, it would be natural to assume that cyber insurance would have become a growth industry. However, that is not true. Relative to cyber risk, there are a dearth of cyber insurance policies as reported by Romanosky et al (2019.) According to Allied Market Research, the North American cyber insurance market (mainly the U.S.) accounted for around 87% of the global cyber insurance market in 2015.[5] While the U.S. market is more advanced than other countries, only around one third of US companies have purchased some sort of cyber insurance. In the case of Europe, the numbers are much smaller.[6] This leaves a huge part of the digital economy dangerously unprotected.
- Additionally, despite attention from academics, policymakers, and industry experts, a recent manifesto (2019) by eighteen researchers, industry experts, and policymakers lamented the lack of progress in cybersecurity research. "Unfortunately, research progress has been modest and has not been sufficient to answer the 'call to action' in many prestigious committee and agency reports."[7]

[4] See Ransomware Attacks Are Testing Resolve of Cities Across America, by New York Times, By Manny Fernandez, David E. Sanger and Marina Trahan Martinez, available at https://www.nytimes.com/2019/08/22/us/ransomware-attacks-hacking.html.

[5] "See https://www.alliedmarketresearch.com/press-release/cyber-insurance-market.html.

[6] Providing detailed information on the state of the cyber insurance markets in the EU and other important markets will be part of the work plan.

[7] A Research Agenda for Cyber Risk and Cyber Insurance," page 1, available at https://weis2019.econinfosec.org/wp-content/uploads/sites/6/2019/05/WEIS_2019_paper_35.pdf.

Cyber-risk is very different from other forms of risk for the following reasons:

1. Long-term historical data do not exist.
2. There are adversaries creating the dangers, and these adversaries behave strategically.
3. There is interdependent security and correlated risks.
4. Cyber-attacks can go undetected for long periods of time

However, it is also true that like other types of risk (think of health,) there are many variables that might affect whether cyber risk leads to security incidents: Such risks include

- Different types of vulnerabilities in computer systems,
- Infrastructure and preventive measures

Further, despite strategic attackers, most cyber-attacks are still based on well-known attacker strategies, such as phishing email attacks, in which a user is tricked into clicking a malicious link.[8] According to the Microsoft 2018 Security Report,[9] phishing indeed remains the most frequently employed cyber-attack. The important takeaways are (I) it is possible to defend against such well-known attack strategies and (II) it is possible to examine the relationship between vulnerabilities, preventive measures, and security incidents.

To the best of our knowledge, little if anything is known about the detailed relationship between specific risks/infrastructure risks, email attacks, and various types of successful attacks (incidents.) Further, to the best of our knowledge, no one has examined this issue empirically at the micro level, that is, the level of the firm.

The "state of the art" regarding how cyber risk factors affect outcomes (incidents) is probably similar to the "state of the art" fifty years ago regarding how health risk factors affect health outcomes. Yes, we can say that it is a good idea to install updates of the operating system. But to the best of our knowledge, there is no research on the relationship between cyber risks and incidents at the level of the firm.

[8] See UK National Security Center 2018 report entitled: "Cyber Security Kits for Boards, available at https://www.ncsc.gov.uk/collection/board-toolkit/introduction-cyber-security-board-memberscyber.

[9] Available at https://info.microsoft.com/rs/157-GQE-382/images/EN-US_CNTNT-eBook-SIR-volume-23_March2018.pdf.

The goal of this paper is to show that such research is possible. We first lay out the key elements of a theoretical model relating vulnerabilities, attempted email attacks, precautions, and firm characteristics, and precautions (security measures.) choices to incidents. Then, using the unique cross-sectional data set, we take a first step towards quantifying the relationship between specific risks, including vulnerabilities and attempted email attacks, and security incidents (breaches.) We show that there are meaningful correlations in the relatively small cross-sectional data set. We then estimate a reduced form model with incidents as the dependent variable to illustrate the potential from employing such data.

We believe that this will enable the start of an analytically based micro approach to measuring cyber-risk, benefits from precautions, and the pricing cyber insurance. There is a long, long way to go, since panel data is necessary to handle endogeneities. Nevertheless, our initial analysis strongly suggests that it will be possible to estimate the relationship between specific risk factors and the probabilities of various cyber incidents. We will begin putting together a panel data set over time (at least one year) to facilitate this. At this stage, we want to show meaningful information can be obtained from a relatively small cross-sectional data set.

The paper proceeds as follows. In the remaining part of this section we discuss the relevant literature. Section two shows the model we will employ when we have panel data on a larger number of firms. Section three describes and explores our unique data set, while section four provides very preliminary (reduced form) results using the data set we put together. Section five offers brief concluding remarks.

1.1 Literature Review

Cyber security issues arise because there are malevolent actors (attackers) and because there are vulnerabilities in information technology systems that can be exploited. Cybersecurity is important because modern society is extremely dependent on information technology for many critical functions.[10] These functions include critical infrastructure, health systems (the WannaCry ransomware attack shut down many hospitals) and our financial system among others.

Over the past few years, researchers of information security have come to realize that cyber security breaches are often not due to the lack of technological solutions, but rather due to the absence of

[10] Clark et al. (2015) is a comprehensive report that provides an overview of the technical and policy issues of cybersecurity; it surveys two decades of research.

appropriate incentives. Humans are often considered the weakest link in internet security. It is now generally accepted that computers will never be completely secure, and that to manage the risks, economics is critical for helping insure improved security. Understanding that behaviour is an important component of cyber security led to the development of research on the economics of cyber security. Anderson and Moore (2006) provide an overview of early work in the field.

There are now many theoretical articles about cyber risk, cybersecurity, and cyber insurance. To date, contributions by economists have primarily been theoretical and have focused on (i) the lack of incentives for individuals, firms or network operators to take adequate security precautions, e.g., Varian (2004) and Camp and Wolfram (2004,) (ii) incentives for firms ex-ante to reduce the number of software vulnerabilities (Arora, et. al 2006,) and (iii) the incentives for firms to disclose information about vulnerabilities (Choi, Fershtman & Gandal, 2010). Despite its importance, cybersecurity has not yet attained widespread attention in the economics discipline.

A primary focus has been to theoretically determine the optimal level of investment in cyber security. Papers in this genre typically aggregate the decisions of a firm that wishes to protect its assets into a single economic choice variable, namely the amount of money the firm will invest in security. The most well-known work in this genre is the pioneering paper by Gordon and Loeb (2002.) They assume that that there is a function that links the cost of security investment to the probability of suffering a well-defined monetary loss. The theoretical literature extends Gordon and Loeb (2002) in several ways. Many papers add interdependent and correlated risk. A large literature theoretically addresses cyber attackers who behave strategically to the framework. Bohme et al. (2017) provide a nice summary of this literature. While these theoretical models illustrate interesting points, they cannot be estimated empirically. Thus these models cannot be used to estimate the relationship between specific risk factors and the probabilities of various cyber incidents.

The empirical research on cyber-risk and cyber-security has focused on estimating the costs of various data breaches. There are two basic approaches:

- Calculate the cost of cyber incidents by analyzing data at the level of the incident: Biener, Eling, and Wirfs (2015) is an example of this genre. They find a mean loss per cyber incident of 36.8 million Euros.
- Use "event study" methodology to estimate the cost of data breaches at publicly traded firms. Several authors (Campbell, Gordon, Loeb, and Zhou 2003; Gatzlaff and McCullough 2011) have taken this approach.

These approaches, while helpful, do not enable us to estimate the relationship between specific vulnerabilities and specific cyber incidents.

2. Theory

Clearly, there is a dynamic process over time, that it, firms invest (or do not invest) in precautions, attacks occur and some of the attacks result in breaches/incidents. Then the cycle repeats. The simplest theoretical model that fits this setting is as follows:

First stage: In the first stage, the firm will choose which investments to make in various types of infrastructure and "precautions" to fix vulnerabilities and improve security.

Second stage: In this stage, incidents (breaches) either occur or do not occur.

The process then repeats itself, i.e., there is a repeated game over time period by period. For this paper, we will focus on the "one period" game, but we will employ a dynamic model when we have panel data.

Empirical strategy:

Stage 1 and stage 2 are sequential. In stage 1, we will estimate the precaution decisions taken by the firm as a function of its characteristics, the expected benefits from precautions (reducing vulnerabilities) and the costs of undertaking such precautions.

However, "precaution" decisions in stage 1 depend on expected outcomes in stage two. The firm takes this into account when choosing actions in stage 1. Hence the precautions are endogenous and we will have to address this issue by finding appropriate instruments. This should be straightforward. Since we do not yet have panel data, we cannot estimate the first stage now.

In the second stage, we will estimate the relationship between incidents (the dependent variable) and firm characteristics, vulnerabilities (that remain,) and infrastructure (precautionary) choices made by the firm in stage 1. This is analogous to determining how risk factors at the individual level affect the probability of having a heart attack.

Since the costs of various infrastructure choices and precautions are publicly available, we will be able to estimate the costs and benefits of these choices.

Variables in the Model

We define Y_1 to be equal to the number of incidents suffered by each firm. We also define a binary variable Y_2, which equals one if the firm suffered a breach or incident.

Let X be characteristics of the firm. For example, whether or not the firm is an ecommerce firm is an example of such a characteristic. The characteristics of firms are exogenous.

Let Z be the set of precautions taken by firm in first stage. These can be investments in infrastructure or choice of infrastructure provider.

Let V be the number of ex-post vulnerabilities that remain following investments in security.

From our cross sectional data, we can use the data we have to estimate the second stage. This can be thought of as a reduced form model. In stage two, we will estimate a regression, where the dependent variable is either Y_1, (the number of incidents) or Y_2 (whether or not the firm suffered an incident.) The independent variables are precautions (Z), firm characteristics (X) and any remaining vulnerabilities (V.)[11] The second stage of the model is then:

1. $Y = f(X, Z, V)$

We will estimates stage two regressions in section four. First we describe and examine our data descriptively in the next section.

3. Data

This section illustrates the rich information that can be obtained from cross-sectional data. The data are unique and critical to the analysis. Hence, it is worth understanding at a micro level what data are available for the analysis. Here, we briefly summarize the categories of variables that are available in the pilot data set. The full set of variables available for the analysis is in the Appendix.

[11] We lack instruments at this stage. Hence, the estimates should be interpreted as correlations rather than causal effects.

The data for our analysis were provided by Kovrr,[12] an Israeli cyber risk modelling firm. One of us (Shalom Bublil) is the Chief Product Officer at Kovrr. The data offer the potential to finally understand the relationship among vulnerabilities, infrastructure, and incidents at the individual firm level. The pilot data Kovrr provided include 990 small and medium UK enterprises. See the appendix for full descriptions of the variables available in the data set. We have the following firm level data for 990 small and midsize firms in the United Kingdom UK.[13] We now give a brief overview of variables for which we have data.

:

- Traditional enterprise characteristics, including data on revenues, employees, industry, whether the firm is engaged in ecommerce, i.e., sales via the Internet.
- Vulnerability data at the level of the enterprise: These data are obtained by Kovrr from online data in the public domain. The "process" is analogous to figuring out what types of locks exist on doors in a University department. Faculty offices are not entered, but it is possible (by observation walking down the hallway) to determine the configuration of the locks (i.e., the infrastructure employed.) Given knowledge about the different types of locks, it is then possible to know whether the locks are secure, i.e., whether they have vulnerabilities. We want to emphasize that all of the data Kovrr collects are in the public domain of the Internet.

  The Kovrr collection process leads to output on thirteen different critical software vulnerabilities. In the pilot data set, firms had average 4.34 vulnerabilities. Fully thirteen percent of the firms had no vulnerabilities, while twenty-five percent of the firms had seven or more vulnerabilities. An example of a vulnerability is an invalid security (SSL) certificate.
- Data on incidents (breaches.) These data come from outside sources (aggregators.) Kovrr obtained these data. An example of an incident in "sensitive data leaked to the web." This is considered a serious incident. Twenty-seven percent of the firms in the pilot data suffered from this incident.

  We also have data on two reputation variables and the data come from outside sources. One of the variables, 'Reputation_domain_malware' takes on the value one if one or more security vendors considered one or more of the company's domains as malicious. Similarly, the variable 'Reputation_ip_spam_or_malware' takes on the value one if one or more

[12] See https://www.kovrr.com/.

[13] We have a full list of all UK registered companies from the UK "Company House portal." In order to construct the dataset, small and medium firms with more than fifteen employees and less than 250 Million pounds in revenue were sampled. We strove for a representative sample across industries and geographical locations.

security vendors associates one of the companies domain as proliferating spam. We consider these variables to be incident variables as well. (These variables likely capture interdependent risk.)

- Overall, 39 percent of the firms in the data set suffered an incident. As noted, the incident that was most prolific was that of “sensitive data leaked to the web.” This obviously is a serious incident, and could compromise consumer data held by the firm. Overall, as noted, 27 percent of the firms suffered from that incident. Only one other incident was suffered by more than five percent of the firms: Fourteen percent of the firms suffered from “Reputation_ip_spam_or_malware.” This is also a serious incident because it potentially causes harm to other firms.
- Precautions Data: These data are can be thought of as security precautions or investments. For example, a CAPTCHA or “Completely Automated Public Turing test to tell Computers and Humans Apart” protects websites against bots by generating “tests” that humans can pass but (current) computer programs cannot. Four variables in the data set represent precautions or security investments.[14] They are collected by Kovrr during the “expedition.”

- We also have data on one variable that measures whether the firm was attacked. The variable “incident_email_infection_attempt” takes on the value one if there was an attempt to attack one of the business employees based on a malicious email. In particular, an email in the particular company was included in the list of email addresses that were targeted. The most common type of email attack is a phishing email attack.

In Appendix A, we include descriptive statistics and correlation among the variables we use in the analysis. In Appendix B, we include the fill list of variables that our available.

Initial Descriptive Exploration of the Data: Vulnerabilities and Incidents

Summary data show that:

- Nearly 87 percent of the firms in the sample (858/990) have at least one vulnerability.
- Overall, 39 percent of the firms (391/990) suffered at least one incident.

The relationship between these variables is shown in Table 1.

[14] We also have data on firm choices among different infrastructure providers.

	At least one incident	No incidents	Total
At least one vulnerability	367	491	858
No vulnerabilities	24	108	135
Total	391	599	990

Table 1: Vulnerabilities and Sensitive data leaked to the web.

Naïve insurance, i.e., the same insurance for all and a competitive industry would set the cost of cyber insurance equal to the (average) expected cost of having at least one security incident. This equals 0.39* L, where L is the expected loss if the firm suffers an incident.

In the raw data, 367 out of the 858 firms (43 percent) with vulnerabilities suffered from at least one incident. On the other hand, only 24 out of 135 firms (18 percent) with no vulnerabilities suffered an incident. Thus the probability of suffering an incident is more than twice as much if a firm has any vulnerabilities. This suggests more efficient pricing of insurance. In the case of a competitive insurance industry, where price equals expected cost,

Price = 0.43*L for firms that have a vulnerability.
Price = 0.18*L for firms without any vulnerabilities.

The difference in probabilities of the incident between those firms with vulnerabilities and those firms without vulnerabilities will provide firms with incentives to eliminate vulnerabilities.

The Relationship between an Attempted Email Attack and the Leaking of Sensitive Data

An attempted email attack is an "in-between" stage event that occurs between stage one and stage two. The highest correlation by far (0.49) among an attempted email attack and one of the incidents is between (i) an attempted email attack and (ii) the leaking sensitive data leaked to the web. In a phishing attack, a user is tricked (through an email) into clicking a malicious link. This can lead to the revealing of sensitive information as well as other problems.[15]

This background suggests a causal pattern: namely firms that suffer from attempted email attacks are more likely to suffer a leak of sensitive data. The relationship between an attack on email and an occurrence of the "data leak" incident are remarkably revealing! See Table 2.

[15] This is what happened in the Pennsylvania incident. Phishing attacks have become very sophisticated.

	Sensitive data leaked to Web	No Sensitive data leaked to Web
Attempted attack on email	87	5
No attempted attack on email	178	720

Table 2: Attempted attack on email and Sensitive data leaked to Web

Table 2 shows that 95 percent (87/92) of the firms that suffered an attempted email attack were found to have leaked sensitive data to the web. On the other hand, only 20 percent (178/898) of the firms that did not suffer an attempted email attack were found to have leaked sensitive data to the web. This table is striking and reveals the power of these data! We had no clue as to the empirical relationship between these variables until we began digging deeply into the data.

There is also a positive correlation between vulnerabilities and attempted attacks: only ten percent of those firms with no vulnerabilities suffered an attempted email attack, while 27 percent of those with at least one vulnerability suffered an attempted email attack.

Since no one has ever analysed such data at this micro level, patterns are not known. This simple analysis suggests that many regularities can be revealed by detailed examination of the data.

4. Analysis: Beyond One Risk Factor: Estimating the Second Stage

Here we run a regression as described above. The left-hand side variable (Y) measures incidents. Right hand variables include firm characteristics, vulnerabilities, whether an attempted email attack occurred and precautionary (security) measures using some of the rich information available in the data set.

Hence we run the following regression for two different dependent variables:

$Y = \beta_0 + \beta_1 X + \beta_2 V + \beta_3 A + \varepsilon$, (ε is a random error term, with expected value 0.)

I. Logit Regression: The dependent variable takes on the value one if the firm suffered a security incident and zero otherwise. The Independent Variables we employ are:

X: This variable takes on the value one if the firm is engaged in ecommerce.[16]

V: The number of vulnerabilities the firm has. (Nothing changes qualitatively in the empirical results if we use a binary variables that takes on the value one if the firm had any vulnerability.

[16] At this stage, we have very limited data on firm characteristics. When we begin collecting the panel data, we will greatly expand on the characteristics of the firm. Hence, this regression should be viewed as illustrative.

A: This variable takes on the value one if there was an attempted email attack

Descriptive statistics and correlations for these variables appear in the Appendix.

This regression appears in the first column of Table 3.

II. Ordinary Least Squares Regression: The dependent variable is the number of security incidents. The Independent Variables we employ are as in I.

This regression appears in the second column of Table 3.

The regressions show that the estimated coefficients on vulnerabilities and an attempted email attack are significant at the 99% level of confidence.[17] The ecommerce coefficient is positive and significant (at the 95% level) in the first regression. The ecommerce coefficient is positive, but not significant in the second regression.

Dependent variable	Incident (Yes, No)	# of Incidents
Independent Variables		
Ecommerce (X)	0.30** (0.15)	0.05 (0.05)
Number of Vulnerabilities (V)	0.16*** (0.028)	0.05*** (0.01)
Attempted email attack (A)	4.10*** (0.60)	1.29*** (0.09)
N=990	Pseudo R^2 = 0.15	Adjusted R^2 = 0.19

Table 3: Regression results[18] (standard errors in parentheses)

[17] Because of potential endogeneities, these coefficients should be interpreted as correlations rather than a causal relationship. It is possible that an attempted email attack could be viewed as an endogenous event. In that case, we could estimate a modified two stage model. First, we could run a regression with attempted email attack on the left hand side and the number of vulnerabilities and whether the firm was an ecommerce firm on the right hand side. Then we could use the residuals from this regression as a right hand variable instead of attempted email attack. When we do that (with the number of incidents as the dependent variable,) we find that both the number of vulnerabilities and the "residuals" are significant in explaining the number of incidents. The estimated coefficients on (I) the number of vulnerabilities and (II) whether the firm is an ecommerce firm are virtually identical to those in the regression in column 2 of Table 3. Hence, the potential endogeneity of an attempted email attack can be handled in our analysis

[18] *=significant at 90% level, **=significant at 95% level, ***=significant at 99% level.

Endogeneity of Precautions:

Our analysis has shown that interesting patterns are revealed in this small cross-sectional data set. A much larger data set will be orders of magnitude more helpful. More importantly, we will need panel data to address the issue of endogeneity of precautions/investments by the firms.

We can see the "endogeneity" problem when we look at the relationship between precautions (which are "first stage" choice variables) and breaches (i.e. successful attacks.) For example, if we were to add an independent variable that measures precautions (either whether the firm has taken any precautions or the number of precautions that it has taken,) to the regressions in Table 3, the coefficients on the "precautions" variable is positive. Obviously, that does not mean that taking precautions leads to incidents, but rather that those who suffered breaches/incidents were likely to install precautions following an incident.

Hence, it will be essential to have panel data at different points in time, so we can see how firms respond to breaches. We are working on generating a panel data set using the 990 firms in this study – and will employ such data in future work.

5. Brief Concluding Thoughts

We envision that going forward, this research will change the way that firms examine their cybersecurity investments, and will enable a robust cybersecurity insurance market. We expect that regulators will help disseminate the knowledge developed in this project. We envision databases being built (and constantly updated) and tools being developed together by industry, academia, and government regulators. Taken together, this long-term project should help reduce the cyber risks society faces and better manage and efficiently insure the cyber risks than cannot be completely eliminated.

Appendix A: Descriptive Statistics and Correlations

	# of Obs.	*Mean*	*Std. Dev*	*Min*	*Max*
Number of Incidents	990	0.60	0.93	0	6
Any Incidents	990	0.39	0.49	0	1
Number of Vulnerabilities	990	4.36	2.66	0	9
Attempted Email Attack	990	0.09	0.29	0	1
Ecommerce	990	0.38	0.49	0	1

Table A1: Descriptive Statistics

	# Incidents	Any Incident	# of Vulnerabilities	Attempted email attack	Ecommerce
Number of Incidents	1				
Any Incidents	0.80	1			
Number of Vulnerabilties	0.17	0.20	1		
Attempted Email Attack	0.41	0.37	0.05	1	
Ecommerce	0.06	0.09	0.10	0.04	1

Table A2: Correlations

Appendix B: Available Data

1. Vulnerability Variables:

These data are obtained by Kovrr during an “expedition,” as described above.

'vulnerability_domain_registration_exposure' - domain is exposed to hijacking, allowing the hijacker to assume control on the hijacked domain.
'vulnerability_email_no_validation' - email doesn't validate sender identity.
'vulnerability_encryption_certificate_invalid' - usage of invalid SSL certificate. A digital certificate is a
digital form of identification.
'vulnerability_encryption_poodle' - critical ssl (secure sockets layer) vulnerability. SSL is a secure protocol for encrypting information sent over the internet.
'vulnerability_encryption_ssl_redirect_non_secure' - unsecured ssl redirection of a webpage
'vulnerability_server_app_cve' - vulnerability in a server. This vulnerability might result in sensitive data being accessible to other users.
'vulnerability_server_cve_availability' - vulnerability that might lead to denial of service (DoS), making the server inaccessible to its intended users.
'vulnerability_server_cve_confidentiality' - vulnerability that might lead to data leakage.
'vulnerability_server_cve_dos' – additional vulnerability that might lead to denial of service.

'vulnerability_server_cve_integrity' - vulnerability that might lead to damage of integrity of data.
'vulnerability_server_exposed_dns' – vulnerability that exposes the DNS (domain name system) server to exploitation or manipulation.
'vulnerability_server_no_port_filter'- server doesn't filter ports. Unfiltered ports are accessible and might by susceptible to vulnerabilities.
'vulnerability_server_os_cve' - operating system of the system is unpatched and vulnerable. Unpatched software refers to a code with known security weaknesses.

All of the above variables are binary, i.e., either equal to one if the firm has such a vulnerability and zero otherwise. But the data will eventually have more delineated data on vulnerabilities: For example, we will know the severity of the vulnerabilities from databases like the NIST vulnerability database.

2. Attempted Attack Variable

'incident_email_infection_attempt' - attempt to attack one of the business employees based on a malicious email. Attackers targeted one of the firm's employees. The data they obtained about attacks showed that an email in the particular company was included in the list of email addresses that were targeted.

3. Incident (Breach) Variables:

'incident_code_data_leak' - indication for sensitive data leaked to the web
'incident_credentials_data_leak' - indication for credentials leaked
'incident_domain_abused_for_phishing' - one of the companies domains was abused to proliferate phishing
'incident_domain_abused_for_ransomware' - one of the companies domains was abused for the proliferation of ransomware. Ransomware is a type of malware that threatens to publish the victim's data or perpetually block access to it unless ransom is paid.
'incident_email_data_leak' - indication for data leak based on email identifier.
'Incident_endpoint_data_theft' - one the computers used by the organization was infected by malware that has stolen sensitive data
'incident_endpoint_ransomware' - one the computers used by the organization was infected by ransomware
'incident_ip_assets_abuse' - Based on IP identifier, one of the firm's assets was abused.
'incident_organization_data_leak' - data connected to the company was leaked to an outside source.

'Reputation_domain_malware' - one or more security vendors associates one of the companies domain as malicious.

'Reputation_ip_spam_or_malware' - one or more security vendors associates one of the companies IP addresses as proliferating spam.

All of the above variables are binary.

4. Precaution Data Variables:

'Infrastructure_service_captcha' - usage of a CAPTCHA security service. CAPTCHA is a Turing test used to tell computers (bots) and humans apart.
'infrastructure_service_email_security' - email security vendor used
'infrastructure_honeypot' - usage of honeypot, a sophisticated security service.

'infrastructure_login_page' - existence of login page at the website.

All of the above variables are binary.

5. Infrastructure Data Variables:

'infrastructure_service_online_video' - usage of online video platform
'infrastructure_service_storage' - storage service used
'infrastructure_service_tag_manager' - website tag manager used
'infrastructure_service_analytics' - analytics service used on the website
'infrastructure_third_party_used' - third party application used
'infrastructure_service_charts_maps' - third party services for maps and charts
'infrastructure_service_hosting' - website hosting provider
'infrastructure_service_email' - email service provider
'infrastructure_saas' – saas (software as a service) provider infrastructure
'infrastructure_service_cms' - content management platform used for the website
'infrastructure_service_content_delivery_network' - content delivery network used to improve loading times of webpages.

All of the above variables are binary.

References:

Anderson, R., and T. Moore, 2006, The economics of information security. Science, 314(5799):610-613.

Arora, A., Caulkins, J.P., and R. Telang, "Sell First, Fix Later: Impact of Patching on Software Quality," Management Science, Vol. 52, No. 3, March 2006, pp. 465–471, available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.165.2653&rep=rep1&type=pdf

Biener, C., Eling, M., and J. Wirfs, Insurability of Cyber Risk: An Empirical Analysis Geneva Papers on Risk and Insurance, Vol. 40, No. 1, 2015 University of St. Gallen, School of Finance Research Paper No. 2015/03, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2577286

Böhme, R., Laube,S., and M. Riek, A Fundamental Approach to Cyber Risk Analysis, Variance Journal, www.variancejournal.org, 2017, available at https://www.variancejournal.org/articlespress/articles/Fundamental-Boehme.pdf

Buldyrev, S., Parshani, R., Paul, G., and H. Stanley, and S. Havlin,2010, "Catastrophic Cascade of Failures in Interdependent Networks," Nature (Letters,) Vol 46,| 15 April 2010, available at doi:10.1038/nature08932, available at http://havlin.biu.ac.il/PS/Catastrophic%20cascade%20of%20failures%20in%20interdependent%20networks.pdf

Camp, L.J., and C. Wolfram, "Pricing Security," in L.J. Camp and S. Lewis, eds., Economics of Information Security, vol. 12, Advances in Information Security. Springer-Kluwer, 2004.

Campbell, K., Gordon, L.A., Loeb, M.P. and Zhou, L. (2003.) 'The economic cost of publicly announced information security breaches: empirical evidence from the stock market', Journal of Computer Security 11(3): 431–448.

Choi, J., Fershtman, C., and N. Gandal, "Network Security: Vulnerabilities and Disclosure Policy," 2010, (with Jay Pil Choi and Chaim Fershtman), *Journal of Industrial Economics*, 58:868-894

Clark, D., Berson, T, Blumenthal, M., Lin, H., and E. Whitaker, 2015, At the Nexus of Cybersecurity and Public Policy: Some Basic Concepts and Issues, The National Academies Press, Washington, DC, https://docs.house.gov/meetings/IF/IF02/20150303/103079/HHRG-114-IF02-20150303-SD006.pdf

Eling, M., and W.Schnell, edited by Fabian Sommerrock, 2016, Ten Key Questions on Cyber Risk and Cyber Risk Insurance by, available at https://www.genevaassociation.org/research-topics/cyber-and-innovation/ten-key-questions-cyber-risk-and-cyber-risk-insurance

Falco, G., Eling, M., Jablanski, D., Miller, V., Gordon, G., Wang, S., Schmit, J., Thomas, R., Elvedi, M., Maillart, T., Donavan, E., Dejung, S., Weber, W., Durand, E., Nutter, F., Scheffer, U., Arazi, G., Ohana, G., and H. Lin, A Research Agenda for Cyber Risk and Cyber Insurance, 2019 Available at https://weis2019.econinfosec.org/wp-content/uploads/sites/6/2019/05/WEIS_2019_paper_35.pdf

Gandal, N., Hamrick, J., Moore, T., and T. Oberman, "Price Manipulation in the Bitcoin Ecosystem, 2018, (with JT Hamrick, Tyler Moore, and Tali Oberman,) Journal of Monetary Economics, https://doi.org/10.1016/j.jmoneco.2017.12.004.

Gandal, N., Riordan , M., and S. Markovich, “Ain’t it “Suite? Bundling in the PC Office Software Market, 2018, (with Sarit Markovich and Michael Riordan,) Strategic Management Journal, https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.2797

Gatzlaff, K., and K.McCullough, 2011. The Effect of Data Breaches on Shareholder Wealth Management, Risk Management and Insurance Review, https://doi.org/10.1111/j.1540-6296.2010.01178.x

Gordon, L., & M. Loeb, (2002.) The Economics of Information Security Investment. ACM Transactions on Information and System Security, 5(4), 438–457.

Romanosky, S., Ablon, L., and A. Kuehn, Therese Jones, Content Analysis of Cyber Insurance Policies: How do carriers write policies and price cyber risk?, Journal of Cybersecurity, Volume 5, Issue 1, 2019, tyz002, https://doi.org/10.1093/cybsec/tyz002

Varian, H., 2004, “System Reliability and Free Riding,” available at http://www.ischool.berkeley.edu/~hal/Papers/2004/reliability.